



Previsão da Evolução do Covid-19 por Região

MARCELO ZEFERINO VIEIRA FERREIRA

Outubro de 2021

Previsão da Evolução do Covid-19 por Região

Marcelo Ferreira

**Dissertação para obtenção do Grau de Mestre em
Engenharia Informática, Área de Especialização em
Sistemas de Informação e Conhecimento**

Orientador: Carlos Ferreira

Júri:

Presidente:

Vogais:

Dedicatória

Esta dissertação é dedicada ao meu avô e a todos as vítimas de covid-19.

Resumo

No ano de 2019 iniciou-se uma pandemia, doença infecciosa que se espalha pela população numa grande escala geográfica, da pneumonia coronavírus 2019 (COVID-19), causada pelo SARS-Cov-2 (severe acute respiratory syndrome coronavirus 2), um auxiliar ao combate da crescente evolução deste vírus seria construir um sistema que agregue informação que está dispersa em várias fontes de informação diferentes e usar essa informação para prever essa evolução usando modelos de Machine Learning.

Através de pesquisas verificou-se que nem todos os modelos são indicados para este tipo de previsão, sendo que modelos lineares como por exemplo a regressão linear não conseguem produzir previsões de qualidade, desta forma os melhores modelos são aqueles que foram projetados para previsões ao longo do tempo, ou transformar os dados de forma que estes possam ser utilizados para este tipo de previsões na maioria dos modelos.

Através de pesquisa foi possível identificar modelos que poderiam obter bons resultados, que medidas utilizar para avaliar estes modelos, assim como variáveis que poderão ajudar os modelos a terem um desempenho melhor nas previsões. Após todo o trabalho de pesquisa o projeto procurou realizar vários testes nas previsões da evolução do covid-19 em Portugal, ou seja, nas previsões do número de casos confirmados, casos ativos, recuperados e óbitos. Os primeiros testes têm como objetivo identificar as características que fornecem aos modelos um melhor desempenho para prever o dia $N+1$, sendo N o dia atual. Após obter as características que produzem melhores resultados para previsões $N+1$, utilizou-se os modelos e essas características para executar previsões até $N+5$, em que se usa as próprias previsões para gerar previsões, desta forma verificou-se que os modelos que eram melhores nas previsões $N+1$, não são tão bons em previsões até $N+5$.

Por fim conseguiu-se implementar um sistema intuitivo, totalmente construído em linguagem R, que permite ao utilizador aceder a várias funcionalidades, como por exemplo analisar os dados atuais da pandemia, analisar a evolução da pandemia desde o seu início até ao dia atual, analisar as previsões em Portugal e suas regiões, assim como comparar as previsões.

Palavras-chave: Machine Learning, Evolução, Previsão, Covid-19, Modelos

Abstract

In 2019 a pandemic began, an infectious disease that spreads throughout the population on a large geographic scale of pneumonia coronavirus 2019 (COVID-19), caused by SARS-Cov-2 (severe acute respiratory syndrome coronavirus 2), an auxiliary to the combat of evolution of this virus would be to build a system that aggregates information that is dispersed across many different information sources and use that information to predict that evolution using Machine Learning models.

Through research it was found that not all models are suitable for this type of prediction, and linear models such as linear regression cannot produce quality predictions, thus the best models are those that were designed for predictions along of time intervals, or transform the data so that they can be used for this type of prediction in most models.

Through research, it was possible to identify models that could obtain good results, which measures to use to evaluate these models, as well as variables that could help the models to have a better performance in the predictions. After all the research work, the project tried to carry out several tests in the predictions of the evolution of covid-19 in Portugal, that is, in the predictions of the number of confirmed cases, recovered cases and deaths. The first tests aim to identify the characteristics that provide the models with a better performance to predict day $N+1$, with N being the current day. After obtaining the characteristics that produce the best results for $N+1$ predictions, we used the models and these characteristics to perform predictions up to $N+5$, in which the predictions themselves are used to generate predictions. were the best models at $N+1$ predictions, didn't perform so good at predictions up to $N+5$.

Finally, we managed to implement an intuitive system, entirely built in R language, which allows the user to access various features, such as analyzing current pandemic data, analyzing the evolution of the pandemic from its beginning to the present day, analyzing forecasts in Portugal and its regions, as well as comparing forecasts.

Keywords: Machine Learning, Evolution, Forecasting, Covid-19, Models

Conteúdo

Lista de Figuras	xiii
Lista de Tabelas	xv
Lista de Algoritmos	xvii
Lista de Código	xvii
1 Introdução	1
1.1 Enquadramento	1
1.2 Problema	1
1.3 Objetivos	2
1.4 Abordagem preconizada	2
1.5 Estrutura	2
2 Estado de Arte	5
2.1 Contexto	5
2.2 Tópicos de Machine Learning	5
2.2.1 O que é Machine Learning e para o que é usado?	5
2.2.2 Cenário de Uso de Machine Learning	6
2.2.3 Modelos mais usados	8
Regressão Linear	8
Naive Bayes	9
k-Nearest Neighbor(KNN)	9
Árvore de Decisão (DT)	10
Redes Neurais (NN)	11
Máquinas de Vetores de Suporte (SVM)	12
2.2.4 Métricas Avaliadoras de Performance	12
Métricas para Classificação	12
Métricas para Regressão	13
2.3 Projetos existentes de previsão de Covid-19	15
2.3.1 Visualization and machine learning for forecasting of COVID-19 in Senegal	15
2.3.2 COVID-19 Future Forecasting Using Supervised Machine Learning Models	18
2.3.3 Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches	20
2.3.4 Análise dos projetos	23
2.4 Soluções Existentes	23
2.4.1 CoronaNet	23
2.4.2 FiveThirtyEight- Covid Forecast	24

2.4.3	Institute for Health Metrics and Evaluation - Covid Projections . . .	26
2.5	Tecnologia/Ferramentas Uteis	28
2.5.1	Shiny	28
2.5.2	Web Scraping	29
3	Análise de Valor	31
3.1	New Concept Development	31
3.1.1	Identificação de Oportunidades	32
3.1.2	Análise de Oportunidades	32
3.1.3	Geração de Ideias	33
3.1.4	Seleção de Ideias	33
3.1.5	Definição de Conceito	34
3.2	Proposta de Valor	34
3.2.1	Segmentação de Mercado	34
3.2.2	Proposta de Valor - Modelo Canvas	34
	Cliente	34
	Produto	35
3.2.3	Modelo Canvas	35
3.3	QFD	36
3.3.1	Requisitos do Cliente	37
3.3.2	Requisitos Técnicos	38
3.3.3	Diagrama	38
3.4	AHP	38
3.4.1	Construção Árvore de Decisão	40
3.4.2	Construção de Matriz de Comparação de Critérios	41
3.4.3	Cálculo da Razão de Consistência	41
3.4.4	Construção da Matriz Comparação de cada critério entre alternativas	42
3.4.5	Cálculo da prioridade de cada alternativas e Escolha da melhor alternativa	43
4	Desenho da Solução	45
4.1	Domínio	45
4.2	Especificação de Requisitos	45
4.2.1	Excerto de Especificação de Requisitos (Funcionais)	46
4.2.2	Excerto de Especificação de Requisitos (Não-Funcionais)	48
4.2.3	Componentes Suplementares	50
4.3	Análise de Alternativas	51
4.3.1	Fontes de Dados	51
4.3.2	Base de Dados	52
4.3.3	Modelos	52
4.3.4	Aplicação	53
4.4	Esboço da Interface Gráfica	54
5	Implementação	57
5.1	Base de Dados	58
5.1.1	Construção da Base de Dados	58
	Dados Regionais	59
	Dados Meteorológicos	60
	Dados da Vacinação	62

	Dados do Índice de Transmissibilidade	63
	Dados sobre a Evolução do Covid-19	64
	Mortalidade	66
	Medida de Atenuação do Covid-19	66
5.1.2	Exploração dos Dados	66
5.1.3	Crítica à Base de Dados	73
5.2	Modelos	73
5.2.1	Transformação de dados	74
5.2.2	Metodologia de Testes	74
5.2.3	Testes identificadores de características	75
	Teste de desempenho com diferentes quantidades de dados históricos	76
	Teste de desempenho com diferentes conjuntos de variáveis	76
	Teste de Teste de desempenho com diferentes intervalos de treino .	79
	Fluxo dos Testes	79
5.2.4	Execução dos Testes	80
	Teste - Desempenho com diferentes quantidades de dados históricos	81
	Teste - Desempenho com diferentes conjuntos de variáveis	84
	Teste - Desempenho com diferentes intervalos de treino	88
5.2.5	Seleção de Modelos	91
5.3	Aplicação	95
5.3.1	Data.R	95
5.3.2	UI.R	96
5.3.3	Server.R	99
6	Avaliação e Experiência	103
6.1	Utilidade	103
6.2	Usabilidade	104
6.3	Comparação com outros sistemas	105
7	Conclusão e Futuro	107
7.1	Conclusão	107
7.2	Trabalho Futuro	108
	Bibliografia	109
A	Especificação de Requisitos	113
A.1	Formato Usado	113
A.2	Especificação de Requisitos	115
A.2.1	Principais exclusões	117
A.2.2	Analisar dados atuais	117
A.2.3	Analisar Previsões	121
A.2.4	Comparar Previsões	127
A.2.5	Excerto de Especificação de Requisitos (Não-Funcionais)	131
A.2.6	Componentes Suplementares	133
B	Implementação	135
B.1	Construção da Base de Dados	135
B.1.1	Lista de Dados	135
	Dados Regionais	135
	Dados Meteorológicos	136

	Dados da Vacinação	136
	Dados do Índice de Transmissibilidade	137
	Dados sobre o Covid-19	137
	Mortalidade	138
	Medida de Atenuação do Covid-19	138
B.1.2	Excertos de Código	139
	Dados Regionais	139
	Dados Meteorológicos	140
	Dados Vacinação	140
	Dados do Índice de Transmissibilidade	141
	Dados sobre evolução Covid-19	141
	Medida de Atenuação do Covid-19	143
B.2	Modelos	143
B.2.1	Excertos de Código	143
B.2.2	Testes	144
	Teste - Desempenho com diferentes quantidades de dados históricos	145
	Teste - Desempenho com diferentes conjuntos de variáveis	146
	Teste - Desempenho com diferentes intervalos de treino	148
B.3	Aplicação	151
B.3.1	Excertos de Código	151

Lista de Figuras

2.1	Diagrama de CRISP-DM, extraído de Wirth e Hipp 2000.	7
2.2	Exemplo de utilização de regressão Linear para a precisão do número de vendas.	9
2.3	Exemplo da distribuição de elementos por um referencial, onde o algoritmo procura encontrar os 7 vizinhos mais próximos do elemento que se pretende classificar ($k=7$), como pode ser observado a maioria dos vizinhos desse elemento está classificado como (A) o que levará o algoritmo a classificá-lo como da classe (A). Extraído de Klaine et al. 2017.	10
2.4	Exemplo de utilização de Árvore de Decisão para a previsão do grau da classificação de um aluno a uma unidade curricular. Extraído de Alloghani et al. 2018.	11
2.5	Representação básica de uma Rede Neuronal. Extraído de Alloghani et al. 2018.	11
2.6	Representação do funcionamento de Support Vector Machine. Extraído de Klaine et al. 2017.	12
2.7	Exemplo de Curva de ROC.	14
2.8	Casos confirmados de COVID-19 por região do Senegal. Extraído de Ndiaye, Balde e Seck 2020.	16
2.9	Previsão de casos confirmados de COVID-19: Prophet (gráfico da esquerda) e Rede Neuronal (gráfico da direita). Extraído de Ndiaye, Balde e Seck 2020.	18
2.10	Importância das variáveis para a execução da previsão. Extraído de Malki et al. 2020.	22
2.11	Exemplos de algumas colunas dos ficheiros CSV presentes no CoronaNet referentes a medidas implementadas em Portugal.	24
2.12	Exemplos das previsões do número de mortos nos Estados Unidos da América. Captura de ecrã do site (<i>Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree</i> <i>FiveThirtyEight</i> 2020).	25
2.13	Exemplo da previsão do número de mortos nos Estados Unidos da América no dia 16 de novembro. Captura de ecrã do site (<i>Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree</i> <i>FiveThirtyEight</i> 2020).	25
2.14	Exemplo da previsão do número de mortos nos estados Alabama, Alaska e Arizona no dia 28 de setembro tendo o modelo desenvolvido pelo exército americano destacado. Captura de ecrã do site (<i>Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree</i> <i>FiveThirtyEight</i> 2020).	26
2.15	Exemplo da previsão do número de mortos no Rio de Janeiro. Captura de ecrã do site (<i>IHME - COVID-19 Projections</i> 2021).	27
2.16	Exemplo da comparação de previsões do número de mortos entre Rio de Janeiro e São Paulo. Captura de ecrã do site (<i>IHME - COVID-19 Projections</i> 2021).	27

2.17	Exemplo de utilização da biblioteca Shiny para criar um formulário.	29
2.18	Exemplo de uso da biblioteca rvest para extrair dados. Extraído de (<i>rvest: easy web scraping with R</i> 2021).	30
3.1	Diagrama do Método New Concept Development.	32
3.2	Modelo de Negócio Canvas do projeto.	36
3.3	Representação da Casa de Qualidade.	37
3.4	Quality Function Deployment do projeto.	38
3.5	Árvore de Hierarquia de Decisão.	39
3.6	Escala de Comparações.	39
3.7	Árvore de Decisão do projeto.	41
3.8	Tabela Superior: Matriz Comparação de Critérios. Tabela Inferior: Matriz Comparação de Critérios Normalizada.	41
3.9	Cálculos para obter a Razão de Consistência.	42
3.10	Matrizes Comparação de cada critério.	42
3.11	Árvore de Decisão com todos os pesos.	43
3.12	Cálculo Final.	43
4.1	Diagrama de casos de uso.	46
4.2	Representação dos componentes do projeto.	51
4.3	Ilustração das alternativas.	52
4.4	Ilustração das alternativas de Arquitetura.	54
4.5	Esboço da interface do projeto que apresenta a situação pandémica atual no país.	55
4.6	Esboço da interface do projeto que apresenta a as previsões da evolução da pandemia no país.	55
4.7	Esboço da interface do projeto que apresenta a comparação entre previsões de 2 regiões.	55
5.1	Implementação do Sistema.	57
5.2	Diagrama da Base de Dados.	58
5.3	Tabelas Adicionais da Base de Dados.	59
5.4	Localizações das cidades representativas de cada região.	61
5.5	Gráfico de cima: Número de confirmados num determinado intervalo de temperatura média. Gráfico de baixo: Número de Óbitos num determinado intervalo de temperatura média.	69
5.6	Evolução do Índice de Transmissão de cada região.	70
5.7	Evolução do número de confirmados diários por faixas etárias e sexo.	71
5.8	Evolução do número de óbitos diários por faixas etárias e sexo.	72
5.9	Fluxo dos testes.	75
5.10	Correlação entre as variáveis que se pretende prever e as restantes.	78
5.11	Componentes do Sistema.	95
5.12	Página Atualidade e Evolução.	97
5.13	Página Previsões.	98
5.14	Página Comparação de Previsões.	98
5.15	Página Ajuda.	99
5.16	Exemplo de gráfico.	102
A.1	Diagrama de casos de uso do projeto.	116
A.2	Representação dos componentes do projeto.	134

Lista de Tabelas

2.1	Exemplo das Métricas mais usadas para avaliar as performances dos modelos de Machine Learning (Jiao e Du 2016)	13
2.2	Comparação de Projetos	15
2.3	Comparação entre os modelos usando RMSE. Extraído de (Ndiaye, Balde e Seck 2020).	17
2.4	Medidas de performance dos diferentes modelos na previsão de novos casos (tabela do topo), na previsão de recuperados (tabela do meio) e na previsão de mortes (tabela do fundo).Extraído de (Rustam et al. 2020).	19
2.5	Performances dos modelos com diferentes intervalos como conjunto de treino. Extraído de (Rustam et al. 2020).	20
2.6	Relação entre Humidade e Temperatura com o número de infetados e taxa de crescimento. Extraído de (Malki et al. 2020).	21
3.1	Tabela Índice de Aleatoriedade	40
4.1	Fluxo de Execução do Caso de Uso: Comparar Previsões	47
4.2	Requisito - Selecionar Regiões para Comparação	47
4.3	Requisito - Gráfico Comparação Casos Confirmados	48
4.4	Requisito Não-Funcional: Gráficos com cores destacadas e interativos	49
4.5	Requisito Não-Funcional: Disponibilidade do Sistema	49
4.6	Requisito Não-Funcional: Servidor	50
4.7	Requisito Não-Funcional: Compatibilidade com várias plataformas e navegadores	50
5.1	Primeiras linhas dos dados Meteorológicos	62
5.2	Primeiras linhas dos dados do Índice de Transmissibilidade	64
5.3	Primeiras linhas dos dados sobre evolução de Covid-19	65
5.4	Características das Regiões.	67
5.5	Médias das condições meteorológicas por região.	67
5.6	Estado do país ao longo do ano.	68
5.7	Médias das condições meteorológicas por região.	74
5.8	Exemplo de como funcionam os testes identificadores de características.	76
5.9	Exemplo de como funcionam os testes de seleção de modelos.	80
5.10	Testes de desempenho com diferentes quantidades de dados históricos a executar em todos modelos.	81
5.11	Resultados do teste de desempenho com diferentes quantidades de dados históricos utilizando o modelo Random Forest.	81
5.12	Resultados do teste de desempenho com diferentes quantidades de dados históricos utilizando o modelo K-Nearest Neighbors.	82
5.13	Resultados do teste de desempenho com diferentes quantidades de dados históricos utilizando o modelo ARIMA.	83

5.14	Testes de desempenho com diferentes conjuntos de variáveis a executar em todos modelos.	84
5.15	Resultados dos testes de desempenho com diferentes conjuntos de variáveis utilizando o modelo Random Forest.	85
5.16	Resultados dos testes de desempenho com diferentes conjuntos de variáveis utilizando o modelo K-Nearest Neighbors.	86
5.17	Resultados dos testes de desempenho com diferentes conjuntos de variáveis utilizando o modelo ARIMA.	87
5.18	Teste de desempenho com diferentes intervalos de treino a executar em todos modelos	88
5.19	Resultados dos testes utilizando o modelo Random Forest.	89
5.20	Resultados dos testes utilizando o modelo K Nearest Neighbors.	90
5.21	Resultados dos testes utilizando o modelo ARIMA.	91
5.22	Quadro resumo do resultado dos testes que tinham como objetivo identificar as caraterísticas que auferem os melhores resultados.	92
5.23	Exemplo de teste de seleção de modelos.	93
5.24	Exemplo de teste de seleção de modelos.	94
5.25	Informações presentes nos retângulos informativos.	100
5.26	Informações presentes nos Gráficos.	101
6.1	Descrição dos utilizadores que irão realizar os testes de usabilidade.	105
6.2	Resultados dos testes de usabilidade.	105
6.3	Comparação de funções com outras soluções.	106
A.1	Formato de descrição de casos de uso	113
A.2	Estados de descrição de casos de uso	114
A.3	Formato de descrição de fluxos de eventos	114
A.4	Formato de descrição de requisitos	115
A.5	Prioridade de requisitos	115
A.6	Atores do sistema de Previsão da evolução do COVID-19 por Regiões	116
A.7	Pacotes de casos de uso do sistema de Previsão da Evolução do COVID-19 por Regiões	117
A.8	Processos não suportados no sistema de Previsão de Evolução do COVID-19 por Região	117
A.9	Descrição do Casos de Uso: Analisar dados atuais	118
A.10	Fluxo de Execução do Casos de Uso: Analisar dados atuais do país	118
A.11	Fluxo de Execução Alternativo do Casos de Uso: Analisar dados atuais do país	118
A.12	Fluxo de Execução do Caso de Uso: Analisar dados atuais de uma região	119
A.13	Requisito - Mapa de Portugal Círculos	119
A.14	Requisito - Apresentar Situação do País	120
A.15	Requisito - Selecionar Região	120
A.16	Requisito - Dados Atuais da Região	120
A.17	Requisito - Alternativa a Selecionar Região	121
A.18	Descrição do Caso de Uso: Analisar previsões	121
A.19	Fluxos de execução do Caso de Uso: Analisar Previsões em Portugal	122
A.20	Fluxo de Execução do Caso de Uso: Analisar Previsões em Região	122
A.21	Requisito - Mapa de Portugal	122
A.22	Requisito - Apresentar Gráfico Casos Confirmados	123
A.23	Requisito - Linha Gráfico Casos Confirmados	123

A.24 Requisito - Apresentar Gráfico Recuperados	123
A.25 Requisito - Linha Gráfico Recuperados	124
A.26 Requisito - Apresentar Gráfico Casos Ativos	124
A.27 Requisito - Linha Gráfico Casos Ativos	125
A.28 Requisito - Apresentar Gráfico óbitos	125
A.29 Requisito - Linha Gráfico óbitos	125
A.30 Requisito - Selecionar Região	126
A.31 Requisito - Previsão de Região	126
A.32 Descrição do Caso de Uso: Comparar Previsões	127
A.33 Fluxo de Execução do Caso de Uso: Comparar Previsões	127
A.34 Requisito - Selecionar Regiões para Comparação	128
A.35 Requisito - Gráfico Comparação Casos Confirmados	128
A.36 Requisito - Gráfico Comparação Recuperados	129
A.37 Requisito - Gráfico Comparação Casos Ativos	130
A.38 Requisito - Gráfico Comparação óbitos	131
A.39 Requisito Não-Funcional: Gráficos com cores destacadas e interativos . . .	132
A.40 Requisito Não-Funcional: Disponibilidade do Sistema	132
A.41 Requisito Não-Funcional: Servidor	133
A.42 Requisito Não-Funcional: Compatibilidade com várias plataformas e navegadores	133
 B.1 Testes de desempenho com diferentes quantidades de dados históricos a executar em todos os modelos.	 145
B.2 Testes de desempenho com diferentes conjuntos de variáveis a executar em todos os modelos.	146
B.3 Continuação - Testes de desempenho com diferentes conjuntos de variáveis a executar em todos os modelos.	147
B.4 Teste de desempenho com diferentes intervalos de treino a executar em todos os modelos	148
B.5 Teste de desempenho com diferentes intervalos de treino a executar em todos os modelos(Continuação)	149
B.6 Teste de desempenho com diferentes intervalos de treino a executar em todos os modelos(Continuação).	150

Capítulo 1

Introdução

Este capítulo introdutório tem como objetivo enquadrar o leitor sobre o contexto e situação em que o documento foi elaborado. Tem também como objetivo apresentar o problema que se pretende resolver, assim como os objetivos do projeto.

Neste capítulo será também apresentada uma ideia de como será o projeto, assim como a estrutura do documento com uma breve explicação do conteúdo de cada capítulo.

1.1 Enquadramento

No ano de 2019 uma doença altamente infecciosa começou a espalhar-se pela população, atingindo o nível de pandemia, ou seja, a doença espalhou-se numa grande escala geográfica, atingindo quase todo o planeta Terra. Pandemia essa denominada de coronavírus 2019 (COVID-19), causada pelo SARS-Cov-2 (Severe Acute Respiratory Syndrome Coronavirus 2)(Dashraath et al. 2020).

Segundo a Organização Mundial de Saúde (WHO) o COVID-19 manifesta-se de forma diferente de paciente para paciente, já que a maioria das pessoas irá desenvolver sintomas ligeiros (como sintomas de tosse e febre ligeiros) enquanto outros poderão experienciar sintomas graves como dificuldade em respirar, dor no peito e perda da voz (World Health Organization 2020).

No presente dia de 23 de fevereiro de 2021, a WHO tem como casos confirmados de coronavírus 111 102 016 e 2 462 911 mortos, dentre os quais Portugal possui 797,525 casos de infeção e 15,962 óbitos(*Portugal: WHO Coronavirus Disease (COVID-19) Dashboard / WHO Coronavirus Disease (COVID-19) Dashboard* 2020).

1.2 Problema

Como se trata de uma doença altamente infecciosa, é cada vez maior o número de infetados e de fatalidades no mundo, levando os governantes de cada país a tomar medidas em resposta ao crescente desenvolvimento do vírus.

Uma técnica que poderia auxiliar a atenuar a crescente evolução do vírus seria a previsão dessa evolução. Existem uns entraves que dificultam a execução destas previsões, como a dispersão dos dados, já que os dados se encontram em diferentes locais, sendo necessário agregá-los. Para além da dispersão existe um problema com as regiões, visto que cada região tem as suas próprias características, o que leva a que o impacto do vírus seja diferente.

Para além das características das regiões existem outros tipos de variáveis que poderão ser importantes para o problema, como por exemplo as condições meteorológicas de cada região, para tal será feita uma análise aos dados da evolução do covid-19 e outras variáveis externas com o objetivo de verificar os seus impactos na evolução da doença e se possível a criação de novas variáveis de interesse para o problema.

Uma solução que disponibiliza informação agregada e correta assim como previsões de qualidade da evolução do covid-19 permitiria auxiliar os interessados em analisar previsões bem como poderia ajudá-los a agir mediante essas previsões, é precisamente uma solução como esta que se pretende desenvolver ao longo deste documento.

1.3 Objetivos

O objetivo deste projeto é o desenvolvimento de uma solução para o problema relatado em 1.2.

Esta solução tem como objetivo extrair e agregar dados de várias fontes, assim como fazer uma análise aos dados recolhidos e se possível gerar novas variáveis úteis para o problema. Pretende-se também a construção de modelos de Machine Learning que irão utilizar os dados que foram previamente mencionados para prever a evolução da covid-19, tendo em conta que as diferentes características de cada região. Para cada região pretende-se prever o número de casos confirmados, de óbitos, recuperados e ativos para os 5 dias a seguir ao dia atual, mas devido a falta de dados as últimas duas previsões anunciadas serão impossíveis de se realizar em todas as regiões.

Os modelos que estarão na solução final serão selecionados dentre um conjunto vasto de modelos após todos serem alvo de testes e avaliações comparativos, usando métricas adequadas, de forma que os modelos com melhor performance sejam selecionados.

É pretendido que a solução apresente os dados que foram agregados das várias fontes de informação, ou seja, os dados reais e as previsões numa aplicação web. Todos os dados apresentados estarão em forma de dashboards, ou seja, através de gráficos onde sejam possíveis a análise e a interação com os dados.

1.4 Abordagem preconizada

Prevê-se que a melhor abordagem seja criar uma aplicação web de forma a ser acessível em todos os dispositivos. O sistema irá extrair dados, por exemplo, dos relatórios diários da Direção Geral de Saúde (que relatam a evolução da pandemia diariamente) e dados do Instituto Português do Mar e da Atmosfera (dados meteorológicos). Estes dados serão armazenados no sistema de forma a serem apresentados ao utilizador e serem utilizados pelos modelos de Machine Learning para executar as previsões.

É também previsto que a extração de dados e as previsões sejam feitas de forma diária numa altura do dia em que os utilizadores não estejam a usar a ferramenta.

1.5 Estrutura

O documento tem a seguinte estrutura:

- Introdução - capítulo onde é feita uma breve descrição do projeto;

- Estado de Arte - capítulo que está dividido em 3 pontos principais:
 1. Tópicos de Machine Learning - onde se procura compreender como este tipo de técnicas funcionam, técnicas mais usadas e que métricas são usadas para avaliar os seus desempenhos;
 2. Projetos Existentes - onde se analisa projetos sobre a previsão da evolução do vírus de forma a analisar quais as técnicas de Machine Learning terão melhor performance e quais as métricas avaliadoras se aplicam melhor ao problema, assim como quais as variáveis de maior relevância para estas previsões;
 3. Soluções Existentes - onde se analisa soluções que estão disponíveis publicamente de forma a perceber como elas funcionam;
- Análise de Valor - capítulo onde se procura entender de que forma uma ferramenta de previsão da evolução do covid-19 pode trazer valor para os seus utilizadores;
- Desenho de Solução - Capítulo onde se irá analisar várias alternativas para os diferentes componentes da solução, de forma a propor uma arquitetura e a especificar os requisitos que constituem o sistema;
- Avaliação e Experiência - capítulo que procura avaliar a solução;
- Conclusões e Futuro - capítulo que apresenta as conclusões que foram obtidos com o desenvolver do documento e se apresenta também os aspetos da solução que ficaram em falta ou poderiam ser adicionados, assim como melhorias que poderiam ser feitas;

Capítulo 2

Estado de Arte

2.1 Contexto

O covid-19 transmite-se principalmente através das gotículas que são geradas quando uma pessoa infetada tosse, espirra ou expira, o que o leva a ser um vírus com elevada taxa de transmissão, ou seja, pode ser transmitido de uma pessoa para outra com grande facilidade. Esta facilidade de transmissão torna-o num vírus com elevada taxa de propagação, tem a capacidade de atingir um elevado número de pessoas em pouco tempo.

Uma forma de auxiliar o combate ao vírus é executar previsões sobre a sua evolução, ou seja, calculando o número de casos confirmados, óbitos e recuperados. Estas previsões poderiam manter a população geral informada, de forma a consciencializar que é preciso manter o foco e cumprir todas as normas impostas, bem como servir de esperança. Estas previsões poderão também ajudar, por exemplo um presidente da Câmara de alguma cidade, a tomar medidas para atenuar a evolução do vírus.

Os dados que mais poderão ter relevância para estas previsões são, por exemplo, dados sobre a evolução da pandemia, dados meteorológicos, dados demográficos do país e das suas regiões. Estes dados encontram-se dispersos já que não pertencem à mesma área, esta dispersão e o facto do covid-19 ainda ser recente faz com que existam poucas soluções credíveis de previsão.

Em todo este capítulo pertence-se analisar informação relacionada com Machine Learning e previsões de covid-19, de forma a compreender melhor o tema.

2.2 Tópicos de Machine Learning

2.2.1 O que é Machine Learning e para o que é usado?

Segundo Alpaydin 2020, as técnicas de Machine Learning podem ser definidas como um conjunto de modelos computacionais que usam a experiência, procurando a semelhança com a aprendizagem humana, para executar previsões. Esta experiência é adquirida através da análise de informações relativas ao passado, normalmente na forma de dados eletrónicos como por exemplo bases de dados (sendo a qualidade e tamanho destes dados fatores chave para o sucesso destes modelos).

Normalmente, o Machine Learning é usado com maior intensidade para as seguintes tarefas (Mohri, Rostamizadeh e Talwalkar 2018):

- Aprendizagem Supervisionada - Prevê valores desconhecidos ou valores futuros de variáveis de interesse, ou seja, orientada por um objetivo:

- Classificação: onde se pretende classificar um determinado item, por exemplo classificar o tópico de um livro, outro exemplo é a classificação de imagens;
- Regressão: onde se pretende prever um valor real, como por exemplo prever o número de unidades de um determinado produto em stock;
- Aprendizagem Não-Supervisionada - Procura encontrar relações que descrevam os dados através de modelos ou seja, não existe um objetivo pré-definido:
 - Ranking(Hierarquia): modelos em que se procura ordenar os itens com base num determinado critério;
 - Clustering: o objetivo destes modelos é agrupar os itens em diferentes subconjuntos com base em algum critério, como por exemplo usar o clustering para agrupar os clientes de uma loja com base nos produtos comprados;

2.2.2 Cenário de Uso de Machine Learning

CRISP-DM, significa Cross-Industry Standard Process for Data Mining e é uma metodologia cíclica que tem como objetivo ajudar a organizar e a implementar um projeto de Machine Learning. Este método é constituído por 6 fases (Wirth e Hipp 2000):

- Business Understanding: fase em que se procura a compreensão dos objetivos e definição de um problema de Machine Learning;
- Data Understanding: esta fase tem como objetivo extrair e organizar os dados das suas fontes de forma se familiarizar com os dados;
- Data Preparation: onde se executam os tratamentos e limpezas aos dados de forma a obter uma base de dados final;
- Modeling: aplicação dos modelos selecionados aos dados tendo em conta o objetivo;
- Evaluation: Antes de se colocar o modelo em uso é necessário avaliar o seu desempenho, muitas vezes comparando com outros modelos;
- Deployment: Colocar o modelo em uso, muitas vezes para gerar relatórios. Notar que esta fase pode não ser a fase final, já que o processo é cíclico e muitas vezes é possível aprimorar os modelos;

7. **Avaliação** - Usando várias métricas como por exemplo a precisão, avalia-se a performance do modelo. Como o processo de aprendizagem de modelos em Machine Learning é um processo cíclico, é possível, caso a performance não seja suficientemente boa podemos alterar a forma como os dados foram partidos ou alterar outras características de treino, alterando o modelo de Machine Learning de forma a aumentar esta performance;
8. **Uso** - Tendo um modelo com uma boa performance é possível a previsão de novos dados, por exemplo a chegada de um novo email que, usando o modelo, é prevista como spam;

Este exemplo possui mais fases do que o método de CRISP-DM pois para uma melhor compreensão a fase de Modelação (Modeling) foi dividida em várias fases.

2.2.3 Modelos mais usados

Nesta secção iremos abordar os modelos de Machine Learning mais utilizados para regressão que é o caso do problema, já que se pretende prever valores numéricos. Alguns dos modelos que serão apresentados foram projetados para classificação, mas também permitem executar regressão. Iremos também fazer uma abordagem mais curta ao clustering pelo fato de ser interessante para o problema descrito em 1.2 para agrupar as regiões com características semelhantes.

Regressão Linear

Os modelos de regressão linear (LR) define que o atributo que se pretende prever é a variável dependente e que os restantes atributos são variáveis independentes, onde o modelo irá usar as variáveis independentes e irá atribuir-lhe diferentes pesos, que podem ser alterados, para construir uma função linear que será usada para executar as previsões, ou seja, a regressão linear assume que a dependência de Y em $f(X)$ é linear e como tal aproxima a distribuição dos dados por uma linha (Hastie, Tibshirani e Friedman 2009).

Através da figura 2.2 é possível observar como os modelos de regressão linear se comportam, eles utilizam os dados disponíveis para calcular uma linha que melhor representa a relação entre as variáveis independentes e a variável dependente, de seguida utiliza essa mesma reta para as previsões.

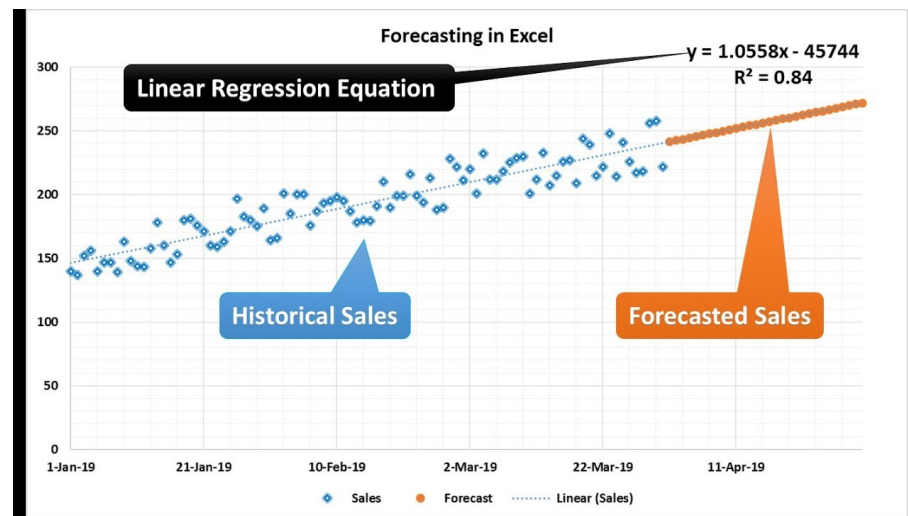


Figura 2.2: Exemplo de utilização de regressão Linear para a precisão do número de vendas.

Naive Bayes

O teorema de Bayes é constituído por regras de probabilidade e análise estatística que permitem calcular probabilidades condicionais, ou seja, permite calcular a probabilidade de uma hipótese (h) ser afetada por uma evidência (e). Esta probabilidade é calculada pela seguinte formula:

$$P(h|e) = \frac{P(e|h) \cdot P(h)}{P(h)} \quad (2.1)$$

Esta fórmula pode ser lida como a probabilidade de h acontecer sabendo que e aconteceu, daí se chamar probabilidade condicional já que a hipótese h depende da evidência e . O algoritmo de Machine Learning Naive Bayes tem como principal recurso esta fórmula probabilística para executar previsões, normalmente de classificação. A expressão "Naive" que em português significa ingênuo, deve-se ao facto deste modelo ser ingênuo no aspeto de considerar que as variáveis que entram no algoritmo são independentes entre elas. Seguindo o exemplo de classificar os emails como spam ou não-spam, os dados de entrada neste modelo poderiam ser o remetente do email, o assunto do email, conteúdo do email e se o email tem ou não anexos, como podemos ver, por exemplo o conteúdo e o assunto de um email estão sempre dependentes um do outro, ao considerar que são independentes pode levar a erros na previsão. Existem algumas medidas e formulas para atenuar este problema (Klaine et al. 2017; Vembandasamy, Sasipriya e Deepa 2015).

Quando se pretende prever um valor numérico as variáveis independentes são normalmente valores numéricos que são normalizados, ou seja, são transformados de forma a estarem na todos mesma escala (Frank et al. 2000).

k-Nearest Neighbor(KNN)

Este algoritmo tem por base a posição dos dados num referencial, daí ser chamado Nearest Neighbor que em português significa "vizinho mais próximo". Este modelo usa um processo muito simples que é classificar um elemento com base nos dados dos k vizinhos (outros

elementos com características parecidas, o que os leva a estarem próximos num referencial) como pode ser observado na figura 2.3, sendo que existem fórmulas matemáticas para calcular o k que melhor se aplica a cada problema, assim como existem fórmulas para calcular as distancias entre o novo elemento e os já existentes, como por exemplo a distância euclidiana. Este modelo é bastante usado para classificação, mas também pode ser usado para regressão (Klaine et al. 2017; Zhang et al. 2017).

No caso da regressão, é feita através da média da variável dependente dos k vizinhos mais próximos, onde se pode considerar que os vizinhos mais próximos possuem um maior peso para a média (Imandoust e Bolandraftar 2013).

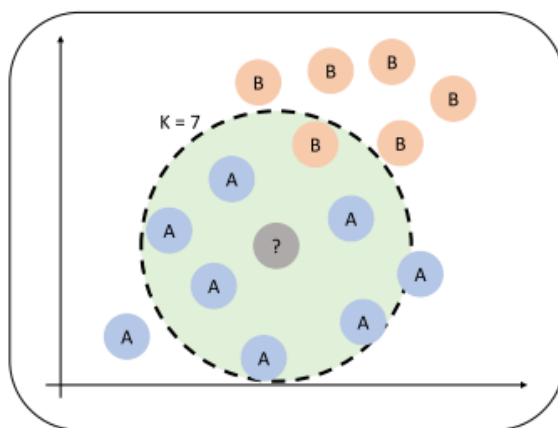


Figura 2.3: Exemplo da distribuição de elementos por um referencial, onde o algoritmo procura encontrar os 7 vizinhos mais próximos do elemento que se pretende classificar ($k=7$), como pode ser observado a maioria dos vizinhos desse elemento está classificado como (A) o que levará o algoritmo a classificá-lo como da classe (A). Extraído de Klaine et al. 2017.

Árvore de Decisão (DT)

Este algoritmo é bastante simples, mas muito poderoso (já que consegue processar texto, dados numéricos assim como dados com erro e/ou ausência de dados). A principal premissa deste algoritmo é dividir os dados fornecidos em subconjuntos tendo em conta as suas variáveis e a dependência entre elas, criando desta forma uma árvore partindo dos dados de treino. Quando se pretende avaliar um novo elemento percorre-se a árvore e irá ser encontrado "ramo" que melhor prevê a sua situação. Existem árvores de classificação e árvores de regressão, no caso da classificação o melhor ramo terá uma variável classificativa enquanto na regressão terá um valor número (Klaine et al. 2017; Somvanshi et al. 2016).

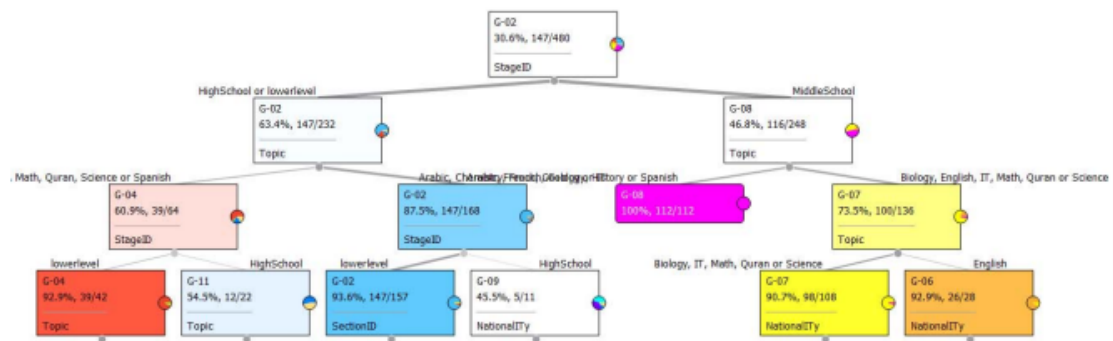


Figura 2.4: Exemplo de utilização de Árvore de Decisão para a previsão do grau da classificação de um aluno a uma unidade curricular. Extraído de Alloghani et al. 2018.

Como podemos observar em 2.4 é uma árvore de decisão para prever a classificação dos alunos a uma determinada unidade curricular (sendo essa classificação feita de G-2a G-11), a árvore começa por se dividir em 2 ramos tendo em conta o ensino, ou seja, se é 1º ciclo, secundário, entre outros, muito provavelmente pela dificuldade de obter notas superiores, a árvore divide-se também em relação á unidade curricular, á nacionalidade dos alunos, tendo como nodo final nos ramos a "melhor decisão para aquela situação" (Alloghani et al. 2018; Klaine et al. 2017).

Redes Neurais (NN)

As redes neurais são inspiradas no cérebro humano que é constituído por uma rede de neurónios que passam informação entre si tomando decisões, acedendo memórias entre outros. Nas redes neurais o equivalente a neurónios são os nodes, estes nodes são responsáveis por receber informação, executa uma determinada função e calcular o output que irá passar a outro node.

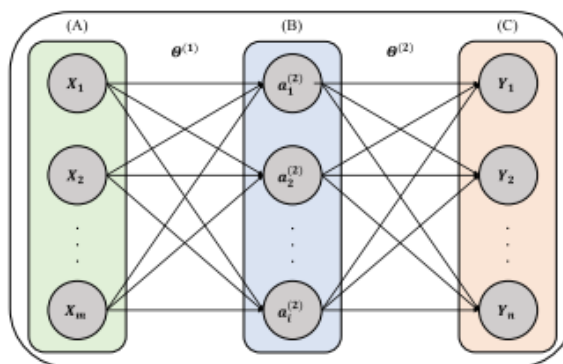


Figura 2.5: Representação básica de uma Rede Neuronal. Extraído de Alloghani et al. 2018.

A figura 2.5 representa a rede neuronal mais básica possível, constituída por uma camada de entrada (A-input layer), uma camada de processamento (B- hidden layer) e outra de saída (C- output layer). Onde X corresponde aos atributos de entrada e o Y aos atributos de saída.

De notar que cada node representa uma variável e que todos os nodes fornecem dados a todos os nodes da camada seguinte. Estes nodes da camada seguinte irão fazer os seus cálculos com base nos dados que recebem e com base no peso da ligação (atributo que caracteriza a ligação entre 2 nodes) (Klaine et al. 2017).

No caso da figura 2.5 existe apenas uma camada de processamento, mas poderão existir mais camadas de forma a melhorar o desempenho do modelo. É possível observar que na figura 2.5 a camada de saída possui várias variáveis, mas ela poderia ter apenas uma.

Máquinas de Vetores de Suporte (SVM)

SVM ou Máquinas de Vetores de Suporte são modelos que representam os dados de entrada/treino na forma de pontos num espaço e procuram encontrar um plano que separe as classes, chamado de hyperspace, de maneira que a distância entre este plano e o ponto mais próximo de cada uma das classes seja a maior possível. Um novo elemento irá ser colocado neste espaço e a previsão será feita com base na sua posição em relação às outras classes e ao hyperspace (Alloghani et al. 2018; Klaine et al. 2017).

Na figura 2.6 podemos observar essa divisão espacial entre as classes (A) e (B), assim como é visível o hyperplane.

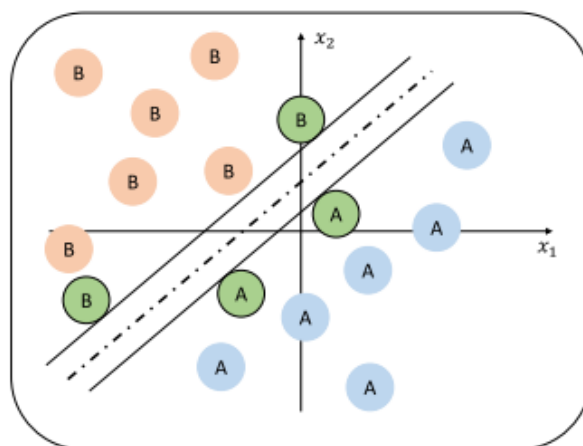


Figura 2.6: Representação do funcionamento de Support Vector Machine.
Extraído de Klaine et al. 2017.

2.2.4 Métricas Avaliadoras de Performance

Métricas para Classificação

A seguinte tabela resume algumas das métricas mais usadas para avaliar a performance dos modelos de Machine Learning de classificação, tendo por base os seguintes pontos (Jiao e Du 2016):

1. TP (Positivos Verdadeiros) - número de casos positivos corretamente classificados;
2. TN (Negativos Verdadeiros) - número de casos negativos corretamente classificados;
3. FP (Positivos Falsos) - número de casos positivos erradamente classificados;
4. FN (Negativos Falsos) - número de casos negativos erradamente classificados;

Tabela 2.1: Exemplo das Métricas mais usadas para avaliar as performances dos modelos de Machine Learning (Jiao e Du 2016)

Métricas Avaliadoras de Performance		
Nome	Fórmula	Descrição
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Mede a capacidade de um modelo identificar corretamente todas as classes, não importa se é positivo ou negativo;
Recall Sensitivity TPR	$\frac{TP}{TP + FN}$	Taxa de acerto na classe Positiva; Mede a capacidade de um modelo na identificação da classe positivas;
Specificity TNR	$\frac{TN}{TN + FP}$	Semelhante a Sensibility, mede a capacidade de um modelo em identificar classe negativas;
Precision PPV	$\frac{TP}{TP + FP}$	Indica a frequência de verdadeiros positivos entre todas as saídas positivas;
F1 Score	$2 \frac{Precision \cdot Recall}{Precision + Recall}$	Mede a performance do modelo não considerando os verdadeiros negativos; Medida usada quando só é importante analisar os outputs positivos

Quando a variável alvo tem mais do que 2 classes, para realizar estes cálculos é necessário fazer uma tabela/matriz para cada classe onde se desconsidera as outras, ou seja, supondo que a variável alvo é a altura de uma árvore que é classificada como grande, média e pequena, é necessário fazer uma tabela para a cada classe onde por exemplo para a altura é grande a classe positiva é "grande" e a classe negativa é "não grande" (média e pequena) (Jiao e Du 2016).

Métricas para Regressão

Outra forma de avaliação importante e muito usada para comparar modelos é a curva de ROC (Receiver Operating Characteristic), uma representação gráfica da relação entre a Taxa dos Verdadeiros Positivos (TPR, eixo das ordenadas) e a Taxa dos Falsos Positivos (FPR = 1-Specificity, eixo das abcissas). É exatamente o que é visível na figura 2.7, onde temos 2 curvas referentes a 2 modelos diferentes (M1 e M2) de onde é possível extrair informação. No caso de a curva passar abaixo da linha diagonal, previsão é oposta à classe correta. Quanto mais próximo do canto superior esquerdo a curva estiver melhor a performance do modelo (Narkhede 2018).

Existe uma medida que pode ser extrair do gráfico de forma a ser possível a comparação entre modelos, onde se reduz a curva ROC a um valor escalar, cálculo da área abaixo da curva ROC (AUC- Area Under the Curve) até a linha diagonal. Na figura 2.7 também podemos observar que o modelo 1 é melhor quando a taxa de positivos é menor, enquanto o modelo 2 é melhor quando a taxa é maior (Narkhede 2018).

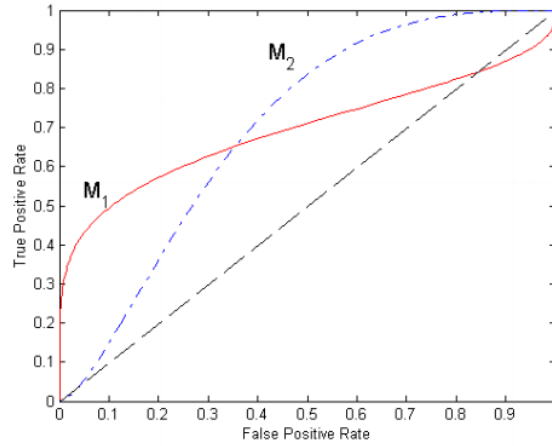


Figura 2.7: Exemplo de Curva de ROC.

