INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Recolha de Informação na Área de Telecomunicações: Sistema Foco no Cliente

RUI FILIPE LIMA PEREIRA
Outubro de 2021

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Recolha de Informação na Área de Telecomunicações: Sistema Foco no Cliente

Rui Filipe Lima Pereira

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

Orientador: Isabel Praça

Porto, Outubro 2021

Resumo

A crescente competitividade no setor das telecomunicações força as operadoras a desenvolver mecanismos para estarem em evolução constante, utilizando tecnologias emergentes de forma a prestar o melhor serviço a um cliente que cada vez tem mais ofertas variadas e personalizadas às suas necessidades.

Para isso, as operadoras necessitam de priorizar o seu foco para os clientes ativos. Mas para isso, surge a necessidade de conhecê-los melhor.

Dai surge a necessidade de estudar uma solução que foque nas interações que o cliente realiza com os canais de uma operadora de telecomunicações.

Esta dissertação tem como objetivo desenvolver um projeto que disponibilize uma visualização focada no cliente. Para esse efeito serão estudados os sistemas que compõem a infraestrutura tecnológica de uma operadora de telecomunicações, as arquiteturas de *Big Data* que sejam referência e capazes de agregar a informação presente nos sistemas que compõem uma operadora, as técnicas de Aprendizagem Automática utilizadas na literatura para melhorar alguns indicadores estratégicos das operadoras.

Com esta análise realizada, esta dissertação culminou na proposta de um sistema, baseada numa arquitetura de Big Data, que processa dados em tempo-real e dados históricos. Para além disso, recorreu-se às técnicas de aprendizagem automática estudadas para desenvolver um modelo que permita combater o abandono de clientes.

Assim esta proposta permite a uma operadora uma completa visão da informação do cliente em mais detalhe, observando não só o detalhe dos serviços, mas também o potencial risco de abandono.

Palavras-chave: Aprendizagem Automática, Arquitetura *Lambda*, Abandono de Clientes, Operadora de Telecomunicações, *Big Data*, Foco no Cliente.

Abstract

The growing trend in the telecommunications sector forces operators to develop mechanisms to be in constant evolution to provide the best service, with emergent technologies, to a customer who increasingly has more varied offers tailored to their needs.

To do this, service providers must prioritize their focus to active customers. But for this, there is a need to know them better.

Hence the need to study a solution that focuses on the interactions that the customer performs with the channels of a telecommunications operator.

This dissertation aims to develop a project that provides a customer-focused view. For this purpose, the systems that are part of a telecommunications operator's technological infrastructure will be studied, Big Data architectures that are reference and capable of aggregating the information present in the operator's systems and Machine Learning's techniques used in the literature to improve some strategic indicators of the operators.

With this analysis performed, this dissertation culminated in the proposal of a system, based on the Big Data architecture, which processes real-time data and historical data. In addition, the studied Machine Learning techniques were used to develop a model to combat customer churn.

Thus, this proposal allows an operator a complete view of customer information in more detail, not only observing the details of the services, but also their potential risk of churning.

Keywords: Machine Learning, Lambda Architecture, Customer Churn, Telecommunications service provider, Big Data, Customer 360

Agradecimentos

Esta tese não seria possível sem a ajuda imprescindível da minha família, principalmente os meus pais. Por isso, os primeiros agradecimentos destinam-se a eles.

Quero também agradecer ao meu irmão pela motivação que sempre me transmitiu para atingir os melhores resultados.

Agradecer aos meus amigos e primos pelo apoio e disponibilidade para me apoiarem e no ânimo que me deram para completar esta etapa final.

Continuar os agradecimentos com professora Isabel Praça que me aceitou desde o primeiro dia e sempre me ajudou com os melhores conselhos e oferecendo dicas e conhecimento para que obtivesse os melhores resultados.

Agradecer também à Celfocus, principalmente ao Rafael Crispim pela ajuda incrível, pelos conhecimentos, pelas conversas que teve comigo para discutir ideias e pelos conselhos que ao longo destes meses me passou.

A todos os colegas de Mestrado, pelo companheirismo ao longo desta etapa.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos

AHP Método da Análise Hierárquica, *Analytic Hierarchy Process.*

AUC Área Sob a Curva, *Area Under the Curve.*

CRISP-DM Processo Padrão Inter-Indústrias para Mineração de Dados, *Cross-Industry Standard Process for Data Mining.*

CRM Gestão do Relacionamento com o Cliente, *Customer Relationship Manager.*

DTO Objeto de Transferência de Dados, *Data Transfer Object.*

EV Centralidade de Autovetor, *Eigenvector Centrality.*

FAST Function Analysis and System Technique.

FFN Rede Neuronal de Alimentação Direta, *Feedforward Neural Network.*

FN Falsos Negativos.

FNR Taxa Acertos Falsos Negativos, *False Negatives Rate.*

FP Falsos Positivos.

FPR Taxa Acertos Falsos Positivos, *False Positive Rate.*

IC Índice de Consistência.

IR Índice Aleatório, *Index Random.*

JSON Notação de Objeto em JavaScript, JavaScript Object Notation.

KPI Indicadores Chave Estratégicos, *Key Performance Indicators.*

MVC Modelo, Vista, Controlador; *Model, View, Controller.*

RC Razão de Consistência.

REST Transferência de estados representacional, *Representational State Transfer.*

RFM Análise Recência, Frequência e Monetária, *Recency Frequency Monetary Analysis.*

RGPD Regulamento Geral de Proteção de Dados.