R^2 é uma medida que mede a força da relação entre a variável dependente e o modelo, significa o quão bem o modelo representa os dados de treino. $R^2_{ajusted}$ tal como o seu antecessor mostra o quão bem os pontos do modelo representam a curva do conjunto de treino da variável dependente, mas tem um pequeno ajuste tendo em conta o número de variáveis independentes (Rustam et al. 2020).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (2.2)$$

$$R^2_{Adjusted} = 1 - \frac{n-1}{n-(k-1)}(1 - R^2) \quad (2.3)$$

O erro médio absoluto (MAE- mean absolute error) é a media dos erros entre os dados previstos na fase de teste (y_i) e os dados reais (x_i). Sendo possível compreender entre 0 e infinito, sendo que quanto menor o erro, melhor a previsão (Rustam et al. 2020).

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (2.4)$$

O erro quadrático médio (MSE- mean square error) é outra medida bastante usada, onde se mede a distância entre os dados valores verdadeiros (Y_i) e o seu par de regressão ($\hat{Y}_i$) e calcula-se o quadrado dessa distancia, para que não haja valores negativos e para dar mais peso às distancias (Rustam et al. 2020).

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (2.5)$$

Raiz do erro quadrático médio (RMSE- root mean square error), baseia-se em ser a raiz quadrada do erro quadrático médio, mede o quão concentrados os pontos de dados reais estão em torno da linha prevista (Rustam et al. 2020).

$$RMSE = \sqrt{MSE} \quad (2.6)$$

2.3 Projetos existentes de previsão de Covid-19

Nesta seção irão ser apresentados projetos existentes sobre a previsão da evolução do Covid-19, de forma a encontrar os melhores modelos de Machine Learning, métricas avaliadoras de performance e variáveis de maior relevância para o problema descrito em 1.2. Foram definidos critérios que foram obrigatoriamente cumpridos para que os resultados da pesquisa estivessem em linha com o desejado. Esses critérios são os seguintes:

- **Fontes:** Google Scholar, Research Gate and b-on.pt;
- **Linguagem de Pesquisa:** Inglês (para ter um maior número de resultados);
- **Palavras-Chave:** Machine Learning, Covid-19, Predictions, Forecasting, Performance;
- **Análise:** Leitura do resumo e da conclusão para ter uma melhor compreensão dos estudos e dos resultados;
- **Referências:** Mais de 20. Estudos com grande número de referências indicam maior relevância e qualidade;

Como a pesquisa gerou um grande número de resultados, foi necessário selecionar aqueles que, de alguma forma, se destacaram dos outros. Além do método de pesquisa, cada projeto foi selecionado por uma razão específica que está descrita na Tabela 2.2.

Tabela 2.2: Comparação de Projetos

Título	Modelos	Destaque
Visualization and machine learning for forecasting of COVID-19 in Senegal (Ndiaye, Balde e Seck 2020)	Support Vector Regression; Linear Regression; Neural Networks;	Previsão no Senegal; Comparação entre Modelos; Previsão dos próximos 5 dias;
COVID-19 Future Forecasting Using Supervised Machine Learning Models (Rustam et al. 2020)	Support Vector Machine; Linear Regression; LASSO Regression;	Previsão Global
Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches (Sajadi et al. 2020)	Support Vector Regression; Decision Trees; Neural Networks; Entre outros;	Uso de 20 Modelos Uso de dados Meteorológicos; Uso de Redes Neurais para previsões em sequências temporais;

2.3.1 Visualization and machine learning for forecasting of COVID-19 in Senegal

Neste projeto os autores usam modelos determinísticos e técnicas de Machine Learning para estudar e prever os casos COVID-19 no Senegal. O projeto divide-se em 2 pontos

importantes, a análise de dados e com recurso a técnicas de Machine Learning construção de modelos para prever a evolução do COVID-19 no Senegal (Ndiaye, Balde e Seck 2020).

Como fonte de dados foram usados os dados divulgados pelo Ministério da Saúde do Senegal entre 2 de março e 26 de julho, a análise dos dados foi feita de forma a explorar se alguns fatores (densidade populacional, religiões / crenças, alimentos, juventude, temperatura, humidade e imunidade cruzada) podem reduzir a taxa de propagação do COVID-19 no Senegal.

Na análise dos dados verificou-se que a grande maioria dos casos positivos se verificavam na capital do país Dakar, como se pode observar na figura 2.8 muito devido á elevada densidade populacional naquela região. Os autores usaram os dados obtidos do Ministério de Saúde e calcularam o grau de crescimento dos dados para cada dia, se o grau fosse igual a 1 não houve diferença entre um dia e outro, se fosse maior que 1 indicava uma tendência no aumento do número de casos diários, caso fosse menor que 1 indicava que o número de casos estava a diminuir, recorrendo ás seguintes fórmulas verificou-se que o grau era, em muitos dias, maior que 1 mas que tinha tendência a descer.

$$\text{Crescimento Casos Confirmados} = \frac{N. \text{ de Casos no Dia}}{N. \text{ de casos no dia Anterior}} \quad (2.7)$$

$$\text{Crescimento Casos Recuperados} = \frac{N. \text{ de Recuperados no Dia}}{N. \text{ de Recuperados no dia Anterior}} \quad (2.8)$$

$$\text{Crescimento Mortes} = \frac{N. \text{ de Mortes no Dia}}{N. \text{ de Mortes no dia Anterior}} \quad (2.9)$$

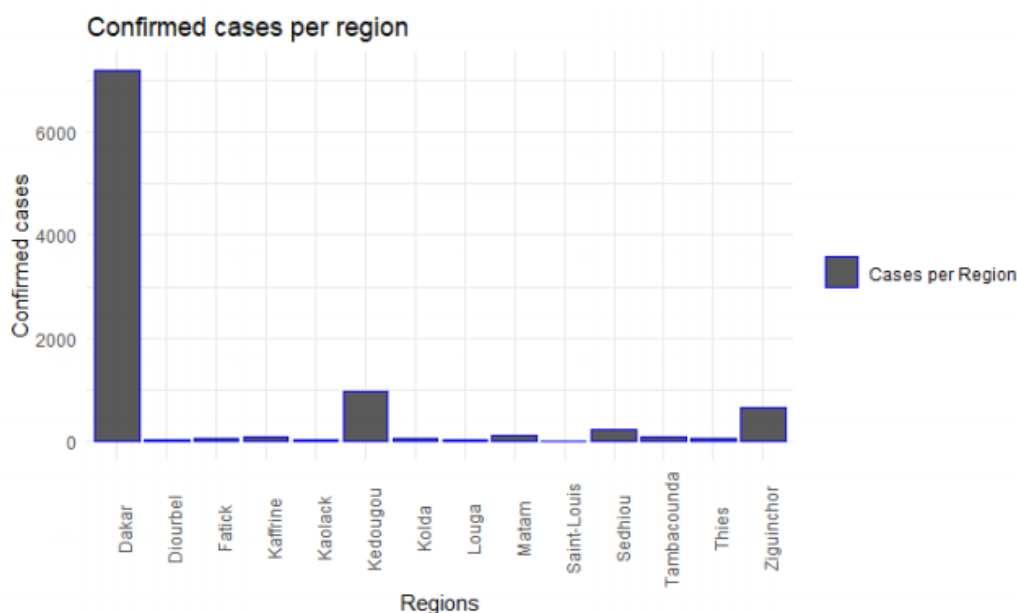


Figura 2.8: Casos confirmados de COVID-19 por região do Senegal. Extraído de Ndiaye, Balde e Seck 2020.

Para uma previsão dos próximos 5 dias foram usados os modelos seguintes: Linear Regression, Polynomial Regression, Support Vector Regression (SVR), Prophet e Redes Neurais, após uma comparação entre os modelos através da métrica RMSE usou-se os 2 melhores modelos para fazer uma previsão a 2 semanas e outra a 40 dias.

Através do modelo de regressão linear descrito na secção 2.2.3 verificou-se que as tendências do número de casos de mortes e infetados não são lineares ou seja no modelo a linha de regressão que melhor define o problema, não foi de encontro com os números de mortes e novos casos, fazendo com que o modelo obtivesse um valor de 1576,9 de RMSE como se pode observar na tabela 2.3.

O próximo modelo a ser construído foi de regressão polinomial (Polynomial Regression) é outra forma de regressão em que a potência máxima da variável independente é mais de um. Com esta técnica ao invés de obtermos uma reta que melhor se ajusta ao problema, temos uma curva proporcionando um melhor ajuste e melhorando a relação de variáveis dependentes e independentes no problema. Dito isto, este modelo obteve uma melhor performance do que a regressão linear, o que pode ser comprovado pelo RMSE, que foi de 132,4.

O modelo de SVM descrito na secção 2.2.3 é usado para classificação e regressão, como neste caso, o objetivo deste modelo é encontrar a função $f(x)$ dependendo das funções do kernel, verificou-se que este método registou o pior resultado deste projeto, tendo 4942.7 de RMSE.

O próximo modelo testado foi o Prophet, modelo este que é usado para previsão de dados de série temporal com base em um modelo aditivo onde tendências não lineares são ajustadas com sazonalidade anual, semanal e diária, além do efeito dos feriados, funcionando bem com ausência de dados, mudanças de tendências e outliers (valores fora do habitual) (Wang et al. 2020). Como é um modelo dedicado a previsões ao longo do tempo teve um bom desempenho tendo 24,3 de RMSE.

O último modelo testado foi a rede neuronal, descrito na secção 2.2.3, no caso foi usada uma rede com múltiplas camadas de perceptrons, foi usado também o pacote de R nnfor (Time Series Forecasting with Neural Networks) que, assim como o modelo Prophet, permite a previsão de séries temporais.

Tabela 2.3: Comparação entre os modelos usando RMSE. Extraído de (Ndiaye, Balde e Seck 2020).

Nome do modelo	Root Mean Squared Error (RMSE)
Linear Regression	1576.974603
Polynomial Regression	132.434659
Support Vector Machine Regressor	49310.716486
Prophet Model	24.332935
Multilayer Perceptron	19.71207

Como já foi dito anteriormente os modelos de Support Vector Machine e de Regressão Linear obtiveram resultados considerados maus, obtendo perto de 1600 e 5000 de RMSE, respetivamente, como se verifica na tabela 2.3. É também observável que os melhores

modelos são os que têm por base previsões em series temporais, no caso o modelo de Redes Neurais (com recurso ao pacote "nnfor") e o modelo Prophet.

A figura 2.9 mostra as previsões efetuadas para os próximos 40 dias, na altura do projeto, verificando-se que estas previsões possuem sempre um intervalo, ou seja, a previsão indica-nos o valor que o modelo prevê de número de casos, mas também um intervalo previsto onde o verdadeiro valor se poderá inserir.

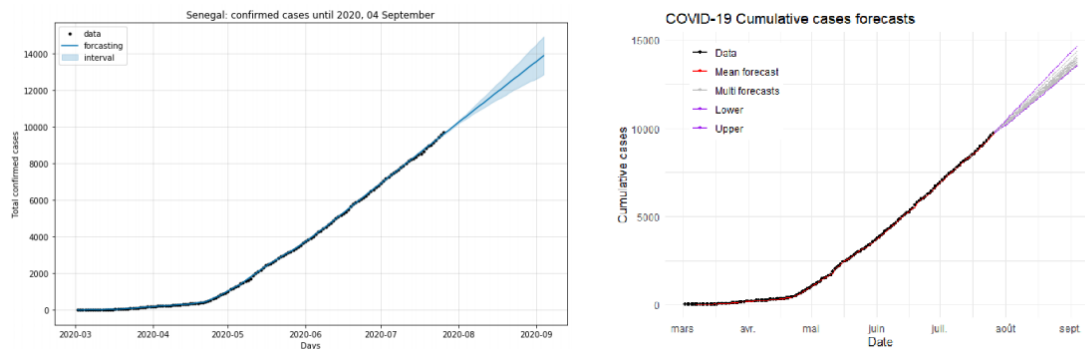


Figura 2.9: Previsão de casos confirmados de COVID-19: Prophet (gráfico da esquerda) e Rede Neuronal (gráfico da direita). Extraído de Ndiaye, Balde e Seck 2020.

2.3.2 COVID-19 Future Forecasting Using Supervised Machine Learning Models

Neste projeto procurou-se prever no número casos confirmados, mortos e de recuperados num espaço de 10 dias após os últimos dados obtidos. O conjunto de dados usado foi fornecido pela Universidade Johns Hopkins, contendo uma tabela para casos confirmados, uma para mortes e outra para recuperados. Em cada uma destas tabelas, cada coluna representa um dia (ou seja, uma data), cada linha representa um país e cada célula representa o número (casos confirmados, mortes ou recuperados) num determinado dia num determinado país (*CSSEGISandData · GitHub* 2020).

Os modelos usados nestas previsões foram os de Regressão Linear e SVM, descritos na secção 2.2.3, bem como os modelos de LASSO e de Suavização Exponencial (ES). O modelo LASSO é um modelo de regressão linear que usa a técnica shrinkage que permite reduzir valores extremos a valores mais centrais, enquanto o modelo de ES baseia-se nos períodos de dados anteriores em que a influencia dos dados diminui exponencialmente à medida que eles envelhecem, usando a seguinte fórmula (Petropoulos e Makridakis 2020):

$$F_t = \alpha A_{t-1} + (1 - \alpha) F_{t-1} \quad (2.10)$$

Procedeu-se da mesma forma para os 3 tipos de dados (casos confirmados, mortes e recuperados): dividiu-se os dados de forma que os conjuntos de treino fossem compostos por 56 dias e os de testes de 10 dias, treinou-se os 4 modelos seleccionados e usou-se cada um dos modelos para fazer uma previsão comparando-a com os dados do conjunto de teste. O último passo seria calcular as métricas de R^2 , R^2 adjusted, MAE, MSE e RMSE para cada modelo em cada uma das 3 previsões.

Tabela 2.4: Medidas de performance dos diferentes modelos na previsão de novos casos (tabela do topo), na previsão de recuperados (tabela do meio) e na previsão de mortes (tabela do fundo). Extraído de (Rustam et al. 2020).

Models performance in confirm cases forecasting					
Model	R^2	R^2 Adjusted	MSE	MAE	RMSE
LR	0.83	0.79	1472986504.96	30279.55	38390.51
LASSO	0.98	0.97	234489560.99	11693.97	15422.11
SVM	0.59	0.47	5760890969.30	60177.90	75911.28
ES	0.98	0.97	283201302.2	8867.43	16828.58
Models performance in recovery rate forecasting					
Model	R^2	R^2 Adjusted	MSE	MAE	RMSE
LR	0.39	0.21	480922814.51	17016.08	21929.95
LASSO	0.29	0.08	1462144344.82	30705.27	38237.99
SVM	0.24	0.02	13121148615.72	106739.82	114547.58
ES	0.99	0.99	5970634.07	1827.85	2443.48
Models performance in death rate forecasting					
Model	R^2	R^2 Adjusted	MSE	MAE	RMSE
LR	0.96	0.95	840240.11	723.11	916.64
LASSO	0.85	0.81	3244066.79	1430.29	1801.12
SVM	0.53	0.39	16016210.98	3129.74	4002.02
ES	0.98	0.97	662228.72	406.08	813.77

A tabela 2.4 representa os desempenhos dos modelos nas 3 previsões que cada um executou, através desta tabela podemos concluir as seguintes performances de acordo com o seguinte tipo de previsão:

- Novos casos - O modelo de LASSO e ES obtiveram as melhores performances, sendo elas bastante semelhantes, onde o modelo de ES se destaca um pouco mais por ter um valor menor de RMSE. O modelo SVM foi o pior modelo obtendo R^2 baixos e valores nas métricas de erros muito elevados comparando com os restantes modelos. Já modelo de regressão linear teve uma performance razoável.
- Recuperados - O ES obteve a melhor performance obtendo 99% de R^2 e valores de erro baixos. Enquanto os outros modelos tiveram performances muito fracas, destacando o modelo de SVM que obteve a pior performance de todos, com um R^2 adjusted de 0.02 e métricas de erro muito elevadas.
- Mortes - O modelo que obteve a melhor performance foi, mais uma vez o ES, tendo o mais elevado R^2 e sendo o modelo com valores de erros (MAE, MSE e RMSE) menores. O SVM foi o que obteve a pior performance, tendo o R^2 mais baixo.

De forma geral o modelo de suavização exponencial (ES) foi o melhor nas 3 diferentes previsões, enquanto o modelo SVM foi sempre o pior, já os modelos de LASSO e LR obtiveram resultados razoáveis tendo o modelo de LASSO obtido uma boa performance na previsão de casos confirmados.

Por fim os modelos foram testados, usando diferentes intervalos de treino e teste para avaliar a performance dos modelos com menores quantidades de dados, para prever os próximos 10 dias de recuperações, obtendo a tabela 2.5.

Tabela 2.5: Performances dos modelos com diferentes intervalos como conjunto de treino. Extraído de (Rustam et al. 2020).

Interval	Dataset Size(days)	Last day of dataset	LASSO Performance	LR Performance	SVM Performance	ES Performance
1	26	12/02/2020	Very poor	Very poor	Very poor	Best
2	41	02/03/2020	Very poor	Very poor	Very poor	Best
3	56	17/03/2020	Poor	Good	Very poor	Best
4	66	27/03/2020	Better	Best	Well Improved	Best

Através da tabela 2.5, podemos observar que com um conjunto de dados de treino com um intervalo de 26 dias, o modelo de ES tem um desempenho muito boa enquanto os restantes modelos têm desempenhos muito fracos. No segundo intervalo, com 41 dias de conjunto de treino, o modelo de ES continua a ter um desempenho muito bom enquanto os restantes têm fracos desempenhos. No terceiro intervalo, com 56 dias de conjunto de treino, o modelo de ES continua a ter um desempenho muito bom enquanto o modelo de LASSO e LR tiveram uma ligeira melhoria. Por fim, no último intervalo verificou-se que os 3 modelos que no primeiro intervalo tiveram fracas performances melhoraram bastante enquanto o modelo ES manteve sempre uma boa performance.

Com isto podemos concluir que o modelo ES tem um bom desempenho com conjuntos de dados mais pequenos, enquanto os outros modelos necessitam de uma quantidade de dados maior para obter melhores resultados.

2.3.3 Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches

Após ser verificada uma forte relação entre as condições meteorológicas, como a temperatura e humidade e a propagação do vírus (Sajadi et al. 2020), este projeto pretende comparar a performance de diferentes modelos de Machine Learning, assim como o projeto 2.3.1, bem como avaliar a importância das diferentes variáveis nas suas previsões, com grande foco em avaliar a importância das condições meteorológicas.

No projeto usou-se uma ferramenta que recolhe e limpa os dados, tendo sido recolhida informação de várias fontes como de relatórios oficiais sobre o COVID-19 e condições meteorológicas do website Meteoblue, criando um conjunto de dados de vários países e que abrange um intervalo de 12 de dezembro a 22 abril, tendo como variáveis as seguintes:

- Data: cada dia linha representa um dia;
- Geográficas: País e Região;
- COVID-19: Casos Confirmados, Mortes, Recuperados e Ativos;
- Demográficas: População, Idade média, Fertilidade, Número de Camas nas Unidades de Cuidados Intensivos por 1000 pessoas e Percentagem de população em áreas urbanas;

- Meteorológicas: Humidade, Precipitação, Horas de Sol, Velocidade do Vento e Média temperatura;

Antes de executar as previsões foi usado o modelo OLS (Ordinary Least Square- Mínimos Quadrados Ordinários) para analisar a relação entre as variáveis meteorológicas (temperatura e humidade) com o número de novos casos e a sua taxa de crescimento á data de 17 de maio, onde o número de novos casos foi transformado em log para seguir uma distribuição de maior facilidade de análise. Este método recorre da seguinte expressão algébrica:

$$y = \alpha(Temperatura - 15) + \beta(Humidade - 75) + erro \quad (2.11)$$

Tabela 2.6: Relação entre Humidade e Temperatura com o número de infectados e taxa de crescimento. Extraído de (Malki et al. 2020).

Model(1/2)	Temp.	Humidity	Alpha	Beta	Temp. Effect	Humidity Effect
Infected cases	15	70	0.861434	0.087399	136.656	9.133222
Growth Rate	15	70	0.347038	0.056481	41.4870	5.810663
Infected cases	15	75	0.890908	0.078644	143.734	8.181937
Growth Rate	15	75	0.374966	0.057369	45.4952	5.904673
Infected cases	15	80	0.90604	0.066578	147.450	6.884406
Growth Rate	15	80	0.401863	0.057235	49.4607	5.890449
Infected cases	20	70	1.450845	0.120595	326.672	12.81685
Growth Rate	20	70	0.630435	0.077356	87.8428	6.042613
Infected cases	20	75	1.398349	0.069157	304.851	7.160394
Growth Rate	20	75	0.655225	0.065638	92.5575	6.783971
Infected cases	20	80	1.220575	0.003514	238.914	0.352016
Growth Rate	20	80	0.637707	0.046733	89.2137	4.784182
Infected cases	25	70	-1.191378	-0.223587	-69.620	-20.03546
Growth Rate	25	70	-0.232620	-0.046521	-20.755	-4.545522
Infected cases	25	75	-1.290774	-0.276363	-72.4943	-24.14623
Growth Rate	25	75	-0.307090	-0.075870	-26.4415	-7.306374
Infected cases	25	80	-1.225998	-0.277625	-70.6535	-24.24193
Growth Rate	25	80	-0.310293	-0.084043	-26.6768	-8.060824
Infected cases	30	70	-1.169891	-0.136057	-68.9600	-12.72069
Growth Rate	30	70	-0.402274	-0.032372	-33.1203	-3.185389
Infected cases	30	75	-1.077191	-0.137452	-65.9450	-12.84234
Growth Rate	30	75	-0.377918	-0.037119	-31.4713	-3.643827
Infected cases	30	80	-0.987054	-0.135130	-62.7327	-12.63975
Growth Rate	30	80	-0.349860	-0.039597	-29.5213	-3.882344
Infected cases	35	70	-0.719808	-0.075775	-51.3154	-7.297566
Growth Rate	35	70	-0.258704	-0.011015	-22.7945	-1.095416
Infected cases	35	75	-0.686391	-0.076581	-49.6611	-7.372178
Growth Rate	35	75	-0.252149	-0.013923	-22.2871	-1.382684
Infected cases	35	80	-0.653181	-0.076517	-47.9612	-7.366293
Growth Rate	35	80	-0.258704	-0.016176	-21.6316	-1.604607

A tabela 2.6 representa os resultados obtidos através do OLS para o número de casos de infectados e para a taxa de crescimento para as diferentes gamas de temperatura (15,20,25,30,25) e as gamas de humidade (70,75,80), onde através dos dados do conjunto de dados (numero de casos, taxa de crescimento, humidade e temperatura) se chegou aos valor de alpha e beta, desta forma sendo possível analisar o efeito que a temperatura e humidade provocam na variável y (numero de casos, taxa de crescimento).

Observando com atenção a tabela 2.6 podemos observar que o efeito da temperatura é sempre maior no número de infectados do que na taxa de infeção e que temperaturas mais altas retratam efeitos negativos e temperaturas mais baixas efeitos positivos, ou seja a temperatura é inversamente proporcional á taxa de crescimento e número de infectados, disto é possível concluir que com temperaturas mais altas o números de casos é menor e com temperaturas mais baixas é maior o número de casos de infeção. No caso da humidade é possível ver que o seu efeito numa mesma temperatura, por exemplo 15, é possível verificar que o seu efeito aumenta com o aumento da humidade, ou seja, seja a humidade é diretamente proporcional ao número de casos e taxa de crescimento.

Como a temperatura e humidade não são suficientes para executar previsões, foram construídos modelos de Machine Learning com todas as variáveis, recorrendo a alguns dos modelos mais usados descritos em 2.2.3, para prever o número de casos e de mortes. Um dos modelos usados é o algoritmo de Random Forest, modelo que tem por base o algoritmo de árvores de decisão, sendo constituído por várias árvores de decisão foi usado por se prever que fosse ter uma boa performance, mas também por possuir uma função que permite verificar quais foram as variáveis mais importantes para executar a previsão.

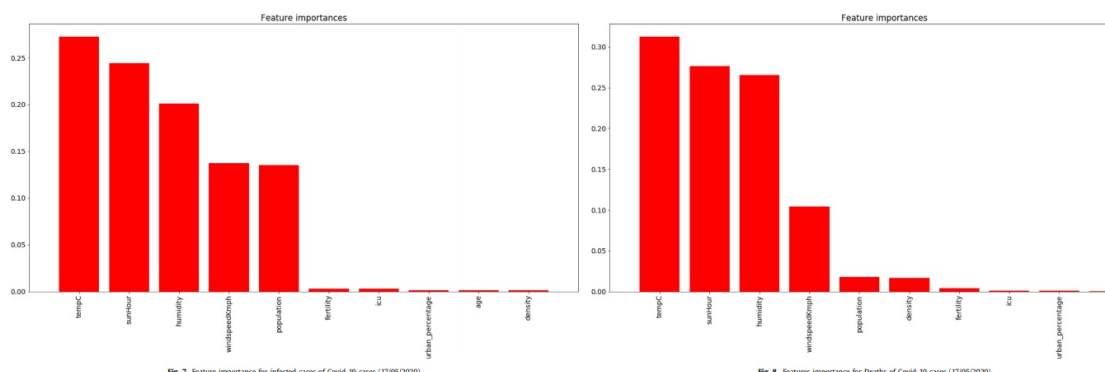


Figura 2.10: Importância das variáveis para a execução da previsão. Extraído de Malki et al. 2020.

Na figura 2.10 podemos observar a importância das variáveis para ambas as previsões, número de casos e número de mortes. É visível que as 4 variáveis meteorológicas (temperatura, humidade, horas de sol e velocidade do vento) são as mais importantes em ambos os casos e enquanto no número de casos a quinta variável mais importante, população, possui uma importância muito próxima da velocidade do vento, no caso das previsões de número de mortes estas duas variáveis possuem uma importância mais baixa.

Recorrendo às métricas de avaliação e comparação RMSE e RMSLE (Root Mean Squared Log Error) verificou-se que os melhores modelos para ambas as previsões foram os de KNN, Decision Trees e Random Forest.

2.3.4 Análise dos projetos

Com o projeto feito com dados do Senegal, percebemos que o modelo de regressão linear e SVM não são os melhores modelos para este problema já que tiveram desempenhos pobres e que modelos que tenham por base previsões temporais como as redes neuronais com recurso a nnfor e o modelo Prophet têm melhor performance. No geral verificou-se a melhor métrica para analisar a performance dos modelos e para realizar comparações é a RMSE.

Com o projeto descrito na secção 2.3.2, podemos concluir que existem modelos que têm boas performances mesmo na presença de uma quantidade de dados reduzida, como o caso do modelo ES, enquanto outros necessitam de um conjunto de dados maior para terem performances, que com um conjunto de dados menor não conseguiam atingir. Verificou-se mais uma vez que o modelo SVM teve fracas performances. Para avaliar os desempenhos dos modelos usou-se R^2 , R^2 adjusted, MAE, MSE e RMSE.

Analisando o projeto 2.3.3 é possível concluir que os dados meteorológicos influenciam a propagação do vírus e que são dados bastante relevantes para a execução de previsões. Uma vez mais as métricas usadas para comparar modelos foram o RMSE e RMSLE, concluindo que os melhores modelos foram KNN, Decision Trees e Random Forest.

Em suma:

- **Melhores modelos:** Previsão de série temporal com redes neuronais (NNfor), modelo Prophet, Suavização Exponencial (ES), K Vizinhos mais próximos (KNN) e árvores de decisão (DT);
- **Melhores métricas:** R^2 , R^2 adjusted, MAE, MSE e RMSE;
- **Variáveis mais relevantes:** Temperatura, humidade, velocidade do vento, densidade populacional, população, grau de urbanização, média de idade e tempo de luz do dia;

2.4 Soluções Existentes

2.4.1 CoronaNet

Os governos de todo o mundo, como foi dito em 1.2, foram implementando medidas para conter o vírus, existindo desta forma um elevado número de leis e políticas que estiveram, estão e/ou estarão em vigor. Devido ao fato destas medidas serem realizadas em vários países diferentes, que usam línguas diferentes e graças ao tempo de vida que estas leis muitas vezes têm, isto é, ficam em vigor por um espaço determinado de tempo podendo posteriormente serem alargados, é difícil ter acesso a dados rigorosos referentes a todas estas medidas (Cheng et al. 2020).

Para combater esta dificuldade de ter todas as medidas implementadas, no mundo para impedir o coronavírus de se espalhar, num só lugar entra o CoronaNet (*CoronaNet* 2020). O CoronaNet é um projeto que agrega numa base de dados as respostas dos governos no combate ao COVID-19, tendo como principal objetivo a recolha de dados sobre a maioria dos países, tendo atualmente 40.000 eventos de cerca de 198 países(Cheng et al. 2020; *CoronaNet* 2020).

O CoronaNet possui um repositório público no github onde, entre outros recursos, possui vários ficheiros CSV onde estão armazenados os dados das medidas políticas, estando

estes dados todos num único ficheiro CSV ou possuindo cada país um ficheiro próprio (*corona_tscs/data/CoronaNet at master · saudiwin/corona_tscs · GitHub* 2020).

policy_id	entry_type	date_start	date_end	country	ISO_A2	init_country_level	province	city	type
5945175	new_entry	2020-03-18	NA	Portugal	PT	Municipal	Aveiro	Ovar	Declaration of Emergency
8533922	new_entry	2020-03-18	2020-05-02	Portugal	PT	National	NA	NA	Declaration of Emergency
8533922	update	2020-03-18	2020-05-02	Portugal	PT	National	NA	NA	Declaration of Emergency
8533922	update	2020-04-02	2020-05-02	Portugal	PT	National	NA	NA	Declaration of Emergency
538924	new_entry	2020-03-13	NA	Portugal	PT	National	NA	NA	External Border Restrictions
538924	new_entry	2020-03-13	NA	Portugal	PT	National	NA	NA	External Border Restrictions
6369772	new_entry	2020-03-13	NA	Portugal	PT	Provincial	Madeira	NA	External Border Restrictions
6369772	new_entry	2020-03-13	NA	Portugal	PT	Provincial	Madeira	NA	External Border Restrictions
8336844	new_entry	2020-03-15	2020-05-14	Portugal	PT	National	NA	NA	External Border Restrictions

Figura 2.11: Exemplos de algumas colunas dos ficheiros CSV presentes no CoronaNet referentes a medidas implementadas em Portugal.

A lista que se segue procura identificar o tipo de dados presentes nos ficheiros *corona_tscs/data/CoronaNet at master · saudiwin/corona_tscs · GitHub* 2020:

- O tipo de medida implementada (por exemplo quarentena ou o fecho de escolas);
- O nível governamental que iniciou a medida (por exemplo nacional (Primeiro Ministro) ou regional (Presidente da Câmara);
- O nível geográfico a que a medida é aplicável, caso seja (por exemplo nacional ou municipal);
- Casos seja aplicável, que alvos materiais ou humanos estão abrangidos pela medida (por exemplo turistas, emigrantes ou máscaras);
- Casos seja aplicável, que meios de transporte estão abrangidos pela medida (por exemplo aviões, comboios e carros);
- A força da medida (obrigatório ou voluntário);
- Entidade que vai fiscalizar o cumprimento da medida (por exemplo polícia ou exército);
- O timing da medida (data de anúncio, data de início e data de fim da atividade da medida);

Este repositório possui também ficheiros com dados da evolução do covid-19.

2.4.2 FiveThirtyEight- Covid Forecast

Neste website (*Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree | FiveThirtyEight* 2020), estão reunidos 13 modelos desenvolvidos em diferentes universidades dos Estados Unidos da América para a previsão do número de mortes. Numa primeira parte do website estão presentes os gráficos que preveem o número de mortes num espaço de um mês de cada modelo, a figura 2.12 mostra as previsões dos modelos construídos na Universidade de Colúmbia e na Northeastern.

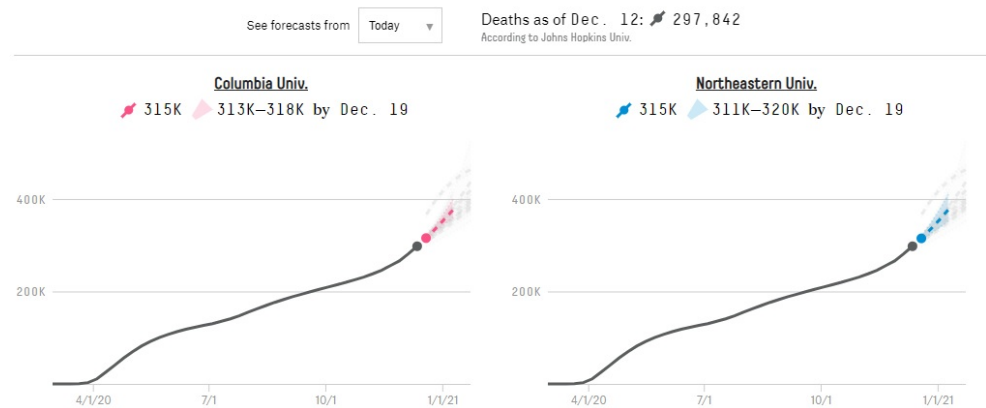
FiveThirtyEight

Figura 2.12: Exemplos das previsões do número de mortos nos Estados Unidos da América. Captura de ecrã do site (*Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree* | FiveThirtyEight 2020).

Como é visível na figura 2.13 no topo da imagem existem um local que permite alterar a previsão para uma mais antiga para comparar a evolução real do número de mortos com o valor projetado pelos modelos.

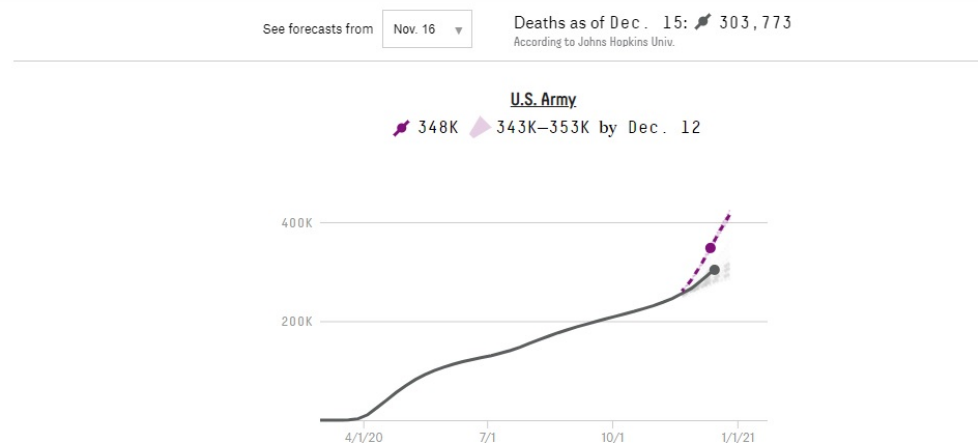
FiveThirtyEight

Figura 2.13: Exemplo da previsão do número de mortos nos Estados Unidos da América no dia 16 de novembro. Captura de ecrã do site (*Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree* | FiveThirtyEight 2020).

Numa zona mais abaixo do website, estas previsões são separadas da forma a que cada gráfico represente um estado e contenha todas as previsões dos diferentes modelos, mais uma vez o site permite alterar a data da previsão assim como salientar um determinado modelo nos gráficos. Um exemplo disso é a figura 2.14 que possui a previsão do número de mortos nos estados Alabama, Alaska e Arizona dia 28 de setembro tendo o modelo desenvolvido pelo exército americano destacado.

FiveThirtyEight

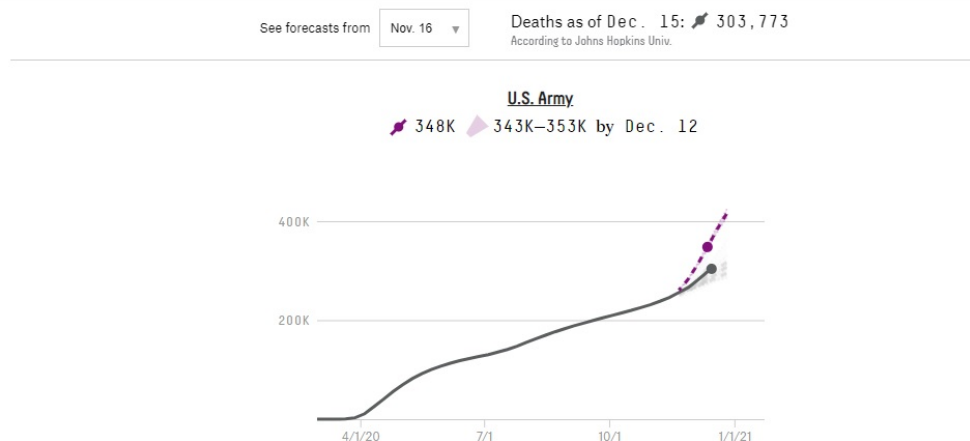


Figura 2.14: Exemplo da previsão do número de mortos nos estados Alabama, Alaska e Arizona no dia 28 de setembro tendo o modelo desenvolvido pelo exército americano destacado. Captura de ecrã do site (*Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree* | FiveThirtyEight 2020).

2.4.3 Institute for Health Metrics and Evaluation - Covid Projections

Esta solução foi desenvolvida pelo Instituto de Métricas e Avaliação em Saúde (Institute for Health Metrics and Evaluation) que consiste num centro de pesquisa independente, fazendo parte da Universidade de Washington que procura medir os problemas mais importantes na área da saúde mundial, assim como avaliar estratégias (IHME - COVID-19 Projections 2021).

Nesta solução que faz várias previsões para o covid-19, como por exemplos o número de mortes, possuem os seguintes cenários:

- Projeção - modelo que executa as previsões tendo em conta uma distribuição de vacinas em 3 meses;
- Uso universal de máscara - modelo que executa as previsões tendo em conta uma distribuição de vacinas em 3 meses e o uso universal de máscara;
- Rápida Distribuição de Vacinas - modelo que executa as previsões tendo em conta uma distribuição de vacinas em 45 dias;
- Rápida distribuição de Vacinas para os grupos de maior risco - modelo que executa as previsões tendo em conta uma distribuição de vacinas em 3 meses, em que se vacina em primeiro lugar os grupos de risco, em seguida trabalhadores essenciais (profissionais de saúde, entre outros) e por fim todos os restantes adultos;
- Relaxamento - modelo executa previsões tendo em conta o caso de os governos se demitirem face a um aumento do número de casos;
- Sem vacina - o modelo executa previsões tendo em conta o facto de não haver distribuição de vacinas;

O sistema do IHME permite observar estas as previsões em vários países, assim como, em alguns casos, observar estas previsões em termos de regiões, função essa que não está

disponível para Portugal. Como podemos observar na figura 2.15, onde temos a previsão do número de mortes de COVID-19 no estado do Rio de Janeiro do Brasil, em que todos os cenários estão selecionados, já que o sistema permite selecionar os que desejarmos. Podemos também observar que o modelo que prevê o menor número de mortes é o uso universal de máscara, o que comprova a importância da mesma.

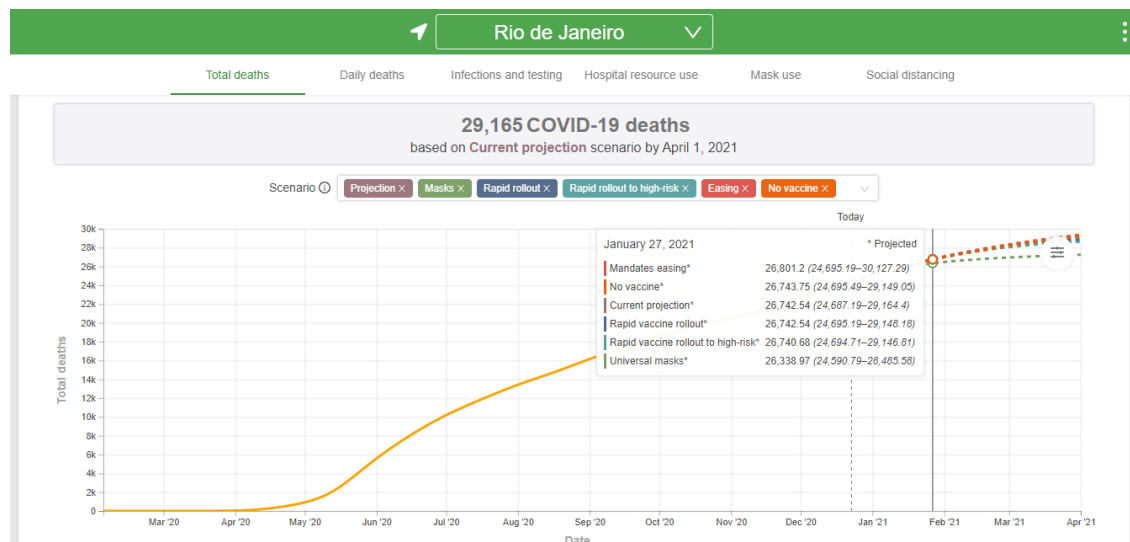


Figura 2.15: Exemplo da previsão do número de mortos no Rio de Janeiro. Captura de ecrã do site (IHME - COVID-19 Projections 2021).

Outra função do sistema criado pelo IHME, é a comparação, esta função permite comparar a previsão de um país ou estado com vários outros, mas, ao contrário da previsão normal, apenas permite selecionar um cenário de previsão, no caso da figura 2.16 o cenário escolhido foi o da rápida distribuição da vacina e estão a ser comparados os estados do Rio de Janeiro e de São Paulo, ambos do Brasil.



Figura 2.16: Exemplo da comparação de previsões do número de mortos entre Rio de Janeiro e São Paulo. Captura de ecrã do site (IHME - COVID-19 Projections 2021).

Para além da previsão do número de mortes, o sistema permite outras previsões, usando os mesmos cenários anteriormente descritos, tais como:

- Mortes diárias - no lugar de se prever o número total de mortes, um número que está sempre a crescer, prevê-se o número diário de mortes, número este que tem tendência a baixar;
- Infecções Diárias - onde se prevê o número de novos casos por dia;
- Testes Diários - onde se prevê o número de testes ao COVID-19 executados por dia;
- Recursos Hospitalares - onde se prevê o número de camas e ventiladores usados, tendo uma linha horizontal que indica o número total de camas e ventiladores que o país ou estado possui;
- Uso de máscara - prevê a percentagem de população que utiliza máscara em publico;
- Distanciamento Social - prevê o distanciamento social, tendo por base dados de mobilidade de dispositivos móveis;

2.5 Tecnologia/Ferramentas Úteis

Nesta seção irão ser estudadas algumas técnicas e ferramentas que poderão ser úteis no desenvolvimento do projeto.

2.5.1 Shiny

Shiny é um pacote que permite a criação de aplicações web através de linguagem R. Este pacote permite integrar temas CSS, ações de JavaScript, bem como componentes de HTML (Beeley 2013).

O *Shiny* possui extensões, como por exemplo *Shiny Server* e *Shinyapps.io*, que permitem tornar pública e online uma aplicação web criada em *Shiny* (Beeley 2013).

A figura 2.17 representa a utilização de *Shiny* para criar uma ferramenta interativa, na parte direita da figura é apresentado o código que irá criar a interface web presente á esquerda com linguagem R.

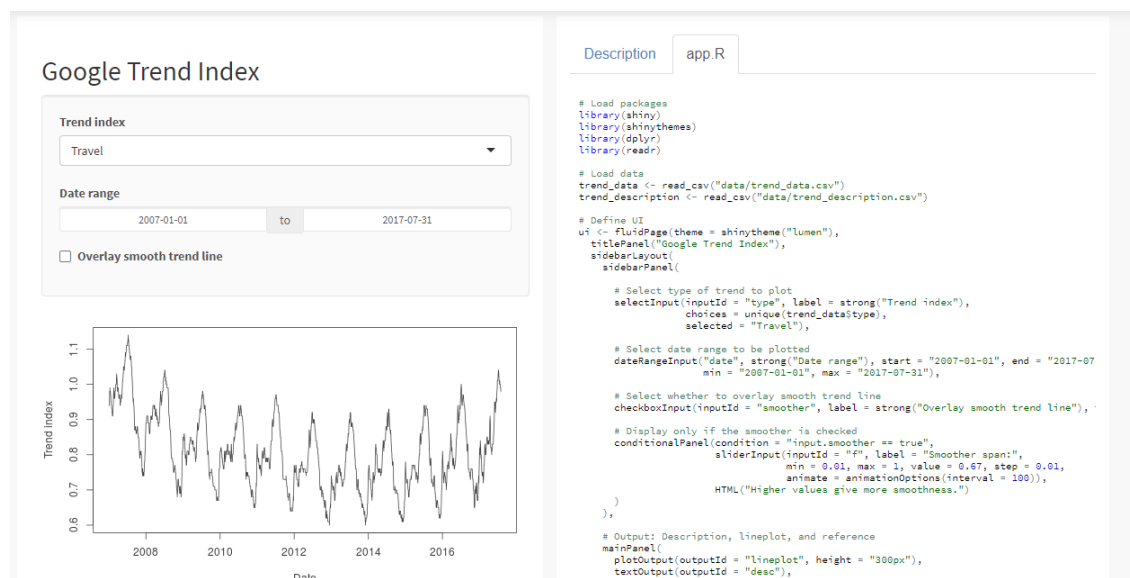


Figura 2.17: Exemplo de utilização da biblioteca Shiny para criar um formulário.

2.5.2 Web Scraping

Web Scraping pode ser definido como o processo de extrair dados da Web, este agente de software irá imitar o comportamento humano, pesquisando os dados, extraíndo-os e organizando-os da forma que se pretende (Glez-Peña et al. 2014).

Estes agentes software são chamados de *Web Scrapers* e pode ser livrarias de determinadas linguagens que permitem construir o algoritmo que irá extrair os dados com por exemplo a livreria "rvest" de linguagem R ou podem ser software que permitem uma interação sem código, como por exemplo "IrobotSoft" (Glez-Peña et al. 2014).

Estes *WebScrapers* têm as seguintes tarefas básicas:

- Acesso ao Site: através do protocolo HTTP o *Web Scrapers*, estabelece uma ligação entre o cliente e o servidor. O *WebScrapers* irá usar métodos GET, que permite pedir ao site acesso a dados, tendo cautela para não executar demasiados pedidos ao servidor do site. Normalmente os sites possuem um ficheiro "robots.txt" que indica quais os recursos que não devem ser acedidos por este tipo de software;
- Extração dos dados: através do código HTML obtido do website, o *Web Scraper* irá extrair os dados que lhe são relevantes;
- Construção do resultado: esta etapa consiste em transformar os conteúdos extraídos em uma representação estruturada mais adequada ao problema, como por exemplo uma folha de cálculo;

```

library(rvest)
lego_movie <- html("http://www.imdb.com/title/tt1490017/")
lego_movie %>%
  html_nodes("#titleCast .itemprop span") %>%
  html_text()
#> [1] "Will Arnett"      "Elizabeth Banks" "Craig Berry"
#> [4] "Alison Brie"      "David Burrows"   "Anthony Daniels"
#> [7] "Charlie Day"      "Amanda Farinos"  "Keith Ferguson"
#> [10] "Will Ferrell"     "Will Forte"      "Dave Franco"
#> [13] "Morgan Freeman"  "Todd Hansen"     "Jonah Hill"

```

Figura 2.18: Exemplo de uso da biblioteca rvest para extrair dados. Extraído de (rvest: easy web scraping with R 2021).

A figura 2.18 representa um exemplo de *webscraping* através da livraria rvest, onde os caracteres pretos representam o código que permite extração de dados e os caracteres cinzentos são os dados extraído, que no exemplo são os atores que participam no filme "Lego".

Capítulo 3

Análise de Valor

Nesta seção será apresentado o New Concept Development, com o objetivo de identificar o conceito do projeto, serão também apresentadas as propostas de valor e benefícios/sacrifícios que o utilizador terá usando o projeto. Também será usado o modelo de Canvas para representar o plano de negócio, o modelo QFD para será usado para definir os requisitos tendo por base os requisitos do cliente e permitirá também definir uma importância/prioridade para cada requisito e por fim o método de AHP para avaliar alternativas e escolher a melhor.

3.1 New Concept Development

O método New Concept Development (NDC) foi construído com o foco de cultivar ideias e conceitos de forma a pensar no front-end do produto. Na figura 3.1 é possível observar os seus constituintes que são (Koen et al. 2001):

- Identificação de Oportunidades: onde se identifica uma oportunidade de negócio;
- Análise de Oportunidades: analisando a oportunidade identificada, com informação adicional de forma a perceber se a oportunidade tem ou não futuro;
- Geração de Ideias: nesta fase deve fazer-se uma lista com todos as ideias de, por exemplo, funcionalidades que poderão estar presentes no produto;
- Seleção de ideias: nesta fase seleciona-se as ideias geradas na fase anterior tendo em conta os objetivos da organização;
- Definição de Conceito: Definir o conceito do produto;
- Fator Externos: fatores do ambiente externo como por exemplo estratégias da organização ou fatores envolvendo a concorrência;

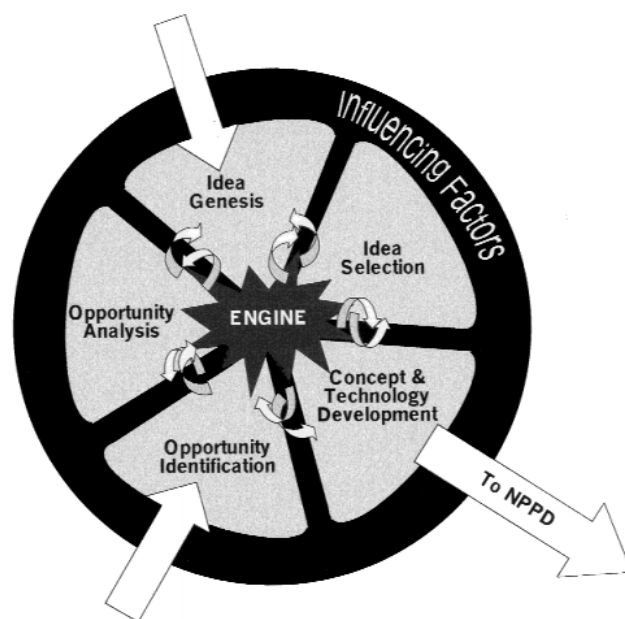


Figura 3.1: Diagrama do Método New Concept Development.