ROC Curva Característica de Operação do Recetor, *Receiver Operating Characteristic.*

SEMMA Amostragem, Explorar, Modificar, Modelação, Avaliar; *Sample, Explore, Modify, Model, and Assess.*

SIM Módulo de Identificação do Subscritor, *Subscriber Identity Module.*

SMOTE Técnica Sintética de Sobreamostragem, *Synthetic Minority Over-Sampling Technique.*

SNA Análise de Rede Social, *Social Network Analysis.*

TNR Taxa Acertos Verdadeiros Negativos, *True Negative Rate.*

USSD Dados de Serviços Suplementares não Estruturados, *Unstructured Supplementary Service Data.*

VN Verdadeiros Negativos.

VP Verdadeiros Positivos.

1 Introdução

O presente capítulo tem como objetivo apresentar ao leitor um enquadramento teórico do projeto a desenvolver na área das telecomunicações, através das motivações que serviram de base para a sua elaboração.

Ainda neste capítulo, é apresentada a abordagem escolhida para a sua realização, descrevendo as etapas mais relevantes.

Por fim, são descritos os resultados esperados e a estrutura da dissertação.

1.1 Contexto

O presente projeto foi desenvolvido na Celfocus, uma organização que se dedica ao desenvolvimento de soluções informáticas para o setor das telecomunicações. Atualmente, a Celfocus detém projetos com vários provedores de serviços de telecomunicações, em diversos países europeus, africanos e na região do médio oriente.

Dados do (ITU, n.d.-b) mostram um crescimento no número de subscrições para telemóveis, em 2019, atingindo 106,5 por cada 100 habitantes, a nível mundial. Esse valor aumenta para 122,4 no espaço europeu e 123,7 nos Estados Unidos.

Também, segundo (ITU, n.d.-a), o número de serviços subscritos de natureza fixa tem aumentado constantemente ao longo do século XXI, passando de 0,8 por cada 100 habitantes, no inicio do século para 14,7 em 2019, a nível mundial. No espaço europeu estes valores vão desde 1,7 até 37,8 no mesmo período, e nos Estados Unidos os valores vão desde 4,7 até aos 34,7.

Com o crescimento das infraestruturas e dispositivos de telecomunicações existem cada vez mais opções disponíveis para um consumidor de telecomunicações, comprovada pela procura elevada, referida anteriormente. Com essa expansão, e o consequentemente aumento da informação, as operadoras terão de desenvolver estratégias, de forma a melhorar a experiência dos seus clientes para criar fontes de receitas alternativas.

1.2 Problema

Indicadores chave de desempenho, *Key Performance Indicators* (KPI) são instrumentos vitais utilizados por gestores das organizações para entender se o negócio caminha numa direção adequada ou não. Um dos principais problemas que os gestores têm hoje em dia é na identificação e compreensão dos indicadores mais relevantes e na elaboração de estratégias para melhorá-los. Na área das telecomunicações, dois dos indicadores mais importantes são o abandono de clientes e a satisfação dos clientes.

Tabela 1 – Índice de satisfação do cliente norte-americano [1]

Setor	**2020**
Serviços de Saúde e Ação Social	71,7
Energia	72,1
Telecomunicações	72,2
Transporte	75,6
Serviços Financeiros	76
Serviços alimentares	77,9
Bens de fabrico de longa duração	78,3
Bens de fabrico de curta duração	79,2
Retalho	Sem informação
Administração pública	Sem informação

De acordo com o índice de satisfação do cliente norte-americano (ASCI, 2020), representado na Tabela 1, o setor das telecomunicações apresenta, em 2020, índices de satisfação baixos comparativamente a outros setores. Estes números pressupõem um desafio permanente no interesse da rentabilidade e da competitividade das provedoras de serviços.

Um outro indicador chave utilizado pelas provedoras de serviços e que está normalmente associado ao índice de satisfação é o abandono de clientes, que mede a quantidade de clientes que deixam de utilizar um produto, um serviço ou deixam de ser clientes.

A TM FORUM – organização que propõe os *standards* para a indústria das telecomunicações – analisou dados de 36 operadoras de 24 países diferentes e concluiu que o abandono de clientes vai desde 14% a 75% para qualquer tipo de clientes e de 5% a 32% para clientes com serviços pós-pago, os mais lucrativos

[1] Retirado de (ASCI, 2020)

(TM FORUM, 2018). Para além do prejuízo financeiro que estes números demonstram, as operadoras sofrem com perdas de valor reputacional.

Por causa desses números, as provedoras de serviços de telecomunicações têm desenvolvido procedimentos para identificar e reter os seus clientes, pois é mais barato do que atrair novos clientes (Idris & Khan, 2012). Isto deve-se aos custos de publicidade, entre outros, que podem ser cinco ou seis vezes maiores quando comparando com os custos de manter os clientes (Verbeke et al., 2011).

Tendo em conta a quantidade de projetos que a Celfocus realiza na área das telecomunicações foi identificado, internamente, que não existe um estudo aprofundado das interações que um cliente realiza com os vários sistemas das telecomunicações e a que a informação recolhida deve ser visualizada numa solução *Customer 360*. Este tipo de solução, inspirada no termo foco no cliente, agrega as informações dos vários sistemas e disponibiliza essa informação numa ótica de melhoria da experiência do cliente.

1.3 Objetivo

Esta dissertação tem como objetivo estudar as interações de um cliente com todos os sistemas de um provedor de serviço de telecomunicações, de forma a criar um sistema de visualização *Customer* 360, centrada nas informações mais relevantes dos clientes, recorrendo a técnicas de *Big Data.*

Esta visualização deverá assentar numa arquitetura de referência das telecomunicações, segundo as melhores práticas da Engenharia de Software, para que o sistema seja aplicado a qualquer provedora consoante a informação disponível.

Para complementar, deverão ser identificadas e aplicadas técnicas, recorrendo a artigos científicos que usem técnicas de aprendizagem automática, para melhorar os indicadores estratégicos das provedoras de serviços de telecomunicações.

A solução final consiste numa prova de conceito que permita validar a implementação de um sistema que agregue e disponibilize a informação mais relevante dos clientes. Para acrescentar ao sistema deverá ser desenvolvido um módulo recorrendo a técnicas de aprendizagem automática que permita melhorar um indicador estratégico relacionado com a experiência do cliente.

O projeto também deverá traçar conclusões relativas aos custos-benefícios em adotar este tipo de sistema.

1.4 Abordagem

Este subcapítulo documenta, resumidamente, a abordagem ao problema, descrevendo as diferentes fases da elaboração deste projeto, para que seja possível compreender por que etapas e com que sequência este projeto se realiza desde o início até à sua conclusão.

De forma a implementar esta solução, realiza-se, numa primeira fase, um contacto com o cliente, neste caso a Celfocus. Foram abordados diversos aspetos relativamente ao tipo de dados a consultar, quais os resultados pretendidos e o objetivo do conhecimento a ser extraído.

Após esta fase, tendo em conta a informação disponível, foi realizado uma análise às abordagens existentes na literatura e às tecnologias utilizadas. Nesta segunda fase, foram estudados artigos científicos, mas também soluções comerciais.

A terceira fase tem como objetivo um estudo aos sistemas pertencentes a um sistema de telecomunicações, mais concretamente que tipo de dados é que estão armazenados, identificando onde é que existe a informação mais relevante. Nesta fase deverão ser estudados os indicadores estratégicos mais importantes. Esta fase possui uma grande importância, pois permite conhecer os dados que existem nos vários sistemas que irão permitir o desenvolvimento do projeto.

Na quarta fase, tendo em conta as abordagens estudadas na fase anterior e o tipo de informação disponível, foram definidos os casos de uso a que este projeto deve responder. Foi também realizada uma proposta de arquitetura tendo em conta a aprendizagem adquirida e os casos de uso identificados.

Na quinta e sexta fase inicia-se a implementação, explorando a recolha, tratamento, leitura e processamento da informação.

Por fim, após a conclusão da solução, é realizada uma análise geral dos resultados.

1.5 Resultados Esperados

Com a realização deste projeto, é esperado que o estudo realizado permita compreender melhor as interações que um cliente das telecomunicações realiza, a informação que é armazenada, os indicadores estratégicos mais relevantes e que a implementação da solução seja utilizada numa provedora de serviços de telecomunicações.

1.6 Estrutura do Documento

O documento está estruturado em sete capítulos. Inicialmente, neste mesmo capítulo, foi apresentada uma fase introdutória, onde foram abordados o problema e o contexto.

No segundo capítulo são analisados os sistemas existentes, os indicadores estratégicos, o mapeamento para casos de uso com o objetivo de compreender as técnicas utilizadas para melhorar os indicadores estratégicos, bem como as soluções de mercado existentes.

No terceiro capítulo é descrita o estado de arte, apresentando conclusões acerca da utilização de técnicas de aprendizagem automática.

Seguidamente, no quarto capítulo, é apresentada a análise e proposta de valor da solução prevista.

No capítulo seguinte, é apresentada a análise dos casos de uso e é criada uma proposta de *design*. É Ainda é definida a arquitetura da solução a implementar.

No sexto capítulo é descrito o sistema desenvolvido, através dos procedimentos, técnicas utilizadas para a implementação do sistema.

No sétimo capítulo é realizada uma análise crítica dos resultados obtidos. Por fim, o documento descreve as conclusões e o trabalho futuro.

2 Contexto

Neste capítulo é contextualizada a informação do negócio das telecomunicações apresentando a descrição dos indicadores estratégicos existentes, dos sistemas que compõem uma provedora de serviços de telecomunicações e dos passos necessários para a construção de um sistema *Customer 360*.

Para além disso, introduz-se o tema de Aprendizagem Automática apresentando as categorias de problemas, terminologia utilizada, cenários de aprendizagem, métricas de avaliação, algoritmos e tecnologias utilizadas.

De seguida, são apresentados os conceitos de *Big Data*, através das suas dimensões, características dos dados, arquiteturas existentes e ferramentas utilizadas.

Seguidamente, contextualizam-se os trabalhos relacionados que permitem conhecer as técnicas utilizadas no desenvolvimento de problemas/oportunidades para melhorar a experiência do cliente.

O capítulo é concluído com a apresentação de um mapeamento entre os indicadores estratégicos mais relevantes e as técnicas utilizadas para os melhorar.

2.1 Análise e Contexto

A ideia na base de uma solução *Customer* 360 é construir uma visualização completa de cada cliente, agregando os dados estruturados e não-estruturados dos vários sistemas dentro de uma empresa.

A área das telecomunicações é composta por diversos sistemas que captam, armazenam informações relativas a clientes, produtos e equipamentos. No entanto, existem poucas soluções capazes de retirar a informação relativa a um cliente, pois os seus dados normalmente residem em várias *stacks* dispersas de

sistemas de apoio aos negócios. Esta falta de clarificação pode levar a altos custos devido a processos ineficientes (TM FORUM, 2019).

Na Tabela 2 são identificados os principais sistemas que compõem a arquitetura de uma operadora das telecomunicações, resultado da análise feita ao sector, tendo por base a experiência dos especialistas da Celfocus, bem como os três anos de experiência do autor nesta indústria.