Nesta secção serão identificados os 5 elementos principais do modelo NCD:

3.1.1 Identificação de Oportunidades

Atualmente vivemos um período conturbado devido á pandemia (COVID-19) que afeta todo mundo quer seja socialmente ou economicamente, já que esta doença é altamente infecciosa e por vezes fatal. Existem vários especialistas que fazem várias previsões manualmente através de cálculos matemáticos em relação á evolução futura da doença (COVID-19), previsões essas que por vezes não são do conhecimento publico. Com os avanços das áreas de Data Mining e Machine Learning é possível analisar grandiosas quantidades de dados e com base nesses dados criar previsões, que seriam mais rápidas e precisas do que as realizadas manualmente por humanos.

Concluindo, projetos semelhantes a este já existem hoje em dia, mas nenhum recorre a estas tecnologias para executar previsões em Portugal tendo em conta as diferentes características de cada região do país.

Como atualmente é uma doença que marca os anos de 2019, 2020 e certamente marcará o ano de 2021 é um produto que poderá ter o timing ideal para integrar o mercado e no futuro, este tipo de solução poderia ser usado em outros países e também para prever a evolução de outras doenças.

3.1.2 Análise de Oportunidades

O coronavírus veio criar uma crise financeira em vários países e empresas como por exemplo as companhias aéreas que por causa de medidas, como encerramento de fronteiras e restrições de voos para determinados países, têm cada vez menos clientes. Outro exemplo são as lojas de roupa que em muitos países estão fechadas ou as pessoas simplesmente por estarem em casa não compram. Verificou-se também aspetos positivos em termos financeiros para algumas farmacêuticas que produzem medicamentos ou produtos que ajudam a combater

o vírus, assim como algumas fábricas de máscaras (*Coronavirus: How the pandemic has changed the world economy* 2020). Tal como os produtos hospitalares este projeto poderia ajudar a combater o vírus.

Esta doença afeta todo o mundo e como tal é muito pesquisada pelas pessoas na Internet, seja para saber os seus sintomas, seja para estar a par da sua evolução. De acordo com o Google Trends (*O ano em Pesquisa Google- Google Trends* 2020) as pesquisas mais realizadas em termos de pesquisa geral e de pesquisa por notícias foi "coronavírus" em Portugal e em todo um mundo. Através do Google Trends verificou-se também que as palavras "covid-19 forecast", que significa previsão de covid-19 em português, são uma expressão bastante pesquisada e que o seu interesse por parte das pessoas tem vindo a aumentar, sendo que estas palavras são maioritariamente pesquisadas nos Estados Unidos, Canadá e Reino Unido (*Covid Forecast- Google Trends* 2020). Já em Portugal o tópico de previsões da evolução do covid-19 não muito pesquisado, o que prova o timing do projeto, já que em Portugal não existem soluções conhecidas com boas previsões.

Em termos de concorrência, existem algumas soluções construídas em outros países, mas que não têm previsões para Portugal e as suas regiões, assim sendo nenhuma é diretamente concorrência já que o público-alvo é o povo português.

3.1.3 Geração de Ideias

Ideias:

- Apresentação dos dados atuais da pandemia - apresentar ao utilizador os dados de casos confirmados de covid, casos ativos, recuperados e óbitos, estes dados deverão estar em 2 grandezas: total e diário. Todos estes dados atuais deverão estar disponíveis para Portugal e todas as suas regiões, ou seja, deverá ser apresentada a situação atual de cada região do país.
- Apresentação de previsões de pandemia - apresentar ao utilizador previsões de casos confirmados de covid, casos ativos, recuperados e óbitos, estes dados deverão estar em 2 grandezas: total e diário. Todas estas previsões deverão estar disponíveis para Portugal e todas as suas regiões, ou seja, deverão ser apresentadas as previsões de cada região do país.
- Comparar previsões - permitir ao utilizador comparar as previsões de 2 regiões.
- Previsão vacinação - prever quando é que uma pessoa será vacinada e/ou o número de vacinados diariamente.
- Previsão estados dos Hospitais- apresentar a previsão de número camas ocupadas, camas usadas nos cuidados intensivos e número de ventiladores em uso diariamente em cada hospital do país.

3.1.4 Seleção de Ideias

Das ideias enumeradas em 3.1.3, para esta fase foram selecionadas as 3 primeiras, ou seja, apresentação dos dados atuais da pandemia, apresentação de previsões de pandemia e comparação de previsões. Estas ideias foram as selecionadas tendo em conta o risco tecnológico, os custos de desenvolvimento e o benefício que trazem para os utilizadores. As restantes ideias ficarão como trabalho futuro a realizar neste projeto.

3.1.5 Definição de Conceito

O principal objetivo deste projeto é o cálculo e disponibilização de previsões precisas da evolução do coronavírus em Portugal. O momento em que nos encontramos é propício ao lançamento deste projeto visto que no país este conceito está pouco enraizado e não existir concorrência relevante. A solução implementada pode vir a ser muito importante para a tomada de decisão por parte dos governantes para atenuar a evolução do vírus.

3.2 Proposta de Valor

O projeto pretende proporcionar aos seus utilizadores uma plataforma de referência com previsões atualizadas referentes à evolução da pandemia de covid-19 de forma auxiliar no combate ao vírus e melhorar o dia-a-dia da população.

- **Produto:** Ferramenta web que prevê evolução do covid-19;
- **Público-Alvo:** Governantes (pessoas com poder de decisão) e público geral com interesse em previsões;
- **Benefício principal:** Previsões atualizadas referentes à evolução da pandemia de covid-19;
- **O que o torna único:** Levar em consideração as diferentes características das regiões de Portugal;

3.2.1 Segmentação de Mercado

Para este projeto é possível identificar dois públicos-alvo:

- Para os governantes que necessitam de tomar decisões para minimizar os danos da pandemia de covid-19 este projeto que é uma plataforma web que fornece previsões da evolução da pandemia tendo em conta as diferentes características das regiões de Portugal ao contrário de outras soluções que não têm em conta este fator bastante influenciador ajudando assim os governantes a tomar decisões.
- Para o utilizador geral que necessita analisar a evolução da pandemia de covid-19 este projeto que é uma plataforma web que fornece previsões da evolução da pandemia tendo em conta as diferentes características das regiões de Portugal ao contrário de outras soluções que não têm em conta este fator bastante influenciador, fornecendo assim informação útil ao utilizador geral que com base na previsão poderá também tomar decisões (usar mais a máscara, ter mais cuidados, exigir melhores medidas aos governantes, etc).

3.2.2 Proposta de Valor - Modelo Canvas

Nesta seção iremos usar o modelo de Canvas para a proposta de valor, que se divide em cliente e produto.

Cliente

Neste modelo a área do cliente divide-se em:

- **Ganhos:** O utilizador terá o benefício de encontrar numa só plataforma dados sobre a situação atual da pandemia de coronavírus assim como previsões sobre a evolução da mesma em Portugal e suas regiões;
- **Sacrifícios:** Os sacrifícios que os utilizadores terão de fazer serão muito poucos, estes prendem-se essencialmente em terem que suportar publicidade;
- **Trabalho a fazer:** Decidir que medidas implementar;

Produto

Neste modelo a área do produto divide-se em:

- **Produto e serviços:** Informações sobre a evolução de covid-19 em Portugal e suas regiões, assim como previsões tendo em conta as diferentes características das regiões;
- **Aliviador de dores:** Informação sempre disponível e de forma gratuita;
- **Criador de ganhos:** Previsões de qualidade e segurança;

3.2.3 Modelo Canvas

De seguida encontra-se o modelo de negócio CANVAS que apresenta uma visão global da estrutura do negócio deste projeto, que se compromete a fornecer informação atualizada, sobre a evolução do covid-19 em Portugal, bem como se compromete a fornecer previsões dessa mesma evolução.

Em termos financeiros este projeto terá como principais gastos os recursos humanos, hardware e possivelmente algumas licenças de software. Como se pretende que o produto seja um website gratuito e disponível para todas as pessoas a principal fonte de rendimento será a publicidade, que estará presente no site de forma a ser possível sustentar o mesmo. Salientar também que se pretende estar em contato com o cliente via redes sociais, e-mail e um serviço de suporte no caso de ser necessário o cliente esclarecer alguma dúvida ou reportar uma anomalia.

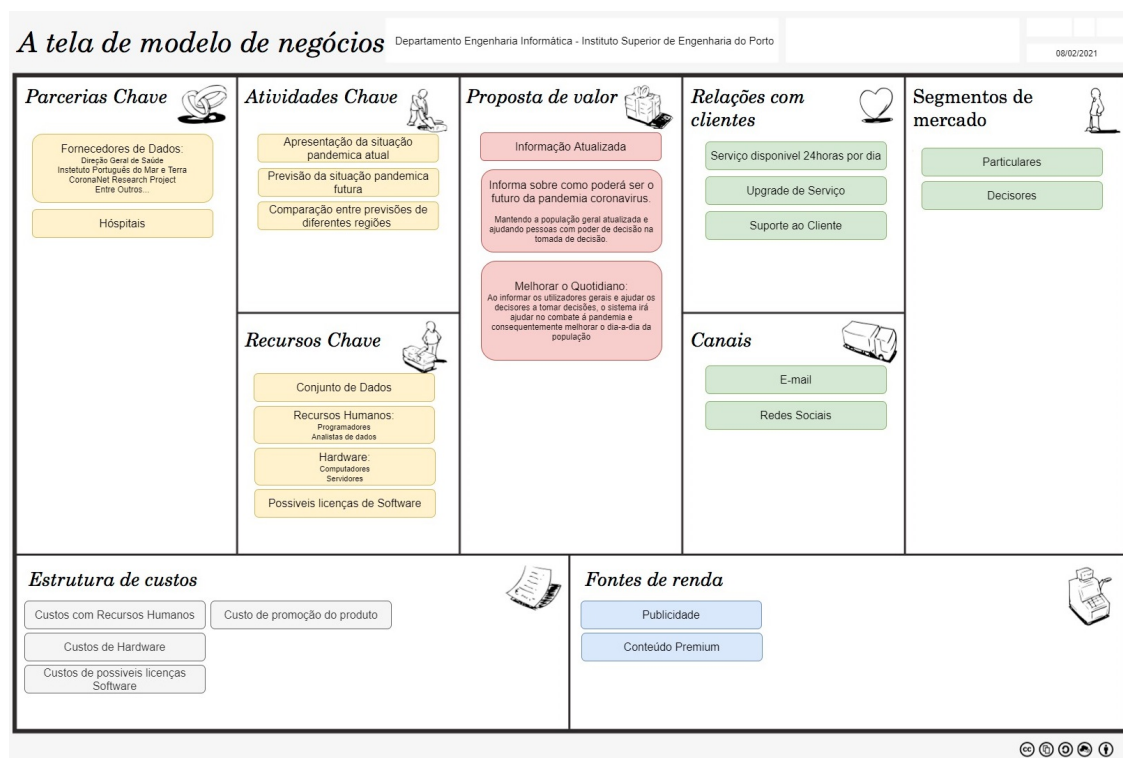


Figura 3.2: Modelo de Negócio Canvas do projeto.

3.3 QFD

O QFD ("Quality Function Deployment") é um método que procura desenhar um produto de qualidade, tendo por base os requerimentos dos clientes, ou seja, o QFD tem 3 objetivos, primeiro desenhar produto com elevada qualidade de forma a chegar ao mercado rapidamente e com baixos custos, segundo produzir um produto derivado do cliente e por fim servir de objeto de monitorização para futuras melhorias no produto (Erdil e Arani 2019).

Este método utiliza o diagrama designado de "House of Quality", em português "Casa de Qualidade", como o exemplo da figura 3.3. Este diagrama é composto por (Erdil e Arani 2019):

- Requisitos do Cliente: consiste numa lista de requisitos que o cliente considera como importantes para o produto;
- Importância dos requisitos para o Cliente: classificação dos requisitos do cliente por ordem de importância para o mesmo;
- Especificação do produto (Requisitos): requisitos técnicos que procuram resolver os requisitos do cliente;
- Matriz de Relação: dentro da matriz central estão os símbolos que mostram a força da relação entre os requisitos técnicos e do cliente, esta relação pode ter várias escalas como por exemplo: muito positiva, positiva, negativa, muito negativa;
- Matriz de Correlação: Mostra a relação entre os requisitos técnicos, usando os mesmos símbolos que na matriz relação;

- Avaliação Técnica: contam a importância de cada coluna (ou seja, de cada requisito técnico), algumas considerações de desenho e uma comparação entre produtos concorrentes em relação aos requisitos técnicos;
- Análise da Concorrência: uma comparação entre produtos concorrentes em relação aos requisitos do cliente;

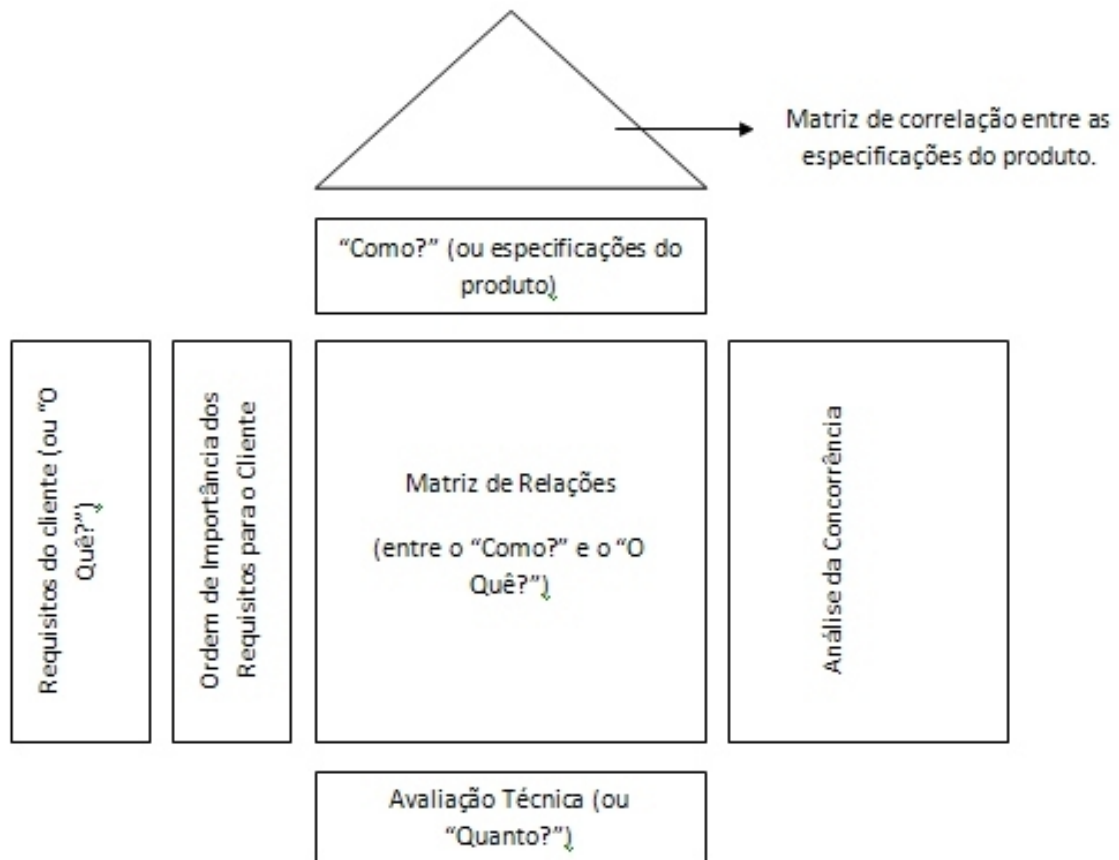


Figura 3.3: Representação da Casa de Qualidade.

3.3.1 Requisitos do Cliente

Os requisitos do cliente identificados são os seguintes:

- Interface gráfica agradável e de fácil utilização - já que será uma ferramenta usado por todas as faixas etárias, com diferentes graus de experiência com tecnologia;
- Design intuitivo - de forma que seja fácil para os utilizadores identificar onde estão e para onde querem ir;
- Utilização em vários tipos de dispositivos - já que se pretende que seja possível aceder tanto em computadores como em dispositivos móveis;
- Previsões nas diferentes regiões do país - o grande foco do projeto, permitir prever a evolução do covid-19 nas regiões de Portugal;

- Previsões de qualidade - prever a evolução do covid-19 nas regiões de Portugal, de forma que as previsões não divirjam muito da realidade;
- Permitir expansão para outras doenças e/ou países - com visto no futuro em que a ferramenta poderá ser usada para prever a evolução noutros países e/ou a evolução de outras doenças;

3.3.2 Requisitos Técnicos

Os requisitos técnicos identificados são os seguintes:

- Desenho Intuitivo - de forma a ser uma experiência de uso agradável e de fácil uso;
- Acesso online - de forma a permitir o acesso em todos os dispositivos;
- Sistema Responsivos - de forma a interagir com o utilizador e a adaptar-se aos diferentes tamanhos de ecrãs;
- Modelos de Machine Learning - de forma a serem capazes de executar as previsões;
- Base de Dados - para armazenar todos os dados necessários para executar as previsões e as próprias previsões;

3.3.3 Diagrama

A figura 3.4 representa o QFD realizado para este projeto, através dele é possível concluir que os requisitos mais importantes e aos quais se deverá dar prioridade são os requisitos dos modelos de Machine Learning e da base de dados, sendo o menos importante o acesso online da aplicação.

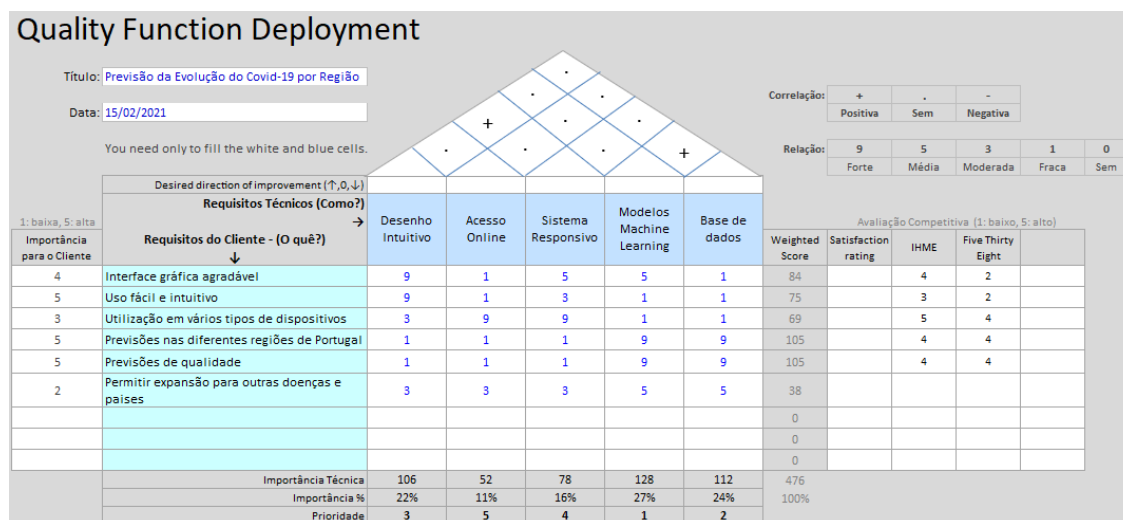


Figura 3.4: Quality Function Deployment do projeto.

3.4 AHP

O modelo AHP ("Analytic Hierarchy Process") é um modelo de decisão multicritério, ou seja, permite avaliar diferentes alternativas tecnológicas de acordo com diferentes critérios.

A primeira fase do método AHP é a construção da árvore de decisão, como o exemplo da figura 3.5, onde se define o objetivo da análise, os critérios que serão usados para avaliar as alternativas e as próprias alternativas (Filho, Atamanczuk e Marçal 2010).

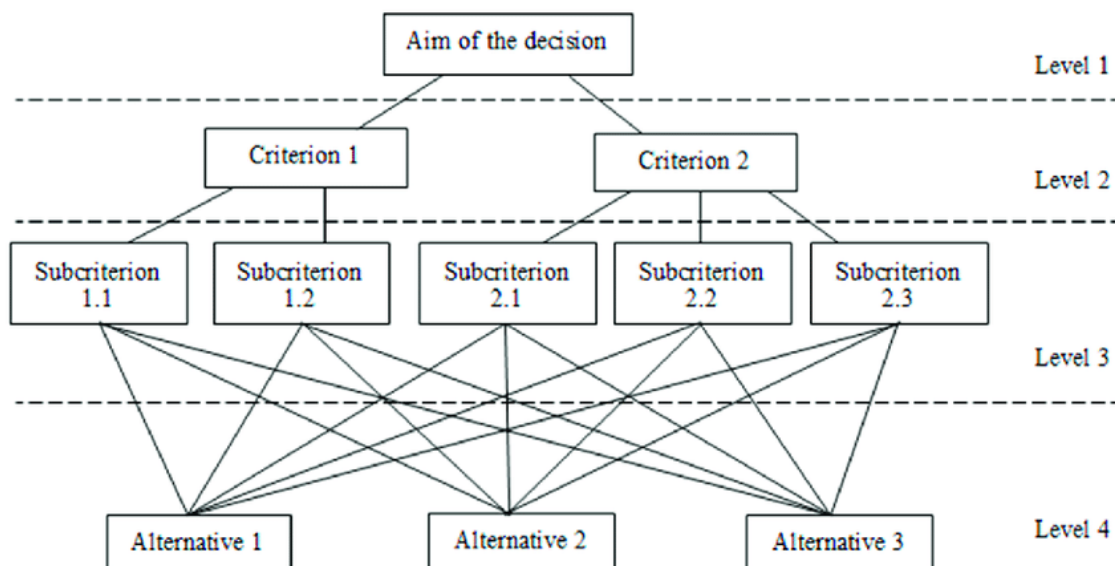


Figura 3.5: Árvore de Hierarquia de Decisão.

A segunda fase prende-se em definir e estabelecer prioridades entre as hierarquias através da escala fundamental presente na figura 3.6, esta avaliação irá gerar uma matriz comparação (Filho, Atamanczuk e Marçal 2010).

Nível de importância	Definição	Explicação
1	Igual importância	As duas atividades contribuem igualmente para o objetivo
3	Fraca importância	A experiência e o julgamento favorecem levemente uma atividade em relação à outra
5	Forte importância	A experiência e o julgamento favorecem fortemente uma atividade em relação à outra
7	Muito forte importância	Uma atividade é muito fortemente favorecida em relação a outra
9	Importância absoluta	A evidência favorece uma atividade em relação a outra com o mais alto grau de certeza
2,4,6,8	Valores intermediários	Quando se procura uma condição de compromisso entre duas definições

Figura 3.6: Escala de Comparações.

Após obter a matriz de comparação, esta matriz deverá ser normalizada, ou seja, igualar todas as prioridades à mesma unidade. Após a normalização da matriz, ter-se-á de identificar a prioridade relativa de cada critério, este cálculo é feito através da média de cada linha da matriz comparação.

A quarta fase é calcular a razão de consistência (RC), para avaliar se as avaliações foram consistentes, para tal deve-se executar a seguinte equação:

$$RC = IC/IR \quad (3.1)$$

Para obter o Índice de consistência (IC) deve usar-se a seguinte equação, onde n é o número de critérios usados e λ_{max} consiste na média da divisão da matriz comparação com os respetivos pesos.

$$IC = \frac{\lambda_{max} - n}{n - 1} \quad (3.2)$$

Para obter o Índice de Aleatoriedade (RI), recorre-se á tabela 3.1 para identificar este valor, sendo n o número de critérios usados.

Tabela 3.1: Tabela Índice de Aleatoriedade

N	1	2	3	4	5	6	7	8	9	10
RI	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

Após obter o valor de RC se este for menor que 0.10 podemos concluir que os valores das prioridades usados são consistentes, caso o RC seja maior que 0,10 tal conclusão não se verifica (Filho, Atamanczuk e Marçal 2010).

A quinta fase do método AHP consiste em fazer uma matriz comparação em que cada linha da matriz será uma alternativa e cada coluna será um critério, estes valores serão avaliados usando a tabela presente na figura 3.6.

A sexta fase é multiplicar a matriz comparação de alternativas pelos pesos dos critérios obtidos na fase 3, assim possibilitando escolher a melhor alternativa no sétimo e último passo.

Para este projeto pretende-se identificar qual o melhor modelo de Machine Learning usando o AHP, tendo em consideração que futuramente serão realizados testes mais afundo para apurar quais os melhores modelos para utilizar. Esta identificação será feita nas secções abaixo.

3.4.1 Construção Árvore de Decisão

A figura 3.7 representa a árvore de decisão do método AHP, onde o problema a resolver é encontrar o melhor modelo de Machine Learning para prever a evolução da pandemia de covid-19, os critérios para avaliar as alternativas são o tempo que o modelo leva para criar previsões, a precisão das previsões e a capacidade do modelo de executar previsões em termos temporais, ou seja, uma previsão ao longo do tempo. As alternativas foram seleccionadas com base no estado de arte e são as seguintes: NNFOR, ES e KNN.

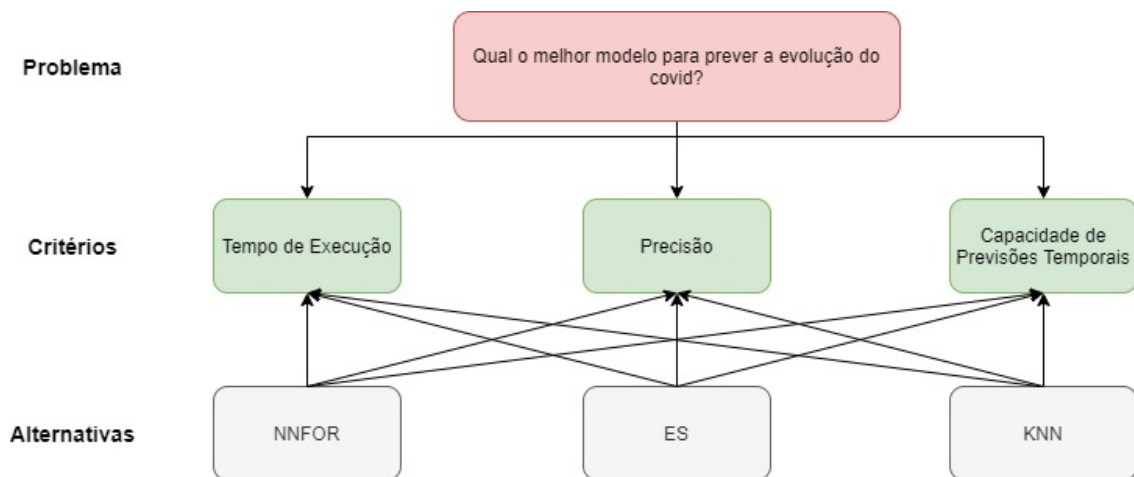


Figura 3.7: Árvore de Decisão do projeto.

3.4.2 Construção de Matriz de Comparação de Critérios

Após a construção da árvore de decisão construiu-se Matriz de Comparação de Critérios com base na escala da figura 3.6 e de seguida normalizou-se a matriz. A matriz comparação e a sua comparação está presente na figura 3.8.

Crítérios	Previsão Temporal	Precisão	Tempo Execução
Previsão Temporal	1	3	5
Precisão	1/3	1	3
Tempo Execução	1/5	1/3	1
Total	23/15	13/3	9

Crítérios	Previsão Temporal	Precisão	Tempo Execução	Peso Critérios
Previsão Temporal	15/23	9/13	5/9	0.6
Precisão	5/23	3/13	4/9	0.3
Tempo Execução	3/23	1/13	1/9	0.1

Figura 3.8: Tabela Superior: Matriz Comparação de Critérios.
Tabela Inferior: Matriz Comparação de Critérios Normalizada.

3.4.3 Cálculo da Razão de Consistência

O quarto passo é calcular a Razão de Consistência, os cálculos para obter esta razão estão presentes na figura 3.9 (o valor de IR é de 0.58 já que o projeto utiliza 3 critérios, este valor é obtido na figura 3.1) e através desta razão foi possível concluir que as prioridades e pesos dos critérios são consistentes já que a razão é inferior a 0.1 (0.04).

$$\begin{array}{c}
 \begin{array}{|c|c|c|} \hline 1 & 3 & 5 \\ \hline 1/3 & 1 & 3 \\ \hline 1/5 & 1/3 & 1 \\ \hline \end{array}
 \end{array}
 \begin{array}{|c|} \hline 0.6 \\ \hline 0.3 \\ \hline 0.1 \\ \hline \end{array}
 \approx \lambda_{\max}
 \begin{array}{|c|} \hline 0.6 \\ \hline 0.3 \\ \hline 0.1 \\ \hline \end{array}$$

$$\begin{array}{|c|} \hline 2 \\ \hline 0.79 \\ \hline 0.32 \\ \hline \end{array}
 \approx \lambda_{\max}
 \begin{array}{|c|} \hline 0.6 \\ \hline 0.3 \\ \hline 0.1 \\ \hline \end{array}$$

$$\begin{aligned}
 \lambda_{\max} &\approx \text{Média}((2/0.6)+(0.79/0.3)+(0.32/0.1)) \\
 &\approx 3.05 \\
 IC &= (3.05-3)/(3-1) \\
 &= 0.025 \\
 RC &= 0.025/0.58 \\
 &= 0.04
 \end{aligned}$$

Figura 3.9: Cálculos para obter a Razão de Consistência.

3.4.4 Construção da Matriz Comparação de cada critério entre alternativas

Neste passo é necessário avaliar cada alternativa nos diferentes critérios obtendo assim uma matriz comparação para cada critério. Calculando o peso de cada alternativa em cada critério, permite avaliar as alternativas em cada critério como é possível analisar na figura 3.10 em que, por exemplo, o modelo KNN é o melhor classificado em termos de tempo de execução.

Previsão Temporal	NNFOR	ES	KNN	Peso Alternativa
NNFOR	1	3	5	0.619
ES	1/3	1	4	0.284
KNN	1/5	1/4	1	0.096
Total	23/15	17/4	10	

Precisão	NNFOR	ES	KNN	Peso Alternativa
NNFOR	1	2	5	0.581
ES	1/2	1	3	0.309
KNN	1/5	1/3	1	0.109
Total	17/10	10/3	9	

Tempo Execução	NNFOR	ES	KNN	Peso Alternativa
NNFOR	1	1/2	1/5	0.115
ES	2	1	1/5	0.182
KNN	5	5	1	0.702
Total	8	13/2	7/5	

Figura 3.10: Matrizes Comparação de cada critério.

Através dos pesos que cada alternativa tem em cada critério é possível construir a matriz final que contem a avaliação de cada alternativa em cada critério, como a matriz presente

na figura 3.11.

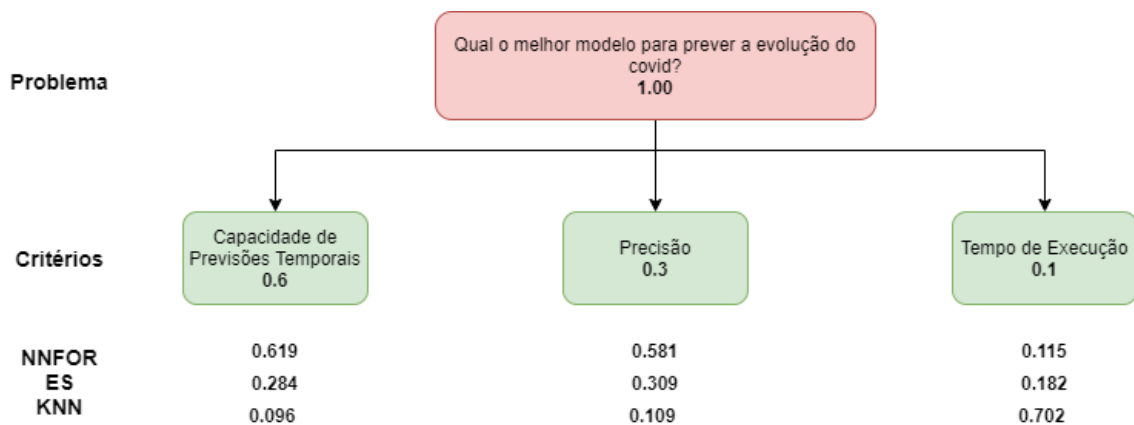


Figura 3.11: Árvore de Decisão com todos os pesos.

3.4.5 Cálculo da prioridade de cada alternativas e Escolha da melhor alternativa

Por fim o último passo é multiplicar a última matriz obtida, presente na figura 3.11 pelos pesos de cada critério também presente na mesma figura. Este cálculo final, presente na figura 3.12, permitiu concluir que a melhor opção seria o modelo NNFOR.

NNFOR	0.619	0.581	0.115	x	<table border="1"> <tbody> <tr> <td>0.6</td> </tr> <tr> <td>0.3</td> </tr> <tr> <td>0.1</td> </tr> </tbody> </table>	0.6	0.3	0.1	=	0.557
0.6										
0.3										
0.1										
ES	0.284	0.309	0.182	0.281						
KNN	0.096	0.109	0.702	0.161						

Figura 3.12: Cálculo Final.

Capítulo 4

Desenho da Solução

Neste capítulo serão analisadas várias alternativas para os diferentes componentes da solução, de forma a propor uma arquitetura, assim como será apresentada como será feita a alternativa selecionada.

4.1 Domínio

Para esta solução pretende-se, como já foi mencionando anteriormente, construir uma aplicação web que disponibilize informação em forma de dashboard, ou seja, pretende-se construir um website que disponibilize gráficos referentes á evolução da covid-19.

O ponto chave desta solução são os modelos de Machine Learning, que irão utilizar os dados presentes no sistema de forma a gerar previsões sobre o covid-19.

Pretende-se que o sistema obtenha todos os dias, de forma automatizada, dados do exterior para a sua base de dados relacional, de forma a fornecer informação atualizada. Quando estes novos dados entram no sistema, pretende-se que o sistema os utilize para atualizar os modelos de Machine Learning e executar novas previsões. Esta automatização será feita durante a madrugada, de forma a não atrapalhar o seu bom funcionamento durante o período em que terá mais acessos, ou seja, durante o dia.

Em suma, diariamente o sistema recolhe e trata de informação do exterior, essa informação é armazenada na base de dados relacional do sistema (que irá conter informação referente á evolução do vírus nas regiões de Portugal). Após o sistema obter esses novos dados, irá utilizar os seus modelos para executar novas previsões. Após o sistema gerar novas previsões, esses novos dados ficam disponíveis na aplicação web.

4.2 Especificação de Requisitos

Nesta seção será retirado um excerto da Especificação de Requisitos que tem por base alguns princípios abordados em Withall 2007, que se encontra completa na seção Anexo A onde numa seção introdutória explica o formato utilizado, de seguida são apresentados descrições, fluxos de execução e requisitos de cada caso de uso. Notar que os requisitos presentes neste excerto possuirão algumas referências a outras requisitos que apenas estão presentes em anexo.

4.2.1 Excerto de Especificação de Requisitos (Funcionais)

Nesta seção será apresentado um excerto dos requisitos funcionais do sistema, que são os requisitos necessários para o sistema realizar as suas funções principais.

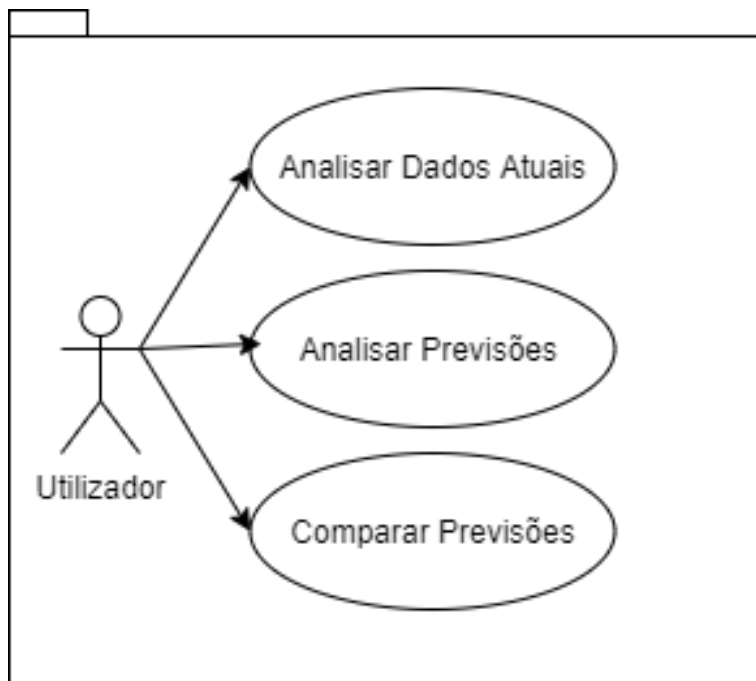


Figura 4.1: Diagrama de casos de uso.

A figura 4.1 representa o diagrama de casos de uso para o sistema pretendido. O sistema deverá possuir os seguintes casos de uso:

- **Analisar Dados Atuais:** tem como objetivo apresentar ao utilizador um dashboard contendo os dados atuais da evolução do covid-19. Onde o utilizador pode seleccionar se pretende ver os dados de Portugal ou de uma das suas regiões;
- **Analisar Previsões:** tem como objetivo apresentar ao utilizador um dashboard contendo gráficos que representam as previsões da evolução do covid-19. Onde o utilizador pode seleccionar se pretende ver as previsões de Portugal ou de uma das suas regiões;
- **Comparar Previsões:** neste caso de uso o utilizador escolherá duas regiões (estando Portugal incluído) e o sistema irá apresentar um dashboard constituído por gráficos que comparam a evolução da covid-19 nas regiões seleccionadas;

Para este excerto será apenas analisado o caso de uso "Comparar Previsões". A tabela 4.1 representa o fluxo das interações entre o utilizador e o sistema para obter o dashboard que permite comparar as previsões entre regiões.

Tabela 4.1: Fluxo de Execução do Caso de Uso: Comparar Previsões

Fluxo básico	
Passo	Descrição
1	O utilizador inicia o caso de uso Comparar previsões
2	O Sistema apresenta uma lista de regiões e pede ao utilizador para seleccionar duas regiões
3	O utilizador escolhe as 2 regiões
4	O sistema apresenta os dados das 2 regiões
	Conclusão do caso de uso atingindo o seu objetivo.

Como para este caso de uso o sistema deverá apresentar um dashboard que permita comparar previsões de diferentes regiões, é necessário permitir o utilizador seleccionar as regiões que pretende, é exatamente isso que descreve o requisito presente na tabela 4.2.

Tabela 4.2: Requisito - Seleccionar Regiões para Comparação

REQ.017	Seleccionar Regiões para Comparação
Prioridade	Essencial
Descrição	O sistema deverá apresentar ao utilizador duas "drop-down list" para o utilizador escolher as duas regiões que pretende analisar as previsões. Após seleccionar uma opção numa das listas, essa opção não deverá constar na outra lista para facilitar a ação do utilizador
Motivação	O utilizador precisa indicar quais as regiões pretende obter informação.
Informação adicional	As regiões disponíveis serão: Portugal (para o caso de o utilizador querer comparar a evolução de Portugal e de uma região de Portugal), Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Açores e Madeira.

Este dashboard terá vários gráficos, como exemplo a tabela 4.3 representa o requisito que permite a apresentação de um gráfico que compara as previsões do número de casos confirmados em diferentes regiões. Todos os gráficos possuem um requisito associado, descrito em Anexo A.

Tabela 4.3: Requisito - Gráfico Comparação Casos Confirmados

REQ.018	Gráfico Comparação Casos Confirmados
Prioridade	Essencial
Descrição	Com as duas regiões selecionadas pelo utilizador, o sistema deverá apresentar as previsões da evolução do vírus, para isso deverá apresentar o gráfico de casos confirmados, onde deverão estar presentes as previsões das 2 regiões selecionadas. Para tal no eixo das abcissas irão ficar os dias e as regiões, de forma a que cada dia possua duas barras distintas cada uma correspondendo a uma região, em que as cores nos 10 primeiros dias que correspondem a dados reais serão cinzento claro para a primeira região selecionada e cinzento escuro para a segunda, os restantes dias correspondem às previsões, onde as barras deverão estar azul claro para a primeira região selecionada e azul escuro as barras correspondentes á segunda região selecionada. No eixo das ordenadas ficará o número de casos confirmados no referente dia.
Motivação	é necessário distinguir os dados de cada uma das regiões.
Informação adicional	No topo de cada barra deverá estar o seu valor. O sistema deverá apresentar dentro do gráfico de barras duas linhas tal como em REQ.008 que representarão o total de casos confirmados de cada região. Estas linhas deverão ter as mesmas cores que as barras dos gráficos (para a primeira região cinzento-claro nos primeiros 10 dias e azul claro nos restantes, para a segunda região cinzento escuro nos primeiros dias e azul escuro nos restantes). Quando o utilizador passar o cursor em cima de uma linha, ser-lhe-á apresentado o número total de casos confirmados naquele dia.

4.2.2 Excerto de Especificação de Requisitos (Não-Funcionais)

Nesta seção será apresentado um excerto dos requisitos não-funcionais do sistema, que são requisitos necessários para o sistema sem ser descreverem casos de uso. Para isso será utilizado o modelo FURPS, que divide os requisitos não-funcionais do sistema em 4 categorias diferentes, para além dos requisitos funcionais:

- Usabilidade: requisitos voltados para os aspetos estéticos e interativos do sistema;
- Confiança: requisitos voltados para a disponibilidade, que geram confiança ao utilizador;
- Desempenho: requisitos voltados para o rendimento do sistema;
- Suporte: requisitos voltados para a configuração do sistema;

A tabela A.39 representa um requisito de usabilidade, visto que procura melhorar a experiência do utilizador, assim como melhorar a sua interação com o sistema. Enquanto o requisito

A.40 especifica o tempo diário que o sistema deverá estar online, garantindo ao utilizador a confiança de que quase a qualquer hora do dia poderá aceder e analisar as informações disponíveis.

Tabela 4.4: Requisito Não-Funcional: Gráficos com cores destacadas e interativos

REQ.NF.01	Gráficos com cores destacadas e interativos
Prioridade	Essencial
Descrição	Todos os gráficos deverão utilizar os esquemas de cores já indicados, assim como possuírem uma legenda e um título, de forma a informar o utilizador do que se trata o mesmo. Os gráficos deverão ser interativos, da forma a ser tornar possível funções como "zoom in", "zoom out", deslizar o gráfico em todas as direções e reiniciar o gráfico (fazer o gráfico voltar às suas condições normais).
Motivação	É necessário criar ao utilizador uma experiência cativante, que o deixe confortável e com uma opinião positiva do sistema.

Tabela 4.5: Requisito Não-Funcional: Disponibilidade do Sistema

REQ.NF.02	Servidor
Prioridade	Essencial
Descrição	O Sistema deverá estar disponível 23 horas e 50 minutos por dia, sendo os 10 minutos diários que sobram para o sistema atualizar a base de dados e gerar novas previsões. Sendo estas pausas de 10 minutos programadas para serem feitas às 15:50
Motivação	É necessário que o sistema estar quase 100% do tempo disponível. A hora selecionada para atualizar a base de dados e consequentemente gerar novas previsões tem a ver com os relatórios diários da pandemia que são publicados entre as 13:00 e 15:30.

A tabela A.41 descreve as características que o servidor deverá, no mínimo, possuir para executar todas as extrações de dados e gerar todas as previsões num espaço de tempo de 10 minutos. Enquanto o requisito A.42 aborda a compatibilidade do sistema, que deverá ser compatível com várias plataformas e navegadores.

Tabela 4.6: Requisito Não-Funcional: Servidor

REQ.NF.03	Servidor
Prioridade	Essencial
Descrição	O sistema deverá ter um servidor com um processador com no mínimo 2.0GHz e memória RAM de no mínimo 6Gb.
Motivação	É necessário que o sistema esteja alojado num servidor bom para que este demore pouco tempo a executar todo o processo de atualizar a base de dados e gerar novas previsões, assim como tenha a capacidade de ter vários utilizadores a utilizar o sistema ao mesmo tempo.

Tabela 4.7: Requisito Não-Funcional: Compatibilidade com várias plataformas e navegadores

REQ.NF.04	Compatibilidade com várias plataformas e navegadores
Prioridade	Essencial
Descrição	O sistema deverá ser feito de forma que seja compatível com navegadores como o Microsoft Edge, Google Chrome, Firefox e Safari. Sendo compatível também com Android e IOS. Em suma, o sistema deverá se adaptar a qualquer tamanho de ecrã e navegador.
Motivação	É necessário permitir o utilizador aceder ao sistema da forma que ele quiser.

4.2.3 Componentes Suplementares

Para complementar o requisito descrito na tabela A.41 segue-se uma descrição dos componentes do sistema presentes no servidor. A figura A.2 representa a estrutura base do sistema, através da figura é possível extrair os seguintes componentes:

- **Algoritmos de Extração:** O Sistema deverá possuir algoritmos para extração de informações das diversas fontes. Como por exemplo algoritmos de Web Scraping para retirar informações de sites, assim como outros algoritmos para outros tipos de fontes de informações. Estes algoritmos irão extrair dados e executar algum tratamento de dados e irá armazenar os dados na base de dados intermédia;
- **Modelos Machine Learning:** O sistema terá modelos capazes de prever o número de casos confirmados, casos ativos, óbitos e recuperados em Portugal e suas regiões;
- **Base de Dados:** O Sistema deverá possuir uma base de dados capaz de suportar os extraídos, assim como os dados após passarem por um tratamento compondo os dados reais de Portugal e suas regiões, assim como suportar as previsões;

Todos estes requisitos não-funcionais, estão presentes em formato de tabela em Anexo A.

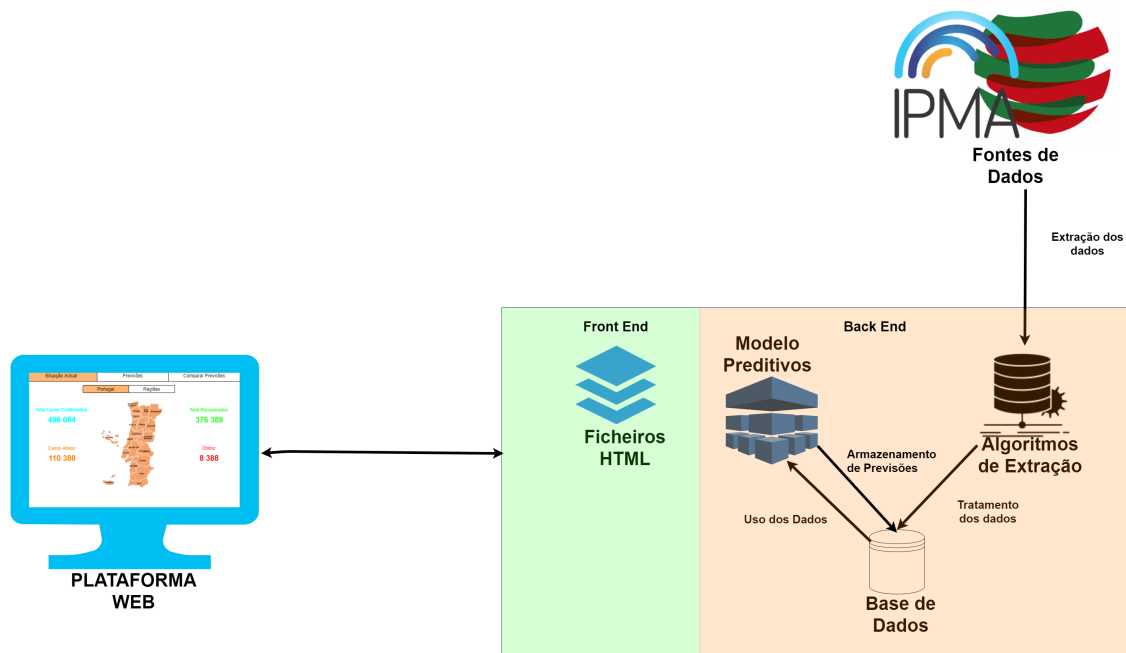


Figura 4.2: Representação dos componentes do projeto.

4.3 Análise de Alternativas

Nesta seção irão ser apresentadas diferentes alternativas para os componentes do sistema, assim como as vantagens e desvantagens de cada opção, tendo também no final a escolha que foi selecionada pelo utilizador.

4.3.1 Fontes de Dados

Existem várias alternativas de fontes de dados que irão abastecer o sistema com dados, dentre as quais se destacam as seguintes 2 soluções:

- **Repositório com bases de dados:** esta solução consiste em fazer o download de ficheiros CSV de repositório públicos como por exemplo o CoronaNet (*CoronaNet* 2020), esta solução tem como vantagem a facilidade de manipulação e tratamento de dados, mas tem a desvantagem do risco de não ter dados atualizados, ou seja, com o exemplo do CoronaNet (*CoronaNet* 2020) por vezes demoram vários dias para atualizar os dados referentes a Portugal, indo contra um dos principais objetivos do projeto, de ter dados atualizados;
- **WebScraping de fontes atualizadas:** esta solução consiste em usar técnicas de webscraping para extrair os dados de fontes atualizadas, como os relatórios diários que a DGS publica. Esta solução tem como principal desvantagem requerer um maior tratamento de dados, mas tem a vantagem de ser possível ter sempre dados atualizados;

Tendo em conta que o foco principal é a qualidade dos dados e o tempo de extração, pois uma alta qualidade dos dados poderá significar previsões de melhor qualidade, desta forma não é possível optar uma das duas opções, já que por vezes a melhor opção poderá ser fazer Web Scraping, outras vezes será melhor utilizar repositórios, por exemplo dados sobre medidas implementadas para atenuação da evolução da pandemia, a melhor fonte de dados

poderá ser um repositório como o CoronaNet (CoronaNet 2020) ou fazer WebScraping do portal Diário da República. Em suma a melhor alternativa é utilizar uma mistura das duas alternativas abordadas anteriormente, já que depende da fonte de dados.

4.3.2 Base de Dados

Existem algumas alternativas para a construção da base de dados, dentre as quais se destacam as duas seguintes:

- **1 tabela:** Nesta solução todos os dados são armazenados em uma tabela, em que teria por uma variável que indicava se os dados são reais ou previsões. A vantagem seria ter todos os dados numa só também enquanto possui como desvantagem pouca organização.
- **Várias tabelas:** Nesta solução os dados seriam armazenados em várias tabelas, ou seja, uma tabela teria os dados reais, noutra tabela as previsões e por fim uma tabela com os dados mais atuais. Tendo por vantagem a organização.

A figura 4.3 representa as alternativas para a base de dados, a alternativa que melhor se enquadra com o problema é a de colocar os dados em várias tabelas, já que serão necessárias várias tabelas auxiliares para gerar as previsões, assim como testar os modelos de Machine Learning. Desta forma existirão 2 tabelas principais, uma que irão conter toda a evolução do covid-19 e outra que irá conter as previsões atuais, as restantes tabelas serão auxiliares, como por exemplo a de treino dos modelos.

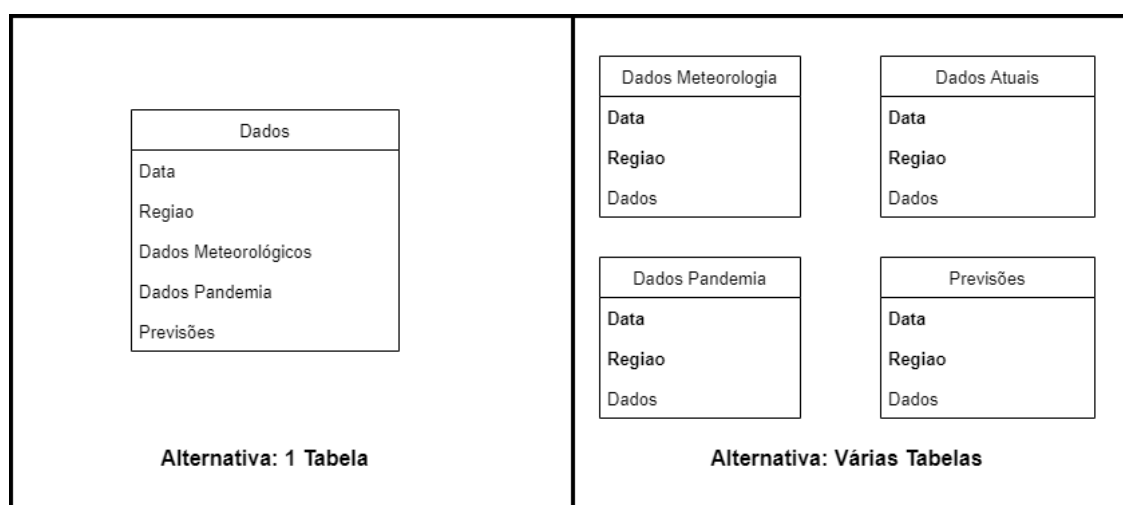


Figura 4.3: Ilustração das alternativas.

4.3.3 Modelos

Como já foi abordado anteriormente o objetivo principal do sistema será apresentar 4 tipos de previsões: número de casos confirmados, número de óbitos, número de recuperados e casos ativos. O sistema deverá executar estas quatro previsões para Portugal e para cada uma das suas regiões, ou seja, o sistema irá prever por exemplo o número de casos confirmados em Algarve, Alentejo, Norte, Centro, R.A. Madeira, R.A. Açores, Lisboa e Portugal.

Existem algumas alternativas para construir estes modelos:

- **Máximo número de modelos** esta solução consiste em construir o máximo de modelos possível, ou seja, para cada território (Portugal e as suas regiões) existirão 4 modelos, um modelo para cada tipo de previsão, o que significaria um total de 32 modelos. Esta solução tem como vantagem que cada modelo representará melhor o seu terreno e tipo de previsão, mas tem como desvantagem a manipulação de dados necessária, assim como o tempo de execução para gerar todas as previsões.
- **Mínimo número de modelos:** esta solução consistem em construir apenas 3 modelos, um modelo para previsão de casos confirmados nas regiões de Portugal, um modelo para previsão de óbitos nas regiões de Portugal e outro para prever o número de recuperados nas regiões de Portugal. Estes modelos irão treinar com os dados de cada região antes de executar as previsões.

Por exemplo, o sistema começa por prever os confirmados no Algarve, o modelo treina com os dados do Algarve e posteriormente gera as previsões, de seguida treina com os dados do Alentejo e gerar as suas previsões, até gerar previsões para todas as regiões de Portugal. As previsões de confirmados em Portugal será obtida através do somatório das previsões das regiões. Este processo é repetido nas outras previsões, enquanto a previsão de casos ativos é obtida através das outras previsões recorrendo á equação:

$$Ativos = Total Confirmados - (Total Recuperados + Total Obitos) \quad (4.1)$$

Esta solução tem a vantagem de, comparativamente com a outra solução, requerer uma manutenção e manipulação dos modelos menor.

Assim como na seção 4.3.2 é necessário ter em conta o baixo poder computacional, desta forma a melhor opção é a que utiliza o mínimo de modelos possível. Estes modelos iram passar por várias etapas tais como:

- Utilizar os modelos tendo por base o estudo do estado da arte realizado no capítulo 2, são eles o modelo K-Nearest Neighbors, ARIMA e Random Forest;
- Selecionar quais são as variáveis mais importantes para cada tipo previsão e modelo, tendo como base da seleção as métricas MSE, RMSE, MAE e R^2 ;
- Selecionar qual o melhor intervalo de dados de treino para cada tipo previsão e modelo, tendo como base da seleção as métricas MSE, RMSE, MAE e R^2 ;

4.3.4 Aplicação

Nesta seção serão focadas as alternativas da aplicação como um todo, ou seja, diferentes combinações de componentes com o objetivo de encontrar o melhor. Alternativas:

- **Solução Centralizada:** Esta solução consiste em ter toda a aplicação numa mesma linguagem e framework, ou seja, o processo de extração dos dados seria feito em linguagem R, todos os dados seriam armazenados em dataframes do projeto em RStudio, os modelos seriam feitos em R e a aplicação web seria feita com recurso ao pacote RShiny que permite criar ferramentas web com linguagem R. A principal vantagem desta solução é permitir uma interação entre os componentes do projeto de forma direta, como desvantagem tem possíveis conflitos entre os mesmos.

- **Solução Descentralizada:** Esta solução consiste em, ao contrário da solução anterior, ter os vários componentes em diferentes linguagens e locais, ou seja, a aplicação seria desenvolvida em HTML e PHP, os modelos em python, a base de dados armazenada em MySQL e a extração dos dados seria feita com algum algoritmo de python ou ferramenta externa. Esta solução tem como vantagem não haver conflitos entre os componentes, mas tem a desvantagem da interação entre os componentes não ser tão direta.

A figura 4.4 ilustra as diferentes arquiteturas, nela é possível verificar que na solução centralizada os elementos surgem num mesmo bloco e interagem diretamente entre si enquanto na solução descentralizada cada componente do sistema está isolado no seu bloco sendo necessário existir uma interoperabilidade entre os blocos para o sistema funcionar. Como este projeto requer uma grande interação entre os dados e os modelos, a melhor escolha é a solução centralizada, já que permite a interação direta entre dados e modelos.

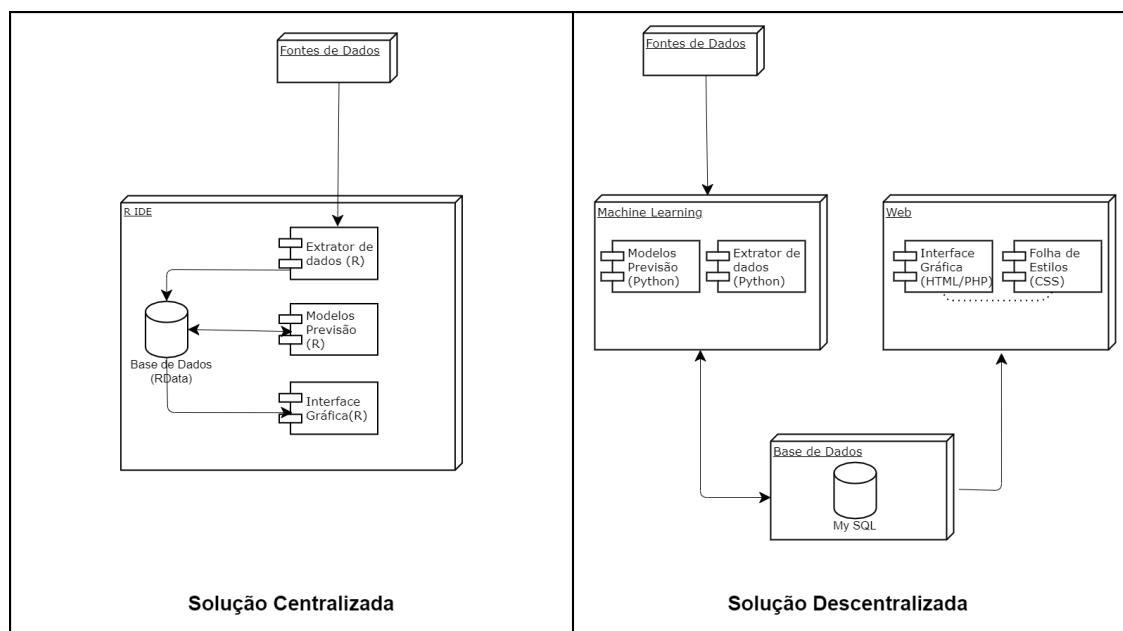


Figura 4.4: Ilustração das alternativas de Arquitetura.

4.4 Esboço da Interface Gráfica

Nesta seção pretende-se demonstrar um esboço da interface gráfica que será apresentada aos utilizadores. Sendo um esboço, a versão final poderá ter outro aspeto, seguem-se os diferentes ecrãs que o sistema deverá possuir. A figura 4.5 representa o primeiro ecrã a ser mostrado ao utilizador, nele será possível analisar os dados atuais de Portugal, assim como escolher uma região e ver os seus dados atuais.

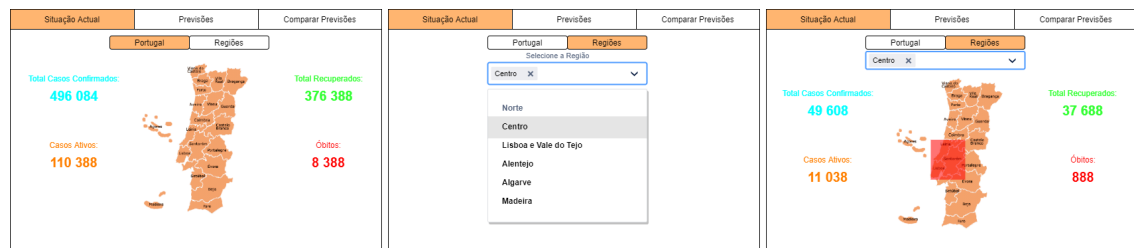


Figura 4.5: Esboço da interface do projeto que apresenta a situação pandémica atual no país.

A figura 4.6 representa uma outra seção do site chamada de "Previsões" nela o utilizador terá acesso a um dashboard contendo um gráfico para cada previsão (casos confirmados, casos ativos, óbitos e recuperados). Sendo que esses gráficos mostrarão os dados dos dias anteriores às previsões e as próprias previsões. Podendo o utilizador alternar as previsões entre Portugal e uma região selecionada.

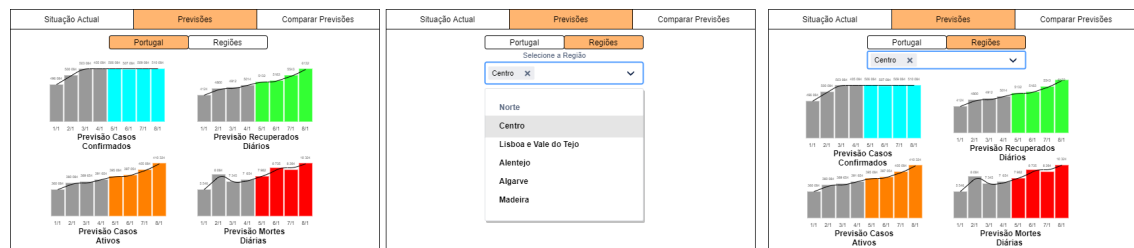


Figura 4.6: Esboço da interface do projeto que apresenta a as previsões da evolução da pandemia no país.

A figura 5.14 representa a seção do site que permitirá ao utilizador, comparar previsões de diferentes regiões, através de um dashboard composto por gráficos de barras contendo os dados das regiões selecionadas.

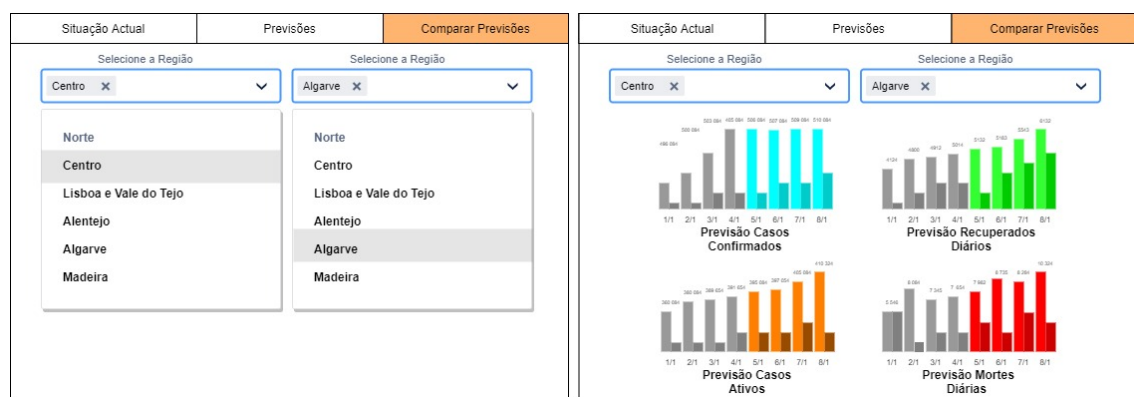


Figura 4.7: Esboço da interface do projeto que apresenta a comparação entre previsões de 2 regiões.

Capítulo 5

Implementação

Este capítulo tem como objetivo explicar como foi feita a implementação do sistema, sendo que a figura 5.1 representa os ficheiros que o compõem. Onde:

- DBase.RData - será o ficheiro que irá armazenar a base de dados;
- Data.R - será o ficheiro responsável por atualizar a base de dados e gerar as previsões;
- ui.R - será responsável por descrever a interface gráfica do sistema;
- server.R - servirá de comunicação entre o Data.R e ui.R, preenchendo os elementos da interface gráfica, com os dados obtidos pelos pedidos do ficheiro Data.R;

Primeiro será apresentada a seção da Base de Dados, que ficará armazenada no ficheiro "DBase.RData", nesta seção será apresentado um diagrama representativo da base de dados, será também explicado de que forma todos os dados foram extraídos, assim como as suas fontes. Por fim a base de dados será alvo de uma análise exploratória e de uma crítica à sua construção.

De seguida será apresentada uma seção sobre os Modelos, nela será apresentada um conjunto de transformações que os dados serão alvo. Será também apresentada um conjunto de testes com diferentes características que os modelos passarão, com o objetivo de encontrar os modelos que irão integrar a aplicação. Por fim uma seção que aborda a implementação da aplicação em si.

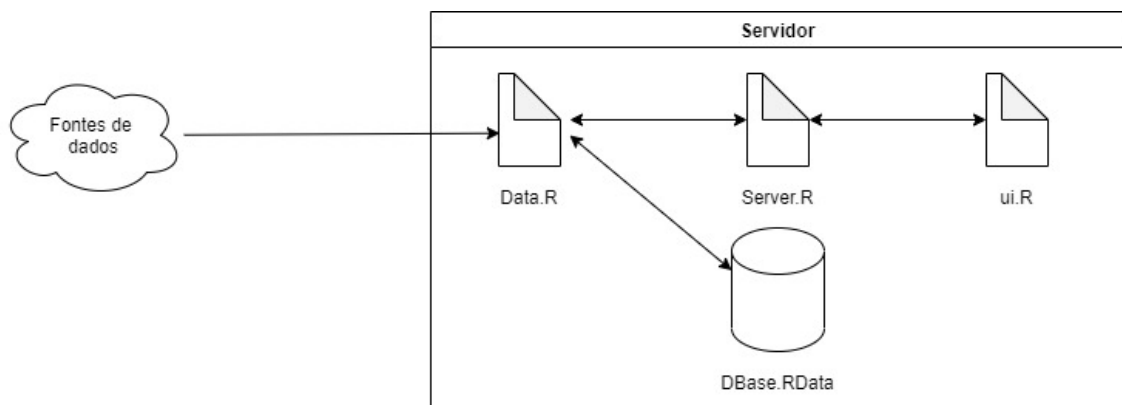


Figura 5.1: Implementação do Sistema.

5.1 Base de Dados

Esta secção é dedicada á implementação da base de dados, nela serão apresentadas as fontes de dados, as variáveis que constituem os dados e como foi feita a extração dos dados. Será também apresentada uma exploração aos dados e por fim uma análise critica aos dados.

5.1.1 Construção da Base de Dados

Na presente subsecção será mostrada como é feita a construção da base de dados. A figura 5.3 representa a base de dados relacional do sistema, onde estarão armazenados todos dados para apresentação ao utilizador e para futura utilização para geração de previsões. A base de dados possui uma tabela central que corresponde a evolução da pandemia tendo como chaves primárias a data e a região ao qual os dados correspondem, estas duas chaves estabelecem uma relação de 1 para 1 com todas as outra tabelas, com a exceção da tabela que caracteriza as regiões, já que os dados não possuem data, assim como a tabela das medidas políticas implementadas para atenuação da evolução do vírus não possui região, já que as medidas são nacionais e portanto para todas as regiões

A figura 5.3 representa duas tabelas auxiliares da base de dados, que são a tabela que possui todos os dados servindo como base para os modelos gerarem as previsões e a tabela das previsões que é composta por dados da evolução covid-19 nos últimos 5 dias assim como por previsões dos próximos 5 dias, deixando de lado tabela auxiliares como as de treino e as de uso dos modelos. De notar que ambas as figuras contém apenas as variáveis mais importantes e que os nomes foram alterados para melhorar a compreensão do leitor.

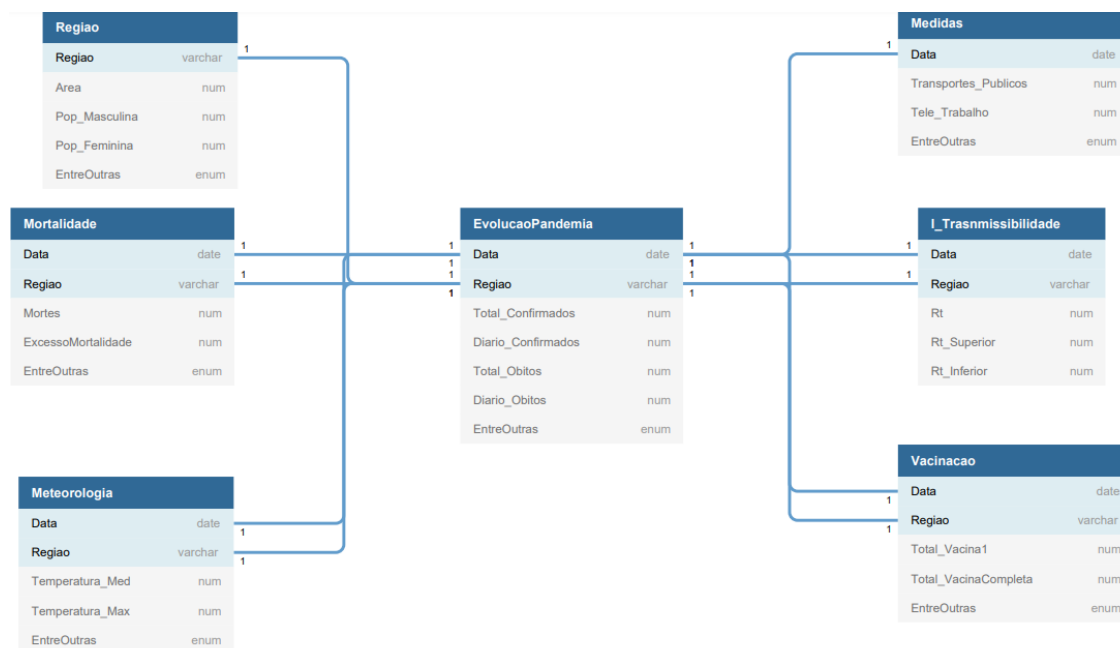


Figura 5.2: Diagrama da Base de Dados.

Dados				Previsões	
Data	date	Reg_Diario_Confirmados	num	Data	date
Regiao	varchar	Reg_Total_Obitos	num	Regiao	varchar
mes	chr	Reg_Diario_Obitos	num	Tipo_Dado	real_previsao
Reg_Area	num	RegMortes	num	Reg_Total_Confirmados	num
Reg_DensidadeP	num	RegMortalidadeBase	num	Reg_Diario_Confirmados	num
Reg_Envelhecimento	num	RegExcessoMortalidade	num	Reg_Total_Obitos	num
RegP_Masculi	num	Portugal_Confirmados_Total	num	Reg_Diario_Obitos	num
RegP_Femini	num	Portugal_Confirmados_Diario	num	Portugal_Confirmados_Total	num
RegP_Total	num	Portugal_Obitos_Total	num	Portugal_Confirmados_Diario	num
Reg_TemperaturaMax	num	Portugal_Obitos_Diario	num	Portugal_Obitos_Total	num
Reg_TemperaturaMin	num	Portugal_Recuperados_Total	num	Portugal_Obitos_Diario	num
Reg_TemperaturaMed	num	Portugal_Recuperados_Diario	num	Portugal_Recuperados_Total	num
Reg_Precipitacao	num	Portugal_Ativos_Total	num	Portugal_Recuperados_Diario	num
Reg_VelocidadeVento	num	Portugal_Ativos_Diario	num	Portugal_Ativos_Total	num
Reg_Humidade	num	Portugal_Vigilancia	num	Portugal_Ativos_Diario	num
Reg_Vac1Semanal	num	Portugal_Vigilancia_Diario	num		
Reg_Vac2Semanal	num	Portugal_Internados	num		
Reg_TotalSemanal	num	Portugal_Internados_Diario	num		
Reg_Vac1Total	num	Portugal_InterdosUCI	num		
Reg_Vac2Total	num	Portugal_InterdosUCI_Diario	num		
Reg_Total	num	Portugal_Mascara	chr		
Reg_CoberturaVac1	num	Portugal_Restaurantes	chr		
Reg_CoberturaVac2	num	Portugal_Ensino	chr		
Reg_Rt	num	Portugal_T_Publicos	chr		
Reg_Total_Confirmados	num	Portugal_Trabalho	chr		

Figura 5.3: Tabelas Adicionais da Base de Dados.

É possível observar que a tabela dos dados da figura 5.3 embora esteja orientada de acordo com a região de Portugal, possui variáveis globais que estão orientadas ao nível do país, isto significa que estes dados estão repetidos numa mesma data independentemente da região, esta opção deve-se ao fato de que poderá ser interessante analisar como se comportam os modelos nas previsões utilizando dados de Portugal.

De seguida será apresentado de onde e como foram extraído os dados, assim como será explicada como é feita a junção de cada tabela á tabela auxiliar dos dados, já esta será utilizada pelos modelos e já que é difícil compreender a sua construção.

Dados Regionais

Para diferenciar as regiões são necessários dados que as caracterizem, para isso foram usados dados presentes em *Pordata - Base de Dados Portugal Contemporâneo 2021*, base de dados esta que se responsabiliza por recolher, organizar e divulgar diversas formas de informação sobre várias áreas da sociedade portuguesa e europeia. Para a utilização dos dados presentes neste portal foram extraídos vários ficheiros CSV contendo os seguintes dados para as regiões de Portugal, usando Nomenclatura das Unidades Territoriais para Fins Estatísticos 2 (NUTS2), a descrição de todas as variáveis utilizadas encontra-se no anexo B.1.1.

Optou-se por utilizar ficheiros CSV desta fonte por ser uma fonte de confiança e porque estes dados não sofrerão alterações, já que serão sempre os mesmos a representar cada região. Cada ficheiro representava uma variável, por exemplo um ficheiro "densidade.csv", possui as densidades populacionais de todas as regiões.

Para a extração conjuntos de dados utilizou-se a função `read_excel`, onde por vezes era necessário especificar em que intervalos de células estavam os dados, já que os ficheiros CSV tinham demasiada informação extra. Posteriormente alterou-se os nomes das colunas dos dados já que estavam incorretas e executou-se algum tratamento de dados.

Por fim fundiu-se todos os conjuntos de dados com base na região, de forma que todas as regiões tivessem todos os dados que as caracterizam. O excerto de código B.1 presente no apêndice B representa o carregamento e tratamento dos dados que foram extraídos de *Pordata - Base de Dados Portugal Contemporâneo 2021*.

Dados Meteorológicos

Dados meteorológicos poderão ser dados importantes para a geração de previsões, como foi visto em 2.3.3. A maioria das fontes de dados meteorológicos ou possui dados sobre Portugal como um todo, ou de todos os distritos do país, ou seja, nenhuma fonte de dados meteorológicos possui dados regionais.

De forma a contornar o problema de não existirem dados para as regiões de Portugal, optou-se por escolher uma cidade que se localizasse próxima ao centro de uma região para representar a mesma. Esta opção deve-se ao fato de existir um baixo poderio computacional, de maneira que outras soluções iriam exigir mais poder computacional, assim como exigir mais tempo de extração e processamento, por exemplo os dados meteorológicos de cada região, serem obtidos através de médias dos dados dos seus distritos.

A figura 5.4 demonstra as cidades selecionadas para representar, em dados meteorológicos, as regiões de Portugal. A região do Norte é representada pela cidade de Vila Real, a região do Centro pela cidade de Coimbra, a região de Lisboa e Vale do Tejo pela cidade de Lisboa, a região do Alentejo por Évora, a região do Algarve pela cidade de Faro e as Regiões Autónomas da Madeira e dos Açores pela cidade de Funchal e de Horta respetivamente.

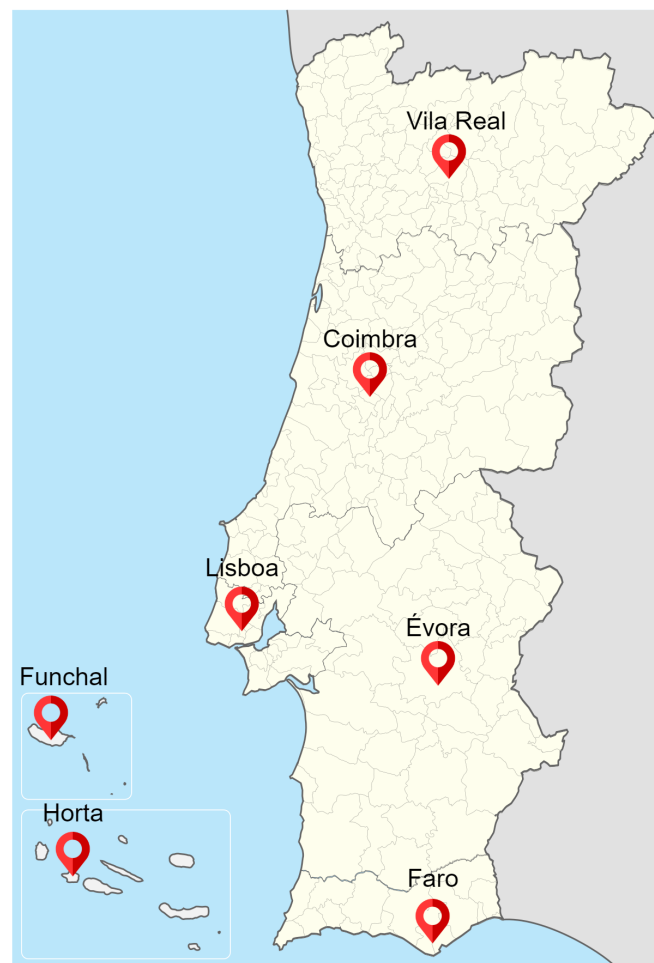


Figura 5.4: Localizações das cidades representativas de cada região.

Como fonte destes dados meteorológicos foi utilizada a base de dados meteorológicos de *Weather Data Weather API | Visual Crossing* 2021, que é uma organização líder e especializada em fornecer soluções a nível mundial de dados e outras formas de visualização de informação meteorológica. Esta base de dados possui dados históricos, ou seja, dados sobre o passado, assim como previsões da meteorologia mundial num espaço de 15 dias. Desta base de dados foram extraídos os dados das cidades representativas de cada região desde o início da pandemia em Portugal, contendo as variáveis descritas no anexo B.1.1.

O excerto de código B.2 presente no anexo B representa o carregamento de dados da região da Madeira, através da função *read.csv* onde *lhe* é passado um URL contendo o pedido de dados, sendo possível definir a data inicial e final do conjunto de dados no caso do exemplo os dados são compreendidos entre dia 2021-03-02 e 2021-03-31, é possível também definir a chave de acesso aos dados (já que cada conta possui a sua chave), bem como a localização dos dados que no caso do exemplo é a cidade de Funchal. Para todos os pedidos realizados executa-se o mesmo processo: traduzir os nomes das colunas (linha 4 no excerto de código presente no anexo B.2), acrescentar a coluna região com a região que os dados representam e o tratamento de dados. Este tratamento de dados passa pela eliminação de dados que não são usados e a alteração da ausência de valores para o valor 0, por exemplo o caso da precipitação que por vezes tinha valores vazios no lugar de 0. O último passo é juntar os

dados de diferentes regiões, para isso junta-se os dados do pedido a um conjunto de dados, utilizando a função *rbind* (linha 11 no excerto de código presente no anexo B.2).

Após a execução de todos os pedidos, utilizou-se a função *arrange*, que ordenou os dados de acordo com a data e a Região, de forma a ficarem como na tabela 5.1, sendo esta a tabela de dados meteorológicos e criando a tabela auxiliar de previsões, contendo os dados meteorológicos e a descrição da região que se repete sempre que uma linha representa uma região.

Tabela 5.1: Primeiras linhas dos dados Meteorológicos

Datas	Região	Temp. Max	Temp. Min	Humidade	Precipitação	V.Vento
2020-03-01	Alentejo	16.6	10.6	92.60	0.10	27.4
2020-03-01	Algarve	18.2	11.6	92.26	0.20	26.3
2020-03-01	Centro	13.8	9.9	94.93	43.05	29.4
2020-03-01	A.M. Lisboa	16.1	14.0	92.37	0.68	38.6
2020-03-01	Norte	14.1	7.2	84.83	40.05	39.5
2020-03-01	R.A. Açores	17.7	15.7	89.04	1.89	42.6
2020-03-01	R.A Madeira	22.6	16.6	70.60	0.00	10.8

Dados da Vacinação

Foram também extraídos dados sobre a vacinação contra o Covid-19 em Portugal do portal da Direção Geral de Saúde, também conhecida como DGS. A DGS é o órgão de maior importância em Portugal na área da saúde, tendo como missão coordenar atividades de progresso da saúde e prevenção de doenças, assegurar a qualidade do sistema de saúde e elaborar e executar o Plano Nacional de Saúde (*Direção-Geral da Saúde* 2021).

No caso da vacinação a Direção Geral de Saúde publica um relatório semanal (*Relatório de Vacinação - COVID-19* 2021) e publica um novo conjunto de dados contendo as variáveis descritas no anexo B em B.1.1.

A DGS atualiza semanalmente, no mesmo endereço, o URL para a sua base de dados da vacinação, já que esta base de dados altera todas as semanas, ou seja, semanalmente é necessário analisar a estrutura da página e identificar o URL que apontar para o novo ficheiro contendo os dados mais atuais, para isso utilizou-se Web Scraping.

O excerto de código B.3, presente no anexo B, representa a extração dos dados da vacinação. Para extrair os dados foi necessário recorrer à biblioteca de *Web Scraping - rvest*, onde no primeiro passo, se usou a função *read_html* que é capaz de extrair todo o código html presente no endereço que lhe foi passado (endereço onde estão presentes os relatórios de vacinação)(linhas 2-4 no excerto B.3). O passo seguinte é localizar o endereço da base de dados, para isso utilizou-se a função *html_attr* que localiza diferentes componentes de html, no caso o componente "href", ou seja, localiza todos os endereços presentes na página (linha 6 no excerto B.3). De seguida procurou-se dentre todos os endereços encontrados na página o único que termina em "CSV", que será o endereço para a base de dados (linha 8 no excerto B.3). Após ser encontrado o endereço, é feita a extração dos dados usando a função *read.csv* e armazenando-os na tabela de vacinação (linha 10 no excerto B.3).

Após a extração dos dados, foi realizado um tratamento de dados onde se corrigiu, principalmente, os nomes das regiões. No exemplo B.4, disponível no anexo B é possível verificar

que se utilizou a função *gsub* capaz de substituir conjunto de caracteres de uma variável, para retirar a sigla "ARS"(Administração Regional de Saúde) da variável região, bem como para corrigir o nome de regiões.

Dados do Índice de Transmissibilidade

O índice de transmissibilidade, também conhecido como $R(t)$, consiste em representar o número de pessoas que um indivíduo pode infectar, estando ele infectado com Covid-19. Sendo um indicador que poderá ser importante para as previsões.

Como fonte de dados optou-se por utilizar o Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA) colabora com a Direção Geral de Saúde, divulgando semanalmente relatórios contendo dados sobre o índice de transmissibilidade (*Covid-19: curva epidémica e parâmetros de transmissibilidade* 2021). Da mesma forma que acontecia com o dados meteorológicos, abordado em 5.1.1, o portal do Instituto atualiza semanalmente a página que contem os endereços URL para os ficheiros de dados mais atualizados, desta forma será necessário recorrer á mesma técnica, ou seja, utilizar Web Scraping para obter o URL correto que aponta para os dados mais atualizados e posteriormente extraí-los.

Na página do Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA) (*Covid-19: curva epidémica e parâmetros de transmissibilidade* 2021), referente ao índice de transmissibilidade, possui relatórios semanais e ficheiros CSV de cada região com as variáveis descritas em B.1.1.

Toda a extração dos dados sobre o índice de transmissibilidade está presente no excerto de código B.5, presente no anexo B. Para obter estes dados foi necessário utilizar a biblioteca de *Web Scraping - rvest*, onde numa primeira instância se utilizou a função *read_html* para guardar todo o código HTML da página que contém os dados pretendidos. De seguida, isolou-se apenas as tabelas existentes no código da página já que os ficheiros contendo os parâmetros de transmissibilidade pretendidos estão presentes numa tabela (linha 2 no excerto B.5). De seguida isola-se a segunda tabela existente, já que a primeira corresponde aos relatórios semanais (linha 3 no excerto de código B.5). Após isolar a tabela que contem as bases de dados dos índices de transmissão de cada região, armazenou-se o valor de cada célula da tabela (nome da região) e o respetivo endereço (linhas 4-8 no excerto de código B.5), sendo necessário converter o formato dos dados para caracteres.

Para extrair os dados dos ficheiros, percorreu-se o conjunto de dados que armazenou o nome da região e o seu respetivo endereço, para cada região extraiu-se os dados para tabelas auxiliares e de seguida acrescentou-se a variável região com os respetivos valores. Por fim juntou-se os dados sobre o índice de transmissibilidade das diferentes regiões numa única tabela, assim como se juntou á tabela auxiliar, que já continha os dados meteorológicos, de vacinação e descrição da região, através da função *rbind*.

A tabela 5.2 representa uma parte da tabela auxiliar para as previsões, tendo os dados do índice de transmissibilidade (rt) e um exemplo de cada um dos outros tipos de dados. Ficando os dados ordenados pela data e região, como é possível observar no exemplo da tabela 5.2.

Tabela 5.2: Primeiras linhas dos dados do Índice de Transmissibilidade

Datas	Região	Temp. Max	Vac Completa	Rt	Rt L. Inferior	Rt L. Superior
2020-03-10	Algarve	16.6	0%	1.671465	1.0660436	2.410899
2020-03-10	Centro	18.2	0%	1.929256	1.6774508	2.197080
2020-03-10	A.M. Lisboa	13.8	0%	1.793020	1.4537736	2.177835
2020-03-10	Norte	16.1	0%	1.821209	1.6300954	2.025740
2020-03-10	R.A. Açores	14.1	0%	1.194052	0.5184661	2.195469
2020-03-10	R.A Madeira	17.7	0%	1.285605	0.5580801	2.301766
2020-03-10	Alentejo	22.6	0%	1.273978	0.5879368	2.234733

Dados sobre a Evolução do Covid-19

Outro conjunto de dados importante para o projeto é o dos dados que retratam a evolução da pandemia, dados estes que são diariamente reportados pela Direção Geral de Saúde, órgão máximo da área de saúde em Portugal.

Numa primeira instância tinha-se pensado em fazer WebScraping do Dashboard da DGS, contudo ele apenas possui os valores do dia, ou seja, não tem o histórico de dados da evolução da pandemia. De seguida pensou-se em fazer Web Scraping dos relatórios diários publicados pela DGS, contudo são mais de 300 relatórios, tendo eles diferentes formatos e vários erros. Então optou-se pela opção mais fácil, recorreu-se a um repositório, *GitHub - Dados relativos à pandemia COVID-19 em Portugal 2021*, repositório esse que possui ficheiros CSV onde os autores diariamente atualiza os dados consoante os dados que são divulgados pela DGS. Os autores tiveram também o cuidado de armazenar os dados desde o início da pandemia, corrigindo também todas as incongruências que os dados possuem.

As variáveis extraídas são as seguintes:

- **Data** - Data a que os dados se referem;
- **Total Confirmados** - Número de confirmados em Portugal;
- **Total Regional Confirmados** - Uma variável para cada região, contendo o respetivo total de confirmados;
- **Confirmados Diário** - Número de confirmados no dia anterior dos dados;
- **Confirmados Total por género** - Uma variável para cada género, contendo o total confirmados em Portugal;
- **Confirmados Total por faixa etária e género** - Uma variável para cada faixa etária e género, contendo o total confirmados em Portugal (por exemplo, número de indivíduos masculinos confirmados entre os 0 e 9 anos);
- **Total Recuperados** - Número total de recuperados em Portugal;
- **Total Recuperados Regional**- Uma variável para cada região, contendo o respetivo total de recuperados (Variáveis que estão vazias porque a DGS não divulga o número de recuperados de cada região);
- **Total Óbitos** - Número total de óbitos em Portugal;

- **Total Óbitos Regional**- Uma variável para cada região, contendo o respetivo total de óbitos;
- **Óbitos Total por género** - Uma variável para cada género, contendo o total de óbitos em Portugal;
- **Óbitos Total por faixa etária e género** - Uma variável para cada faixa etária e género, contendo o total de óbitos em Portugal (por exemplo, número de indivíduos masculinos que faleceram entre os 0 e 9 anos);
- **Internados** - Número total de internados em Portugal;
- **Internados UCI**- Número total de internados em unidades de cuidados intensivos em Portugal;
- **Vigilância**- Número de casos sob vigilância pelas autoridades de saúde;
- **Ativos** - Número de casos ativos em Portugal;

O primeiro passo foi carregar os dados, fornecendo á função *read.csv* o endereço dos dados. Após os dados serem carregados foram criadas novas variáveis que retratam a evolução diária do Covid-19, ou seja, através das variáveis totais (já que apenas os dados de confirmados total possui um valor diário) criou-se variáveis diárias.

Observando a tabela 5.3 podemos verificar de que forma os dados ficaram organizados, ou seja, os dados ficarão organizados segundo a data e região a que pertencem, contudo existem variáveis globais, que retratam a evolução do covid-19 em Portugal, que se repetirão de acordo com a data, já que elas não estão ligadas á região. Por exemplo a coluna "Ativos" representa o total de casos ativos em Portugal, sendo o seu valor o mesmo sempre que a data é "2020-03-10".

Outro aspeto que é possível observar na tabela 5.3 é a criação das variáveis diárias, observando a coluna "Ativos Total" é possível observar que na data "2020-03-09" era de 39 casos ativos e que no dia seguinte passou a ser de 41, o que corresponde a um aumento de 2 novos casos ativos. Através desta diferença obteve-se a nova variável "Ativos Diário", que retrata os casos ativos diários. O excerto de código B.6, presente no anexo B, representa a criação de alguma destas variáveis, através de um ciclo.

Tabela 5.3: Primeiras linhas dos dados sobre evolução de Covid-19

Datas	Região	Reg. Confir- mados Diário	Reg. Confir- mados Total	Ativos Total	Ativos Diário
2020-03-09	R.A Madeira	0	0	39	1
2020-03-10	Alentejo	0	0	41	2
2020-03-10	Algarve	0	2	41	2
2020-03-10	Centro	1	2	41	2
2020-03-10	A.M. Lisboa	1	10	41	2
2020-03-10	Norte	0	27	41	2
2020-03-10	R.A. Açores	0	0	41	2
2020-03-10	R.A Madeira	0	0	41	2

Foi necessário criar um ciclo para armazenar estes dados na tabela auxiliar á geração de previsões, ciclo este que percorre as duas tabelas de forma a identificar em que linhas são

iguais, ou seja, tem como objetivo identificar onde é que as tabelas possuem uma data e uma região iguais. Sempre que estas condições se verifiquem, o ciclo passa os dados para a linha da tabela principal que eles representam. O excerto de código B.7, presente no anexo B, representa a transferência de dados para a tabela principal, através de um ciclo.

Mortalidade

De forma a ter mais dados sobre os óbitos em Portugal, recorreu-se à plataforma de Vigilância da mortalidade em Portugal (*Sistema de Informação dos Certificados de Óbito - Vigilância de Mortalidade* 2021), de onde se retirou as seguintes variáveis:

- **Mortes Registadas** - mortes registadas num dia e local específico independentemente da causa;
- **Mortalidade Base** - média de mortes esperadas num dia e local específico independentemente da causa;
- **Excesso de Mortalidade** - excesso de mortes num dia e local;

Medida de Atenuação do Covid-19

Outro conjunto de dados que poderá ser importante para as previsões, são as medidas de atenuação do covid-19, medidas estas que, por norma, são impostas pelo governo. Como fonte de dados utilizou-se o CoronaNet, descrito em 2.4.1 e de forma a complementar estes dados utilizou-se outra fonte de dados, fonte de dados essa que consiste numa base de dados de medidas criada pela universidade de Oxford (*COVID-19 GOVERNMENT RESPONSE TRACKER* 2021). Esta base de dados está disponível no repositório *Github - Oxford Covid-19 Government Response Tracker (OxCGRT)* 2021. As medidas obtidas destas fontes encontram-se descritos no anexo B.1.1.

Com estas variáveis criou-se a tabela com medidas que foram implementadas, contendo as datas de início e fim de cada medida, por exemplo as escolas do 2º e 3º ciclos estiveram fechadas de 2021-01-22 a 2021-04-05. De forma passar esta informação para a tabela principal, criou-se um ciclo que preenche, por exemplo os valores "Fechada" entre o início e o fim desta medida na coluna correspondente. Este ciclo preenche todas as medidas extraídas e um excerto de código está disponível em B.8, tendo em conta que no excerto existem linhas que correspondem a correções nos dados. Estes dados estão orientados apenas em relação à data visto que correspondem a medidas implementadas em Portugal e desta forma afetam todas as regiões. Estes dados foram também adicionados à tabela auxiliar de geração de previsões.

5.1.2 Exploração dos Dados

Nesta secção serão feitas várias análises aos dados recolhidos, com recurso a tabelas e gráficos. O conjunto de dados no total possui 2772 linhas e cerca de 109 colunas, o que representa dados das 7 regiões em 396 dias, desde 2020-03-01 até 2021-03-31.

A tabela 5.4 representa algumas das características das regiões de Portugal, através dela é possível observar que as regiões com maior área são o Alentejo e o Centro, já as menores são as Regiões Autónomas de Madeira e dos Açores. É possível observar que a densidade populacional da Área Metropolitana de Lisboa é muito superior às restantes, tendo uma densidade de 946.8 enquanto a segunda maior densidade é de 317. Em termos de índice

de envelhecimento, é possível observar que Portugal como um todo possui uma população envelhecida, destaque para as regiões do Alentejo e Centro que possuem índices acima dos 200, esta constatação é suportada pela relação entre a população total e idosa das regiões. Por fim, em relação ao turismo, é possível observar que as regiões do Algarve, Lisboa e Madeira são mais procuradas pelos turistas.

Tabela 5.4: Características das Regiões.

Regiao	Área	D. Populacio- nal	I.Envelhecimento	P. Total	P. Idosa	Turismo
Alentejo	31605	22.3	204.6	705018	179861	34.1%
Algarve	4997	87.8	145.4	438635	95503	70.9%
Á.M. Lisboa	3015	946.8	138.2	2854802	626946	72.9%
Centro	28199	78.6	201.4	2216927	541464	39.7%
Norte	21286	167.9	162.7	3573961	739987	52.8%
R.A. Ma- deira	801	317.0	126.5	254100	42738	78.2%
R.A. Açores	2322	104.6	95.1	242821	35827	49.6%

Neste espaço de 396 dias, o país passou 189 deles em estado de Emergência, 190 dias em estado de Calamidade e 18 em Normalidade. Na tabela 5.6 é possível observar o número de dias e em que estado esteve o país em cada mês, sendo possível observar que país esteve em estado de calamidade nos meses mais quentes, ou seja, desde metade da primavera até metade de outono, consequentemente este no estado de emergência nos meses mais frios, ou seja, desde a segunda metade do outono até a meio da primavera. A possível, assim estabelecer uma ligação entre o tempo mais quente e o estado do país. Examinando a coluna dos casos confirmados por mês, é possível retirar que nos meses mais frios existem muito mais casos do que nos meses quentes, um exemplo disso é o mês de janeiro que teve 306.838 casos enquanto o mês de agosto teve menos de um terço disso (6940), o mesmo pode ser observado no número de óbitos.

Através da tabela 5.5 é possível concluir que Portugal como um todo é um país com um ar bastante húmido. É possível também concluir que as regiões do Norte e Centro possuem temperaturas mais baixas durante o ano do que as restantes regiões do país, sendo o Algarve e a Madeira as zonas mais quentes do país.

Tabela 5.5: Médias das condições meteorológicas por região.

Região	T. Máxima	T. Média	T. Mínima	V. Velocidade	Humidade
Alentejo	22.7	16.5	10.7	21.5	73.3
Algarve	21.5	18.2	15	23.8	71.3
Lisboa	21.4	17	13.8	21.6	72.8
Centro	19.9	15	11.2	19.0	77.7
Norte	19.0	13.8	9.5	16.5	73.2
Madeira	22.7	19.8	17.6	12.9	66.6
Açores	19.5	17.3	15.4	30.1	79.5

Tabela 5.6: Estado do país ao longo do ano.