Tabela 2 – Descrição dos sistemas

Sistema	**Descrição**
Customer Relationship Manager (CRM)	Este sistema tem como objetivo centralizar o registo das contas de um cliente, histórico de compras e diferentes interações como por exemplo, queixas ou feedback. Neste sistema, é ainda armazenado o catálogo de produtos e serviços disponíveis para um cliente particular ou empresarial.
Provisioning	Onde é realizada a ativação na infraestrutura de rede dos serviços móveis e fixos que um cliente possui.
Billing	Responsável por armazenar dados de consumos móveis e fixos, calcular e produzir as informações de faturação, processar os pagamentos e gerir a cobrança de dívidas.
Field Service Management	Sistema responsável por gerir a afetação dos colaboradores que se encontram a realizar trabalhos fora das instalações, bem como dos recursos necessários.
Order Management	Gere os processos e ações necessárias para a correta entrega de um produto/serviço a um cliente.
Resources Management Inventory	Sistema responsável por gerir os equipamentos físicos.
Fault Management	Sistema responsável por detetar, isolar e corrigir eventuais falhas nos serviços dos clientes.

Para além destes sistemas, os clientes ainda interagem com diversos canais, nomeadamente os *call centers,* lojas físicas, redes sociais, vendedores e dados de serviços suplementares não estruturados (*Unstructured Supplementary Service Data*, USSD) - protocolo utilizado por telemóveis para comunicar com os fornecedores de serviços através de mensagens de texto.

A agregação da informação permite perceber o que é mais importante para os clientes, e aplicar esse conhecimento para melhorar as suas experiências, bem como melhorar os objetivos da empresa. Por isso importa perceber que tipo de indicadores é que as provedoras de serviço utilizam para medir a experiência do cliente e a perceção dos clientes acerca dos serviços que estão a utilizar.

A (TM FORUM, 2020) desenvolveu, juntamente com seis empresas ligadas às telecomunicações, os trinta indicadores estratégicos de referência para a indústria das telecomunicações. Os indicadores estão divididos em seis dimensões:

- Cliente
- Estratégia
- Tecnologia
- Operações
- Cultura
- Dados

Na Tabela 3, é apresentada a descrição e a unidade de medida de cada indicador estratégico relacionado com os clientes das provedoras de serviços.

Tabela 3 – KPIs da Dimensão Cliente

Dimensão Cliente		**Descrição**	**Unidade de medida**
Indicador	*Customer Lifetime Value*	Representa a diferença entre a quantidade expectável de dinheiro gasto por um cliente e o custo de aquisição.	Dólar/Euro
	Net Promoter Score	Mede a satisfação e lealdade dos clientes baseado na probabilidade de recomendação de um produto.	Numérico
	Churn Rate	Mede a taxa de abandono.	Percentagem
	Customer Acquisition Cost	Custo que uma organização teve para adquirir um cliente.	Dólar/Euro
	First Response Time	Duração para que a organização responda a um pedido de um cliente.	Tempo

Na Tabela 4, é apresentada a descrição e a unidade de medida de cada indicador relacionado com a estratégia, sejam em determinados negócios, ou para com a estratégia da imagem da marca.

Tabela 4 – KPIs da Dimensão Estratégia

Dimensão Estratégia		**Descrição**	**Unidade de medida**
Indicador	*Return on Investment*	Métrica relacionada com o ganho ou perda de um investimento consoante o custo.	Percentagem

Dimensão Estratégia		Descrição	Unidade de medida
	Revenue by Digital Investment	Receita adicional gerada através de investimentos digitais.	Percentagem
	Brand Affinity Score	Representa o valor de sentimento por parte de clientes e potenciais clientes.	Numérico
	Revenue by Ecosystem Partners	Medida de proporção da receita gerada de serviços prestados em conjunto com um ou mais parceiros digitais do ecossistema.	Percentagem
	Customer self-service success	Mede a proporção de solicitações de serviço por parte do cliente que são completadas com sucesso em canais de auto atendimento sem a necessidade de suporte humano.	Percentagem

Na Tabela 5, apresenta-se a descrição e unidade de medida dos indicadores da dimensão tecnologia. Estes indicadores relacionam-se com processos digitais e tecnológicos.

Tabela 5 – KPIs da Dimensão Tecnologia

Dimensão Tecnologia		Descrição	Unidade de medida
Indicador	*Digital User Journeys*	Indicador da proporção de interações do cliente que podem ser totalmente realizadas em canais digitais sem interação humana.	Percentagem
	App Market Performance	Reflete a qualidade e usabilidade de uma aplicação baseada em indicadores como: Classificação por estrelas da aplicação, contagem total de downloads / instalações, número total de avaliações e o sentimento médio dos utilizadores.	Numérico
	Adoption of DevOps	Adoção de DevOps nas operações e tecnologias dentro da organização.	Percentagem

Dimensão Tecnologia		Descrição	Unidade de medida
	Process Automation Rate	Quantidade de processos dentro da organização que são automatizados.	Percentagem
	Net Carbon Footprint	Diferença entre as emissões e a absorção de carbono por cliente, num determinado período de tempo.	Kg/Cliente

Na Tabela 6, estão descritos os indicadores da dimensão Operações, bem como a unidade de medida. Estes indicadores têm como objetivo avaliar as operações internas das empresas de telecomunicações.

Tabela 6 – KPIs da Dimensão Operações

Dimensão Operações		Descrição	Unidade de medida
Indicador	*Time-to-Market*	Mede o tempo necessário para implementar um produto novo ou serviço desde o conceito até à chegada ao mercado.	Tempo
	Average Response Time	Quantidade de tempo para fornecer resultados de uma solicitação feita por um utilizador ou cliente.	Tempo
	Average Usage Rate on digital	Mede a proporção de utilizadores que comunicam com um canal digital (*Chatbots, Website).*	Percentagem
	Cost-to-Serve	Custo total de servir utilizadores ou clientes.	Dólar/Euro

De seguida, na Tabela 7, estão descritos os indicadores da dimensão Cultura e a respetiva unidade de medida. Estes indicadores relacionam-se com a cultura organizacional interna.