Mês	Normali- dade	Calamidade	Emergência	Confirmados	Óbitos
Janeiro	0	0	31	306838	5576
Fevereiro	0	0	28	84046	3835
Março	17	0	45	24603	691
Abril	0	0	30	17249	829
Maio	0	29	2	7808	421
Junho	0	30	0	9671	166
Julho	0	31	0	8901	159
Agosto	0	31	0	6940	87
Setembro	0	30	0	17530	149
Outubro	0	31	0	65737	536
Novembro	0	8	22	156782	1998
Dezembro	0	0	31	115617	2401

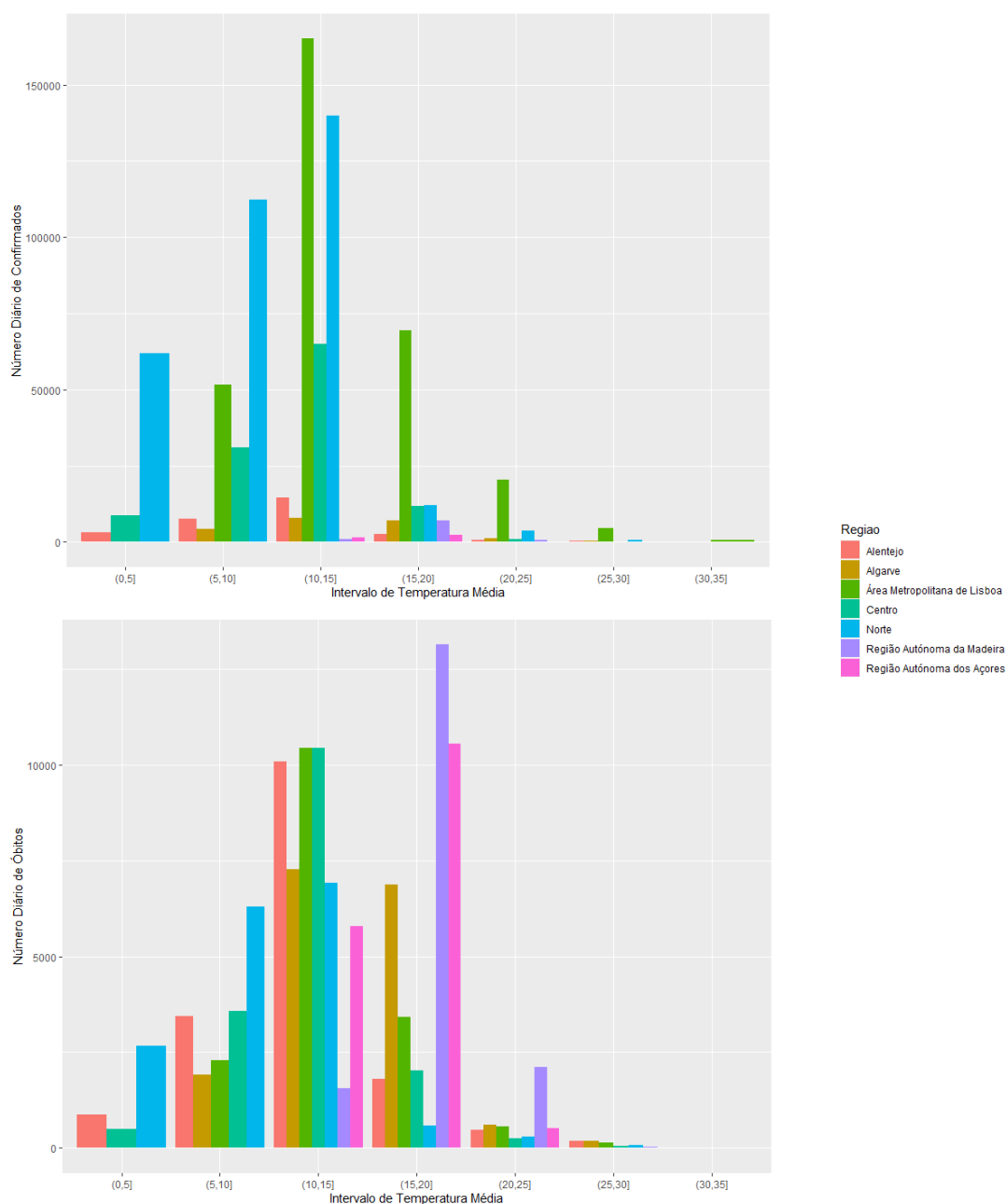


Figura 5.5: Gráfico de cima: Número de confirmados num determinado intervalo de temperatura média.

Gráfico de baixo: Número de Óbitos num determinado intervalo de temperatura média.

A figura 5.5 representa as relações de diferentes intervalos de temperatura média com o número de confirmados e com o número de óbitos nas diferentes regiões. Observando a figura 5.5 é possível analisar que as únicas regiões que registaram, tanto ao nível de confirmados como de óbitos, no intervalo de temperaturas entre 0°C e 5°C foram o Centro, Norte e Alentejo, isso deve-se, possivelmente, ao facto de que a temperatura média não chegou a este intervalo nas outras regiões, que são mais quentes.

Atentando em ambos os gráficos se verifica que a grande maioria das ocorrências entre

os 5°C e os 20°C, enquanto nas gamas entre 20°C e os 35°C se encontram em menor quantidade. Em todos os intervalos de temperatura no gráfico dos casos confirmados as regiões de Lisboa e Norte são as que têm mais casos, este facto poderá estar relacionado com a grande densidade destas zonas do país.

O número de casos nas regiões autónomas é muito inferior às restantes regiões, contudo ao nível dos óbitos no intervalo 15°C e 20°C o oposto se verifica.

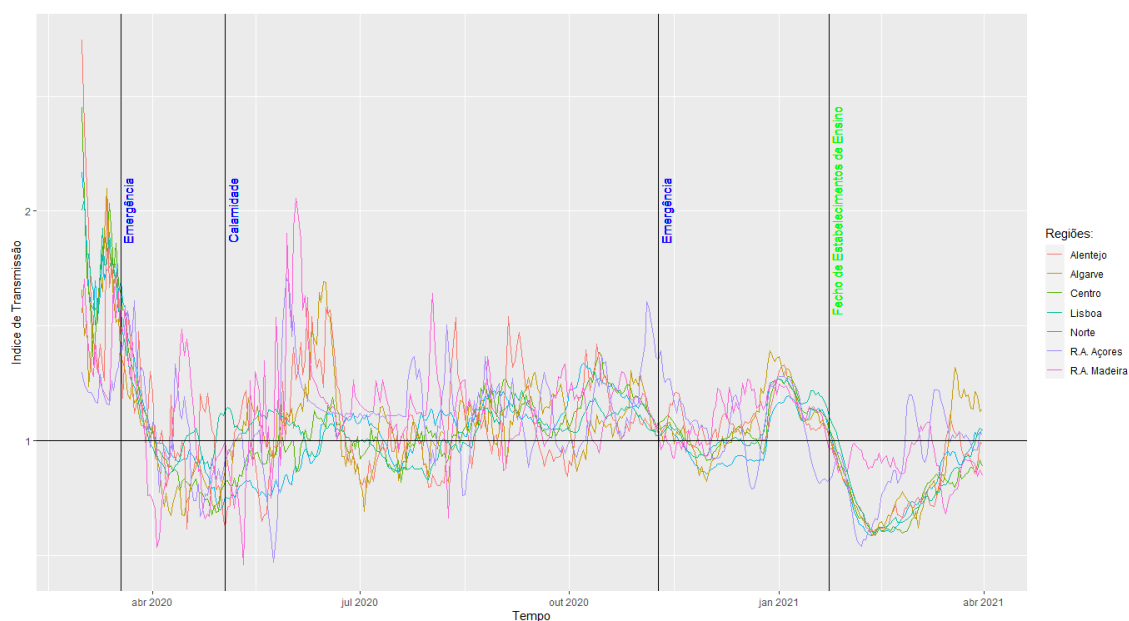


Figura 5.6: Evolução do Índice de Transmissão de cada região.

A figura 5.6 representa a evolução dos índices de transmissibilidade de cada região ao longo do intervalo de dados. Observando atentamente o gráfico, com foco especial nas linhas verticais, é possível verificar que quando o país entra em estado de Emergência (estado este que, traz consigo medidas mais duras) é eficaz para baixar os índices de transmissibilidade, o mesmo acontece na linha vertical de "Fecho de estabelecimentos de ensino" que representa mais do que o fecho das escolas, representa um endurecimento extra às medidas que o normal estado de emergência acarretava até então.

É observável também que nos meses com menores números de casos confirmados, é o período de tempo que se verificam mais ocorrências de índices abaixo do 1 comparativamente a meses com maiores números de casos (por exemplo, novembro).

Outro aspeto possível de extrair do gráfico é que o índice de transmissibilidade da região Norte é o índice com menor variância, ou seja, o seu valor mantém-se próximo de 1 sem nunca ter subidas muito bruscas, ao contrário de por exemplo a Região Autónoma da Madeira em maio de 2020. Este facto pode significar que as características das regiões, como por exemplo a densidade populacional, poderão ter maior impacto na transmissão do vírus do que o próprio índice de transmissibilidade, já que a região Norte é sempre das regiões com maior número de casos confirmados e é das regiões que possui o índice mais controlado.

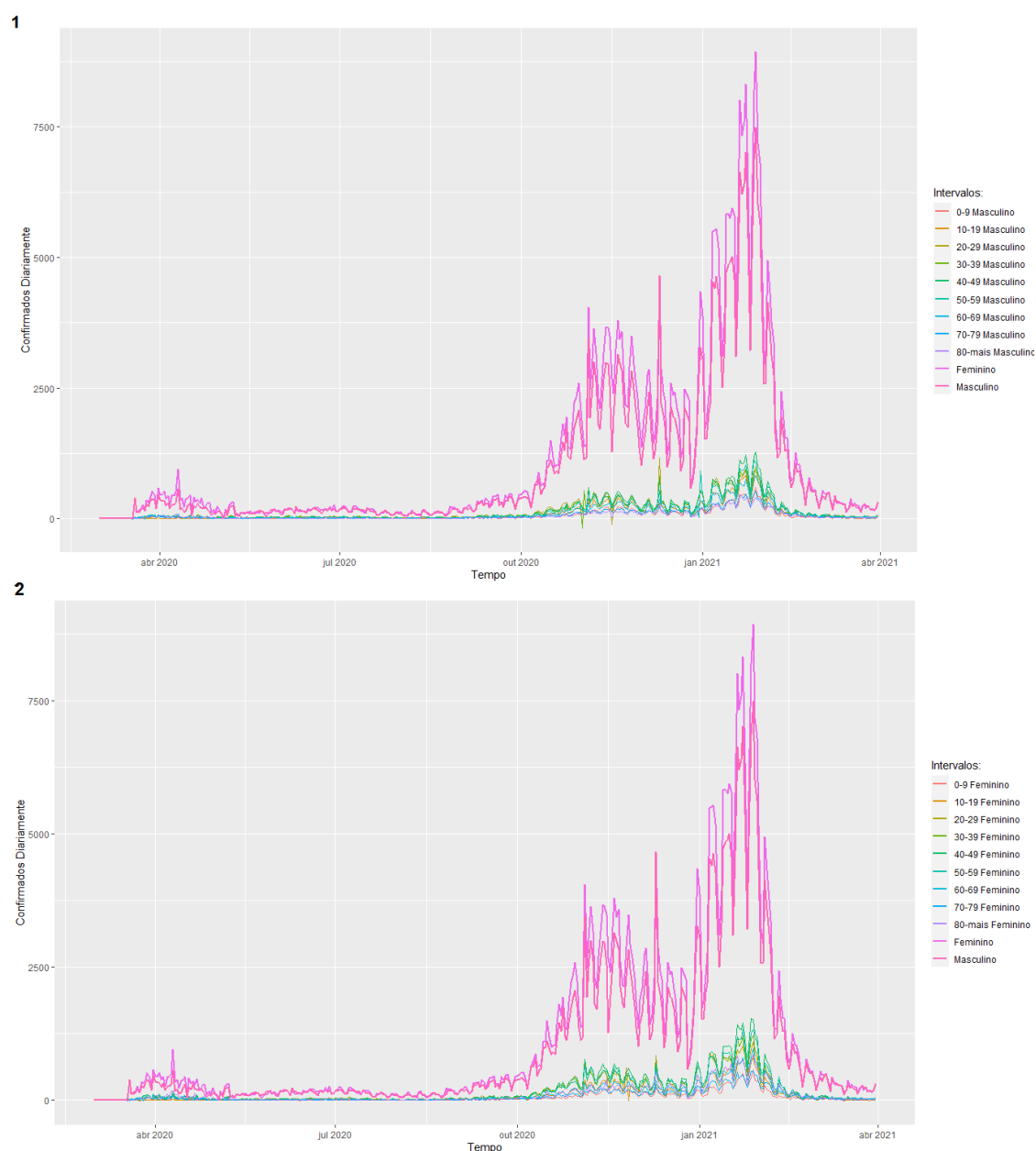


Figura 5.7: Evolução do número de confirmados diários por faixas etárias e sexo.

A figura 5.7 é constituída por 2 gráficos, no gráfico 1 estão presentes as evoluções do número de casos confirmados por faixa etária do sexo masculino e no gráfico 2 estão presentes as evoluções do número de casos confirmados por faixa etária do sexo feminino. Em ambos os gráficos estão também presentes duas linhas gerais que representam o sexo feminino e masculino. Atentando nas linhas gerais é visível que o sexo feminino possui, ao longo de quase todo o gráfico, um maior número de casos confirmados, o que pode ser justificado com base nos dados obtidos das regiões em que existe uma diferença entre a população feminina e masculina de quase 1 milhão, sendo a população feminina superior à masculina em todas as regiões.

Em ambos os gráficos é visível que as faixas etárias dos 0 aos 9 e as faixas com mais de 70

anos possuem, ligeiramente, menos ocorrências de casos comparativamente com as faixas dos 10 aos 59, isto deve-se provavelmente devido á atividade profissional.

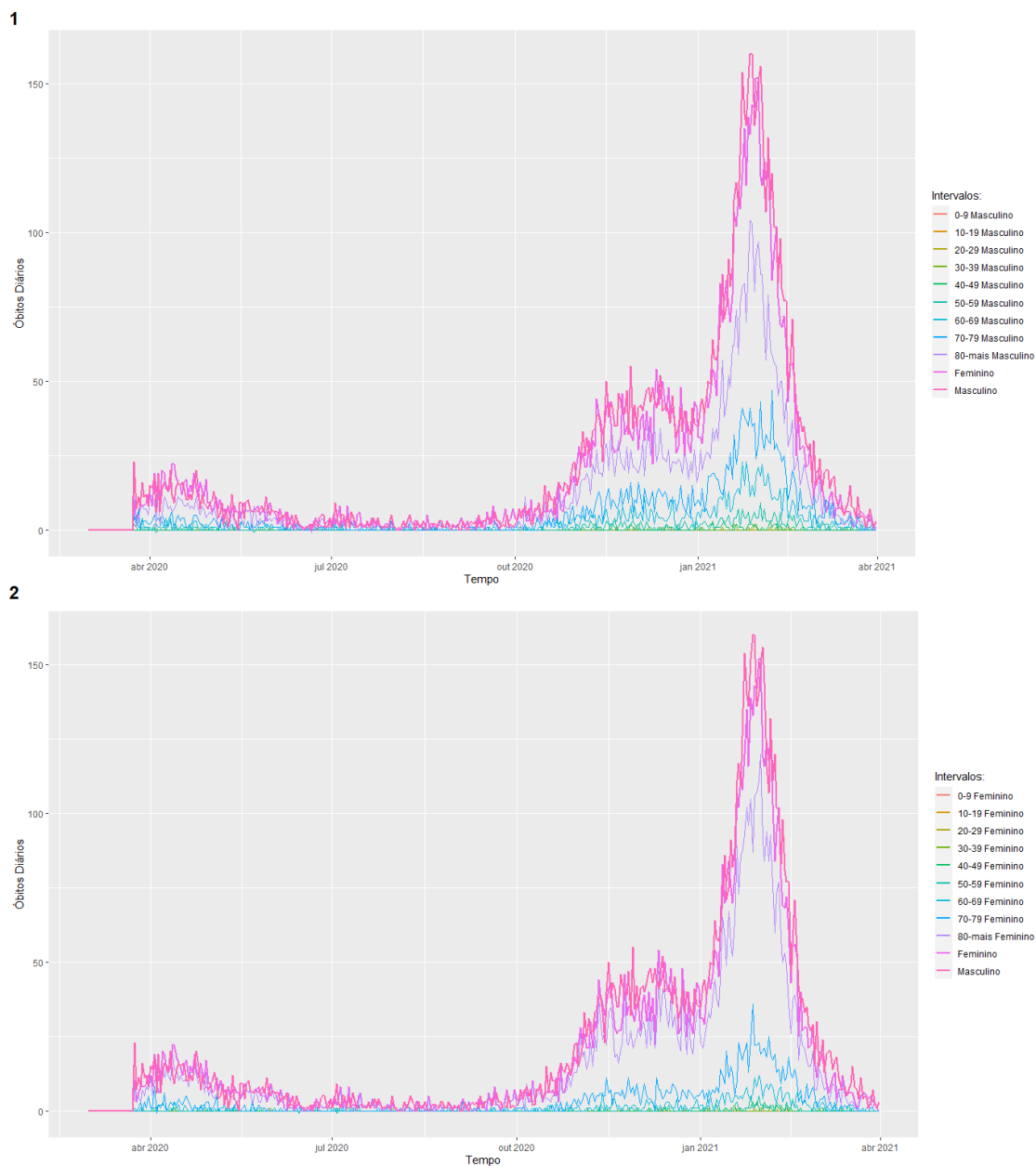


Figura 5.8: Evolução do número de óbitos diários por faixas etárias e sexo.

A figura 5.8 é constituída por 2 gráficos, no gráfico 1 estão presentes as evoluções do número de casos óbitos por faixa etária do sexo masculino e no gráfico 2 estão presentes as evoluções do número de casos óbitos por faixa etária do sexo feminino. Em ambos os gráficos estão também presentes duas linhas gerais que representam o sexo feminino e masculino. Atentando nas linhas gerais é possível analisar que muitas vezes o sexo feminino possui mais óbitos, a justificação para este facto já foi anteriormente descrita.

Ao contrário dos gráficos da figura 5.7 em que as diferentes faixas etárias tinham ligeiras diferenças, no caso dos gráficos da figura 5.8 existem diferenças amplas.

No gráfico 2 a faixa etária com mais de 80 anos é a faixa com maior número de ocorrências de óbitos e representa, quase sempre, mais de metade dos óbitos femininos. De seguida as faixas etárias dos 70 aos 79 e 60 aos 69 são as que possuem maior número de óbitos diários, enquanto as restantes faixas etárias possuem pouquíssimos casos diários. Já no gráfico 1 a faixa etária dos 80 mais é, mais uma vez, a que possui maior número de óbitos ao longo do gráfico, sendo possível verificar que as restantes faixas etárias possuem mais ocorrências do que no gráfico feminino. Observando os gráficos da figura 5.8 e aos dados que nos indicam que, ao dia 2021-03-31, que o número de falecidos do sexo masculino são 8845 enquanto que os do sexo feminino são de 8003, é possível concluir que esta doença pode ser mais mortal para faixas etárias mais idosos por serem indivíduos mais frágeis, assim como poderá ter maior impacto no sexo masculino, que mesmo tendo um menor número de indivíduos que o sexo feminino, possui um número maior de óbitos.

5.1.3 Crítica à Base de Dados

Os dados que caracterizam as regiões, possuem dados muito interessantes e muitos deles não foram utilizados, mas estes dados não estão atualizados, já que muitos são do ano de 2019, uma solução para estes dados ultrapassados seria utilizar os dados dos Censos de 2021. Outro problema com os dados das regiões é a diferença entre as regiões usadas pela DGS e os dados que caracterizam as regiões, ou seja, nos dados das regiões existe uma região chamada "Área Metropolitana de Lisboa" que abrange apenas o distrito de Lisboa, enquanto a DGS utiliza a denominação "Lisboa e Vale do Tejo" que é constituída pelo distrito de Lisboa e um pouco da área do distrito de Setúbal, uma solução para resolver este problema seria utilizar dados dos concelhos no lugar das regiões para que não houvesse esta discrepância.

Relativamente aos dados meteorológicos, talvez a opção de utilizar os dados de uma cidade no centro da região não seja a melhor forma de representar a região, talvez a média de várias cidades da região representasse melhor a região. Outro entrave é a própria fonte de dados que limite os resultados dos pedidos, em que cada pedido só pede ter 100 resultados e existe um limite de 1000 resultados diários, a solução seria optar por outra fonte de dados ou pagar pelo serviço ilimitado.

Os dados que retratam a evolução do Covid-19 possuem alguns erros referentes a correções feitas pela DGS, em dias em que por exemplo alguns laboratórios não reportaram o número de casos confirmados. Outro problema dos dados é o facto da DGS não disponibilizar os dados do número de recuperados e casos ativos das regiões, impossibilitando um dos objetivos do projeto, que seria prever estes dados. Não existe, até á data, forma de contornar estes problemas já que todas as soluções encontradas se baseiam nos dados disponibilizados pela DGS. Enquanto noutros países os dados são todos fornecidos ao público e constantemente atualizados como o caso do Github de Itália, Portugal tem apenas os relatórios diários em formato de PDF, de onde é difícil extrair os dados. Relatório este que durante meses foi alternado o seu formato, quase de forma diária, sem manter as mesmas variáveis e/ou um dicionário que explicasse o significado das mesmas. Tudo isto leva a que os dados da evolução do covid-19 não sejam 100% confiáveis.

5.2 Modelos

Esta seção descreve tudo o que envolve os modelos de Machine Learning, desde a transformação à seleção dos modelos.

5.2.1 Transformação de dados

É um facto de que nem todos os modelos tem perceção da evolução do tempo, ou seja, nem todos têm a capacidade de compreender o passar do tempo e como é que os elementos anteriores afetam um determinado elemento.

A solução encontrada para este problema foi forçar esta compreensão temporal, através da transformação dos dados, onde o objetivo era que cada linha tivesse dados dos dias anteriores.

Esta transformação é obtida através da linha de código presente no excerto 5.1, que cria novas colunas através da deslocação da coluna para cima.

```
1 treino<-cbind(treino , treino[, shift(.SD, 1:5, NA, "lag", TRUE), .SDcols
  =1:5])
```

Listing 5.1: Código que permite a transformação dos dados (R).

A tabela 5.7 representa um exemplo desta transformação, em que na primeira coluna está presente o valor do índice de transmissibilidade no dia que a linha representa, a segunda coluna representa o valor do índice no dia anterior, da mesma forma que a última coluna, representa o valor do quarto dia anterior.

Tabela 5.7: Médias das condições meteorológicas por região.

Rt	Rt_lag_1	Rt_lag_2	Rt_lag_3	Rt_lag_4
0.9788019	0.9909465	0.9884942	0.9888724	1.01137.96
0.9711932	0.9788019	0.9909465	0.9884942	0.9888724
0.9418180	0.9711932	0.9788019	0.9909465	0.9884942

Desta forma os dados transformados estão num formato adequado ao processo e exploração utilizando modelos, já que não irão analisar linhas anteriores, mas sim as variáveis que se repetem representando os dias anteriores.

5.2.2 Metodologia de Testes

Nesta subsecção serão apresentados os testes pelas quais os modelos passarão, de forma a obter o melhor modelo para cada tipo de previsão. No caso dos testes com previsões de casos confirmados e óbitos optou-se por utilizar apenas a região Norte devido ao baixo poderio computacional e esta escolha deve-se às características da região. Quanto á previsão de recuperados, devido á falta de dados, apenas é possível fazer previsões de recuperados em Portugal, quanto aos casos ativos optou-se por não realizar nenhum teste com este tipo de previsão, já que pode ser obtida através de um cálculo com os outros tipos de previsões.

A figura 5.9 compõe um resumo dos testes que serão executados, existem 2 tipos de teste diferentes:

1. Testes identificadores de caraterísticas: testes que na figura 5.9 possuem cubos cor-de-rosa (que representam os modelos), estes testes tem o objetivo de identificar as caraterísticas que traduzem melhores resultados, ou seja, identificar para cada tipo de previsão e modelo, as melhores caraterísticas;

2. Teste para seleção de modelos: teste que na figura 5.9 está representado com cubos dourados, tem como objetivo, após serem identificadas as melhores características, identificar o melhor modelo para cada tipo de previsão.

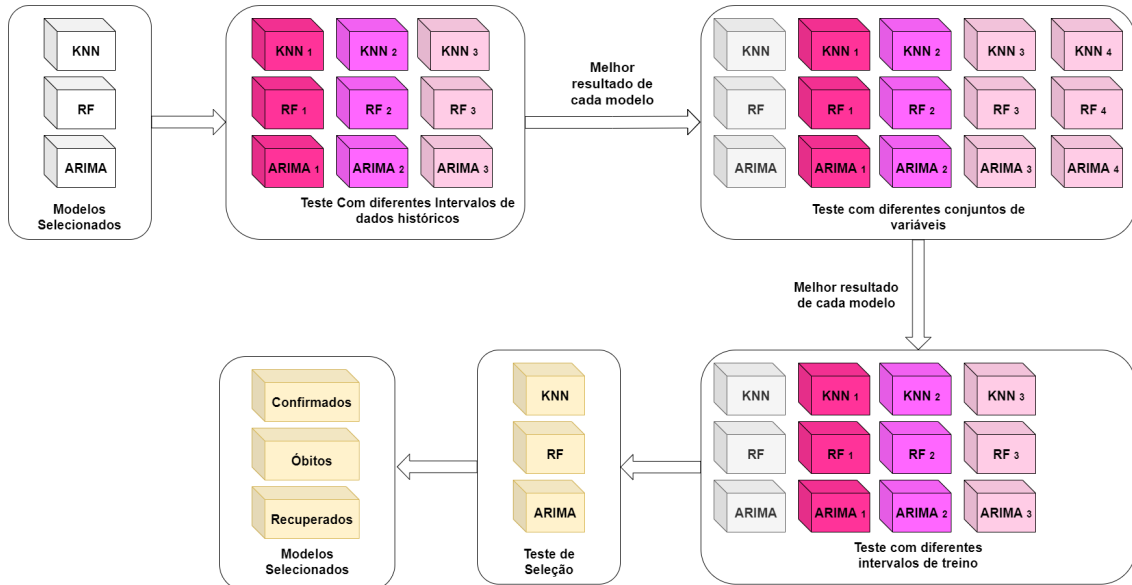


Figura 5.9: Fluxo dos testes.

Os modelos utilizados neste testes são o Random Forest, modelo abordado em 2.2.3 e 2.3.3, K-Nearest Neighbors, que foi abordado em 2.2.3 e 2.3.3 e Autoregressive Integrated Moving Average (ARIMA) - modelo que inclui os modelos Autorregressive (AR) e o modelo Moving Avera (MA). Esta opção deriva do estudo do estado de arte 2 Embora as redes neuronais tenham sido bastante mencionadas anteriormente, não serão utilizadas neste projeto, sendo mencionadas como possível trabalho futuro.

Já as métricas avaliadoras utilizadas dependerão do tipo de teste.

5.2.3 Testes identificadores de caraterísticas

O fluxo 5.9 será repetido para cada tipo de previsão, onde os modelos seleccionados irão passar pelos seguintes testes:

- **Teste de desempenho com diferentes quantidades de dados históricos** - testar diferentes conjuntos de dados passados, ou seja, testar diferentes conjuntos de dias anteriores ao dia da previsão, com o objetivo de identificar qual fornece ao modelo um melhor desempenho;
- **Teste de desempenho com diferentes conjuntos de variáveis** - testar o desempenhos dos modelos adicionando-lhes novas variáveis relacionadas com a variável dependente, com o objetivo de identificar qual fornece ao modelo um melhor desempenho;
- **Teste de desempenho com diferentes intervalos de treino** - testar fornecer ao modelo diferentes quantidades de dados de treino, com o objetivo de identificar qual fornece ao modelo um melhor desempenho;

O ideal seria realizar estes testes em todos os modelos como combinações de vários intervalos de dados históricos, vários intervalos de dados de treino e várias variáveis, com o objetivo de

encontrar o conjunto de características que tem o melhor desempenho para cada uma das previsões. Este elevado número de testes requer um enorme poder computacional, já que estas diferentes características (número de variáveis utilizadas, número de dias do passado e intervalo de dados de treino) podem fazer o volume de dados aumentar exponencialmente, assim como o tempo que demora a execução de uma previsão. Desta forma, devido a um problema de baixo poderio computacional, o número de testes será reduzido.

Estes testes, para identificar as características que fornecem ao modelo um melhor desempenho, serão feitos de forma a que os modelos tentem prever o valor de $N+1$, ou seja, os modelos tem como objetivo prever sempre o dia seguinte ao dia atual.

A tabela 5.8 serve para clarificar como funcionam este tipo de testes, os modelos irão utilizar os dados do dia atual N e dos dias anteriores, no caso do exemplo são 3 dias de dados anteriores, de forma a prever o próximo elemento. No caso do exemplo são previstos 3 dias, mas no caso dos testes que futuramente serão executados serão previstos 30 dias, sempre utilizando os dados dos dias anteriores. Este período de previsões será de dia 01/03/2021 até dia 31/03/2021.

Tabela 5.8: Exemplo de como funcionam os testes identificadores de características.

N-2	N-1	N	Previsão: $N+1$
02/03/2021	03/03/2021	04/03/2021	05/03/2021
03/03/2021	04/03/2021	05/03/2021	06/03/2021
04/03/2021	05/03/2021	06/03/2021	07/03/2021

Estes testes serão avaliados segundo a comparação entre o total previsto no período de 30 dias em questão e o total real, assim como através do erro médio, do erro médio absoluto, MSE, RMSE e R^2 .

Teste de desempenho com diferentes quantidades de dados históricos

Neste teste, tem como base testar diferentes conjuntos de dados históricos, com o objetivo de identificar para cada modelo e tipo de previsão qual fornece melhor desempenho.

A tabela exemplar 5.8 é um exemplo de teste utilizando 3 dias como dados históricos, ou seja usando os 3 dias anteriores ao dia que se pretende prever para prever esse mesmo dia. Neste teste será testado o desempenho dos modelos com 3 dias de dados históricos, 5 dias de dados históricos e 7 dias de dados históricos.

Teste de desempenho com diferentes conjuntos de variáveis

Este teste terá como objetivo identificar com que conjuntos de variáveis os modelos produzem melhores resultados. Para definir estes conjuntos de variáveis recorreu-se ao estudo do estado da arte 2, onde se levantam várias hipóteses, que serão utilizadas nestes testes, tais como:

- **Hipóteses para previsão de Confirmados**

- Variáveis meteorológicas - como foi visto em 2.3.3, existe a possibilidade de que o vírus teria menos efeitos nos períodos de temperaturas quentes, desta forma

será testada a hipótese de as variáveis temporais melhorarem o desempenho dos modelos;

- Variáveis de atenuação pandémica - existem várias medidas que tentam atenuar os efeitos do vírus, como as que estão presentes em 5.1.1. Ao utilizar estas variáveis será testada a hipótese de as variáveis atenuadoras dos efeitos do vírus melhorarem o desempenho dos modelos;
- Variáveis de maior correlação - com esta hipótese será testar se as 4 variáveis com maior relação com a variável que se pretende prever terão um bom desempenho;

- **Hipóteses para previsão de Óbitos**

- Variáveis sobre mortalidade - no conjunto de dados estão presentes os dados sobre a mortalidade geral em Portugal, nesta hipótese serão utilizadas de forma a avaliar se o modelo terá um bom desempenho com as mesmas;
- Variáveis sobre internamentos - existe a hipótese de que a maioria dos óbitos de covid-19 acontecem nos internamentos, desta forma serão utilizadas as variáveis sobre internamento para avaliar se o modelo terá um bom desempenho com as mesmas;
- Variáveis de maior correlação - com esta hipótese será testar se as 4 variáveis com maior relação com a variável que se pretende prever terão um bom desempenho;

- **Hipóteses para previsão de Recuperados**

- Variáveis sobre internamentos - existe a hipótese de que os recuperados de covid-19 estão diretamente ligados aos internamentos, desta forma serão utilizadas as variáveis sobre internamento para avaliar se o modelo terá um bom desempenho com as mesmas;
- Variáveis sobre atualidade - existe a hipótese de que os recuperados de covid-19 estão diretamente ligados ao número de casos ativos e confirmados, desta forma serão utilizadas as variáveis sobre a situação pandémica atual para avaliar se o modelo terá um bom desempenho com as mesmas;
- Variáveis de maior correlação - com esta hipótese será testar se as 4 variáveis com maior relação com a variável que se pretende prever terão um bom desempenho;

Como foi dito anteriormente para cada tipo de previsão, uma das hipóteses é utilizar as 4 variáveis que mais se relacionam com a variável que se pretende prever. Esta relação designa-se por correlação, que se traduz no grau ou força da relação entre duas variáveis, podendo esta correlação ter valores entre -1 e 1, o sinal do valor da correlação significa se a relação é positiva ou negativa (Taylor 1990). O excerto de código B.10 em anexo representa o cálculo da correlação para cada variável que se pretende prever. O primeiro passo foi converter algumas colunas de caractere para numérico, já que esta medida avalia a relação entre variáveis numéricas. De seguida utilizou-se a função *cor* para calcular a correlação, nestes cálculos retirou-se algumas variáveis como por exemplo a data e variáveis que não tinham variância, ou seja, eram constantes já que não produziam nenhuma correlação.

A figura 5.10 representa as 4 variáveis com maior correlação para cada tipo de previsão. Para os testes de previsão dos confirmados diários que utilizem as variáveis de maior correlação, será utilizada a variável que representa o número diário de casos confirmados em Portugal, o número diário de óbitos da região, o número diário de mortes registadas na

região independentemente da causa e o número de casos ativos de Portugal, que são as 4 variáveis de maior correlação. Para os testes previsão dos óbitos diários que utilizem as variáveis de maior correlação, será utilizada a variável que representa o número de internados em Portugal, o número diário de óbitos em Portugal, número de casos ativos de Portugal e o número de internados na UCI em Portugal, que são as 4 variáveis de maior correlação. Quanto aos testes de previsão dos recuperados que utilizem as variáveis de maior correlação será utilizada a variável que representa o número de internados em Portugal, número de casos ativos de Portugal, o número diário de óbitos em Portugal e o número de indivíduos em vigilância em Portugal, que são as 4 variáveis de maior correlação.

	Cor-Confirmados		Cor-Óbitos		Cor-Recuperados
Confirmados_Total_Diario	0.919471040	Internados	0.892973752	Internados	0.86501175
RegDiario_Obitos	0.819350057	Obitos_Diario	0.871268326	Ativos_Total	0.85707502
RegMortes	0.748484086	Ativos_Total	0.860769144	Obitos_Diario	0.84821391
Ativos_Total	0.741154286	InternadosUCI	0.854921936	Vigilancia	0.83497831

Figura 5.10: Correlação entre as variáveis que se pretende prever e as restantes.

Em suma, para cada tipo de previsão as variáveis utilizadas são:

- **Confirmados**

- Variáveis Meteorológicas - temperatura média (regional), precipitação (regional), humidade (regional) e velocidade do vento (regional);
- Variáveis de atenuação pandémica - estado do país (Portugal), trabalho (Portugal), chegada de estrangeiros (Portugal) e estratégia de testagem (Portugal);
- Variáveis de maior correlação - confirmados diários (Portugal), óbitos diários (regional), mortes registadas (regional), total ativos (Portugal);

- **Óbitos**

- Variáveis sobre mortalidade - mortes registadas (regional), mortalidade base (regional) e excesso de mortalidade (regional);
- Variáveis sobre internamentos - total de internados (Portugal), total de internados nos UCI (Portugal), internamentos diários (Portugal) e internamentos diários UCI (Portugal);
- Variáveis de maior correlação - internados (Portugal), óbitos diários (Portugal), total ativos (Portugal), internados UCI (Portugal),

- **Recuperados**

- Variáveis sobre internamentos - total de internados (Portugal), total de internados nos UCI (Portugal), internamentos diários (Portugal) e internamentos diários UCI (Portugal);
- Variáveis sobre atualidade - confirmados diários (Portugal), total de ativos (Portugal), ativos diários (Portugal) e total de confirmados (Portugal);
- Variáveis de maior correlação - internados (Portugal), total ativos (Portugal), óbitos diários (Portugal) e vigilância (Portugal);

Teste de Teste de desempenho com diferentes intervalos de treino

Este teste tem como objetivo identificar com que conjuntos de treino os modelos têm melhor desempenho são eles:

- 1 ano antes das previsões: os modelos treinam com dados que estão num intervalo de 1 ano antes da primeira previsão, ou seja, de 01/03/2020 a 28/02/2021.
- 9 meses antes das previsões: os modelos treinam com dados que estão num intervalo de 9 meses antes da primeira previsão, ou seja, de 01/06/2020 a 28/02/2021.
- 6 meses antes das previsões: os modelos treinam com dados que estão num intervalo de 6 meses antes da primeira previsão, ou seja, de 01/09/2020 a 28/02/2021.
- 3 meses antes das previsões: os modelos treinam com dados que estão num intervalo de 3 meses antes da primeira previsão, ou seja, de 01/12/2020 a 28/02/2021.

Fluxo dos Testes

Atentando novamente na figura 5.9 que representa todo o processo que irá ser executado até se encontrar os melhores modelos para cada tipo de previsão. Observado a figura podemos ver que os modelos são representados como cubos e que inicialmente esses cubos estão sem cor visto que ainda não passaram por nenhum conjunto/bateria de testes. O primeiro conjuntos de testes procuram identificar o número de dias de dados históricos que retira o melhor desempenho do modelo. Nesse conjunto de testes, os modelos são testados com 3, 5 e 7 dias de dados históricos sendo cada um destes testes representado por um dos cubos coloridos da área que representa os testes de dados históricos. De seguida, o modelo com melhor desempenho passa para o conjunto de testes seguinte (teste de desempenho com diferentes conjuntos de variáveis) e irá representar o teste sem variáveis já que não se utilizou nenhum conjunto de variáveis para além da que se pretende prever, assim como a característica encontrada (número de dias de dados históricos) que passará a ser sempre utilizada por esse modelo, por exemplo, no teste com diferentes intervalos de dados históricos verifica-se que o modelo KNN tem um melhor desempenho utilizando 7 dias de dados históricos para previsões de confirmados, todos os restantes testes usando o modelo KNN para previsões de confirmados serão utilizados os 7 dias de dados históricos.

No teste com diferentes variáveis, os cubos cinzentos representam os melhores modelos do teste anterior, assim como representam um teste que não utiliza variáveis. Os outros cubos coloridos representam os testes com as 3 hipóteses mencionadas anteriormente que cada tipo de previsão tem, assim como a representação de um quarto teste que reúne as variáveis das 3 hipóteses num teste só, ou seja, utiliza-se as variáveis de todas. O melhor modelo passará para o próximo conjunto de testes (teste de desempenho com diferentes intervalos de treino), onde irá representar o teste com 1 ano de treino, já que até então se utiliza sempre um ano de treino. A característica encontrada, o conjunto de variáveis que produz o melhor desempenho, passará a ser utilizada, tendo em conta que o teste com melhor desempenho poderá ser o teste sem variáveis, por exemplo, verificou-se que o modelo KNN utilizando variáveis meteorológicas (e 7 dias de dados históricos, característica obtida previamente) para previsão de casos confirmados produz um desempenho melhor que os outros conjuntos de variáveis, todos os testes seguintes com o modelo KNN para prever o número de confirmados irão utilizar o conjunto de variáveis meteorológicas (assim como os 7 dias de dados históricos).

No último conjunto de teste testa-se com 9 meses, 6 meses e 3 meses de dados de treino, já que o teste com 1 ano, é o mesmo que o melhor resultado do teste anterior, no final deste teste teremos obtido o conjunto de características que melhor definem os modelos.

Em suma, cada teste identifica uma característica para cada modelo, de forma que no final termos o melhor desempenho possível em cada um dos modelos selecionados. O fluxo representado na figura 5.9 deverá ser repetido para cada tipo de previsão, sendo que este fluxo foi construído de forma a minimizar o tempo de processamento dos modelos, já que poderiam haver várias combinações das características selecionadas, mas devido ao baixo poderio computacional, optou-se por testar isoladamente cada uma.

De realçar ainda a necessidade da normalização dos dados durante os testes com o modelo K-Nearest Neighbors. Esta necessidade deve-se à natureza deste tipo de modelo, já que o modelo se baseia no cálculo das distâncias entre os dados, dados estes que têm que estar na mesma escala, ou contrário do que acontece no caso deste projeto que tem variáveis com grandes escalas, como por exemplo o total de casos confirmados que possui valores que rondam 1 milhão, assim como existem variáveis com valores pequenos como por exemplo o índice de transmissibilidade que ronda valores entre 0 e 2. Esta normalização transforma os dados, de forma a todas as variáveis possuírem valores em uma mesma escala (Pandey e Jain 2017).

Após identificar as características que produzem melhores resultados, o resultado obtido é uma lista de 9 modelos, já que cada um dos modelos selecionados (KNN, RandomForest e ARIMA) terá um conjunto de características para cada tipo de previsão.

Como o objetivo do projeto é prever a evolução do covid-19 nos próximos dias, é necessário testar estes modelos de uma forma diferente, visto que até então foram testados de forma a saberem sempre os dias anteriores e só tinham que prever o próximo dia. Desta forma os modelos irão utilizar as próprias previsões para prever os dias a seguir ao segundo dia, como é possível observar na tabela 5.9 em que na segunda linha se utiliza os dados da previsão anterior. Supondo que um modelo utiliza uma temperatura para gerar previsões, é impossível saber a temperatura que estava no dia da previsão 1, então para executar a previsão 2, a temperatura da previsão 1 será a média dos dias anteriores.

Tabela 5.9: Exemplo de como funcionam os testes de seleção de modelos.

N-2	N-1	N	Previsão:N+1
02/03/2021	03/03/2021	04/03/2021	Previsão 1
03/03/2021	04/03/2021	Previsão 1	Previsão 2
04/03/2021	Previsão 2	Previsão 3	Previsão 3

Nestes testes de seleção serão calculados vários intervalos de previsão diferentes, sempre com o objetivo de prever os próximos 5 dias, de forma a que os modelos sejam avaliados pelo média dos erros que cometem na previsão por cada dia do intervalo.

5.2.4 Execução dos Testes

Nesta seção serão apresentados os resultados de todos os conjuntos de testes, assim como a seleção do melhor modelo para cada tipo de previsão.

Os testes terão sempre uma identificação que agrega um número identificador do teste e uma letra que identifica o tipo de variável que se pretende prever.

Teste - Desempenho com diferentes quantidades de dados históricos

Como foi referido anteriormente este teste tem como objetivo avaliar o desempenho dos modelos perante a alteração do número de dias de dados históricos, sendo as variáveis utilizadas apenas as que se pretende prever, assim como o intervalo de dados de treino é de 1 ano.

A tabela 5.10, representa um resumo dos diferentes testes que serão abordados nesta bateria de testes, tendo a tabela em anexo B.1 uma descrição mais detalhada dos testes.

Tabela 5.10: Testes de desempenho com diferentes quantidades de dados históricos a executar em todos modelos.

Teste	Previsão	Intervalo T.	Variáveis	Histórico
1-C	Confirmados	1 ano	2	3
2-C	Confirmados	1 ano	2	5
3-C	Confirmados	1 ano	2	7
4-O	Óbitos	1 ano	2	3
5-O	Óbitos	1 ano	2	5
6-O	Óbitos	1 ano	2	7
7-R	Recuperados	1 ano	2	3
8-R	Recuperados	1 ano	2	5
9-R	Recuperados	1 ano	2	7

Tabela 5.11: Resultados do teste de desempenho com diferentes quantidades de dados históricos utilizando o modelo Random Forest.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
1-C	4369	8227,3	124,46	18155,61	134,74	0,0029	124,46
2-C	4369	8323,7	127,57	18819,70	137,18	0,0001	127,57
3-C	4369	155776,2	124,72	17792,75	133,39	0,1768	124,72
4-O	95	242,5	4,76	27,81	5,27	0,4619	4,86
5-O	95	245,8	4,86	28,48	5,34	0,6613	4,91
6-O	95	253,1	5,10	31,09	5,58	0,6327	5,14
7-R	59233	98977,4	1282,08	4106065	2026,34	0,0997	1904,63
8-R	59233	98486,7	1266,25	4169479	2041,93	0,0534	1911,22
9-R	59233	95648,8	1174,70	4025555	2006,38	0,0234	1866,00

A tabela 5.11 representa o resultado destes testes utilizando o modelo Random Forest, atendendo nas primeiras 3 linhas, que representam os teste com previsões de casos confirmados, utilizando respectivamente 3,5 e 7 dias de histórico nas previsões, é possível observar que os modelos possuem resultados bastante semelhantes, já que os seus erros MSE, RMSE e MAE têm valores bastante próximos, observando a coluna do total previsto e comparando com o total real é possível concluir que o teste 3-C que utiliza 7 dias possui um total previsto mais distante do real, como tal não representa o melhor resultado. Analisando os valores dos erros dos outros 2 testes é possível concluir que o modelo obtém um resultado melhor

para a previsão de confirmados utilizando 3 dias de dados históricos, já que o teste 1-C possui erros mais baixos.

Observando agora as linhas que correspondem aos testes que executam previsões do número de óbitos (4-O, 6-O e 7-O), é possível observar que ao longo dos testes o total previsto foi se distanciando do real estando sendo que os totais previsto estão todos a cima de 240 óbitos nos 30 dias e o valor real é de 95. Assim os valores dos erros são altos já as previsões prevêem um total que é mais do que o dobro do real. Observando atentamente os resultado verificamos que o melhor resultado é o 4-O que utiliza 3 dias de dados históricos, tendo eles um erro médio absoluto de 4,86.

Os últimos testes representam as previsões do número de recuperados (7-R, 8-R e 9-R), é possível observar que do primeiro teste para o segundo, se observou um declínio de desempenho, já que o segundo teste possui erros mais elevados do que o primeiro, assim como um R^2 mais baixo e um total previsto muito distantes dos valores ideais, isto significa que para prever o número de óbitos utilizando 3 dia de dados históricos obtém-se melhor desempenho do que usando 5 dias. Observando o terceiro e último teste, teste este que utiliza 7 dias de dados históricos, é possível observar melhorias em relação aos outros testes, visto que os seus erros são os menores, será esta a característica utilizada no seguinte conjunto de testes.

Em suma para o modelo Random Forest, serão utilizados 3 dias de dados históricos nos testes de previsões de confirmados e óbitos, assim como se utilizarão 7 dias nas previsões de recuperados.

Tabela 5.12: Resultados do teste de desempenho com diferentes quantidades de dados históricos utilizando o modelo K-Nearest Neighbors.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
1-C	4369	4804,7	14,1	3737,9	61,1	0,0041	44,4
2-C	4369	4861,7	15,9	2657,6	51,6	0,1068851	38,1
3-C	4369	5525,3	37,3	3284,3	57,3	0,2824	44,7
4-O	95	102,0	0,2	4,7	2,2	0,4374	1,8
5-O	95	103,6	0,3	3,8	2,0	0,5295771	1,6
6-O	95	114,0	0,6	4,5	2,1	0,5243675	1,7
7-R	59233	69873,3	343,2	2861660	1691,6	0,1801	1308,8
8-R	59233	69354,6	326,5	3446331	1856,4	0,0605	1348,7
9-R	59233	79546,7	655,3	4055131	2013,7	0,0257	1613,6

A tabela 5.12 representa o resultado destes testes utilizando o modelo K-Nearest Neighbors, atentando nas primeiras 3 linhas, que representam os teste com previsões de casos confirmados, utilizando respectivamente 3,5 e 7 dias de histórico nas previsões, é possível identificar desde logo que o teste 3-C, produz o pior resultado, sendo o seu erro médio de 37,3 enquanto que os outros testes possuem erros muito menores, dos 2 testes restantes o teste 1-C tem um erro médio ligeiramente menor que o teste 2-C, mas como o teste 2-C, que utiliza 5 dias de dados de histórico, possui valor de MSE, RMSE e MAE bastante mais baixos essa será a característica identificada.

Analisando os testes de previsão de óbitos (4-O, 5-O e 6-O) verifica-se que os resultados são bastante bons em qualquer um dos testes, já que os valores dos erros são bastante baixos e os totais previstos estão bastante próximos do real. Tendo por base o modelo com menor MAE, MSE, RMSE e maior R^2 , o modelo que produz um melhor desempenho ao modelo KNN para previsão de óbitos é o 5-O, 5 dias de dados de histórico.

Observando agora as linhas que correspondem aos testes que executam previsões do número de recuperados (7-R, 8-R e 9-R), é possível observar que com o aumento dos dados de histórico os resultados vão piorando já que os erros vão aumentando e o total previsto se distanciando do real, desta forma é visível que o modelo que utiliza 3 dias de dados históricos (7-R) é o melhor.

Em suma, o modelo K-Nearest Neighbors obtém melhores resultados usando 5 dias de dados de histórico para previsões de confirmados e óbitos, assim como obtém melhores resultados utilizando 3 dias para previsão de recuperados.

Tabela 5.13: Resultados do teste de desempenho com diferentes quantidades de dados históricos utilizando o modelo ARIMA.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
1-C	4369	5502,7	36,57	4334,03	65,83	0,1966	53,78
2-C	4369	4543,5	5,63	3211,16	56,67	0,1958	45,06
3-C	4369	4034,6	-10,79	4337,99	65,86	0,1545	58,87
4-O	95	137,4	1,37	13,56	3,68	0,1341	2,91
5-O	95	105,5	0,34	3,05	1,75	0,6252	1,48
6-O	95	90,6	-0,14	3,97	1,99	0,4952	1,54
7-R	59233	122829,2	2051,49	10148463	3185,67	0,1088	2809,15
8-R	59233	61552,6	74,82	3073161	1753,04	0,0968	1226,58
9-R	59233	51329,1	-254,97	3973024	1993,25	0,0069	1257,25

A tabela 5.13 representa o resultado dos testes com diferentes quantidades de dados históricos, atentando nas primeiras 3 linhas, que representam os testes com previsões de casos confirmados, utilizando respectivamente 3, 5 e 7 dias de histórico nas previsões, é possível descartar a opção de utilizar 7 dias de dados históricos, já que o modelo obtém o pior resultado no teste 3-C. Analisando os testes 1-C e 2-C é observável que, embora possua R^2 bastante próximos o modelo 2-C obtém valores menores em todas as métricas avaliadoras de erros, assim como possui um total previsto mais próximo do real.