Tabela 7 – KPIs da Dimensão Cultura

Dimensão Cultura		Descrição	Unidade de medida
Indicador	*Internal Net Promoter Score*	Avaliar até que ponto os funcionários estão dispostos a recomendar a empresa e os seus produtos / serviços.	Numérico

Dimensão Cultura		Descrição	Unidade de medida
	Digital Match of Skills	Disponibilidade de funcionários com as habilidades necessárias para entregar iniciativas digitais (incluindo produtos e serviços).	Percentagem
	Employee Effort Score	Mede a facilidade com que os funcionários conseguem realizar as suas tarefas.	Numérico
	Workforce involved in Digital initiatives	Proporção do tempo total gasto no desenvolvimento, entrega ou apoia a iniciativas digitais.	Percentagem
	Training budget for Digital initiatives	Mede a percentagem entre o orçamento alocado a formações e as iniciativas digitais.	Percentagem

Concluindo, na Tabela 8, encontra-se descritos os indicadores da dimensão Dados, relacionados com a gestão de informação e dados.

Tabela 8 – KPIs da Dimensão Dados

Dimensão Dados		Descrição	Unidade de medida
Indicador	*Economic Value of Data Assets*	Valor monetário de todos os ativos relacionados com dados de uma organização.	Dólar/Euro
	Data Democratization	Mede a proporção média de dados/informação a que os stakeholders sentem que são imediatamente capazes de acesso num formato utilizável.	Percentagem
	Data Integrity	Proporção de cada tipo de dados realizada pela organização que é “adequado para o propósito".	Percentagem
	Revenue from Data Monetization	Proporção da receita obtida da rentabilização de dados externa.	Percentagem
	Key-Data Assets	Proporção dos principais dados ativos que são adequadamente descritos e catalogados com meta-dados.	Percentagem

Dimensão Dados		Descrição	Unidade de medida
	Compliance to Data Regulation and Policies	Proporção de processos que utilizam dados totalmente compatíveis com regulamentos relacionados a dados e políticas.	Percentagem

2.1.1 Soluções *Customer* 360

Na Figura 1 é possível visualizar todos os passos necessários para a construção de uma solução *Customer 360*. O primeiro passo tem como objetivo recolher os dados dos vários sistemas e canais, por exemplo dos sistemas e canais referenciados. Para isso deverão ser aplicadas técnicas que permitam recolher os dados das várias fontes de dados existentes. Ainda neste passo é identificada a periocidade da recolha de dados: momentânea, em períodos definidos de hora ou diária.

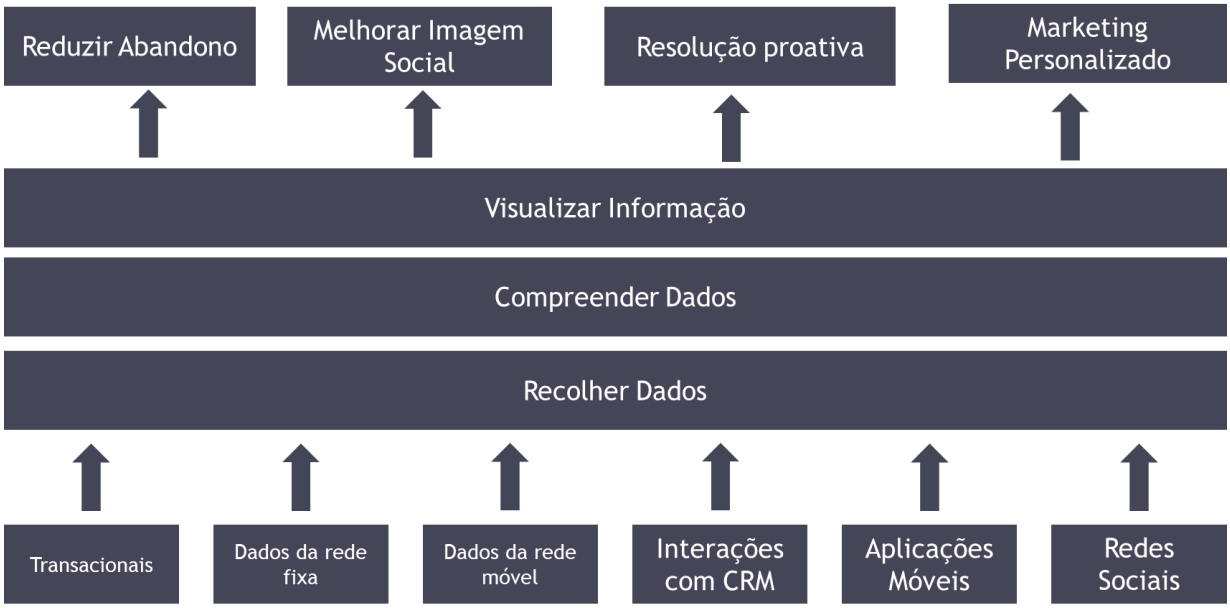


Figura 1 – Esquematização de uma solução *Customer 360*

O segundo passo passa por identificar os dados mais relevantes recolhidos no passo anterior. Através desta informação, são implementadas soluções para melhorar a experiência do cliente e consequentemente, os indicadores estratégicos.

De seguida, é criada a visualização da informação mais relevante, incluindo, por exemplo, interações realizadas nos vários canais, histórico de faturação, atividade relativa aos produtos e serviços.

2.2 Soluções de Mercado

Neste subcapítulo são apresentadas algumas soluções que têm como objetivo auxiliar as operadoras de telecomunicações a melhorar a experiência do cliente.

No entanto, tendo em conta que as soluções são empresariais não foi possível obter uma versão de demonstração para testar as suas capacidades, pelo que é apresentada uma descrição das principais funcionalidades retiradas a partir dos seus *websites*.

2.2.1 PEGA

O sistema PEGA [2] é utilizado em vários provedores de serviços de telecomunicações, entre os quais a Orange e Vodafone para complementar os sistemas de gestão do relacionamento dos clientes.

Este sistema possui uma tecnologia baseada em regras que permite direcionar determinados produtos ou serviços para segmentos de clientes (Chen, 2012) e através de algoritmos de aprendizagem automática, permitindo aos clientes uma oferta de produtos e serviços mais contextualizada, tendo em conta o seu perfil (PEGA, n.d.)

2.2.2 aia

A solução *aia* [3] da AMDOCS consiste na utilização de Inteligência Artificial para disponibilizar dados em tempo-real. Estes dados são continuamente utilizados pelo sistema *aia* para criar previsões, automatizar decisões e gerir conversas diretamente com os clientes.

Para além da visualização *Customer 360*, esta ferramenta também disponibiliza um conjunto de dados retirados de forma inteligente, tais como o risco do abandono de cliente ou o valor *Net Promoter Score*. Na Figura 2, está representada a interface principal desta ferramenta.

[2] https://www.pega.com/customers/vodafone-marketing acedido a 01/02/2021
[3] https://www.amdocs.com/media-room/introducing-aia-amdocs-bringing-real-time-intelligence-heart-communications-business, acedido a 01/02/2021

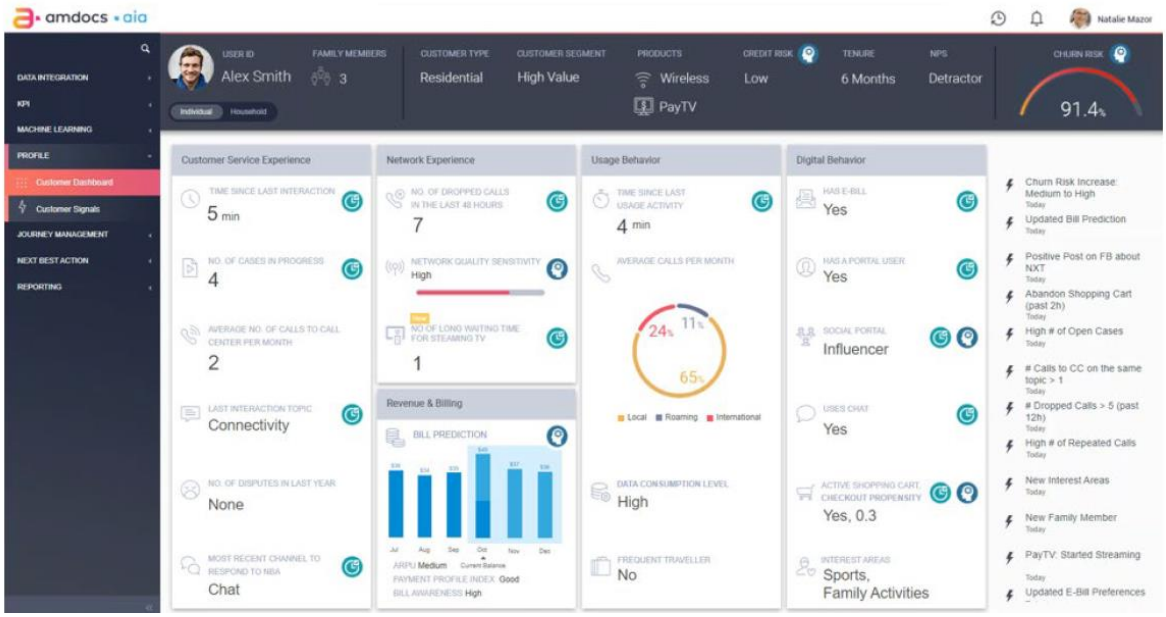


Figura 2 – Solução *Customer* 360 aia [4]

2.2.3 Beesion

A solução Beesion [5] consiste num portal onde as empresas podem visualizar as interações dos clientes num único local. Entre as funcionalidades principais destaca-se a visualização da utilização de consumos relativamente a produtos ou serviços e a determinação das ofertas e produtos para reter clientes.

2.3 Conclusões

Pela análise realizada neste capítulo, percebe-se a complexidade do negócio das telecomunicações, a diversidade de sistemas existentes e da respetiva informação armazenada. Assim de forma, a desenvolver soluções que permitam agregar a informação disponibilizada nos sistemas, técnicas de *Big Data* devem ser consideradas para extrair os dados, tendo em conta o volume existente e a variedade de dados. O conhecimento extraído sobre os clientes, recorrendo a técnicas de Aprendizagem Automática, permite às operadoras visualizar e melhorar, de forma inteligente e esclarecedora, os indicadores estratégicos mais relevantes.

[4] Retirado de (TM FORUM, 2019)
[5] https://beesion.com/telecom-crm-with-a-360-degree-view/ acedido a 01/02/2021

3 Estado da Arte

Neste capítulo descrevem-se as principais tecnologias e técnicas utilizadas para lidar com os desafios que as soluções *Customer* 360 exigem.

Este capítulo tocará em pontos chaves relativamente à importância da utilização de *Big Data* na área das telecomunicações, mas também o ecossistema de tecnologias, bibliotecas e ferramentas existentes.

O capítulo faz ainda um levantamento teórico da extração de conhecimento dos dados e de aprendizagem automática.

No final, são apresentados os trabalhos presentes na literatura, mais relevantes para o projeto em questão.

3.1 Big Data

(Dumbill, 2013) define *Big Data* como "…dados que excedem a capacidade de processamento dos sistemas tradicionais de base de dados. Os dados são em grande quantidade, movem-se rapidamente e não se adequam nas restrições das arquiteturas de base de dados tradicionais. Para extrair o conhecimento destes dados, é necessária uma maneira alternativa de os processar".