Observando agora as linhas que correspondem aos testes que executam previsões do número de óbitos (4-O, 5-O e 6-O) é possível observar que o primeiro teste possui o pior desempenho de todos tendo valores de erros muito acima dos restantes testes, como por exemplo o valor do erro médio quadrático que é de 13,56 enquanto nos restantes testes está à volta dos 3. Do teste 5-O e 6-O embora o teste 6-O possua um total previsto mais próximo do real, o teste 5-O que utiliza 5 dias de dados, possui valores de erros mais baixos e como tal será considerado com a característica identificada neste teste.

Os últimos testes representam as previsões do número de recuperados (7-R, 8-R e 9-R), é possível observar que o teste que utiliza 3 dias de dados de histórico (7-R) possui um resultado bastante pior que os restantes, sendo o seu erro médio absoluto mais de 2 vezes

maior que os restantes. Dos restantes testes embora, o 8-R possua um valor total previsto mais próximo do real, o teste 8-R é o teste que apresenta menores erros e um maior R^2 comparativamente com o teste 9-R, com isto o melhor resultado obtém-se utilizando 5 dias de dados históricos.

Em suma, para todos os tipos de previsão o modelo ARIMA obtém o seu melhor resultado utilizando 5 dias de dados de histórico.

Teste - Desempenho com diferentes conjuntos de variáveis

A tabela 5.14 representa um resumo deste conjunto de testes que tem como objetivo identificar que conjunto de variáveis traduz um melhor desempenho aos modelos. Na tabela a coluna que representa os dados históricos foi omitida por depender do modelo e tipo de previsão identificados na bateria de teste anterior, estando esta informação nas tabelas que descrevem, com maior detalhe, esta bateria de teste em anexo B.2 e B.3.

Os testes identificados como os melhores na bateria de testes anteriores serão utilizados nesta bateria, já que eles foram executados variáveis (utilizando apenas a variável que se pretende prever) e estarão identificados como 0-C no caso do teste de previsão de confirmados, 0-O no caso dos óbitos e 0-R no caso dos recuperados.

Tabela 5.14: Testes de desempenho com diferentes conjuntos de variáveis a executar em todos modelos.

Teste	Previsão	Intervalo T.	Variáveis
0-C	Confirmados	1 ano	Sem Variáveis
1-C	Confirmados	1 ano	Variáveis Meteorológicas
2-C	Confirmados	1 ano	Variáveis de Atenuação Pandémica
3-C	Confirmados	1 ano	Variáveis de maior correlação
4-C	Confirmados	1 ano	Variáveis Meteorológicas, Atenuação Pandémica e de maior correlação
0-O	Óbitos	1ano	Sem Variáveis
5-O	Óbitos	1ano	Variáveis sobre mortalidade
6-O	Óbitos	1ano	Variáveis sobre internamentos
7-O	Óbitos	1ano	Variáveis de maior correlação
8-O	Óbitos	1ano	Mortalidade, Internamentos e de Maior Correlação
0-R	Recuperados	1 ano	Sem Variáveis
9-R	Recuperados	1ano	Internamentos
10-R	Recuperados	1ano	Atualidade
11-R	Recuperados	1ano	Maior Correlação
12-R	Recuperados	1ano	Internamentos, Atualidade e de Maior Correlação

Tabela 5.15: Resultados dos testes de desempenho com diferentes conjuntos de variáveis utilizando o modelo Random Forest.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
0-C	4369	8227,3	124,46	18155,61	134,74	0,0029	124,46
1-C	4369	6149,3	57,43	6995,42	83,64	0,0012	69,12
2-C	4369	6840,0	79,71	12187,92	110,40	0,0774	90,42
3-C	4369	5830,5	47,14	6645,77	81,52	0,0124	64,28
4-C	4369	6205,3	59,23	8282,47	91,01	0,0047	75,01
0-O	95	242,5	4,76	27,81	5,27	0,4619	4,86
5-O	95	263,3	5,43	36,01	6,00	0,1959	5,51
6-O	95	313,4	7,04	54,92	7,41	0,4338	7,04
7-O	95	274,4	5,79	37,31	6,11	0,5045	5,79
8-O	95	337,1	7,81	65,75	8,11	0,3986	7,81
0-R	59233	95648,8	1174,70	4025555,00	2006,38	0,0234	1866,00
9-R	59233	66120,4	222,17	2743148,00	1656,25	0,0393	1337,49
10-R	59233	66756,3	242,69	2910981	1706,16	0,0065	1367,64
11-R	59233	66229,0	225,68	3143701	1773,05	0,0024	1427,29
12-R	59233	68908,4	312,11	3067593	1751,46	0,0008	1425,00

A tabela 5.15 representa os resultados obtidos nesta bateria de testes com o modelo Random Forest. Observando as linhas que correspondem aos testes de previsão de confirmados (0-C a 4-C) é possível observar que todos os testes com variáveis possuem melhores resultados do que o teste sem variáveis executado na outra bateria de testes. Atentando nos quatro testes de previsões de confirmados com variáveis, é possível apontar como os melhor o 1-C e o 3-C, que usam respectivamente um conjunto de variáveis meteorológicas e de maior correlação com a variável que se pretende prever. Deste 2 teste o 3-C, possui um total previsto mais próximo de real, um erro médio mais baixo, assim como as outras métricas medidoras de erros, possuindo também um R^2 ligeiramente superior e como tal é o melhor resultado, fazendo com que o melhor resultado de testes de previsão de confirmados utilizando o modelo Random Forest, seja obtido utilizando as variáveis de maior correlação,

Analizando as linhas que correspondem aos testes de previsões de óbitos (0-O a 8-O) é possível verificar que todos os testes com variáveis possuem resultados piores do que o teste sem variáveis, executado anteriormente, já que este é o teste que possui menores erros, um total previsto mais próximo do real, assim como um R^2 bom.

Atentando nas linhas dos testes de previsões de recuperados (0-R, 9-R, 10-R, 11-R E 12-R) é possível observar que os testes com variáveis produzem melhores resultados do que o teste executado na bateria anterior sem variáveis, já que produzem previsões mais próximas dos valores reais, ou seja, com erros mais baixos. Estes teste possuem resultados bastante próximos, sendo que o teste 12-R, que utiliza um conjunto das 3 hipóteses como variáveis possui os piores resultados já que o seu total previsto está á volta dos 68900 enquanto que os outros testes estão mais baixos e tendo este teste também erros um pouco maiores que os outros testes. Dentre os restantes testes o teste que utiliza as variáveis de maior correlação (11-R) possui erros mais elevados, como por exemplo o erro médio absoluto que é de 1427,29 enquanto os outros teste possuem erros na casa dos 1300. Entre o teste

9-R e 10-R, é possível verificar que o 9-R é ligeiramente melhor possuindo erros menores. Significando que o melhor desempenho para previsão de recuperados com o modelo Random Forest é obtido utilizando o conjunto de variáveis sobre os internamentos.

Em suma, utilizando o modelo Random forest, para previsão de casos confirmados obtém-se o melhor desempenho utilizando o conjunto de variáveis de maior correlação, para previsão de recuperados obtém-se o melhor desempenho utilizando o conjunto de variáveis sobre os internamentos, enquanto que no casos dos óbitos o melhor desempenho é verificado não utilizando variáveis.

Tabela 5.16: Resultados dos testes de desempenho com diferentes conjuntos de variáveis utilizando o modelo K-Nearest Neighbors.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
0-C	4369	4861,667	15,892	2657,603	51,55	0,1068	38,05
1-C	4369	9947,3	179,94	78111,40	279,48	0,0704	195,49
2-C	4369	23658,2	622,23	865818,10	930,49	0,0858	637,42
3-C	4369	2174,6	-70,78	12216,99	110,53	0,0511	99,91
4-C	4369	8717,2	140,26	81456,79	285,41	0,0327	192,82
0-O	95	103,63	0,27	3,83	1,96	0,5295	1,60
5-O	95	272,7	5,73	39,34	6,27	0,2412	5,77
6-O	95	247,1	4,90	32,77	5,72	0,3099	4,91
7-O	95	243,6	4,79	30,64	5,54	0,0001	5,01
8-O	95	250,3	5,01	31,44	5,61	0,2080	5,05
0-R	59233	69873,3	343,237	2861660	1691,64	0,1801	1308,85
9-R	59233	29435,4	-961,2	4433055	2105,48	0,0089	1477,97
10-R	59233	97340,1	1229,26	4292519	2071,84	0,0077	1913,29
11-R	59233	34792,2	-788,41	3255686	1804,35	0,0676	1228,26
12-R	59233	95358,7	1165,34	4138430	2034,31	0,0052	1842,28

A tabela 5.16 representa os resultados dos testes feitos com diferentes conjuntos de variáveis utilizando o modelo K-Nearest Neighbors. Observando as primeiras 5 linhas da tabela, que correspondem aos testes que tinham como objetivo prever o número de confirmados, É possível observar que os teste com variáveis produzem todos resultados piores do que o teste sem variáveis, ao contrário do ocorrido nos testes com o modelo Random Forest. Observando os valores é possível verificar que o teste 2-C, que utiliza com conjunto de variáveis medidas de atenuação do efeito da pandemia, produz o pior resultado de todos, já que o valor, por exemplo de RMSE sem variáveis é de 51,55 e com a adição destas variáveis passa para 930,49. O teste com variáveis que possui melhor resultado é o 3-C que utiliza o conjunto de variáveis de maior correlação, ainda que tenham um MAE de 99,91 enquanto o teste sem variáveis possui um MAE de 38,05, o que significa que para a previsão de confirmados o modelo KNN obtém o melhor resultado utilizando apenas o histórico da própria variável,

Observando os testes realizados com o objetivo de testar vários conjuntos de variáveis na previsão de óbitos, 0-O,5-O,6-O,7-O e 8-O, é possível notar que os resultados dos testes com variáveis possuem valores piores do que o teste sem variáveis. Estes testes com variáveis possui resultados bastante semelhantes, sendo o melhor o teste 6-O, que utiliza variáveis

sobre os internamentos. Deste forma o teste O-O é o que possui um melhor desempenho, o que significa que para a previsão de óbitos o modelo KNN obtém o melhor resultado utilizando apenas o histórico da própria variável,

Por fim, analisando os restantes testes, que têm como objetivo testar vários conjuntos de variáveis na previsão de recuperados (O-R, 9-R, 10-R, 11-R e 12-R) é possível verificar que existe uma grande variedade de resultados. Comparando o teste 9-R, que utiliza as variáveis sobre os internamentos, com o teste sem resultados é observável que o teste com variáveis possui um resultado pior que o sem variáveis, já que possui erros maiores. Comparando o teste 10-R, que utiliza dados sobre a atualidade, com todos os outros resultados é possível verificar que, junto com o teste 12-R, são os piores resultados dos testes de previsão de recuperados, tendo, por exemplo, MAE superiores a 1842,28, enquanto os restantes tem valores muito abaixo. O melhor resultado é obtido utilizando as características do teste 11-R, ou seja, com as variáveis de maior correlação.

Em suma, utilizando o modelo K-Nearest Neighbors, para previsão de recuperados obtém-se o melhor desempenho utilizando o conjunto de variáveis de maior correlação, enquanto que no casos dos óbitos e no caso dos recuperados o melhor desempenho é verificado não utilizando variáveis.

Tabela 5.17: Resultados dos testes de desempenho com diferentes conjuntos de variáveis utilizando o modelo ARIMA.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
0-C	4369	4543,5	5,63	3211,16	56,67	0,196	45,06
1-C	4369	5792,5	45,92	8972,62	94,72	0,064	73,25
2-C	4369	4562,4	6,24	4710,23	68,63	0,078	55,80
3-C	4369	4849,1	15,49	12028,45	109,67	0,080	90,09
4-C	4369	5255,8	28,61	18946,92	137,65	0,068	100,73
0-O	95	105,5	0,34	3,05	1,75	0,625	1,48
5-O	95	97,3	0,07	6,03	2,46	0,296	2,01
6-O	95	118,6	0,76	6,36	2,52	0,338	2,01
7-O	95	93,7	-0,04	8,87	2,98	0,053	2,44
8-O	95	95,8	0,03	11,78	3,43	0,001	2,70
0-R	59233	61552,6	74,82	3073161	1753,04	0,097	1226,58
9-R	59233	57402,5	-59,05	3178225	1782,76	0,058	1156,68
10-R	59233	61140,6	61,53	2745395	1656,92	0,118	1323,89
11-R	59233	60336,5	35,60	2231190	1493,72	0,341	1069,89
12-R	59233	43563,8	-505,46	2933358	1712,71	0,058	1093,64

A tabela 5.17 representa os resultados dos testes feitos com diferentes conjuntos de variáveis utilizando o modelo ARIMA. Analisando os testes de previsão de confirmados, mais propriamente os testes que utilizam variáveis, verifica-se que existe uma grande diversidade de resultados, sendo uns mais próximos e outros mais distantes do teste se variáveis, mas nunca sendo melhores. De notar que os testes 3-C e 4-C, que utilizam respetivamente um conjunto de variáveis de maior correlação e um conjunto das 3 hipóteses, produzem os piores resultados, tendo valores de MSE acima de 112028,45 enquanto que o teste sem variáveis

possui um valor de 3211,16. O teste com variáveis que possui o melhor resultado corresponde ao teste que utiliza as medidas de atenuação da evolução da pandemia, não sendo um resultado bom o suficiente, já que o resultado do teste sem variáveis possui melhores atributos, nomeadamente um MAE de 45,06, que é mais baixo que o valor de 55,80 do teste 3-C,

Analisado as linhas que representam os resultados dos testes feitos com diferentes conjuntos de variáveis na previsão de óbitos é nítido que, embora os teste com variáveis possuam um total previsto mais próximo do real, do que o teste sem variáveis, este possui os melhores atributos, já que apresenta os valores de MSE, RMSE e MAE mais baixos e o valor de R^2 mais elevado.

Observando as linhas 0-R, 9-R, 10-R, 11-R e 12-R, que correspondem aos testes de previsão de recuperados com diferentes conjuntos de variáveis, existe um teste que se destaca pela negativa,

Em suma, o modelo ARIMA, obtém o melhor resultado para a previsão de recuperados utilizando o conjunto de variáveis de maior correlação, enquanto que no caso das previsões de confirmados e óbitos, o melhor resultado é obtido sem uso de variáveis adicionais.

Teste - Desempenho com diferentes intervalos de treino

Tabela 5.18: Teste de desempenho com diferentes intervalos de treino a executar em todos modelos

Teste	Previsão	Intervalo Treino
0-C	Confirmados	1 ano
1-C	Confirmados	9 meses
2-C	Confirmados	6 meses
3-C	Confirmados	3 meses
0-O	Óbitos	1 ano
4-O	Óbitos	9 meses
5-O	Óbitos	6 meses
6-O	Óbitos	3 meses
0-R	Recuperados	1 ano
7-R	Recuperados	9 meses
8-R	Recuperados	6 meses
9-R	Recuperados	3 meses

Este conjunto de testes tem como objetivo identificar que intervalos de treino proporcionam aos modelos e tipos de previsão um melhor desempenho. A tabela 5.18 representa um resumo dos testes que serão executados, sendo que as tabelas em anexo B.4, B.5 e B.6 representam os testes de forma mais detalhada.

Durante os testes os intervalos que serão utilizados serão o ano anterior aos dias que se pretendem prever, 9 meses, 6 meses e 3 meses. Em que os resultados que utilizam 1 ano de dados de treino estarão identificados com um 0 (0-C, 0-O e 0-R), visto serem os melhores obtidos na bateria de testes anteriores (em que se utilizou sempre 1 ano de dados de treino).

Estes testes irão utilizar o número de dias de histórico identificados nos testes anteriores para cada modelo e diferentes tipos de previsões, assim como utilizarão os conjuntos de variáveis anteriormente identificados.

Tabela 5.19: Resultados dos testes utilizando o modelo Random Forest.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
0-C	4369	5 830,5	47,14	6645,77	81,52	0,0124	64,28
1-C	4369	5 681,7	42,34	6090,55	78,04	0,0453	65,78
2-C	4369	6 825,3	79,23	8540,56	92,42	0,1301	81,69
3-C	4369	20880,0	532,61	303514,00	550,92	0,0366	532,61
0-O	95	242,5	4,76	27,81	5,27	0,4619	4,86
4-O	95	246,6	4,89	30,39	5,51	0,6765	5,01
5-O	95	241,5	4,73	28,33	5,32	0,7238	4,84
6-O	95	306,4	6,82	54,14	7,36	0,3166	6,82
0-R	59233	66120,4	222,17	2743148	1656,25	0,0393	1337,49
7-R	59233	61720,7	80,25	2731889	1652,84	0,0408	1269,36
8-R	59233	78968,3	636,62	3332166	1825,42	0,0001	1551,55
9-R	59233	86713,2	886,46	3997082	1999,27	0,0188	1729,25

A tabela 5.19 representa os resultados dos testes utilizando diferentes intervalos de treino com o modelo Random Forest. Analisando a tabela como um todo é possível concluir que a diminuição do intervalo de dados não traduz sempre uma melhoria ou uma piora dos resultados, já que existem teste com intervalos curtos com resultados melhores que alguns intervalos mais longos e vice-versa, ou seja, cada caso é um caso.

No caso dos testes de previsão de confirmados é possível notar que o teste 0-C que utiliza 1 ano de dados de treino é o que possui um melhor resultado, erros menores. O teste mais com um desempenho mais próximo do teste 0-C é o 1-C, que utiliza 9 meses de dados de treino.

Observando os teste 0-O, 4-O, 5-O, 6-O, teste que prevêm o número de óbitos, é possível perceber que os resultados são bastante próximos tendo apenas diferenças de décimas. Os dois melhores resultados são o teste que utiliza 1 ano de dados (0-O) e o o que utiliza 6 meses de dados (5-C), onde o teste 0-C se destaca pelos seus valores de MSE, RMSE e MAE que são ligeiramente inferiores ao teste 5-C. Desta forma se conclui que o modelo Random Forest obtém o melhor desempenho utilizando 1 ano de dados de treino.

Examinando os testes que prevêm o número de recuperados é justificável excluir o teste 9-R, que usa 3 mês de dados de treino, como o melhor o teste com melhor resultado, já que apresenta os valor de MAE, RMSE e MSE mais elevados, assim como a diferença entre o total real e total previsto. O teste 7-R, que utiliza 9 meses de dados para treino é o que possui o melhor resultado, visto os seus valores de erros.

Tabela 5.20: Resultados dos testes utilizando o modelo K Nearest Neighbors.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
0-C	4369	4861,7	15,89	2657,60	51,55	0,1069	38,05
1-C	4369	4972,4	19,47	3034,42	55,09	0,0891	41,24
2-C	4369	5518,4	37,08	3516,14	59,30	0,1353	47,15
3-C	4369	8017,1	117,68	16756,07	129,45	0,0022	117,96
0-O	95	103,6	0,28	3,83	1,96	0,5296	1,60
4-O	95	114,5	0,63	3,08	1,76	0,6653	1,44
5-O	95	131,9	1,19	4,31	2,08	0,6690	1,65
6-O	95	296,5	6,50	49,90	7,06	0,3166	6,53
0-R	59233	34792,2	-788,41	3255686	1804,35	0,0676	1228,26
7-R	59233	31982,4	-879,05	3498131	1870,33	0,1388	1275,26
8-R	59233	41061,8	-586,17	3051697	1746,91	0,1818	1184,42
9-R	59233	109226,2	1612,68	6578744	2564,91	0,1015	2276,69

A tabela 5.20 representa os resultados dos testes utilizando diferentes intervalos de treino com o modelo K-Nearest Neighbors. Explorando os resultados dos testes de previsão de confirmados (0-C, 1-C, 2-C e 3-C) é possível perceber que com a diminuição da quantidade de dados de treino os resultados vão piorando, já que do teste 0 para o 3 os erros aumentam e o total previsto se afasta do real. Desta forma é possível concluir que o melhor intervalo de dados de treino para prever casos confirmados utilizando o modelo K-Nearest Neighbors é 1 ano.

Olhando para as linhas 0-O, 4-O, 5-O e 6-O, que correspondem aos testes de previsão óbitos, é possível notar que com a redução do intervalo de dados de 1 ano para 9 meses, o modelo obteve um resultado melhor, baixando os seus erros e aumentando ligeiramente o R^2 . Nos restantes testes verifica-se que com a restante diminuição do intervalo de dados de treino os resultados pioram, com o MAE do teste 4-O a passar de 1,44 para 6,53.

Observando as linhas que correspondem aos testes de previsão de recuperados é possível verificar que com a diminuição da quantidade de dados de treino, os resultados vão oscilando. Já que com a diminuição do intervalo de 1 ano para 9 meses (7-R) o modelo apresenta um desempenho pior com o RMSE passando de 1804,35 para 1870,33, enquanto que o teste que utiliza 6 meses possui um RMSE de 1746,91. O teste 9-R possui o pior resultado, sendo desta forma possível concluir que o modelo KNN obtém o seu melhor desempenho para previsão de recuperados, utilizando 6 meses de dados de treino.

Tabela 5.21: Resultados dos testes utilizando o modelo ARIMA.

Teste	Total Real	Total Pre-visto	Erro Médio	MSE	RMSE	R^2	MAE
0-C	4369	4543,5	5,63	3211,16	56,67	0,196	45,06
1-C	4369	4821,8	14,61	12303,06	110,92	0,057	95,74
2-C	4369	5841,2	47,49	15647,07	125,09	0,075	96,98
3-C	4369	4901,4	17,17	12963,75	113,86	0,068	96,22
0-O	95	105,5	0,34	3,05	1,75	0,625	1,48
4-O	95	94,4	-0,02	3,68	1,92	0,528	1,57
5-O	95	104,9	0,32	3,09	1,76	0,620	1,49
6-O	95	103,5	0,27	3,09	1,76	0,615	1,50
0-R	59233	60336,5	35,60	2231190	1493,72	0,341	1069,89
7-R	59233	45604,2	-439,64	2818982	1678,98	0,093	1169,52
8-R	59233	70889,3	376,01	3767082	1940,90	0,144	1446,40
9-R	59233	118846,0	1923,00	15830537	3978,76	0,073	3073,33

A tabela 5.21 representa os resultados dos testes utilizando diferentes intervalos de treino com o modelo ARIMA. Analisando as linhas que representam os teste de previsão de confirmados, é possível verificar que os resultados vão piorando com a diminuição do intervalo de dados de treino, sendo possível concluir que o modelo ARIMA obtém um melhor resultado utilizando 1 ano de dados de treino.

Atentando as linhas que correspondem aos testes de previsão de óbitos, é possível verificar que os resultados vão piorando com a diminuição do intervalo de dados de treino, sendo possível concluir que o modelo ARIMA obtém um melhor resultado utilizando 1 ano de dados de treino. O mesmo acontece com os testes de previsão de recuperados, podendo assim concluir que o modelo ARIMA tem um desempenho melhor utilizando 1 ano de dados de treino em todos os tipos de previsão.

5.2.5 Seleção de Modelos

Nesta seção será feita uma análise aos resultados finais obtidos dos testes que tinham como objetivo identificar as características que tornam os modelos mais eficazes de prever a variáveis alvo no dia seguinte, assim como será feito um teste de previsão dos próximos 5 dias.

Tabela 5.22: Quadro resumo do resultado dos testes que tinham como objetivo identificar as características que auferem os melhores resultados.

Modelo	Previsão	Histórico	C.Variáveis	Intervalo Treino	Total Real	Total Previsto	MSE	RMSE	MAE
RF	C	3	Maior Correlação	1 ano	4369	5830,5	6645,8	81,5	64,3
KNN	C	5	Sem Variáveis	1 ano	4369	4861,7	2657,6	51,6	38,1
ARIMA	C	5	Sem Variáveis	1 ano	4369	4543,5	3211,2	56,7	45,1
RF	O	3	Sem Variáveis	1 ano	95	114,5	3,1	1,8	1,4
KNN	O	5	Sem Variáveis	9 meses	95	242,5	27,8	5,3	4,9
ARIMA	O	5	Sem Variáveis	1 ano	95	105,5	3,1	1,7	1,5
RF	R	7	Inter-namentos	9 meses	59233	61720,7	2731889	1652,8	1269,4
KNN	R	3	Maior Correlação	6 meses	59233	41061,8	3051697	1746,9	1184,4
ARIMA	R	5	Maior Correlação	1 ano	59233	60336,5	2231190	1493,7	1069,9

A tabela 5.22 representa um resumo dos modelos obtidos nos testes anteriores. Na tabela podemos observar que independentemente do tipo de previsão o modelo ARIMA obtém os melhores resultados utilizando 5 dias de dados históricos, enquanto os outros modelos variam de acordo com o tipo de previsão. Atentando a coluna das variáveis, é possível verificar que a maioria dos modelos obtém um melhor resultado utilizando apenas o histórico das variáveis que se pretende prever, como é o caso das previsões de óbitos onde os modelos obtém melhores desempenhos sem utilizarem variáveis extras. Observando a coluna dos intervalos de dados de treino é possível concluir que a maioria dos modelos utiliza 1 ano, como por exemplo o modelo ARIMA, que utiliza sempre 1 ano.

Analisando os desempenhos dos modelos, nomeadamente os modelos de previsão de confirmados, é possível verificar que o modelo K-Nearest Neighbors possui um desempenho melhor que os restantes modelos, possuindo valores nas métricas de medidas de erros menores. Enquanto que nos modelos de previsão de óbitos existem 2 modelos com desempenhos bastante semelhantes, é possível identificar o modelo KNN como o pior.

No caso das previsões de recuperados, o modelo ARIMA possui um desempenho bastante melhor que os restantes, já que possui, por exemplo, um RMSE de 1493,7 enquanto que o mais próximo é de 1652,8. Analisando os modelos identificados como melhores é possível verificar que possuem características semelhantes, possuindo todos 5 dias de dados de histórico e 1 ano de intervalo de treino,

Toda esta análise tem por base a previsão do dia seguinte ao último dia dos dados, contudo o objetivo do projeto é prever mais do que o "amanhã", então a seleção do modelo terá por base outro teste, onde todos os modelos presentes na tabela 5.22 irão prever os próximos 5 dias, utilizando as próprias previsões para gerar outras previsões a partir do segundo dia. A tabela 5.23 demonstra um exemplo do que será feito, são utilizados vários intervalos diferentes de março de 2021 de forma a calcular o erro de previsão de cada um dos dias de previsão. Como à medida que se vai prevendo, vai-se utilizando as previsões dos dias anteriores como input nos modelos para prever os dias seguintes, quanto mais longe do dia atual a previsão está mais provável é do erro ser maior, ou seja, o erro na previsão do dia $N+1$ será provavelmente menor que o erro de previsão de $N+5$. De notar que a tabela é um exemplo serão utilizados 15 intervalos de previsões diferentes.

Tabela 5.23: Exemplo de teste de seleção de modelos.

Data Inicio	Data Fim	Erro N+1	Erro N+2	Erro N+3	Erro N+4	Erro N+5
01/03/2021	05/03/2021	1,714	3,857	5,143	4,714	4,875
02/03/2021	06/03/2021	1,571	4,000	2,714	3,857	3,417
03/03/2021	07/03/2021	0,143	0,286	0,429	2,429	5,571
04/03/2021	08/03/2021	0,000	0,000	1,143	5,143	5,500
05/03/2021	09/03/2021	0,571	1,857	2,857	5,143	2,500
06/03/2021	10/03/2021	3,571	0,286	3,714	1,714	4,778
07/03/2021	11/03/2021	0,286	3,143	1,000	4,778	2,765

Tabela 5.24: Exemplo de teste de seleção de modelos.

Tipo	Histórico	Vars	I.Treino	MSE N+1	MSE N+2	MSE N+3	MSE N+4	MSE N+5	MSE TO-TAL
C-RF	3	Maior Correção	1ano	73,00	81,02	98,37	118,05	142,1	512,5
C-KNN	5	Sem Variáveis	1ano	39,90	91,84	143,33	133,79	128,5	537,4
C-ARIMA	5	Sem Variáveis	1ano	65,07	58,68	70,70	85,03	102,4	381,8
O-RF	3	Sem Variáveis	1ano	4,31	4,33	4,60	5,08	5,5	23,8
O-KNN	5	Sem Variáveis	9meses	1,09	1,60	2,19	3,54	3,5	11,9
O-ARIMA	5	Sem Variáveis	1ano	1,55	2,21	2,60	3,28	3,8	13,47
R-RF	7	Internamentos	9meses	1373,1	1611,8	1766,8	1772,6	1812,6	8336,9
R-KNN	3	Maior Correção	6meses	1687,3	2363,8	2426,8	2448,8	2118,6	11045,3
R-ARIMA	5	Maior Correção	1ano	1130,7	2424,3	16001,6	57451	51134,8	128142

A tabela 5.24 representa os resultados deste teste que tem como objetivo selecionar para cada tipo de previsão o modelo que será utilizado. Na tabela a coluna tipo, identifica que tipo de previsão é (C para previsões de confirmados, O para previsões de óbitos e R para previsões de recuperados) e o modelo utilizando. Observando os resultados dos modelos que executam previsões de confirmados, presentes nas primeiras três colunas, é possível verificar que na maioria das vezes quanto mais o valor de $N+i$, sendo i o número de dias de distancia para o dia atual, maior é o erro com a exceção do erro de $N+2$ no caso do modelo ARIMA que em vez de aumentar, diminuiu. Em qualquer dos dias que se pretende prever o modelo ARIMA é o que apresenta erros menores, desta forma será selecionado para executar as previsões de confirmados. Enquanto que nas previsões do dia seguinte o modelo KNN tinha um desempenho melhor, para prever num horizonte de 5 dias, não consegue manter o desempenho e os seu resultados pioram.

Observando os testes de previsão de óbitos, verifica-se que os modelos, quanto mais distantes do dia atual estão mais dificuldade têm em prever. Sendo que o modelo que possui menores erros ao longo de todos os dias futuros é o modelo KNN, enquanto que anteriormente tinha

sido identificado como o modelo com pior desempenho, sendo o modelo identificado para executar previsões de óbitos.

Atentando nos testes de previsão de recuperados, verifica-se que para $N+1$ o modelo com menores erros é o mesmo que tinha sido anteriormente identificado como o melhor para previsões $N+1$, ou seja, o modelo ARIMA. Contudo o modelo apenas se destaca para previsões de $N+1$, tendo um enorme aumento dos erros com o passar dos dias. O modelo que possui menos erros em maioria dos dias e como tal o selecionado para ser utilizado é o modelo Random Forest.

Com este teste foi possível concluir que embora uns modelos sejam melhores para previsões de $N+1$, não significa que sejam os melhores para previsões em intervalos de tempo no futuro maiores.

5.3 Aplicação

A figura 5.11 representa os componentes da aplicação, onde 4 ficheiros compõem o servidor e o único componente externo à aplicação, são as fontes de dados, que como foi apresentado em 5.1.1 são um conjunto de repositórios online e websites.

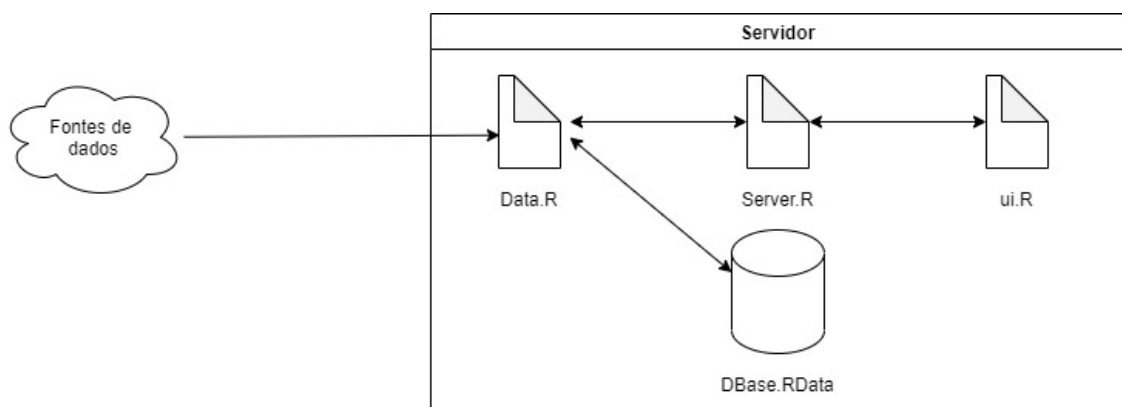


Figura 5.11: Componentes do Sistema.

5.3.1 Data.R

O ficheiro *Data.R* é o ficheiro mais importante do sistema, pois ele é responsável pelas seguintes funções:

- **Carregar dados** - Nesta função o ficheiro é capaz de carregar para o ambiente do RStudio todos os dados presentes na base de dados *Dados.RData*;
- **Atualizar a Base de Dados** - O ficheiro deve atualizar a base de dados interna da aplicação, obtendo os dados mais recentes das suas fontes, da mesma forma que foi descrito em 5.1.1;
- **Previsões** - O ficheiro é o responsável por executar as todas as previsões;
- **Criar uma tabela para as previsões** - O ficheiro armazena todas as previsões numa mesma tabela, assim como os 5 dias anteriores ao início das previsões;

O ficheiro é responsável por obter atualizar a base de dados e gerar previsões. No que toca á base de dados, o sistema executa o mesmo processo descrito em 5.1.1, no que se

refere a previsões, o ficheiro *Data.R* utiliza os modelos seleccionados em 5.2.5 e cria um modelo para cada região do país e tipo de previsão. No caso das previsões de confirmados o ficheiro faz as previsões individuais de cada região do país e a previsão de confirmados em Portugal é a soma das previsões das regiões. No caso das previsões de óbitos, têm o mesmo procedimento, o ficheiro prevê os óbitos de cada região nos próximos 5 dias e no final faz o somatório diário das previsões, calculando assim os óbitos de Portugal. No caso dos recuperados, como só existem dados suficientes para executar a previsão ao nível do país, o ficheiro executa apenas a previsão de recuperados para Portugal nos seguintes 5 dias. Por fim, o número de casos ativos, como é uma variável dependente do número de recuperados, é também uma variável que só pode ser prevista a nível nacional, para calcular os número de casos ativos nos próximos 5 dias, optou-se por utilizar as outras previsões para calcular esta, já que o número de casos ativos pode ser obtido pelos totais das outras variáveis, o que foi obtido somando o total já existente às previsões que foram obtidas, ou seja, o número de casos ativos é obtido pela seguinte equação:

$$\text{CasosAtivos} = \text{TotalCasosConfirmados} - (\text{TotalRecuperados} + \text{TotalObitos}) \quad (5.1)$$

O excerto e código 5.2 contem a criação da tabela que será utilizada para mostrar aos utilizadores as previsões, onde se pretende mostrar os dados dos 5 dias anteriores às previsões e as previsões. O primeiro passo foi juntar todas as previsões numa só tabela (linha 1 a linha 6), de seguida organizou-se a tabela tendo em conta a data e a região (linha 7). Por fim criou-se uma variável chamada "tipo" e atribui-se o valor de "previsao" já que estes dados correspondem a previsões. De seguida utilizou-se uma *query* para carregar apenas as variáveis que se pretende dos dados reais (linha 10) e criou-se uma variável chamada "tipo" e atribui-se o valor de "real" já que estes dados correspondem a dados reais. De seguida realizou-se uma nova *query* que altera a tabela apenas para os dados dos últimos 5 dias e por fim junta-se as duas tabelas numa só (linha 12 e linha 14).

```

1 dados_previsto<-rbind(alentejo , algarve)
2 dados_previsto<-rbind(dados_previsto , centro)
3 dados_previsto<-rbind(dados_previsto , norte)
4 dados_previsto<-rbind(dados_previsto , lisboa)
5 dados_previsto<-rbind(dados_previsto , madeira)
6 dados_previsto<-rbind(dados_previsto , acores)
7 dados_previsto<-arrange(dados_previsto , datas , Regiao)
8 dados_previsto$tipo <- "previsao"
9
10 dados_reais<-dados[, c(1,5,7:18)]
11 dados_reais$tipo <- "real"
12 dados_reais<-dados_reais[(nrow(dados_reais)-34):nrow(dados_reais),]
13
14 dados_previsao<-rbind(dados_reais , dados_previsto)
```

Listing 5.2: Criação da tabela com dados previstos e dados reais (R).

5.3.2 UI.R

O ficheiro "UI.R" é responsável por criar a interface gráfica utilizando a biblioteca "*RShiny*", a função "*dashboardPage*" foi utilizada para criar uma aplicação com um menu lateral que

permite navegar entre diferentes páginas da aplicação. Nesse menu lateral criou-se várias páginas, tais como:

- **Atualidade e Evolução** - página que possui os números mais atuais da pandemia assim como gráficos com a evolução da pandemia;
- **Previsões** - página onde são apresentadas previsões da evolução do covid-19;
- **Comparação de Previsões** - página onde é possível comparar previsões de diferentes regiões;
- **Ajuda** - página onde está presente uma explicação de como utilizar a aplicação;

A figura 5.12 representa a página "Atualidade e Evolução", onde é possível observar uma zona para seleção da região ao qual os dados pertencem, esta zona de seleção foi criada através da função *selectInput*. É possível observar também que os dados sobre a atualidade pandémica estão presentes em retângulos coloridos criados com a função *valueBoxOutput*. Quanto aos gráficos que representam a evolução foram criados com a função *plotlyOutput* que apenas indica que naquele lugar existirá um gráfico.

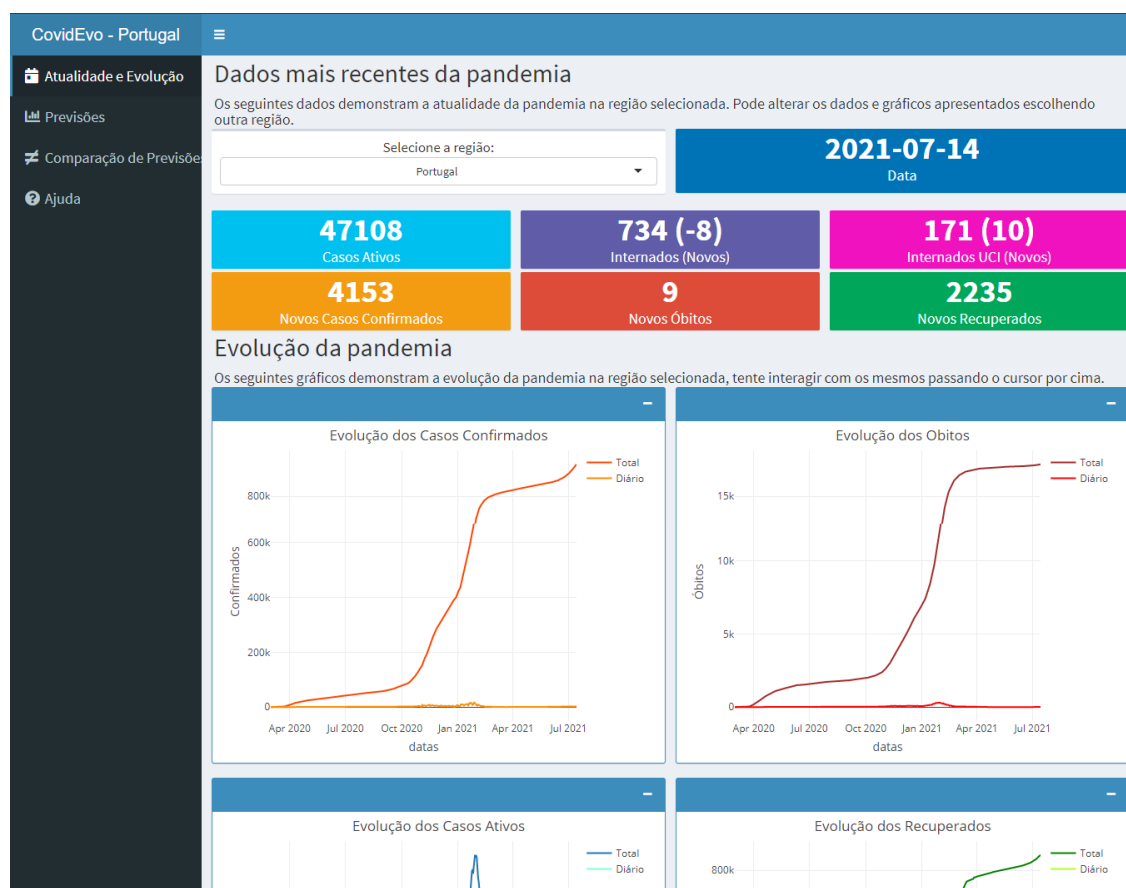


Figura 5.12: Página Atualidade e Evolução.

A figura 5.13 apresenta a página das previsões, nela existe uma zona de seleção, para o utilizador selecionar que região pretende observar as regiões, assim como também existem gráficos de barras inicializados com a função *plotlyOutput*.

A figura 5.14 representa a página "Comparação de Previsões", nela existem duas zonas de seleção e 2 zonas com gráficos, estas zonas de seleção contêm apenas as regiões de Portugal porque verificou-se que uma comparação entre Portugal e uma região é de difícil compreensão e representação devido á sua diferença de grandeza.



Figura 5.13: Página Previsões.

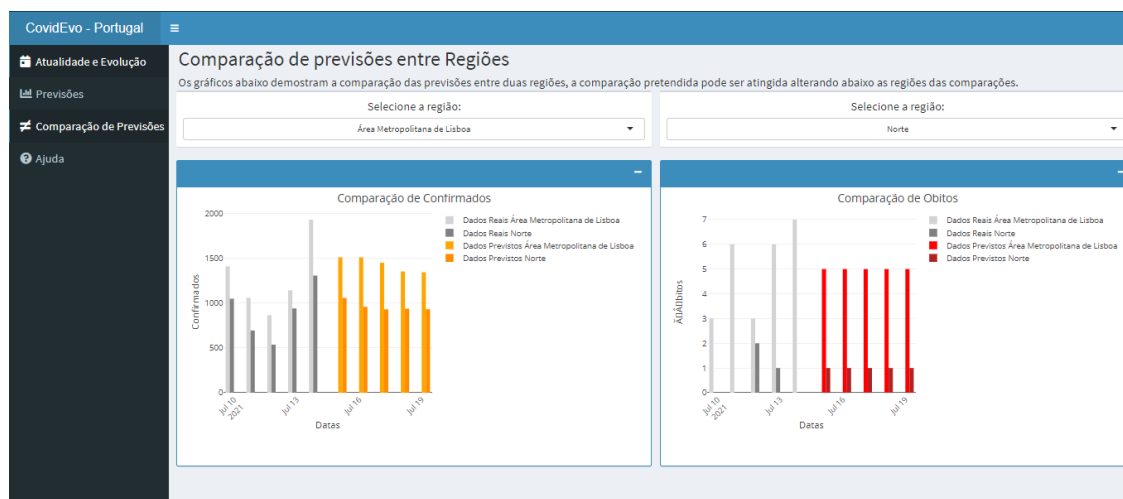


Figura 5.14: Página Comparação de Previsões.

Por fim a página "Ajuda", representada na figura 5.15, consiste em possuir os vários elementos da aplicação assim como uma explicação de como utilizar. Um conjunto de explicações

aborda as funções que a biblioteca *plotly* oferece para uma melhor exploração dos seus gráficos, funções essas que serão exploradas em 5.3.3.



Figura 5.15: Página Ajuda.

5.3.3 Server.R

O ficheiro *"server.r"* tem duas funções controlar a informação que é apresentada nos retângulos informativos e nos gráficos, assim como estar sempre "à escuta" para uma possível alteração de região nas zonas de seleção.

O excerto de código B.11 presente em anexo, representa o controlo feito sobre um dos retângulos de informação, como é possível observar, existe uma função *renderValueBox* que é capaz de controlar o que acontece dentro do retângulo e outra *valueBox* onde é definido o que estará escrito no retângulo. No caso do excerto existe uma condição em que no caso de a região selecionada ser Portugal, o retângulo irá apresentar o número de internados nos cuidados intensivos, caso seja uma região a ser selecionada a informação que será apresentada é a percentagem de população totalmente vacinada na região.

Esta necessidade de ter diferentes informações nos mesmos retângulos deve-se ao fato de que certas informações apenas existem para Portugal como um todo. Estes retângulos são utilizados apenas na página "Atualidade e Evolução", pelo que existem 6 retângulos no total. Na tabela 5.25 está descrita que informação é apresentada ao utilizado no caso de selecionar Portugal ou uma das suas regiões, tendo em conta que a informação apresentada se refere à região selecionada.

Tabela 5.25: Informações presentes nos retângulos informativos.

Retângulo	Portugal	Regiões de Portugal
1	Casos Ativos	Total de Confirmados
2	Internados (Internados Diários)	Total de Óbitos
3	Internados UCI (Internados Diários UCI)	% de vacinação completa
4	Confirmados Diários	Confirmados Diários
5	Óbitos Diários	Óbitos Diários
6	Recuperados Diário	% de vacinação iniciada

O excerto de código B.12, presente em anexo, representa um exemplo de controlo sobre os gráficos, em que da mesma forma que para os retângulos, existe uma condição que verifica se a região selecionada é Portugal, caso se confirme são apresentados os valores de Portugal, caso contrário são apresentados os dados da região selecionada. No caso do exemplo utiliza-se as funções da biblioteca *plotly* para definir como será o gráfico. A biblioteca *plotly* permite adicionar ao gráfico mais conjuntos de dados com a função *add_trace*, permite também customizar toda a aparência do gráfico com a função *layout* e permite também configurar algumas funções extra com a função *config*.

No que diz respeito a gráficos a aplicação possui 11 gráficos, sendo que na página "Atualidade e Evolução" possui 4 gráficos de linhas para apresentar aos utilizadores a evolução da pandemia desde o início até ao dia mais recente. Os restantes gráficos são de barras e apresentam sempre previsões.

A tabela 5.26 representa que informação é apresentada nos gráficos, no caso de o utilizador selecionar Portugal ou uma das suas regiões. Analisando a tabela 5.26 é possível observar que os gráficos 1 e 2 possuem as mesmas informações, respetivamente a evolução dos casos confirmados e de óbitos, já no gráfico 3 e 4 as informações apresentados são diferentes, se o utilizador selecionar Portugal será-lhe apresentada a evolução de casos ativos e de recuperados, se optar por uma região de Portugal ser-lhe-á apresentada a evolução do Índice de Transmissibilidade e de vacinação, isto, como foi dito anteriormente, deve-se ao fato de não existir dados sobre casos ativos e recuperados para as regiões de Portugal. Quanto aos gráficos 5 e 6, apresentam respetivamente as previsões de casos confirmados e de óbitos para ambas as opções, já os gráficos 7 e 8 que apresentam as previsões de casos ativos e recuperados apenas existem para a região nacional, já que não existem dados que permitem prever estes dados para as regiões de Portugal.

Os gráficos 9 e 10 consistem nas comparações de previsões e como foi dito anteriormente, optou-se por retirar a opção de Portugal, permitindo apenas comparar previsões entre as regiões do país. Por fim, o gráfico 11 consiste na repetição do gráfico 5 e tem como intuito servir de teste para os utilizadores testarem as funções da aplicação.

Tabela 5.26: Informações presentes nos Gráficos.