Na área das telecomunicações, as provedoras de serviço criam e analisam grandes quantidades de dados sobre clientes, operações e transações. Ao longo dos anos, as provedoras usaram uma variedade de técnicas para analisar esses dados.

Nesta área, destaca-se a utilização de *Big Data*, nomeadamente em:

- Monitorização de tráfego de rede.
- Analisar registos de comunicações para identificar atividade fraudulenta.
- Personalizar planos de consumo baseado em padrões.
- Utilização de dados de redes sociais para otimizar campanhas de *marketing.*

Algumas das fontes de *Big Data* incluem chamadas telefónicas, *emails*, mensagens de texto, informação geoespacial, dados de redes sociais. Os dados que se encontram nessas fontes apresentam algumas características distintas.

3.1.1 Dimensões de *Big Data*

Para além da definição de *Big Data,* descrita no capítulo 3.1, existem três dimensões que geralmente acompanham a definição descrita anteriormente. As três dimensões são as seguintes:

- Volume
  Refere-se à quantidade ou dimensão dos dados.

- Velocidade
  Refere-se à velocidade com que os dados são gerados

- Variedade
  Refere-se à heterogeneidade das fontes de informação e às características dos dados existentes.

Estas dimensões são dimensões consensuais surgindo na generalidade das referências. Para além destas dimensões alguns autores ainda acrescentam uma outra dimensão, divergindo entre veracidade, variabilidade e valor.

3.1.2 Características dos Dados

As fontes de dados encontram-se organizadas de várias formas e apresentam características entre si.
De seguida, são apresentadas as características mais comuns dos dados existentes:

- Estruturado
  Qualquer dado que possa ser armazenado, acedido e processado num formato fixo é, geralmente, classificado como estruturado. Alguns exemplos são dados que se encontram em base de dados relacionais.

- Não-estruturado
  Dados de um formato aleatório são classificados como não-estruturados. Um exemplo de dados não-estruturados são os dados que combinam imagens e texto.

- Semiestruturado
  Dados que combinam as duas características anteriores. Normalmente, são dados que combinam uma estrutura não-estruturada e meta dados associado.

3.1.3 Arquiteturas de *Big Data*

Neste subcapítulo são descritas duas arquiteturas utilizadas em projetos de *Big Data*.

Inicialmente, são apresentados os detalhes da arquitetura *Lambda*.

Bahga e Madisetti (2016) descreveram esta arquitetura como "sendo escalável, tolerante a falhas e adequado para aplicações que envolvem fluxos muito rápidos a partir dos dados de origem." A arquitetura é composta pelas seguintes camadas:

- Camada *Batch*
  Responsável por guardar os dados extraídos da origem. Esta camada é responsável por processar a totalidade em períodos pré-definidos.

- Camada *Real-Time*
  Responsável por disponibilizar uma maneira eficiente para processar os dados mais recentes. Esta camada também tem como objetivo preencher eventuais dados que não foram processados pela camada *batch*.

- Camada *Serving*
  Responsável por criar a indexação dos dados para permitir pesquisas rápidas nos dados.

A arquitetura encontra-se representada na Figura 3.

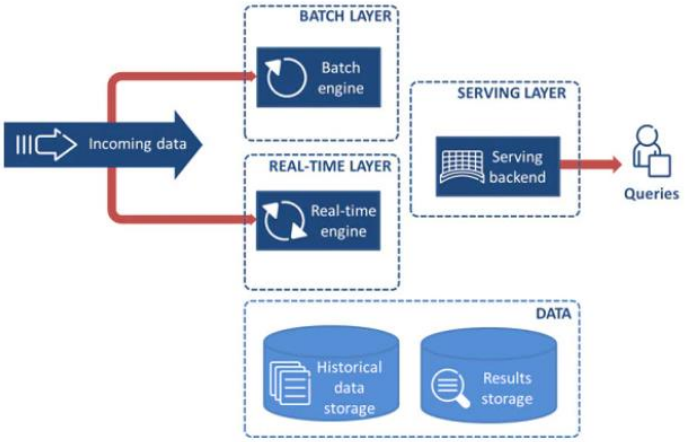


Figura 3 – Arquitetura *Lambda* [6]

Em 2014, Jay-Kreps, propôs uma alternativa à arquitetura *Lambda*, chamada *Kappa*. A principal diferença desta para a arquitetura apresentada anteriormente reside na remoção da camada *batch*, ficando a camada *real-time* como principal camada de processamento e armazenamento. A arquitetura encontra-se representada na Figura 4.

[6] Retirado de https://www.ericsson.com/en/blog/2015/11/data-processing-architectures--lambda-and-kappa

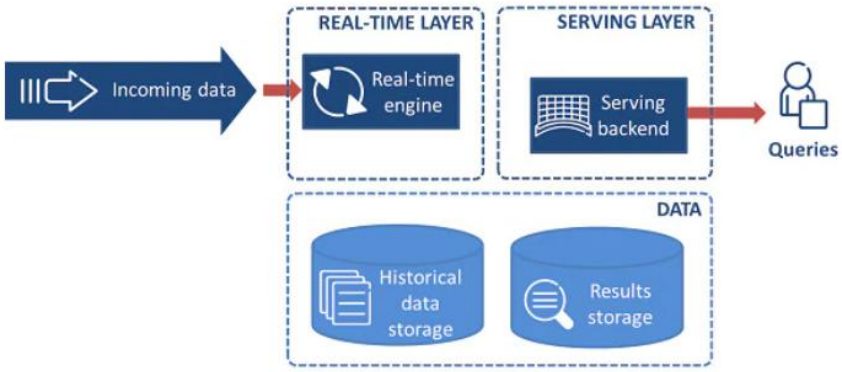


Figura 4 – Arquitetura *Kappa 6*

3.1.4 Ferramentas de *Big Data*

Neste subcapítulo são apresentadas algumas das ferramentas presentes no ecossistema de *Big Data:*

- Apache Nifi
  Ferramenta open-source utilizada para extração de dados de vários tipos de fontes, processamento de dados.