Gráf.	Local na Aplicação	Portugal	Regiões de Portugal
1	Atualidade e Evolução	Evolução de casos confirmados	Evolução de casos confirmados
2	Atualidade e Evolução	Evolução de Óbitos	Evolução de Óbitos
3	Atualidade e Evolução	Evolução dos casos	Evolução do Rt
4	Atualidade e Evolução	Evolução dos Recuperados	Evolução da Vacinação
5	Previsões	Previsão de Confirmados	Previsão de Confirmados
6	Previsões	Previsão de Óbitos	Previsão de Óbitos
7	Previsões	Previsão de Ativos	
8	Previsões	Previsão de Recuperados	
9	Comparação de Previsões		Comparação de Previsões de Confirmados
10	Comparação de Previsões		Comparação de Previsões de Óbitos
11	Ajuda	Previsão de Confirmados	Previsão de Confirmados

Uma das vantagens da utilização da biblioteca *plotly* é o fato de ser possível observar os dados ao passar o cursor em cima do gráfico, esta vantagem permite retirar mais informação dos gráficos principalmente os gráficos de evolução pandémica que representam vários dias. Esta funcionalidade está presente na figura 5.16 em que no dia 19 de julho de 2021 se prevê existirem 2925 novos casos confirmados em Portugal, para além desta funcionalidade os gráficos *plotly* possuem um menu de ferramentas que se encontra no canto superior direito dos gráficos, como é visível na figura, estas ferramentas estão numeradas e permitem:

1. **Descarregar gráfico** - a primeira ferramenta, que possui como ícone uma câmara permite fazer download do gráfico em formato PNG;
2. **Zoom quadrado** - a segunda ferramenta, que possui como ícone uma lupa com um quadrado no seu interior permite fazer selecionar um quadrado do gráfico e o gráfico automaticamente irá se transformar no quadrado selecionado;
3. **Deslizar** - a terceira ferramenta, que possui como ícone um conjunto de quatro setas permite ao utilizador "agarrar" o gráfico e deslizar a sua informação para onde quiser;
4. **Ampliação** - a quarta ferramenta, que possui como ícone com um sinal positivo permite ao utilizador fazer "zoom in" no gráfico, ou seja, aproximar todo o gráfico tendo como referência o centro do mesmo;
5. **Distanciação** - a quinta ferramenta, que possui como ícone com um sinal negativo permite ao utilizador fazer "zoom out" no gráfico, ou seja, distanciar todo o gráfico tendo como referência o centro do mesmo;
6. **Ajuste automático** - a sexta ferramenta, que possui como ícone 4 setas dentro de um quadrado permite escalar automaticamente o gráfico da forma que a biblioteca *plotly* considera melhor;
7. **Reiniciação** - a sétima ferramenta, que possui como ícone uma casa permite ao gráfico voltar ao seu estado inicial;

8. **Apresentação dos dados ao passar cursor** - a oitava ferramenta, que possui como ícone uma seta permite ao gráfico mostrar os dados que correspondem à posição onde está o rato, esta ferramenta está sempre ativa;
9. **Apresentação de comparação dos dados ao passar cursor** - a nona ferramenta, que possui como ícone duas setas permite ao gráfico mostrar comparação de dados, ou seja, nos gráficos que correspondem a comparações, ao passar o rato numa coluna de uma região, serão apresentados os dados dessa região assim como os dados que correspondem ao mesmo dia, mas a outra região, permitindo assim uma melhor comparação. Esta opção está sempre ativa nos gráficos de comparação;
10. **Suporte** - a décima ferramenta, que possui como ícone o símbolo da biblioteca *plotly* permite aceder ao website da biblioteca;

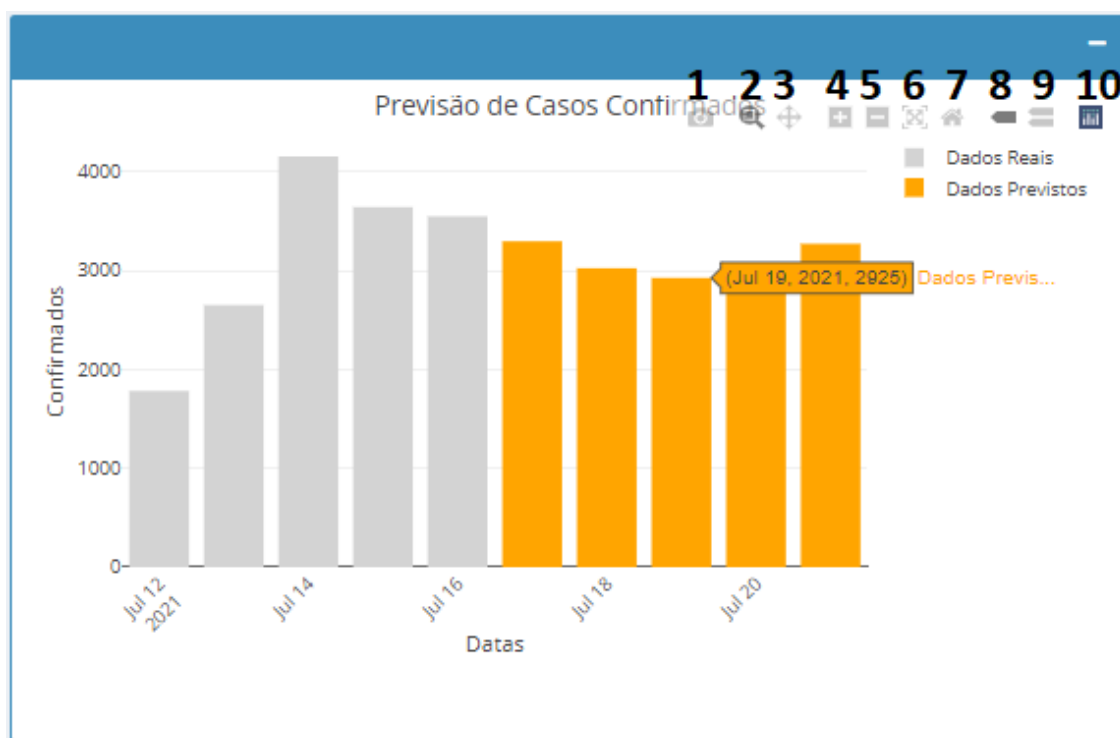


Figura 5.16: Exemplo de gráfico.

Capítulo 6

Avaliação e Experiência

Esta secção tem como objetivo avaliar a aplicação em vários aspetos.

6.1 Utilidade

No capítulo 3, foi analisado o valor que esta aplicação poderia ter para os utilizadores, contudo este valor foi analisado do ponto de vista do autor, tendo isto em conta será agora a vez do público em geral de avaliar a utilidade da aplicação.

O termo "utilidade" traduz-se na qualidade de um bem ou serviço que o torna apropriado para satisfazer os desejos dos seus intervenientes, ou seja, com a utilidade procura-se apurar se a aplicação será útil para os seus futuros utilizadores.

Para avaliar a utilidade da aplicação foi criado um questionário no Google Forms com questões que pretendem avaliar se o público em geral acharia a aplicação útil, assim como pretende avaliar de que forma os utilizadores utilizariam a aplicação, este formulário foi partilhado pelo utilizador em alguns fórum e redes sociais.

Nesse questionário o utilizador irá responder a questões que o caracterizam como o género e faixa etária, de seguida irá responder a questões direcionadas á aplicação como por exemplo o número de vezes semanais que o questionado acederia o sistema, assim como em que dispositivo acederia. Por fim o questionário tem questões mais voltadas para as funções do sistema:

- Utilizaria um web site que permitisse analisar a situação atual de Portugal e suas regiões;
- Utilizaria um web site que disponibilizasse previsões sobre a evolução da pandemia Portugal e suas regiões;
- Utilizaria um web site que permitisse comparar previsões sobre a evolução da pandemia nas regiões de Portugal;
- Considera importante um só web site agregar vários conjuntos de dados como a situação atual pandémica, situação atual vacinal, evolução da pandemia, a evolução da vacinação e previsões da evolução pandémica;

Das 31 respostas obtidas, 55% eram do género masculino e 45% do género feminino, onde 62% eram da faixa etária 20-40 e 26% da faixa etária de 40-60. Das 31 respostas, 27 responderam que utilizaria um web site que permitisse analisar a situação atual de Portugal e suas regiões, 25 utilizariam um web site que disponibilizasse previsões sobre a evolução da pandemia Portugal e suas regiões e 23 acham importante a comparação entre regiões.

Por fim maioria utilizaria um sistema como o descrito neste documento mais de 4 vezes por semana, sendo que 67% acederia ao sistema através de um telemóvel, comprovando desta forma a utilidade do sistema, o formulário e as respostas estão disponíveis ao leitor em *Formulário - Avaliação da Utilidade do Sistema de Previsões de Covid-19 2021*.

6.2 Usabilidade

Hoje em dia os produtos tendem a ser cada vez mais complexos, com cada vez mais funções, o que pode trazer limitações aos utilizadores e um produto que seria para ajudar o utilizador numa tarefa pode acabar não tendo esse efeito. A usabilidade procura medir com que facilidade os utilizadores usam um determinado produto (Jordan et al. 1996).

O que se pretende é que a ferramenta seja bastante usável, ou seja, que permita ao utilizador realizar as suas tarefas da forma mais rápida e eficaz possível. No final pretende-se que o utilizador tenha uma boa experiência com a utilização da aplicação de forma que se sinta confortável desde a primeira utilização, garantindo que apenas ao olhar para a aplicação seja possível perceber que caminho terá que fazer para realizar determinadas ações.

Para avaliar a usabilidade da ferramenta serão utilizados cobaias para testar a ferramenta, ou seja, recorrendo a vários utilizadores de características diferentes serão lhes dadas tarefas a realizar dentro da aplicação e serão medidos o tempo que os utilizadores levam a realizar uma tarefa e o número de cliques feitos até à realização da tarefa. Como ao utilizarem a aplicação os utilizadores irão ganhando conhecimento sobre a aplicação a ordem da realização das tarefas é diferente de utilizador para utilizador e cada tarefa tem sempre início na página inicial da aplicação, caso contrário os testes não seriam precisos pois utilizadores em diferentes páginas poderiam obter uma cronometragem e número de cliques diferentes. As tarefas a realizar pelo utilizador para avaliar a usabilidade são:

1. Identificar o número de Internados no dia atual em Portugal;
2. Identificar o número de óbitos no dia atual na região do Algarve;
3. Identificar o número total de óbitos no dia 10/10/2020 na região do Algarve;
4. Identificar o número de confirmados diário no dia 10/10/2020 na região do Alentejo;
5. Identificar o número previsto de recuperados em Portugal para daqui a 4 dias;
6. Identificar o número de confirmados diário previsto para a região Autónoma da Madeira daqui a 2 dias;
7. Identificar entre Norte e Centro, qual possui uma previsão no número de confirmados mais baixa;

A tabela 6.1 representa os utilizadores que irão realizar estes testes de usabilidade, sendo eles de vários géneros, idades e possuindo diferentes graus de experiência utilizando este tipo de ferramentas.

Tabela 6.1: Descrição dos utilizadores que irão realizar os testes de usabilidade.

Utilizador	Género	Faixa Etária	Experiência Tecnológica
1	Feminino	20-30 anos	Muita
2	Masculino	10-20 anos	Muita
3	Masculino	40-50 anos	Pouca
4	Feminino	30-40 anos	Alguma

A tabela 6.2 representa os resultados dos testes de usabilidade, observando os utilizadores 1 e 2, portadores de muita experiência na utilização de sites, é possível observar que ambas executaram as tarefas em poucos segundos e poucos cliques, com a exceção da tarefa 7, onde se pretende comparar previsões, em que os utilizadores demoraram um pouco mais. Analisando os utilizadores 3 e 4, portadores de pouca e alguma experiência na utilização de sites, verifica-se um aumento do tempo para a conclusão das tarefas, assim como em algumas tarefas existe um aumento do número de cliques o que pode significar que os utilizadores se enganaram ou estavam a conhecer a aplicação. Ainda observando os utilizadores 3 e 4, é possível verificar que a tarefa 7 e a que leva mais tempo a ser concluída.

Em suma, a aplicação tem um alto nível de usabilidade já que os utilizadores conseguiram realizar as tarefas em pouco tempo e poucos passos, com a exceção da tarefa de comparar previsões, isto pode significar que o desenho da página está pouco intuitivo. Estes testes foram executados sem o utilizador recorrer à página de ajuda, desta forma essa página confere um pouco mais de usabilidade já que explica como utilizar o sistema.

Tabela 6.2: Resultados dos testes de usabilidade.

	Utilizador 1		Utilizador 2		Utilizador 3		Utilizador 4	
Tarefa	Tempo	Cliques	Tempo	Cliques	Tempo	Cliques	Tempo	Cliques
1	7	0	4	0	20	0	12	0
2	11	1	12	1	21	3	13	1
3	24	1	22	1	33	1	31	1
4	20	1	21	1	30	1	35	3
5	13	1	14	1	29	3	15	1
6	11	2	12	2	19	2	14	2
7	36	3	43	3	63	4	54	4

6.3 Comparação com outros sistemas

Nesta seção será feita uma comparação com outros sistemas semelhantes, com o objetivo de identificar a unicidade do sistema, analisando se os outros sistemas possuem as funções presentes neste sistema.

A tabela 6.3 contém uma comparação entre o sistema desenvolvido e duas soluções, uma que consiste em ser uma ferramenta global que possui previsões para todos os países e a outra uma ferramenta disponibilizada por um jornal português em colaboração com o Instituto de Saúde Pública da Universidade do Porto. Analisando a tabela é possível verificar que enquanto a ferramenta do IHME se limite a apresentar os dados e previsões de Portugal, o

jornal possui várias funcionalidades ligadas às regiões de Portugal. A ferramenta do jornal é a que mais se assemelha ao sistema em termos de funcionalidades, não possuindo previsões de recuperados, nem de casos ativos e não permite comparações entre as previsões de diferentes regiões.

Em suma, estas duas soluções não têm as mesmas funcionalidades presentes no sistema, comprovando assim a utilidade e unicidade do sistema. De notar também, que estas soluções possuem outras funcionalidades que o sistema não tem, como por exemplo manter um histórico de previsões passadas, funcionalidades esta que poderá ser adicionada no futuro.

Tabela 6.3: Comparação de funções com outras soluções.

Função	Sistema	IHME Covid-19 Projections	Jornal
Apresentar os dados atuais de Portugal	Sim	Sim	Sim
Apresentar os dados atuais das Regiões de Portugal	Sim	Sim	Sim
Apresentar a evolução pandémica em Portugal	Sim	Sim	Sim
Apresentar a evolução pandémica nas regiões de Portugal	Sim	Não	Sim
Apresentar as previsões de confirmados e óbitos em Portugal	Sim	Não	Sim
Apresentar as previsões de confirmados e óbitos nas regiões de Portugal	Sim	Sim	Sim
Apresentar as previsões de recuperados e casos ativos em Portugal	Sim	Não	Não
Comparação de previsões entre regiões de Portugal	Sim	Não	Não

Capítulo 7

Conclusão e Futuro

7.1 Conclusão

Através deste trabalho foi possível concluir que, apesar de existirem muitos modelos de Machine Learning que possibilitam a regressão, os modelos projetados para previsões temporais, ou seja, usam por exemplo dados de dias anteriores para executar uma previsão, terão melhor desempenho do que outros modelos neste tipo de problemas.

O estudo do estado de arte foi um ponto fundamental para a realização deste documento, visto que foi possível desde logo identificar modelos que poderiam ter um bom desempenho neste tipo de previsões, foram eles o modelo Random Forest, o modelo KNN e o modelo ARIMA. Foi possível identificar também que algumas variáveis poderão ter maior impacto no desempenho dos modelos de Machine Learning, como por exemplo as medidas de prevenção ou a temperatura e a velocidade de vento que estão muito relacionadas com a evolução do vírus (Tosepu et al. 2020).

Através do estudo de estado de arte foi possível concluir também que existem várias métricas de avaliação, dentre as quais umas se destinam a classificação e outras a regressão, assim sendo, as métricas que se irão aplicar a este projeto são as que avaliam o desempenho de um modelo de regressão, sendo elas: R^2 , R^2_{adjusted} , MAE, MSE, RMSE.

Para a construção do sistema foi possível concluir que a melhor opção seria construir todos os seus componentes em R, ou seja, utilizar a linguagem de programação R para a construção dos modelos de Machine Learning que irão executar as previsões, bem como utilizar as suas livrarias para extração de informação das suas fontes e construção da aplicação web.

A partir dos testes realizados para obter as melhores características de cada modelo para cada tipo de previsão, verificou-se que certos modelos possuem um melhor desempenho com as mesmas características, como o caso do modelo ARIMA que em todos os tipos de previsão utilizou 5 dias como dados históricos e 1 ano de intervalo de dados, enquanto em outros modelos o comportamento altera de acordo com o tipo de previsão como o caso do modelo KNN.

No teste de variáveis, verificou-se que cada tipo de previsão influencia os modelos de forma diferente, já que por exemplo para as previsões de óbitos os modelos obtiveram os seus melhores resultados sem recorrer a variáveis extra, enquanto que nos outros tipos de previsões utilizaram.

Com o teste de seleção de modelos verificou-se que dos modelos eram melhores para prever o dia seguinte ao dia atual, nenhum desses modelos obteve um bom resultado nas previsões de 5 dias, podendo assim concluir que as previsões são diferentes e que uns modelos são

melhores para prever o dia seguinte enquanto outros são melhores em prever intervalos de tempo maiores. De nota que as conclusões obtidas têm por base os testes realizados, provavelmente com outros testes os resultados seriam diferentes.

No final foi possível desenvolver um aplicação que cumpre o seu objetivo de agregar dados de várias fontes e disponibiliza-las ao utilizador, assim como disponibiliza as melhores previsões, tendo em conta os testes realizados.

7.2 Trabalho Futuro

Ao longo do documento vários problemas foram sendo identificados e o futuro passa por corrigir esses problemas. O ponto mais importante seria realizar mais testes aos modelos, utilizando mais hipóteses do que as que foram utilizadas, ou seja, usar mais intervalos de dados históricos, utilizar mais intervalos de treino e mais variáveis, já que muitas das que compõem a base de dados não foram utilizadas, assim como testar várias combinações destas características. Outro ponto importante seria a utilização de outros modelos nestes testes, como por exemplo as redes neuronais que chegaram a ser identificadas como uma boa opção para este tipo de previsões, mas acabaram por não ser utilizadas. Outra opção seria utilizar um modelo para prever cada um dos dias seguintes, ou seja, um modelo para $N+1$, outro para $N+2$, entre outros.

Outro aspeto a ter em conta para a produção de melhores previsões é a base de dados do sistema, onde é possível e necessário elaborar várias melhorias. Em primeiro lugar existe um problema com os dados de Portugal, que são repetidos ao longo das várias regiões, a solução poderia ser tornar a base de dados um DataWarehouse e usar definir Portugal como uma região da dimensão regiões, visto que repetir os dados de Portugal em todas as regiões não é a forma mais acertiva de representar os dados. Existe também um problema com os dados meteorológicos, já que estas variáveis são obtidas através da conceção de que uma cidade pode representar uma região, visto que por exemplo os dados meteorológicos da região Norte são obtidas da cidade Vila Real, para resolver este problema poderia encontrar-se uma outra fonte que possua dados meteorológicos das regiões de Portugal, ou definir que a meteorologia de uma região é obtida através da média dos distritos que a compõe. Outro problema com a base de dados são os dados da vacinação, visto que por vezes o relatório da DGS possui colunas diferentes e/ou ordens diferentes, seria melhor criar um mecanismo que detetasse estas mudanças.

Em termos da aplicação em si o seu futuro passa por melhorar a interface gráfica, melhorando a página direccionada para a comparação de previsões visto que ser pouco intuitiva e melhorando os gráficos que demonstram a evolução da pandemia desde o seu início para que seja mais fácil encontrar uma data. O caminho a seguir deveria passar por implementar novas funcionalidades, como por exemplo o sistema alertar aos utilizadores quando é que os novos dados diários e consequentemente as novas previsões estavam disponíveis no sistema, e implementar as funcionalidades que não foram possíveis implementar, como o mapa interativo que permitia o utilizador seleccionar uma região apenas ao clicar no mapa.

Por fim, seria importante automatizar o processo de actualização da base de dados e geração de previsões já que o sistema ficou com a capacidade de, ou ser reiniciado, obter os dados mais recentes e gerar as previsões, mas não existe um mecanismo automático para fazer com que ele se reinicie, nem o melhor timing para este reinício, a melhor solução seria criar um mecanismo que fica "à escuta" para quando os novos dados estiverem disponíveis, o sistema reiniciar, obtendo-os e gerando novas previsões.

Bibliografia

- Alloghani, M. et al. (2018). «Application of Machine Learning on Student Data for the Appraisal of Academic Performance». Em: *2018 11th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 157–162. doi: 10.1109/DeSE.2018.00038.
- Alpaydin, Ethem (2020). *Introduction to machine learning*. MIT press.
- Beeley, Chris (2013). *Web application development with R using Shiny*. Packt Publishing Ltd.
- Cheng, Cindy et al. (2020). «Covid-19 government response event dataset (coronanet v. 1.0)». Em: *Nature human behaviour* 4.7, pp. 756–768.
- corona_tscs/data/CoronaNet at master · saudiwin/corona_tscs · GitHub (2020). url: https://github.com/saudiwin/corona%7B%5C_%7Dtscs/tree/master/data/CoronaNet (acedido em 18/11/2020).
- CoronaNet (2020). url: <https://www.coronanet-project.org/> (acedido em 18/11/2020).
- Coronavirus: How the pandemic has changed the world economy (2020). url: <https://www.bbc.com/news/business-51706225> (acedido em 16/12/2020).
- Covid Forecast- Google Trends (2020). url: <https://trends.google.pt/trends/explore?q=covid%5C%20forecast> (acedido em 16/12/2020).
- COVID-19 GOVERNMENT RESPONSE TRACKER (2021). url: <https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker> (acedido em 06/05/2021).
- Covid-19: curva epidémica e parâmetros de transmissibilidade (2021). url: <http://www.insa.min-saude.pt/category/areas-de-atuacao/epidemiologia/covid-19-curva-epidemica-e-parametros-de-transmissibilidade/> (acedido em 06/05/2021).
- CSSEGISandData · GitHub (2020). url: <https://github.com/CSSEGISandData> (acedido em 15/12/2020).
- Dashraath, Pradip et al. (jun. de 2020). «Coronavirus disease 2019 (COVID-19) pandemic and pregnancy». Em: *American Journal of Obstetrics and Gynecology* 222.6, pp. 521–531. issn: 10976868. doi: 10.1016/j.ajog.2020.03.021.
- Direção-Geral da Saúde (2021). url: <https://www.dgs.pt/> (acedido em 06/05/2021).
- Erdil, Nadiye Ozlem e Omid M Arani (2019). «Quality function deployment: more than a design tool». Em: *International Journal of Quality and Service Sciences*.
- Filho, Jaime, Mauricio Atamanczuk e Rui Marçal (mar. de 2010). «Seleção de técnicas de manutenção para processo de armazenagem através do Método de Análise Hierárquica». Em: *Revista Produção Online* 10. doi: 10.14488/1676-1901.v10i1.375.
- Formulário - Avaliação da Utilidade do Sistema de Previsões de Covid-19 (2021). url: <https://forms.gle/caeUYvVbrA3zD28F9> (acedido em 06/05/2021).
- Frank, Eibe et al. (2000). «Naive Bayes for regression». Em: *Machine Learning* 41.1, pp. 5–25.
- GitHub - Dados relativos à pandemia COVID-19 em Portugal (2021). url: <https://github.com/dssg-pt/covid19pt-data> (acedido em 06/05/2021).
- Github - Oxford Covid-19 Government Response Tracker (OxCGRT) (2021). url: <https://github.com/OxCGRT/covid-policy-tracker> (acedido em 06/05/2021).

- Glez-Peña, Daniel et al. (2014). «Web scraping technologies in an API world». Em: *Briefings in bioinformatics* 15.5, pp. 788–797.
- Hastie, Trevor, Robert Tibshirani e Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- IHME - COVID-19 Projections (2021). url: <https://covid19.healthdata.org/global?view=total-deaths&tab=trend> (acedido em 23/02/2021).
- Imandoust, Sadegh Bafandeh e Mohammad Bolandraftar (2013). «Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background». Em: *International Journal of Engineering Research and Applications* 3.5, pp. 605–610.
- Jiao, Yasen e Pufeng Du (2016). «Performance measures in evaluating machine learning based bioinformatics predictors for classifications». Em: *Quantitative Biology* 4.4, pp. 320–330.
- Jordan, Patrick W et al. (1996). *Usability evaluation in industry*. CRC Press.
- Klaine, Paulo Valente et al. (2017). «A survey of machine learning techniques applied to self-organizing cellular networks». Em: *IEEE Communications Surveys & Tutorials* 19.4, pp. 2392–2431.
- Koen, Peter et al. (2001). «Providing clarity and a common language to the “fuzzy front end”». Em: *Research-Technology Management* 44.2, pp. 46–55.
- Malki, Zohair et al. (2020). «Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches». Em: *Chaos, Solitons & Fractals* 138, p. 110137.
- Mohri, Mehryar, Afshin Rostamizadeh e Ameet Talwalkar (2018). *Foundations of machine learning*. MIT press.
- Narkhede, Sarang (2018). «Understanding AUC-ROC Curve». Em: *Towards Data Science* 26.
- Ndiaye, Babacar Mbaye, Mouhamadou A. M. T. Balde e Diaraf Seck (ago. de 2020). «Visualization and machine learning for forecasting of COVID-19 in Senegal». Em: arXiv: 2008.03135. url: <http://arxiv.org/abs/2008.03135>.
- O ano em Pesquisa Google- Google Trends (2020). url: <https://trends.google.pt/trends/yis/2020/GLOBAL/> (acedido em 16/12/2020).
- Pandey, Amit e Achin Jain (2017). «Comparative analysis of KNN algorithm using various normalization techniques». Em: *International Journal of Computer Network and Information Security* 11.11, p. 36.
- Petropoulos, Fotios e Spyros Makridakis (2020). «Forecasting the novel coronavirus COVID-19». Em: *PloS one* 15.3, e0231236.
- Pordata - Base de Dados Portugal Contemporâneo (2021). url: <https://www.pordata.pt/> (acedido em 06/05/2021).
- Portugal: WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease (COVID-19) Dashboard (2020). url: <https://covid19.who.int/region/euro/country/pt> (acedido em 18/11/2020).
- Relatório de Vacinação - COVID-19 (2021). url: <https://covid19.min-saude.pt/relatorio-de-vacinacao/> (acedido em 06/05/2021).
- Rustam, F. et al. (2020). «COVID-19 Future Forecasting Using Supervised Machine Learning Models». Em: *IEEE Access* 8, pp. 101489–101499. doi: 10.1109/ACCESS.2020.2997311.
- invest: easy web scraping with R (2021). url: <https://www.bbc.com/news/business-51706225> (acedido em 20/01/2021).
- Sajadi, Mohammad M et al. (2020). «Temperature and latitude analysis to predict potential spread and seasonality for COVID-19». Em: *Available at SSRN* 3550308.

- Sistema de Informação dos Certificados de Óbito - Vigilância de Mortalidade* (2021). url: <https://evm.min-saude.pt> (acedido em 06/05/2021).
- Somvanshi, M. et al. (2016). «A review of machine learning techniques using decision tree and support vector machine». Em: *2016 International Conference on Computing Communication Control and automation (ICCUBE)*, pp. 1–7. doi: 10.1109/ICCUBE.2016.7860040.
- Taylor, Richard (1990). «Interpretation of the correlation coefficient: a basic review». Em: *Journal of diagnostic medical sonography* 6.1, pp. 35–39.
- Tosepu, Ramadhan et al. (2020). «Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia». Em: *Science of The Total Environment* 725, p. 138436.
- Vembandasamy, K, R Sasipriya e E Deepa (2015). «Heart diseases detection using Naive Bayes algorithm». Em: *International Journal of Innovative Science, Engineering & Technology* 2.9, pp. 441–444.
- Wang, Peipei et al. (2020). «Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics». Em: *Chaos, Solitons & Fractals* 139, p. 110058.
- Weather Data Weather API | Visual Crossing* (2021). url: <https://www.visualcrossing.com/> (acedido em 06/05/2021).
- Where The Latest COVID-19 Models Think We're Headed — And Why They Disagree | FiveThirtyEight* (2020). url: <https://projects.fivethirtyeight.com/covid-forecasts/> (acedido em 16/12/2020).
- Wirth, Rüdiger e Jochen Hipp (2000). «CRISP-DM: Towards a standard process model for data mining». Em: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Springer-Verlag London, UK.
- Withall, Stephen (2007). «Software requirement patterns». Em:
- World Health Organization (2020). *Coronavirus*. url: https://www.who.int/health-topics/coronavirus%7B%5C#%7Dtab=tab%7B%5C_%7D3 (acedido em 18/11/2020).
- Zhang, Shichao et al. (2017). «Learning k for knn classification». Em: *ACM Transactions on Intelligent Systems and Technology (TIST)* 8.3, pp. 1–19.

Apêndice A

Especificação de Requisitos

A.1 Formato Usado

Nesta secção serão especificados os os requisitos, onde estes serão enquadrados em casos de uso. Cada caso de uso é apresentado de acordo com o formato da tabela A.1.

Tabela A.1: Formato de descrição de casos de uso

Identificador	Nome do caso de uso
Objetivo	Curta descrição do que é pretendido com o caso de uso.
Descrição sumária	Pequeno texto com três ou quatro frases no máximo que descrevam de forma genérica o caso de uso.
Pré-condições	Condições que devem verificar-se no início do fluxo básico do caso de uso.
Pós-condições	Condições que devem verificar-se na conclusão do fluxo básico do caso de uso.
Estado	Descoberto / Descrito sumariamente / Esboçado / Delineado / Descrito detalhadamente / Descrito completamente

O estado indica a forma como está descrito o caso de uso. Os estados possíveis são os indicados na tabela A.2.

Tabela A.2: Estados de descrição de casos de uso

Estados	Característica	Objetivo
Descoberto	Indicação apenas do nome e dos Atores envolvidos. Ausência de descrição.	Identificar a existência do caso de uso
Descrito sumariamente	Inclui uma breve descrição do que o Ator pretende fazer com o caso de uso.	Identificar o propósito do caso de uso
Esboçado	Identificados os passos do fluxo de eventos do cenário principal. Identificados os cenários alternativos e passos em que ocorrem divergências.	Identificar a forma e extensão do caso de uso
Delineado	Descreve o cenário principal sob ponto de vista externo com enfoque deliberado na usabilidade.	Resumir a essência do caso de uso
Descrito detalhadamente	Acrescentar detalhes à descrição do cenário principal e detalhar também os cenários alternativos.	Permitir que o detalhe seja adicionado incrementalmente
Descrito completamente	Inclui também a lista de todos os requisitos no seu contexto, possibilitando a sua utilização em documentos de especificação de requisitos.	Incluir em especificação de requisitos

Os cenários de utilização são descritos através de fluxos de eventos. Estes podem ser de quatro tipos:

1. Fluxo básico – sequência de eventos considerada mais comum para atingir o objetivo do caso de uso (existe um só fluxo básico);
2. Fluxos alternativos – outras sequências de eventos para realizar o caso de uso;
3. Fluxos opcionais – sequências de eventos que estendem o fluxo básico do caso de uso;
4. Fluxos de exceção – sequências de eventos que normalmente levam a que o caso de uso não atinja o seu objetivo.

Os fluxos de eventos são descritos de acordo com o formato da tabela A.3.

Tabela A.3: Formato de descrição de fluxos de eventos

Nome do fluxo		
Passos	Descrição	Pontos de extensão
Nº do passo	Frase que descreve a interação entre o Ator e o sistema	Indicação opcional do nome do passo

Os pontos de extensão são necessários para referenciar o início e conclusão dos fluxos alternativos, opcionais ou de exceção no fluxo básico.

Após a indicação dos fluxos de eventos segue-se a lista de requisitos do caso de uso. Cada requisito é apresentado no formato da tabela A.4.

Tabela A.4: Formato de descrição de requisitos

Identificador	Sumário
Prioridade	Essencial / Condicional / Opcional
Descrição	Texto que descreva o requisito utilizando, se necessário, linguagem formal para que seja compreensível de forma unívoca por todo o público-alvo.
Motivação	Principais razões que levaram à inclusão do requisito. O texto pode ser extenso.
Informação adicional	Detalhes adicionais relacionados com o requisito, mas não essenciais para a sua implementação. O texto pode ser extenso e fazer uso de diagramas, figuras ou outros elementos de apoio.
Sugestões de implementação	Conjunto de sugestões que podem auxiliar a forma de implementação do requisito. O texto pode ser extenso.

A indicação de todos os elementos é obrigatória excepto os de informação adicional e de sugestões. A prioridade dos requisitos é a indicada na tabela A.5.

Tabela A.5: Prioridade de requisitos

Grau	Significado
Essencial	O produto não pode ser aceite enquanto o requisito não for implementado da forma descrita
Condicional	O requisito melhora o produto, mas não o torna inaceitável se não for implementado
Opcional	Representa uma função que pode ser interessante, dando ao fabricante a oportunidade de propor funcionalidades não previstas na especificação inicial.

A.2 Especificação de Requisitos

Na figura A.1 representa-se o modelo genérico de casos de uso do sistema de previsão da Evolução do COVID-19 por região sob a forma de um diagrama de pacotes. Cada pacote agrega uma ou mais partes do sistema que se destinam a suportar processos da organização e/ou a reunir um conjunto de funcionalidades. Em cada pacote incluem-se alguns exemplos Atores e casos de uso desenvolvidos para o sistema. São representadas igualmente as relações de dependência entre os pacotes. Os Atores são somente incluídos nos pacotes em que surgem pela primeira vez. Assume-se que os pacotes que não têm Atores, na realidade incluem os Atores dos pacotes de que dependem. Os pacotes para os quais não são indicadas dependências são pacotes com casos de uso transversais a todo o sistema.

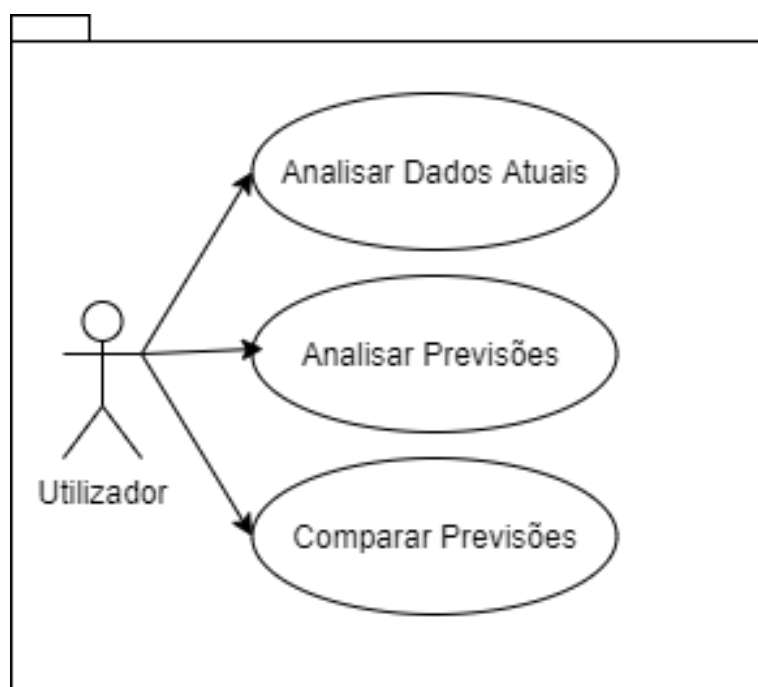


Figura A.1: Diagrama de casos de uso do projeto.

Na tabela A.6 apresenta-se uma descrição de cada Ator do sistema.

Tabela A.6: Atores do sistema de Previsão da evolução do COVID-19 por Regiões

Ator	Descrição resumida
Utilizador	O sistema terá apenas 1 tipo de ator, o utilizador, mas este poderá se incluir numa das 2 classes seguintes: Utilizador Geral – serão utilizadores normais, que usam a ferramenta para se informar como poderá ser o futuro, sendo que o sistema tem como objetivo incentivar os utilizadores a seguir as medidas que estejam a ser aplicadas (caso a previsão seja negativa, incentivar os seus utilizadores por exemplo ao uso de máscara, caso positiva, continuarem que estão num bom caminho); Utilizador Decisor – este tipo de utilizador irá usar a ferramenta para não só se informar como poderá ser o futuro, mas também, como base nessas previsões, tomarem ações. Esta classe descreve utilizadores que podem tomar decisões que definem a evolução da doença, como por exemplo diretores de empresas/fábricas até presidentes e políticos;

Na tabela A.7 descrevem-se sumariamente cada um dos pacotes do sistema de Previsão da

Evolução do COVID-19 por Regiões Os capítulos seguintes deste documento destinam-se à descrição em detalhe de cada pacote, dos casos de uso que inclui e da forma como cada Ator interage com o sistema.

Tabela A.7: Pacotes de casos de uso do sistema de Previsão da Evolução do COVID-19 por Regiões

<i>Pacote</i>	<i>Descrição resumida</i>
1. Dados Atuais	Permite ao utilizador analisar os dados atuais e reais assim como navegar entre eles
2. Previsões	Permite ao utilizador analisar os dados reais e as previsões num mesmo local ao nível do país e da região, assim como comparar os dados entre diferentes regiões.

A.2.1 Principais exclusões

Na tabela A.8 indicam-se os processos que não são apoiados pelo sistema de Previsão de Evolução do COVID-19 por Região. A justificação para as exclusões relaciona-se com o facto de o foco ser apenas fornecer previsões da evolução do vírus em Portugal e nas suas diferentes regiões.

Tabela A.8: Processos não suportados no sistema de Previsão de Evolução do COVID-19 por Região

<i>Processo</i>	<i>Descrição resumida</i>
Previsão da mortalidade do vírus	Prever com base em dados demográficos se uma determinada pessoa pode acabar por falecer se contrair o vírus.
Previsão mundial ou externa a Portugal	Prever a evolução do vírus no mundo ou em diferentes países.

A.2.2 Analisar dados atuais

Na tabela A.9 está presente uma descrição do caso de uso. Na restante seção serão apresentados os seus fluxos de execução e requisitos.

Tabela A.9: Descrição do Casos de Uso: Analisar dados atuais

DadosAtuais.01	Analisar dados atuais do país
Objetivo	Mostrar os dados mais recentes em relação ao vírus em Portugal e suas regiões
Descrição sumária	O sistema irá apresentar um mapa de Portugal contendo círculos consoante o número de casos confirmados da região, apresentado também o número atual de casos confirmados, casos ativos, de recuperados e de óbitos. Sendo possível alterar os dados apresentados para uma determinada região
Pós-condições	Abrir a Plataforma Web
Estado	Descrito completamente

A tabela A.10 representa o fluxo das interações entre o utilizador e o sistema para obter os dados atuais de Portugal. A tabela A.12 representa o fluxo das interações entre o utilizador e o sistema para obter os dados atuais de uma região de Portugal, já a tabela A.11 representa um percurso alternativo.

Tabela A.10: Fluxo de Execução do Casos de Uso: Analisar dados atuais do país

Fluxo básico		
Passo	Descrição	Ponto de extensão
1	O utilizador abre a plataforma	PE_1
2	O sistema apresenta os dados mais recentes	PE_2
	Conclusão do caso de uso atingindo o seu objetivo.	

Tabela A.11: Fluxo de Execução Alternativo do Casos de Uso: Analisar dados atuais do país

Fluxo alternativo 1:		Utilizador vindo de outra aba
Passo	Descrição	Ponto de extensão
	Início em PE_1 .	
1	O utilizador clica na aba "Situação Atual"	
	Retomar o fluxo básico em PE_2 .	

Tabela A.12: Fluxo de Execução do Caso de Uso: Analisar dados atuais de uma região

Fluxo básico		
Passo	Descrição	Ponto de extensão
1	O utilizador abre a plataforma ou a aba/página "Situação Atual"	
2	O sistema apresenta os dados referentes á situação atual de Portugal	
3	O utilizador seleciona que pretende obter dados referentes a uma Região	
4	O utilizador seleciona a região que pretende obter dados	
5	O sistema apresenta os dados da região pretendida	
	Conclusão do caso de uso atingindo o seu objetivo.	

Lista de requisitos

Tabela A.13: Requisito - Mapa de Portugal Círculos

REQ.001	Mapa de Portugal Círculos
Prioridade	Essencial
Descrição	Na aba de situação atual deverá ser apresentado o mapa de Portugal, contendo círculos consoante o número de casos confirmados na região, ou seja, quanto mais o número de infetados, maior será este círculo.
Motivação	Isto deve-se ao fato da necessidade de mostrar ao utilizador onde existe maior incidência do vírus.
Informação adicional	Este mapa deverá conter delimitadas as diferentes regiões do país: Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Açores e Madeira.

Tabela A.14: Requisito - Apresentar Situação do País

REQ.002	Apresentar Situação do País
Prioridade	Essencial
Descrição	Os dados mais atuais de total de casos confirmados, total de casos ativos, recuperados e óbitos deverá ser apresentado, assim como o número diário de casos, recuperados e óbitos juntamente com a data.
Motivação	Informar os utilizadores do ponto de situação mais atual.
Informação adicional	Os dados deverão ser agrupados de forma a que se perceba que o número total de casos confirmados contém o número diário de infetados, ou seja, o número total de casos infetados deverá estar junto com o diário, assim como o número total de óbitos deverá estar junto ao número diário de óbitos etc.

Tabela A.15: Requisito - Selecionar Região

REQ.003	Selecionar Região
Prioridade	Essencial
Descrição	Na aba de "Situação Atual", descrito no requisito REQ.001 e REQ.002, deverá existir um botão em que troca os dados atuais entre "Portugal" e Regiões", que como padrão está na opção "Portugal". Quando o utilizador alterar a opção para regional, todo o layout será ocultado, aparecendo apenas uma lista com as regiões, dentre as quais o utilizador irá escolher qual pretende
Motivação	Para serem apresentadas ao utilizador as informações sobre a região que ele pretende;
Informação adicional	As regiões disponíveis serão: Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Açores e Madeira.

Tabela A.16: Requisito - Dados Atuais da Região

REQ.004	Requisito Dados Atuais da Região
Prioridade	Essencial
Descrição	Após o utilizador selecionar qual a região que pretende verificar os dados atuais, o sistema deverá apresentar-lhe os dados.
Motivação	Para serem apresentadas ao utilizador as informações sobre a região que ele pretende;
Informação adicional	Os dados sobre a região são serão: data, total de casos confirmados, casos diários confirmados, total de recuperados, recuperados diários, total de óbitos, óbitos diários e total de casos ativos.

Tabela A.17: Requisito - Alternativa a Selecionar Região

REQ.005	Requisito Alternativa a Selecionar Região
Prioridade	Condicional
Descrição	Na aba de “Situação Atual”, descrito no requisito REQ.001 e REQ.002, o utilizador deverá clicar no mapa na região que pretende. Após selecionar a região os dados, que até então seria sobre o país, atualizam para os dados sobre a região selecionada.
Motivação	Será mais fácil para o utilizador em termos de usabilidade, mas mais complicado em termos de desenvolvimento;
Informação adicional	Os dados sobre a região são serão: data, total de casos confirmados, casos diários confirmados, total de recuperados, recuperados diários, total de óbitos, óbitos diários e total de casos ativos. Caso se use esta alternativa, o mapa deverá ter uma opção para as previsões voltarem a ser ao nível do país no lugar das previsões ao nível da região.

A.2.3 Analisar Previsões

Na tabela A.18 está presente uma descrição do caso de uso. Na restante seção serão apresentados os seus fluxos de execução e requisitos.

Tabela A.18: Descrição do Caso de Uso: Analisar previsões

Previsoes.01	Analisar previsões
Objetivo	Mostrar as previsões relação ao vírus em Portugal
Descrição sumária	O sistema irá apresentar gráficos em relação às previsões da evolução do vírus em Portugal ou suas regiões, sendo que é possível o utilizador selecionar quais previsões pretender ver
Pós-condições	Abrir a aba “Previsões”
Estado	Descrito completamente

A tabela A.19 representa o fluxo das interações entre o utilizador e o sistema apresentar o dashboard com as previsões de Portugal. A tabela A.20 representa o fluxo das interações entre o utilizador e o sistema para obter o dashboard com as previsões de uma região de Portugal.

Tabela A.19: Fluxos de execução do Caso de Uso: Analisar Previsões em Portugal

Fluxo básico		
Passo	Descrição	Ponto de extensão
1	O utilizador tem a plataforma web aberta	
2	O utilizador clica em “Previsões”	
3	O sistema apresenta gráficos com as previsões para Portugal	
	Conclusão do caso de uso atingindo o seu objetivo.	

Tabela A.20: Fluxo de Execução do Caso de Uso: Analisar Previsões em Região

Fluxo básico		
Passo	Descrição	Ponto de extensão
1	O utilizador tem a plataforma web aberta	
2	O utilizador clica em “Previsões”	
3	O sistema apresenta gráficos com as previsões para Portugal	
4	O utilizador escolhe uma região de Portugal	
5	O sistema apresenta gráficos com as previsões para a região selecionada	
	Conclusão do caso de uso atingindo o seu objetivo.	

Lista de requisitos

Tabela A.21: Requisito - Mapa de Portugal

REQ.006	Mapa de Portugal
Prioridade	Essencial
Descrição	Na aba de situação atual deverá ser apresentado o mapa de Portugal.
Motivação	Isto deve-se ao fato da necessidade de mostrar ao utilizador onde existe maior incidência do vírus.
Informação adicional	Este mapa deverá conter delimitadas as diferentes regiões do país: Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Açores e Madeira.

Tabela A.22: Requisito - Apresentar Gráfico Casos Confirmados

REQ.007	Requisito Apresentar Gráfico Casos Confirmados
Prioridade	Essencial
Descrição	O sistema deverá apresentar um gráfico de barras onde no eixo das estarão os dias (dia/mês) e no eixo das ordenadas o número de casos confirmados por dia. As primeiras 10 barras estarão a cinzento e representaram os dados reais, enquanto que as últimas 10 barras estarão a azul ciano e representam os dados das previsões.
Motivação	Permitir o utilizador analisar os dados reais e previsões no mesmo gráfico, assim como distingui-los.
Informação adicional	No topo de cada barra deverá estar o número de casos confirmados daquele dia, para facilitar a precessão do utilizador.

Tabela A.23: Requisito - Linha Gráfico Casos Confirmados

REQ.008	Requisito Linha Gráfico Casos Confirmados
Prioridade	Essencial
Descrição	O sistema deverá apresentar dentro do gráfico de barras descrito em REQ.007 uma linha, que representará o total de casos confirmados. Esta linha estará cinzenta nos 10 primeiros dias para representar os dados reais e azul ciano nos restantes. Quando o utilizador passar o cursor em cima da linha, ser-lhe-á apresentado o número total de casos confirmados naquele dia.
Motivação	Permitir o utilizador analisar a relação entre o número diário de casos confirmados e o respetivo total.
Informação adicional	Devido á desproporcionalidade entre o número total e diário, esta linha não estará no mesmo referencial que as barras de casos diários.

Tabela A.24: Requisito - Apresentar Gráfico Recuperados

REQ.009	Requisito Apresentar Gráfico Recuperados
Prioridade	Essencial
Descrição	O sistema deverá apresentar um gráfico de barras onde no eixo das estarão os dias (dia/mês) e no eixo das ordenadas o número de recuperados por dia. As primeiras 10 barras estarão a cinzento e representaram os dados reais, enquanto que as últimas 10 barras estarão verde e representam os dados das previsões.
Motivação	Permitir o utilizador analisar os dados reais e previsões no mesmo gráfico, assim como distingui-los.
Informação adicional	No topo de cada barra deverá estar o número de recuperados daquele dia, para facilitar a precessão do utilizador.

Tabela A.25: Requisito - Linha Gráfico Recuperados

REQ.010	Requisito Linha Gráfico Recuperados
Prioridade	Essencial
Descrição	O sistema deverá apresentar dentro do gráfico de barras descrito em REQ.009 uma linha, que representará o total de recuperados. Esta linha estará cinzenta nos 10 primeiros dias para representar os dados reais e verde nos restantes. Quando o utilizador passar o cursor em cima da linha, ser-lhe-á apresentado o número total de recuperados naquele dia.
Motivação	Permitir o utilizador analisar a relação entre o número diário de casos confirmados e o respetivo total.
Informação adicional	Devido á desproporcionalidade entre o número total e diário, esta linha não estará no mesmo referencial que as barras de recuperados diários.

Tabela A.26: Requisito - Apresentar Gráfico Casos Ativos

REQ.011	Requisito Apresentar Gráfico Casos Ativos
Prioridade	Essencial
Descrição	O sistema deverá apresentar um gráfico de barras onde no eixo das estarão os dias (dia/mês) e no eixo das ordenadas o número de casos ativos por dia. As primeiras 10 barras estarão a cinzento e representaram os dados reais, enquanto que as últimas 10 barras estarão a laranja e representam os dados das previsões.
Motivação	Permitir o utilizador analisar os dados reais e previsões no mesmo gráfico, assim como distingui-los.
Informação adicional	No topo de cada barra deverá estar o número de casos ativos daquele dia, para facilitar a precessão do utilizador.

Tabela A.27: Requisito - Linha Gráfico Casos Ativos

REQ.012	Requisito Linha Gráfico Casos Ativos
Prioridade	Essencial
Descrição	O sistema deverá apresentar dentro do gráfico de barras descrito em REQ.011 uma linha, que representará o total de casos ativos. Esta linha estará cinzenta nos 10 primeiros dias para representar os dados reais e laranja nos restantes. Quando o utilizador passar o cursor em cima da linha, ser-lhe-á apresentado o número total de casos ativos naquele dia.
Motivação	Permitir o utilizador analisar a relação entre o número diário de casos confirmados e o respetivo total.
Informação adicional	Devido á desproporcionalidade entre o número total e diário, esta linha não estará no mesmo referencial que as barras de recuperados diários.

Tabela A.28: Requisito - Apresentar Gráfico óbitos

REQ.013	Requisito Apresentar Gráfico óbitos
Prioridade	Essencial
Descrição	O sistema deverá apresentar um gráfico de barras onde no eixo das estarão os dias (dia/mês) e no eixo das ordenadas o número de óbitos por dia. As primeiras 10 barras estarão a cinzento e representaram os dados reais, enquanto que as últimas 10 barras estarão vermelhas e representam os dados das previsões.
Motivação	Permitir o utilizador analisar os dados reais e previsões no mesmo gráfico, assim como distingui-los.
Informação adicional	No topo de cada barra deverá estar o número de óbitos daquele dia, para facilitar a precessão do utilizador.

Tabela A.29: Requisito - Linha Gráfico óbitos

REQ.014	Requisito Linha Gráfico óbitos
Prioridade	Essencial
Descrição	O sistema deverá apresentar dentro do gráfico de barras descrito em REQ.013 uma linha, que representará o total de óbitos. Esta linha estará cinzenta nos 10 primeiros dias para representar os dados reais e vermelho nos restantes. Quando o utilizador passar o cursor em cima da linha, ser-lhe-á apresentado o número total de recuperados naquele dia.
Motivação	Permitir o utilizador analisar a relação entre o número diário de casos confirmados e o respetivo total.
Informação adicional	Devido á desproporcionalidade entre o número total e diário, esta linha não estará no mesmo referencial que as barras de recuperados diários.

Tabela A.30: Requisito - Selecionar Região

REQ.015	Requisito Selecionar Região
Prioridade	Essencial
Descrição	Na aba de “Previsões”, descrita do requisito REQ.006 e REQ.014 , deverá existir um botão em que troca os dados atuais entre “Portugal” e Regiões”, que como padrão está na opção “Portugal”. Quando o utilizador alterar a opção para regional, todo o layout será ocultado, aparecendo apenas uma lista com as regiões, dentre as quais o utilizador irá escolher qual pretende analisar os dados.
Motivação	Permitir ao utilizador obter previsões sobre uma determinada região.
Informação adicional	Assim como descrito em REQ.005 caso se verifique uma maior usabilidade, no lugar de se realizar este requisito, poderá se criar um mapa clicável, onde o utilizador clica na região que pretende, sem ocultar o layout de previsão de país e sem ter um botão para trocar de previsão. Caso se use esta alternativa, o mapa deverá ter uma opção para as previsões voltarem a ser ao nível do país no lugar das previsões ao nível da região.

Tabela A.31: Requisito - Previsão de Região

REQ.016	Previsão de Região
Prioridade	Essencial
Descrição	Após o utilizador escolher a região em REQ.015 deverão ser apresentados ao utilizador os gráficos com as previsões de: número de casos confirmados, casos ativos, recuperados e óbitos.
Motivação	Regiões diferentes têm características diferentes, o que irá levar a diferentes evoluções do vírus.
Informação adicional	Estes gráficos já foram descritos anteriormente, apenas no lugar de apresentar dados relativos ao país deverão apresentar dados da região selecionada: O gráfico de casos confirmados está descrito em REQ.007 e REQ.008 . O gráfico de recuperados está descrito em REQ.009 e REQ.010 . O gráfico de recuperados está descrito em REQ.011 e REQ.012 . O gráfico de casos óbitos está descrito em REQ.013 e REQ.014 .

A.2.4 Comparar Previsões

Na tabela A.32 está presente uma descrição do caso de uso. Na restante seção serão apresentados os seus fluxos de execução e requisitos.

Tabela A.32: Descrição do Caso de Uso: Comparar Previsões

Previsões.03	Analisar Comparar Previsões
Objetivo	Mostrar as previsões relação ao vírus em 2 regiões selecionadas pelo utilizador
Descrição sumária	O sistema irá apresentar gráficos em relação às previsões da evolução do vírus em Portugal. Estes gráficos irão conter previsões de duas regiões selecionadas pelo utilizador, para que este possa comparar a evolução do vírus nas 2 regiões.
Pós-condições	Abrir a aba "Comparar Previsões"
Estado	Descrito completamente

A tabela A.33 representa o fluxo das interações entre o utilizador e o sistema para obter o dashboard que permite comparar as previsões entre regiões.

Tabela A.33: Fluxo de Execução do Caso de Uso: Comparar Previsões

Fluxo básico		
Passo	Descrição	Ponto de extensão
1	O utilizador tem a plataforma web aberta	
2	O utilizador clica em "Comparar Previsões"	
3	O sistema fornece 2 listas com as regiões	
4	O utilizador escolhe as 2 regiões	
5	O sistema apresenta os dados das 2 regiões	
	Conclusão do caso de uso atingindo o seu objetivo.	

Lista de requisitos

Tabela A.34: Requisito - Selecionar Regiões para Comparação

REQ.017	Selecionar Regiões para Comparação
Prioridade	Essencial
Descrição	O sistema deverá apresentar ao utilizador duas "drop-down list" para o utilizador escolher as duas regiões que pretende analisar as previsões. Após selecionar uma opção numa das listas, essa opção não deverá constar na outra lista para facilitar a ação do utilizador
Motivação	O utilizador precisa indicar quais as regiões pretende obter informação.
Informação adicional	As regiões disponíveis serão: Portugal (para o caso de o utilizador querer comparar a evolução de Portugal e de uma região de Portugal), Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Açores e Madeira.

Tabela A.35: Requisito - Gráfico Comparação Casos Confirmados

REQ.018	Gráfico Comparação Casos Confirmados
Prioridade	Essencial
Descrição	Com as duas regiões selecionadas pelo utilizador, o sistema deverá apresentar as previsões da evolução do vírus, para isso deverá apresentar o gráfico de casos confirmados, onde deverão estar presentes as previsões das 2 regiões selecionadas. Para tal no eixo das abcissas irão ficar os dias e as regiões, de forma a que cada dia possua duas barras distintas cada uma correspondendo a uma região, em que as cores nos 10 primeiros dias que correspondem a dados reais serão cinzento claro para a primeira região selecionada e cinzento escuro para a segunda, os restantes dias correspondem às previsões, onde as barras deverão estar azul claro para a primeira região selecionada e azul escuro as barras correspondentes à segunda região selecionada. No eixo das ordenadas ficará o número de casos confirmados no referente dia.
Motivação	é necessário distinguir os dados de cada uma das regiões.
Informação adicional	No topo de cada barra deverá estar o seu valor. O sistema deverá apresentar dentro do gráfico de barras duas linhas tal como em REQ.008 que representarão o total de casos confirmados de cada região. Estas linhas deverão ter as mesmas cores que as barras dos gráficos (para a primeira região cinzento claro nos primeiros 10 dias e azul claro nos restantes, para a segunda região cinzento escuro nos primeiros dias e azul escuro nos restantes). Quando o utilizador passar o cursor em cima de uma linha, ser-lhe-á apresentado o número total de casos confirmados naquele dia.

Tabela A.36: Requisito - Gráfico Comparação Recuperados

REQ.019	Gráfico Comparação Recuperados
Prioridade	Essencial
Descrição	Com as duas regiões selecionadas pelo utilizador, o sistema deverá apresentar as previsões da evolução do vírus, para isso deverá apresentar o gráfico de recuperados, onde deverão estar presentes as previsões das 2 regiões selecionadas. Para tal no eixo das abcissas irão ficar os dias e as regiões, de forma a que cada dia possua duas barras distintas cada uma correspondendo a uma região, em que as cores nos 10 primeiros dias que correspondem a dados reais serão cinzento claro para a primeira região selecionada e cinzento escuro para a segunda, os restantes dias correspondem às previsões, onde as barras deverão estar verde claro para a primeira região selecionada e verde escuro as barras correspondentes á segunda região selecionada. No eixo das ordenadas ficará o número de recuperados no referente dia.
Motivação	é necessário distinguir os dados de cada uma das regiões.
Informação adicional	No topo de cada barra deverá estar o seu valor. O sistema deverá apresentar dentro do gráfico de barras duas linhas tal como em REQ.010 que representarão o total de recuperados de cada região. Estas linhas deverão ter as mesmas cores que as barras dos gráficos (para a primeira região cinzento claro nos primeiros 10 dias e verde claro nos restantes, para a segunda região cinzento escuro nos primeiros dias e verde escuro nos restantes). Quando o utilizador passar o cursor em cima de uma linha, ser-lhe-á apresentado o número total de recuperados naquele dia.

Tabela A.37: Requisito - Gráfico Comparação Casos Ativos

REQ.020	Gráfico Comparação Casos Ativos
Prioridade	Essencial
Descrição	Com as duas regiões selecionadas pelo utilizador, o sistema deverá apresentar as previsões da evolução do vírus, para isso deverá apresentar o gráfico de casos ativos, onde deverão estar presentes as previsões das 2 regiões selecionadas. Para tal no eixo das abcissas irão ficar os dias e as regiões, de forma a que cada dia possua duas barras distintas cada uma correspondendo a uma região, em que as cores nos 10 primeiros dias que correspondem a dados reais serão cinzento claro para a primeira região selecionada e cinzento escuro para a segunda, os restantes dias correspondem às previsões, onde as barras deverão estar laranja claro para a primeira região selecionada e laranja escuro as barras correspondentes á segunda região selecionada. No eixo das ordenadas ficará o número de casos ativos no referente dia.
Motivação	é necessário distinguir os dados de cada uma das regiões.
Informação adicional	No topo de cada barra deverá estar o seu valor. O sistema deverá apresentar dentro do gráfico de barras duas linhas tal como em REQ.012 que representarão o total de casos ativos de cada região. Estas linhas deverão ter as mesmas cores que as barras dos gráficos (para a primeira região cinzento claro nos primeiros 10 dias e laranja claro nos restantes, para a segunda região cinzento escuro nos primeiros dias e laranja escuro nos restantes). Quando o utilizador passar o cursor em cima de uma linha, ser-lhe-á apresentado o número total de casos confirmados naquele dia.

Tabela A.38: Requisito - Gráfico Comparação óbitos

REQ.021	Gráfico Comparação Casos óbitos
Prioridade	Essencial
Descrição	Com as duas regiões selecionadas pelo utilizador, o sistema deverá apresentar as previsões da evolução do vírus, para isso deverá apresentar o gráfico de óbitos, onde deverão estar presentes as previsões das 2 regiões selecionadas. Para tal no eixo das abcissas irão ficar os dias e as regiões, de forma a que cada dia possua duas barras distintas cada uma correspondendo a uma região, em que as cores nos 10 primeiros dias que correspondem a dados reais serão cinzento claro para a primeira região selecionada e cinzento escuro para a segunda, os restantes dias correspondem às previsões, onde as barras deverão estar vermelho claro para a primeira região selecionada e vermelho escuro as barras correspondentes á segunda região selecionada. No eixo das ordenadas ficará o número de óbitos no referente dia.
Motivação	é necessário distinguir os dados de cada uma das regiões.
Informação adicional	No topo de cada barra deverá estar o seu valor. O sistema deverá apresentar dentro do gráfico de barras duas linhas tal como em REQ.012 que representarão o total de óbitos cada região. Estas linhas deverão ter as mesmas cores que as barras dos gráficos (para a primeira região cinzento claro nos primeiros 10 dias e vermelho claro nos restantes, para a segunda região cinzento escuro nos primeiros dias e vermelho escuro nos restantes). Quando o utilizador passar o cursor em cima de uma linha, ser-lhe-á apresentado o número total de casos confirmados naquele dia.

A.2.5 Excerto de Especificação de Requisitos (Não-Funcionais)

Nesta seção será apresentado um excerto dos requisitos não-funcionais do sistema, que são requisitos necessários para o sistema sem ser descreverem casos de uso. Para isso será utilizado o modelo FURPS, que divide os requisitos não-funcionais do sistema em 4 categorias diferentes, para além dos requisitos funcionais:

- Usabilidade: requisitos voltados para os aspetos estéticos e interativos do sistema;
- Confiança: requisitos voltados para a disponibilidade, que geram confiança ao utilizador;
- Desempenho: requisitos voltados para o rendimento do sistema;
- Suporte: requisitos voltados para a configuração do sistema;

A tabela A.39 representa um requisito de usabilidade, visto que procura melhorar a experiência do utilizador, assim como melhorar a sua interação com o sistema. Enquanto o requisito

A.40 especifica o tempo diário que o sistema deverá estar online, garantindo ao utilizador a confiança de que quase a qualquer hora do dia poderá aceder e analisar as informações disponíveis.

Tabela A.39: Requisito Não-Funcional: Gráficos com cores destacadas e interativos

REQ.NF.01	Gráficos com cores destacadas e interativos
Prioridade	Essencial
Descrição	Todos os gráficos deverão utilizar os esquemas de cores já indicados, assim como possuírem uma legenda e um título, de forma a informar o utilizador do que se trata o mesmo. Os gráficos deverão ser interativos, da forma a ser tornar possível funções como "zoom in", "zoom out", deslizar o gráfico em todas as direções e reiniciar o gráfico (fazer o gráfico voltar às suas condições normais).
Motivação	É necessário criar ao utilizador uma experiência cativante, que o deixe confortável e com uma opinião positiva do sistema.

Tabela A.40: Requisito Não-Funcional: Disponibilidade do Sistema

REQ.NF.02	Servidor
Prioridade	Essencial
Descrição	O Sistema deverá estar disponível 23 horas e 50 minutos por dia, sendo os 10 minutos diários que sobram para o sistema atualizar a base de dados e gerar novas previsões. Sendo estas pausas de 10 minutos programadas para serem feitas às 15:50
Motivação	É necessário que o sistema estar quase 100% do tempo disponível. A hora selecionada para atualizar a base de dados e consequentemente gerar novas previsões tem a ver com os relatórios diários da pandemia que são publicados entre as 13:00 e 15:30.

A tabela A.41 descreve as características que o servidor deverá, no mínimo, possuir para executar todas as extrações de dados e gerar todas as previsões num espaço de tempo de 10 minutos. Enquanto o requisito A.42 aborda a compatibilidade do sistema, que deverá ser compatível com várias plataformas e navegadores.

Tabela A.41: Requisito Não-Funcional: Servidor

REQ.NF.03	Servidor
Prioridade	Essencial
Descrição	O sistema deverá ter um servidor com um processador com no mínimo 2.0GHz e memória RAM de no mínimo 6Gb.
Motivação	É necessário que o sistema esteja alojado num servidor bom para que este demore pouco tempo a executar todo o processo de atualizar a base de dados e gerar novas previsões, assim como tenha a capacidade de ter vários utilizadores a utilizar o sistema ao mesmo tempo.

Tabela A.42: Requisito Não-Funcional: Compatibilidade com várias plataformas e navegadores

REQ.NF.04	Compatibilidade com várias plataformas e navegadores
Prioridade	Essencial
Descrição	O sistema deverá ser feito de forma que seja compatível com navegadores como o Microsoft Edge, Google Chrome, Firefox e Safari. Sendo compatível também com Android e IOS. Em suma, o sistema deverá se adaptar a qualquer tamanho de ecrã e navegador.
Motivação	É necessário permitir o utilizador aceder ao sistema da forma que ele quiser.

A.2.6 Componentes Suplementares

Para complementar o requisito descrito na tabela A.41 segue-se uma descrição dos componentes do sistema presentes no servidor. A figura A.2 representa a estrutura base do sistema, através da figura é possível extrair os seguintes componentes:

- **Algoritmos de Extração:** O Sistema deverá possuir algoritmos para extração de informações das diversas fontes. Como por exemplo algoritmos de Web Scraping para retirar informações de sites, assim como outros algoritmos para outros tipos de fontes de informações. Estes algoritmos irão extrair dados e executar algum tratamento de dados e irá armazenar os dados na base de dados intermédia;
- **Modelos Machine Learning:** O sistema terá modelos capazes de prever o número de casos confirmados, casos ativos, óbitos e recuperados em Portugal e suas regiões;
- **Base de Dados:** O Sistema deverá possuir uma base de dados capaz de suportar os extraídos, assim como os dados após passarem por um tratamento compondo os dados reais de Portugal e suas regiões, assim como suportar as previsões;

Todos estes requisitos não-funcionais, estão presentes em formato de tabela em Anexo A.

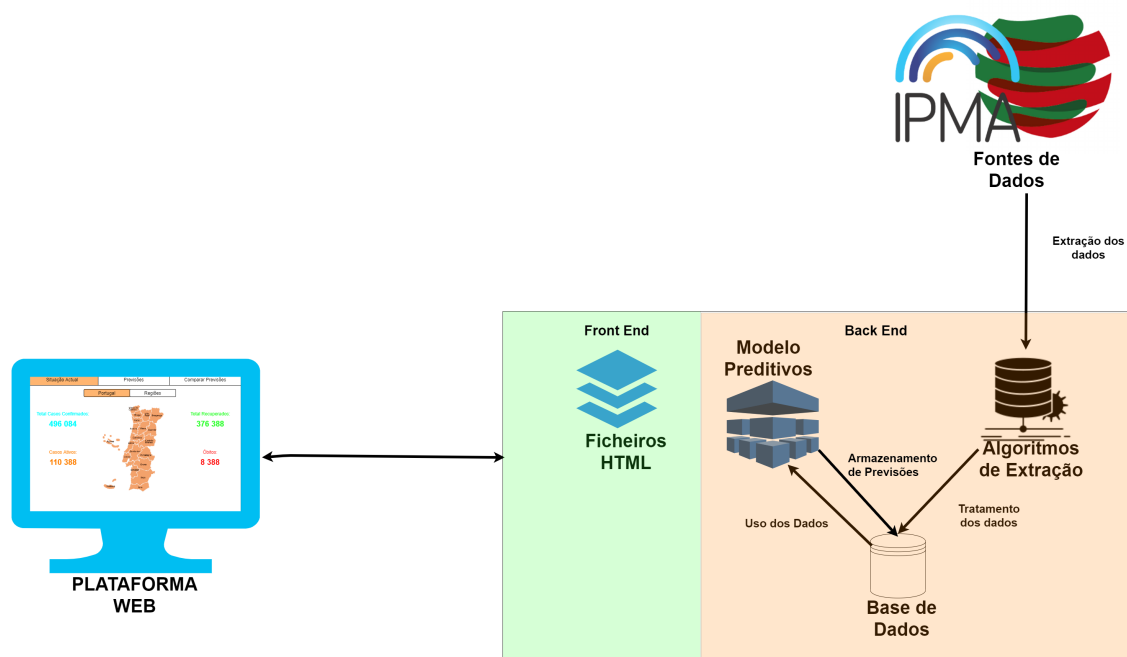


Figura A.2: Representação dos componentes do projeto.

Apêndice B

Implementação

B.1 Construção da Base de Dados

B.1.1 Lista de Dados

Neste apêndice serão apresentadas e definidas todas as variáveis que compõem a tabela principal da base dados, tabela esta de onde se extrai a informação para gerar previsões.

Dados Regionais

- **Região** - Região ao qual os restantes dados caracterizam;
- **Área** - Área da região em quilómetros quadrados;
- **População Total** - Número de indivíduos na região;
- **População Masculina** - Número de indivíduos do sexo masculino na região;
- **População Feminina** - Número de indivíduos do sexo feminino na região;
- **População 0 aos 14 anos** - Número de indivíduos na região com idades entre 0 e 14 anos;
- **População 15 aos 64 anos** - Número de indivíduos na região com idades entre 15 e 64 anos;
- **População mais de 65 anos** - Número de indivíduos na região com mais de 65 anos;
- **Densidade Populacional** - Média de indivíduos por quilometro quadrado na região;
- **População Sector Primário** - Número de indivíduos empregados no sector primário (extracção e produção de materiais);
- **População Sector Secundário** - Número de indivíduos empregados no sector secundário (transformação de materiais);
- **População Sector Terciário** - Número de indivíduos empregados no sector terciário (venda de produtos e serviços);
- **Índice de Envelhecimento** - Número de idosos na região por cada 100 jovens;
- **Número de Médicos** - Número de médicos na região;
- **Número de Enfermeiros** - Número de enfermeiros na região;
- **Número de Farmacêuticos** - Número de farmacêuticos na região;

- **Percentagem de Turismo** - Percentagem de estrangeiros a ocupar os estabelecimentos hoteleiros na região;

Dados Meteorológicos

- **Data** - Dia em que as outras variáveis se verificam;
- **Local** - Cidade onde as outras variáveis se verificam, variável esta que ser substituída pela respectiva região;
- **Temperatura Máxima** - Temperatura máxima registada no dia e local específico;
- **Temperatura Mínima** - Temperatura mínima registada no dia e local específico;
- **Temperatura Média** - Temperatura média registada no dia e local específico;
- **Precipitação** - Quantidade de precipitação de água em milímetros no dia e local específico;
- **Velocidade do Vento** - Velocidade do vento em quilómetros por hora no dia e local específico;
- **Nuvens** - Percentagem de céu coberto por nuvens no dia e local específico;
- **Humidade** - Quantidade de vapor de água no ar, no dia e local específico;
- **Condições** - Condições meteorológicas no dia e local específico;

Dados da Vacinação

- **Data** - Data de publicação do relatório;
- **Semana** - Semana que os restantes dados representam;
- **Região** - Região que os restantes dados representam;
- **Pessoas Vacinadas com 1ª Dose Semanal** - Número de pessoas vacinadas com a primeira dose de uma vacina na referida semana e região;
- **Pessoas Vacinação Completa Semanal** - Número de pessoas que ficaram com a vacinação completa (duas doses da vacina) na referida semana e região;
- **Pessoas vacinadas com dose em fase de registo Semanal**- Número de pessoas que é desconhecida a dosagem de vacina que possuem na referida semana e região;
- **Total Vacinação Semanal**- Total de vacinas administradas na referida semana e região;
- **Pessoas Vacinadas com 1ª Dose Total** - Número total de pessoas vacinadas com a primeira dose na referida região até à referida semana;
- **Pessoas Vacinação Completa Total** - Número total de pessoas vacinadas com a dose completa na referida região até à referida semana;
- **Pessoas vacinadas com dose em fase de registo Total**- Número total de pessoas vacinadas ainda em fase de registo na referida região até à referida semana;
- **Total Vacinação** - Total de Vacinas administradas até à referida semana;

- **Cobertura 1ª Dose** - Percentagem de pessoas da região com a 1ª dose da vacina até á referida semana;
- **Cobertura Vacinação Completa** - Percentagem de pessoas da região com a dose completa da vacina até á referida semana;

Dados do Índice de Transmissibilidade

- **Data** - Dia em que se verificaram os restantes dados na região do ficheiro;
- **Índice de Transmissibilidade** - Número de pessoas que um infetado pode infectar na região do ficheiro;
- **Índice de Transmissibilidade (Índice Confiança 95%- Limite Inferior)** - Índice de Transmissibilidade com o seu verdadeiro valor entre o limite inferior e o superior com uma confiança de 95%;
- **Índice de Transmissibilidade (Índice Confiança 95%- Limite Superior)** - Índice de Transmissibilidade com o seu verdadeiro valor entre o limite inferior e o superior com uma confiança de 95%;

Dados sobre o Covid-19

- **Data** - Data a que os dados se referem;
- **Total Confirmados** - Número de confirmados em Portugal;
- **Total Regional Confirmados** - Uma variável para cada região, contendo o respectivo total de confirmados;
- **Confirmados Diário** - Número de confirmados no dia anterior dos dados;
- **Confirmados Total por género** - Uma variável para cada género, contendo o total confirmados em Portugal;
- **Confirmados Total por faixa etária e género** - Uma variável para cada faixa etária e género, contendo o total confirmados em Portugal (por exemplo, número de indivíduos masculinos confirmados entre os 0 e 9 anos);
- **Total Recuperados** - Número total de recuperados em Portugal;
- **Total Recuperados Regional**- Uma variável para cada região, contendo o respectivo total de recuperados (Variáveis que estão vazias porque a DGS não divulga o número de recuperados de cada região);
- **Total Óbitos** - Número total de óbitos em Portugal;
- **Total Óbitos Regional**- Uma variável para cada região, contendo o respectivo total de óbitos;
- **Óbitos Total por género** - Uma variável para cada género, contendo o total óbitos em Portugal;
- **Óbitos Total por faixa etária e género** - Uma variável para cada faixa etária e género, contendo o total óbitos em Portugal (por exemplo, número de indivíduos masculinos que faleceram entre os 0 e 9 anos);

- **Internados** - Número total de internados em Portugal;
- **Internados UCI**- Número total de internados em unidades de cuidados intensivos em Portugal;
- **Vigilância**- Número de casos sob vigilância pelas autoridades de saúde;
- **Ativos** - Número de casos ativos em Portugal;

Mortalidade

- **Mortes Registadas** - mortes registadas num dia e local específico independentemente da causa;
- **Mortalidade Base** - média de mortes esperadas num dia e local específico independentemente da causa;
- **Excesso de Mortalidade** - excesso de mortes num dia e local;

Medida de Atenuação do Covid-19

- **Uso de Máscara** - Variável que informa se é obrigatório o uso de máscara;
- **Restaurantes** - Variável que informa o estado dos restaurantes, ou seja, se estão abertos ou fechados (tendo em conta que embora fechados, poderia ser permitido entregas);
- **Ensino - 1ºCiclo**- Variável que informa se as escolas de 1º ciclo estão fechadas ou abertas (ou simplesmente sem aulas, no caso das férias de verão);
- **Ensino - 2º e 3º Ciclo**- Variável que informa se as escolas de 2º e 3º ciclo estão fechadas ou abertas (ou simplesmente sem aulas, no caso das férias de verão);
- **Ensino Secundário**- Variável que informa se as escolas de ensino secundário estão fechadas ou abertas (ou simplesmente sem aulas, no caso das férias de verão);
- **Ensino Superior**- Variável que informa se as universidades estão fechadas ou abertas (ou simplesmente sem aulas, no caso das férias de verão);
- **Estado do País** - Estado do país: Emergência, Calamidade ou Normal;
- **Transportes Públicos** - Medidas implementados sobre os transportes públicos: funcionamento normal, funcionamento reduzido ou sem funcionamento;
- **Confinamento** - Medidas sobre ficar em casa: sem medidas, recomendado ficar em casa, proibido sair exceto saídas essenciais ou proibido sair, a não ser uma vez por semana ;
- **Trabalho** - Medidas sobre o trabalho: normal funcionamento, encerramento/teletrabalho recomendado, encerramento/tele trabalho para alguns sectores ou encerramento/teletrabalho, exceto essenciais;
- **Viagens Internas** - Medidas aplicadas a viagens entre cidades: sem restrições, não recomendo viajar entre cidades ou proibido viajar entre cidades;

- **Chegadas de Estrangeiros** - Medidas para a entrada de estrangeiros no país: Sem Restrições, Chegadas analisadas, Quarentena obrigatória á chegada de algumas regiões, Proibidas chegadas de certas regiões e/ou+ Proibidas chegadas de qualquer região;
- **Estratégia de Testagem** - Estratégia adotada para testar a população: sem estratégia, testar mediante um certo critério, testar apenas pessoas com sintomas de covid-19 e/ou teste público (qualquer pessoa pode ser testada);

B.1.2 Excertos de Código

Neste apêndice serão apresentados todos os excertos de código que ajudam na compreensão de como foi construída a base de dados.

Dados Regionais

```

1 #Carregamento de dados
2 area <- read_excel("dados/area.xlsx", range = "B12:C21")
3 densidade <- read_excel("dados/densidade.xlsx", range = "B12:C21")
4 envelhecimento <- read_excel("dados/Indice_envelhecimento.xlsx", range =
  "B12:C21")
5 pessoal_saude <- read_excel("dados/pessoal_saude.xlsx", range = "B12:E21
  ")
6 popula <- read_excel("dados/popula_faixaetaria.xlsx", range = "B12
  :F21")
7 sexo <- read_excel("dados/sexo.xlsx", range = "B12:D21")
8 emprego <- read_excel("dados/setor_atividade.xlsx", range = "B12:F21")
9 turismo <- read_excel("dados/turismo.xlsx", range = "B12:C21")
10
11 #Tratamento de Dados (Corre dos nomes das colunas)
12 colnames(area) <- c("Regiao", "rea")
13 colnames(densidade) <- c("Regiao", "Densidade P.")
14 colnames(envelhecimento) <- c("Regiao", "I. Envelhecimento")
15 colnames(pessoal_saude) <- c("Regiao", "N. Médicos", "N. Enfermeiros", "N.
  Farmaceuticos")
16 colnames(popula) <- c("Regiao", "P. Total", "P.0-14", "P.15-64", "P.65
  Mais")
17 colnames(sexo) <- c("Regiao", "P. Masculina", "P. Feminina")
18 colnames(emprego) <- c("Regiao", "Total", "P. Setor_Primario", "P.
  Setor_Secundario", "P. Setor_Terciario")
19 colnames(turismo) <- c("Regiao", "Perc. Turismo")
20
21 #Altera da grandeza (Milhares -> Indivíduos)
22 emprego$Total <- NULL
23 emprego$P.Setor_Primario <- emprego$P.Setor_Primario * 1000
24 emprego$P.Setor_Secundario <- emprego$P.Setor_Secundario * 1000
25 emprego$P.Setor_Terciario <- emprego$P.Setor_Terciario * 1000
26
27 # Juntar todos os dados das regiões numa única tabela
28 df_list <- list(area, sexo, densidade, emprego, envelhecimento,
  pessoal_saude, popula, turismo)
29 data <- merge_recurse(df_list)

```

Listing B.1: Carregamento e Tratamento dos dados sobre as Regiões (R).

Dados Meteorológicos

```

1 #Carregamento de dados de 2021
2 ### Madeira
3 x<-read.csv("https://weather.visualcrossing.com/
  VisualCrossingWebServices/rest/services/weatherdata/history?
  aggregateHours=24&combinationMethod=aggregate&startDateTime
  =2021-03-02T00%3A00%3A00&endDateTime=2021-03-31T00%3A00%3A00&
  maxStations=-1&maxDistance=-1&contentType=csv&unitGroup=metric&
  locationMode=single&key=94M6F3KYV4UZACB4MKT7XR2AF&dataElements=
  default&locations=Funchal")
4 colnames(x) <- c("Local", "datas", "TemperaturaMax", "TemperaturaMin", "
  TemperaturaMed", "VentoFrio",
5           "IndiceCalor", "Precipitacao", "Neve", "VelocidadeVento", "
  RajadaVento", "Visibilidade",
6           "Nuvens", "Humidade", "TipoTempo", "Condicoes")
7 x$Regiao<-"Regio Autnoma da Madeira"
8 x$datas<-anytime::anydate(x$datas)
9 x <- x[, -c(1,6,7,9,11,12,15)]
10 x[is.na(x)] <- 0
11 dados_meteo<- rbind(dados_meteo,x)

```

Listing B.2: Exemplo de Extracção e Tratamento de dados meteorológicos (R).

Dados Vacinação

```

1 #Define o URL onde se encontra os dados sobre a vacina
2 URL <- "https://covid19.min-saude.pt/relatorio-de-vacinacao/"
3 #Extrai dados do URL (Web Scraping)
4 dados_extraidos <- read_html(URL)
5 #Obtem todos os URLs presentes na página indicada
6 urls_extraidos<-html_attr(html_nodes(dados_extraidos, "a"), "href")
7 #Encontra o URL que direciona para o ficheiro csv com o Dataset mais
  atualizado da vacina
8 url_dados_vacina<-urls_extraidos[grep(".csv$", urls_extraidos)]
9 #Extrai os dados de vacina presentes no CSV encontrado
10 dados_vacina<-read.csv2(url_encode(url_dados_vacina))
11 #Limpeza de variáveis que não serão utilizadas
12 rm(URL, dados_extraidos, urls_extraidos, url_dados_vacina)

```

Listing B.3: Extração dos dados da Vacinação (R).

```

1 dados_vacina$Regiao<-gsub("ARS ", "", dados_vacina$Regiao)
2 dados_vacina$Regiao<-gsub("Lisboa e Vale do Tejo", "rea Metropolitana
  de Lisboa", dados_vacina$Regiao)
3 dados_vacina$Regiao<-gsub("RA_Acores", "Regio Autnoma dos Aores ",
  dados_vacina$Regiao)

```

Listing B.4: Exemplo correção dos dados da Vacinação (R).

Dados do Índice de Transmissibilidade

```

1 URL<-"http://www.insa.min-saude.pt/category/areas-de-atuacao/
  epidemiologia/covid-19-curva-epidematica-e-parametros-de-
  transmissibilidade/"
2 dados_extraidos <- read_html(URL)
3 tabelas <- html_nodes(dados_extraidos, "table")
4 tabelas<-tabelas[[2]]
5 a<-html_nodes(tabelas, "td")
6 urls_extraidos<-html_attr(html_nodes(a, "a"), "href")
7 texto<-as.character(html_text(html_nodes(a, "a")))
8 RegioesRt <- data.frame(urls_extraidos, texto)
9 RegioesRt$urls_extraidos<-as.character(RegioesRt$urls_extraidos)
10 rm(URL, tabelas, a, urls_extraidos, dados_extraidos, texto)

```

Listing B.5: Obtenção dos dados Índice Transmissibilidade (R).

Dados sobre evolução Covid-19

```

1 for(i in 2:nrow(x)){
2   #Confirmados
3   x$confirmados_Centro_diario[i]<-(x$confirmados_arscentro[i]-
4     x$confirmados_arscentro[i-1])
5   x$confirmados_Norte_diario[i]<-(x$confirmados_arsnorte[i]-
6     x$confirmados_arsnorte[i-1])
7   x$confirmados_Lisboa_diario[i]<-(x$confirmados_arslvt[i]-
8     x$confirmados_arslvt[i-1])
9   x$confirmados_Alentejo_diario[i]<-(x$confirmados_arsalentejo[i]-
10     x$confirmados_arsalentejo[i-1])
11   x$confirmados_Algarve_diario[i]<-(x$confirmados_arsalgarve[i]-
12     x$confirmados_arsalgarve[i-1])
13   x$confirmados_Acores_diario[i]<-(x$confirmados_acores[i]-
14     x$confirmados_acores[i-1])
15   x$confirmados_Madeira_diario[i]<-(x$confirmados_madeira[i]-
16     x$confirmados_madeira[i-1])
17
18   # tos
19   x$obitos_Centro_diario[i]<-(x$obitos_arscentro[i]-x$obitos_arscentro[
20     i-1])
21   x$obitos_Norte_diario[i]<-(x$obitos_arsnorte[i]-x$obitos_arsnorte[i
22     -1])
23   x$obitos_Lisboa_diario[i]<-(x$obitos_arslvt[i]-x$obitos_arslvt[i-1])
24   x$obitos_Alentejo_diario[i]<-(x$obitos_arsalentejo[i]-
25     x$obitos_arsalentejo[i-1])
26   x$obitos_Algarve_diario[i]<-(x$obitos_arsalgarve[i]-
27     x$obitos_arsalgarve[i-1])
28   x$obitos_Acores_diario[i]<-(x$obitos_acores[i]-x$obitos_acores[i-1])
29   x$obitos_Madeira_diario[i]<-(x$obitos_madeira[i]-x$obitos_madeira[i
30     -1])

```

Listing B.6: Ciclo que cria variáveis diária da evolução do Covid-19 (R).

```

1 for (j in 1:nrow(x)) {
2   for (i in 1:nrow(dados_pandemia)) {
3     if(dados_pandemia$Regiao[i] == "Alentejo" && dados_pandemia$datas[i]
4       ==x$data[j]){
5       dados_pandemia$RegTotal_Confirmados[i]<-x$confirmados_arsalentejo[j]
6       dados_pandemia$RegTotal_Obitos[i]<-x$obitos_arsalentejo[j]
7       dados_pandemia$RegDiario_Confirmados[i]<-
8         x$confirmados_Alentejo_diario[j]
9       dados_pandemia$RegDiario_Obitos[i]<-x$obitos_Alentejo_diario[j]
10
11      dados_pandemia$Confirmados_Total[i] <- x$confirmados[j]
12      dados_pandemia$Confirmados_Total_Diario[i]<-x$confirmados_novos[j]
13      dados_pandemia$Obitos_Total[i]<- x$obitos[j]
14      dados_pandemia$Obitos_Diario[i]<- x$obitos_diario[j]
15      dados_pandemia$Recuperados_Total[i]<- x$recuperados[j]
16      dados_pandemia$Recuperados_Diario[i]<- x$recuperados_diario[j]
17      dados_pandemia$Ativos_Total[i]<- x$ativos[j]
18      dados_pandemia$Ativos_Diario[i]<- x$ativos_diario[j]
19      dados_pandemia$Vigilancia[i]<- x$vigilancia[j]
20      dados_pandemia$Vigilancia_Diario[i]<- x$vigilancia_diario[j]
21      dados_pandemia$Internados[i]<- x$internados[j]
22      dados_pandemia$Internados_Diario[i]<-x$internados_diario[j]
23      dados_pandemia$InternadosUCI[i]<- x$internados_uci[j]
24      dados_pandemia$InternadosUCI_Diario[i]<- x$internadosuci_diario[j]
25      dados_pandemia$Confirmados_Diario_F[i]<- x$confirmados_f_diario[j]
26      dados_pandemia$Confirmados_Diario_M[i]<- x$confirmados_m_diario[j]
27      dados_pandemia$Obitos_Diario_F[i]<- x$obitos_diario_f[j]
28      dados_pandemia$Obitos_Diario_M[i]<- x$obitos_diario_m[j]

```

Listing B.7: Ciclo que armazena as variáveis da evolução do Covid-19 (R).

Medida de Atenuação do Covid-19

```

1 #Inicia da variavel
2 dados_pandemia$EnsinoB1_Pre<-"Abertos"
3 dados_pandemia$EnsinoB23<-"Abertos"
4 dados_pandemia$EnsinoSecundario<-"Abertos"
5 dados_pandemia$EnsinoSuperior<-"Abertos"
6
7 for (j in 1:nrow(dados_pandemia)) {
8   for(i in 1:nrow(ensino)){
9     if(dados_pandemia$datas[j]>=ensino$Data_Inicio[i] &
10      dados_pandemia$datas[j]<=ensino$Data_Fim[i]){
11       if (ensino$Instituicoes[i]=="Preschool or childcare facilities
12        closed/locked down") {
13         dados_pandemia$EnsinoB1_Pre[j]<-"Fechado"
14       }
15       if (ensino$Instituicoes[i]=="Primary Schools closed/locked down
16        ") {
17         dados_pandemia$EnsinoB23[j]<-"Fechado"}
18       if (ensino$Instituicoes[i]=="Secondary Schools closed/locked
19        down") {
20         dados_pandemia$EnsinoSecundario[j]<-"Fechado"}
21       if (ensino$Instituicoes[i]=="Higher education institutions
22        closed/locked down") {
23         dados_pandemia$EnsinoSuperior[j]<-"Fechado"}}
24       if(dados_pandemia$datas[j]<=anydate("2020-09-13") &
25        dados_pandemia$datas[j]>=anydate("2020-06-26")){
26         dados_pandemia$EnsinoB1_Pre[j]<-"Sem Aulas"
27         dados_pandemia$EnsinoB23[j]<-"Sem Aulas"
28         dados_pandemia$EnsinoSecundario[j]<-"Sem Aulas"}
29       if (dados_pandemia$datas[j]<=anydate("2020-10-01") &
30        dados_pandemia$datas[j]>=anydate("2020-07-21")){
31         dados_pandemia$EnsinoSuperior[j]<-"Sem Aulas"}}}
32 rm(i,j,ensino)

```

Listing B.8: Exemplo da criação de variáveis (R).

B.2 Modelos

B.2.1 Excertos de Código

Neste apêndice serão apresentados alguns excertos de código que ajudam na compreensão de como foram construídos ou modelos.

```

1 for(i in 1:5){
2
3   pred$val[i]<-forecast(arima, level=c(95), xreg = teste_x[,c(1:10,15:25)
4     ][["mean"]][2])
5   pred$Linf[i]<-forecast(arima, level=c(95), xreg = teste_x[,c(1:10,15:25)
6     ][["lower"]][2])
7   pred$Lsup[i]<-forecast(arima, level=c(95), xreg = teste_x[,c(1:10,15:25)
8     ][["upper"]][2])
9
10  teste_x[1,]<-teste_x[2,]
11  teste_x[2,c(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25)]<-
12    teste_x[2,c(1,2,3,4,6,7,8,9,11,12,13,14,16,17,18,19,21,22,23,24)]
13  teste_x[2,c(1,6,11,16,21)]<-NA
14  teste_x[2,c(1)]<-sum(teste_x[2,c(2,3,4,5)])/4
15  teste_x[2,c(6)]<-teste_x[2,7]+pred$val[i]
16  teste_x[2,c(11)]<-pred$val[i]
17  teste_x[2,c(16)]<-sum(teste_x[2,c(17,18,19,20)])/4
18  teste_x[2,c(21)]<-sum(teste_x[2,c(22,23,24,25)])/4
19 }

```

Listing B.9: Exemplo das transformações entre previsões (R).

```

1 #Calcular Correla das variáveis que se pretende prever
2 str(dados)
3 #Converter todas as variáveis de caracteres para nmericos
4 dados[sapply(dados, is.character)] <- lapply(dados[sapply(dados, is.
5   character)], as.factor)
6 dados[sapply(dados, is.factor)] <- lapply(dados[sapply(dados, is.factor)
7   ], as.numeric)
8
9 #Analisar quais variáveis possuem pouco variancia
10 nearZeroVar(dados)
11 confirmadosVARS<-cor(dados[c(-1,-8,-29:-65,-nearZeroVar(dados))],
12   dados$RegDiario_Confirmados)
13 obitosVARS<-cor(dados[c(-1,-10,-29:-65,-nearZeroVar(dados))],
14   dados$RegDiario_Obitos)
15 recuperadosVARS<-cor(dados[c(-1,-16,-29:-65,-nearZeroVar(dados))],
16   dados$Recuperados_Diario)

```

Listing B.10: Correlação entre as variáveis que se pretende prever e as restantes (R).

B.2.2 Testes

Neste apêndice serão apresentadas tabelas que descrevem os testes realizados com modelos.

Teste - Desempenho com diferentes quantidades de dados históricos

Tabela B.1: Testes de desempenho com diferentes quantidades de dados históricos a executar em todos os modelos.

Teste	Região	Previsão	Intervalo Treino	Intervalo Previsão	Variáveis	Histórico
1-C	Norte	Confirmados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	3
2-C	Norte	Confirmados	01/03/2020 a 28/03/2021	01/03/2021 a 05/03/2021	2	5
3-C	Norte	Confirmados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	7
4-O	Norte	Óbitos	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	3
5-O	Norte	Óbitos	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	5
6-O	Norte	Óbitos	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	7
7-R	Portugal	Recuperados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	3
8-R	Portugal	Recuperados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	5
9-R	Portugal	Recuperados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	2	7

Teste - Desempenho com diferentes conjuntos de variáveis

Tabela B.2: Testes de desempenho com diferentes conjuntos de variáveis a executar em todos os modelos.

Teste	Região	Previsão	Intervalo Treino	Intervalo Previsão	Variáveis	Histórico
0-C	Norte	Confirmados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Sem Variáveis	RF-3 KNN-5 ARIMA-5
1-C	Norte	Confirmados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Variáveis Meteorológicas	RF-3 KNN-5 ARIMA-5
2-C	Norte	Confirmados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Variáveis de Atenuação Pandêmica	RF-3 KNN-5 ARIMA-5
3-C	Norte	Confirmados	01/03/2020 a 28/03/2021	01/03/2021 a 05/03/2021	Variáveis de maior correlação	RF-3 KNN-5 ARIMA-5
4-C	Norte	Confirmados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Variáveis Meteorológicas, Atenuação Pandêmica e de maior correlação	RF-3 KNN-5 ARIMA-5
0-O	Norte	Óbitos	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Sem Variáveis	RF-3 KNN-5 ARIMA-5
5-O	Norte	Óbitos	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Variáveis sobre mortalidade	RF-3 KNN-5 ARIMA-5
6-O	Norte	Óbitos	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Variáveis sobre internamentos	RF-3 KNN-5 ARIMA-5
7-O	Norte	Óbitos	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Variáveis de maior correlação	RF-3 KNN-5 ARIMA-5

Tabela B.3: Continuação - Testes de desempenho com diferentes conjuntos de variáveis a executar em todos os modelos.

Teste	Região	Previsão	Intervalo Treino	Intervalo Previsão	Variáveis	Histórico
8-O	Norte	Óbitos	01/02/2021 a 28/02/2021	01/03/2021 a 05/03/2021	Mortalidade, Internamentos e de Maior Correlação	RF-3 KNN-5 ARIMA-5
0-R	Portugal	Recuperados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Sem Variáveis	RF-7 KNN-3 ARIMA-5
9-R	Portugal	Recuperados	01/06/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Internamentos	RF-7 KNN-3 ARIMA-5
10-R	Portugal	Recuperados	01/09/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Atualidade	RF-7 KNN-3 ARIMA-5
11-R	Portugal	Recuperados	01/12/2020 a 28/02/2021	01/03/2021 a 05/03/2021	Maior Correlação	RF-7 KNN-3 ARIMA-5
12-R	Portugal	Recuperados	01/02/2021 a 28/02/2021	01/03/2021 a 05/03/2021	Internamentos, Atualidade e de Maior Correlação	RF-5 KNN-7 ARIMA-5

Teste - Desempenho com diferentes intervalos de treino

Tabela B.4: Teste de desempenho com diferentes intervalos de treino a executar em todos os modelos

Teste	Região	Previsão	Intervalo Treino	Intervalo Previsão	Variáveis	Dias Pas-sado
0-C	Norte	Confirmados	01/03/2020 a 28/02/2021	01/03/2021 a 05/03/2021	RF-Maior Correlação; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
1-C	Norte	Confirmados	01/06/2020 a 28/02/2021	01/03/2021 a 05/03/2021	RF-Maior Correlação; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
2-C	Norte	Confirmados	01/09/2020 a 28/02/2021	02/03/2021 a 05/03/2021	RF-Maior Correlação; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
3-C	Norte	Confirmados	01/12/2020 a 28/02/2021	01/03/2021 a 05/03/2021	RF-Maior Correlação; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
4-C	Norte	Confirmados	01/02/2021 a 28/02/2021	01/03/2021 a 05/03/2021	RF-Maior Correlação; KNN-Sem Variáveis; ARRIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7

Tabela B.5: Teste de desempenho com diferentes intervalos de treino a executar em todos os modelos(Continuação)

Teste	Região	Previsão	Intervalo Treino	Intervalo Previsão	Variáveis	Dias Pas-sado
0-O	Norte	Óbitos	01/02/2021 a 01/03/2020	01/03/2021 a 05/03/2021	RF-Sem Va- riáveis; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
5-O	Norte	Óbitos	01/02/2021 a 01/06/2020	01/03/2021 a 05/03/2021	RF-Sem Va- riáveis; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
6-O	Norte	Óbitos	01/02/2021 a 01/09/2020	01/03/2021 a 05/03/2021	RF-Sem Va- riáveis; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
7-O	Norte	Óbitos	01/02/2021 a 01/12/2020	01/03/2021 a 05/03/2021	RF-Sem Va- riáveis; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7
8-O	Norte	Óbitos	01/02/2021 a 01/02/2021	01/03/2021 a 05/03/2021	RF-Sem Va- riáveis; KNN-Sem Variáveis; ARIMA- Sem Variá- veis;	RF-7 KNN-7 ARIMA-7

Tabela B.6: Teste de desempenho com diferentes intervalos de treino a executar em todos os modelos(Continuação).

Teste	Região	Previsão	Intervalo Treino	Intervalo Previsão	Variáveis	Dias Pas-sado
0-R	Norte	Recuperados	01/02/2021 a 01/03/2020	01/03/2021 a 05/03/2021	RF-Internamentos; KNN-Maior Correlação; ARIMA-Maior Correlação;	RF-5 KNN-7 ARIMA-5
9-R	Norte	Recuperados	01/02/2021 a 01/06/2020	01/03/2021 a 05/03/2021	RF-Internamentos; KNN-Maior Correlação; ARIMA-Maior Correlação;	RF-5 KNN-7 ARIMA-5
10-R	Norte	Recuperados	01/02/2021 a 01/09/2020	01/03/2021 a 05/03/2021	RF-Internamentos; KNN-Maior Correlação; ARIMA-Maior Correlação;	RF-5 KNN-7 ARIMA-5
11-R	Norte	Recuperados	01/02/2021 a 01/12/2020	01/03/2021 a 05/03/2021	RF-Internamentos; KNN-Maior Correlação; ARIMA-Maior Correlação;	RF-5 KNN-7 ARIMA-5
12-R	Norte	Recuperados	01/02/2021 a 01/02/2021	01/03/2021 a 05/03/2021	RF-Internamentos; KNN-Maior Correlação; ARIMA-Maior Correlação;	RF-5 KNN-7 ARIMA-5

B.3 Aplicação

B.3.1 Excertos de Código

```

1 output$valueBox3 <- renderValueBox({
2   if(input$select_region=="Portugal"){
3     valueBox(paste(dados$InternadosUCI[nrow(dados)], " (",
4       dados$InternadosUCI_Diario[nrow(dados)], ")", sep = " "), "Internados UCI
5     (Novos)", color = 'fuchsia')
6   }else{
7     a<-dados[dados$Regiao == input$select_region ,]
8     valueBox(paste(round((a$RegCoberturaVac2[nrow(a)]*100), digits = 2)
9       , " %", sep=""), "Vacina o Completa", color = 'green')
10    })

```

Listing B.11: Exemplo do controlo sobre os retângulos de informação da aplicação (R).

```

1 output$confirmados <- renderPlotly({
2   if(input$select_specie=="Portugal"){
3     fig <- plot_ly(dados, x = ~datas, y = ~Confirmados_Total, name = '
4     Total', type = 'scatter', mode = 'lines', line=list(color="orangered")
5     )
6     fig <- fig %>% add_trace(y = ~Confirmados_Total_Diario, name = '
7     Di rio', mode = 'lines', line=list(color="darkorange"))
8     fig <- fig %>% layout(hovermode = "x unified", yaxis=list(title = "
9     Confirmados"), title = "Evolu o dos Casos Confirmados")
10    fig <- fig %>% config(modeBarButtonsToRemove = c("lasso2d", "
11    select2d", "toggleSpikelines"))
12    fig
13  }else{
14    data <- dados[dados$Regiao == input$select_specie ,]
15    fig <- plot_ly(data, x = ~datas, y = ~RegTotal_Confirmados, name =
16    'Total', type = 'scatter', mode = 'lines', line=list(color="orangered")
17    )
18    fig <- fig %>% add_trace(y = ~RegDiario_Confirmados, name = '
19    Di rio', mode = 'lines', line=list(color="darkorange"))
20    fig <- fig %>% layout(hovermode = "x unified", yaxis=list(title = "
21    Confirmados"), title = "Evolu o dos Casos Confirmados")
22    fig <- fig %>% config(modeBarButtonsToRemove = c("lasso2d", "
23    select2d", "toggleSpikelines"))
24    fig
25  }
26 })

```

Listing B.12: Exemplo do controlo sobre os gráficos da aplicação (R).