- Apache Kafka
  Ferramenta *open-source* utilizada para processamento de eventos em tempo-real.

- Apache Spark
  Ferramenta de processamento de tarefas em grandes conjuntos de dados.

- Elasticsearch
  Permite armazenar, procurar e analisar grandes conjuntos de dados rapidamente e em tempo-real.

- Google BigQuery
  W*eb-service* de *Big Data* disponível na *Cloud* para processar grandes conjuntos de dados.

- Hadoop
  Framework para armazenar grandes quantidades de dados e executar aplicações em *clusters.*

- Kibana
  Ferramenta de visualização de dados armazenados Elasticsearch.

- Microsoft Power BI
  Ferramenta de visualização que permite criar relatórios e *dashboards.*

3.2 Extração de Conhecimento e Aprendizagem Automática

Neste subcapítulo será abordado o tópico de extração de conhecimento e aprendizagem automática. Para isso, foi feito um estudo das metodologias existentes, categorias e terminologia utilizada.

Para além disso foram estudados os algoritmos de aprendizagem automática mais populares, as métricas mais relevantes que permitem avaliar o resultado de um algoritmo. E por fim, são apresentadas tecnologias, bibliotecas e ferramentas que permitem implementar um modelo de aprendizagem automática.

3.2.1 Metodologias Existentes

Neste subcapítulo serão apresentadas duas metodologias utilizadas para a implementação de processos que utilizem técnicas de Aprendizagem Automática.

3.2.1.1 CRISP-DM

Começando pela metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Essa metodologia, representada na Figura 5, foi concebida para catalogar e orientar as etapas mais comuns de projetos de mineração de dados e aprendizagem automática. Rapidamente tornou-se, segundo Marbán et al. (2009), "no *standard* de projetos de mineração de dados".

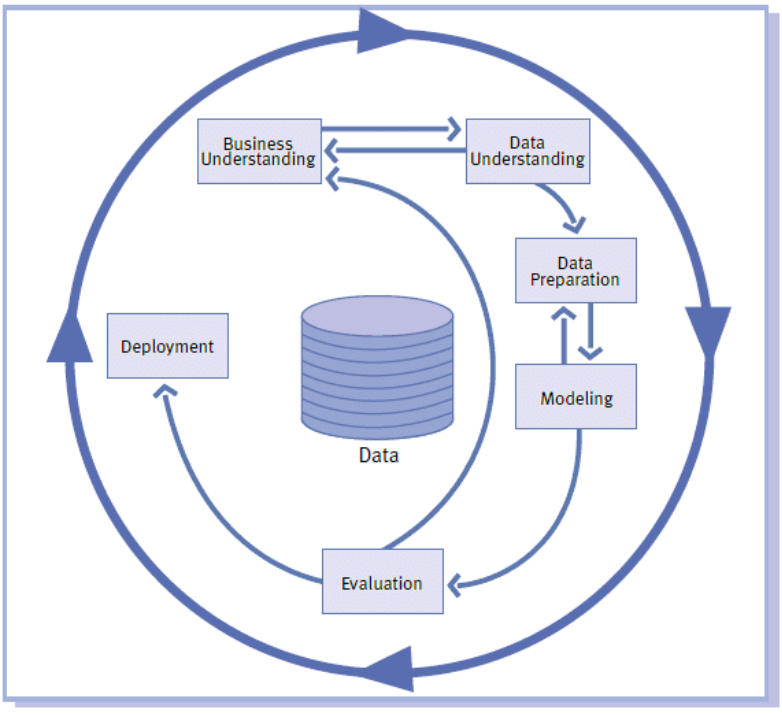


Figura 5 - Metodologia CRISP-DM [7]

Esta metodologia consiste em seis fases:

[7] Retirado de https://www.researchgate.net/figure/Figura-2-Ciclo-das-fases-da-metodologia-CRISP-DM-CRISP-DM-1996_fig2_233843302

- Entendimento do Problema:
  Nesta fase inicial, pretende-se compreender os objetivos e requisitos numa perspetiva de negócio convertendo para um problema de mineração de dados ou de aprendizagem automática e num plano inicial para atingir esses objetivos.

- Entendimento dos Dados:
  Esta fase inicia-se com a recolha dos dados e com as atividades para familiarizar-se com os mesmos, de forma a identificar problemas de qualidade. Esta fase é ainda utilizada para encontrar informações relevantes que possam originar hipóteses.

- Preparação dos Dados:
  A fase de preparação dos dados envolve as atividades para construir o *dataset* a partir dos dados iniciais. As principais tarefas prendem-se com a seleção de atributos chaves, limpeza dos dados, construção dos novos atributos e a transformação dos dados para serem utilizados na fase seguinte.

- Modelação:
  Nesta fase são identificados e aplicados vários algoritmos e técnicas de modelação, e os parâmetros são calibrados de forma a atingir os valores ótimos.

- Avaliação:
  Na fase de avaliação são analisados os resultados obtidos por parte dos modelos desenvolvidos na fase de modelação.

- Implantação:
  Nesta última fase o conhecimento adquirido é disponibilizado às partes interessadas no projeto.

3.2.1.2 SEMMA

A outra metodologia estudada denomina-se SEMMA (Amostragem, Explorar, Modificar, Modelação, Avaliar; *Sample*, *Explore*, *Modify*, *Model*, *Assess*), foi desenvolvida pelo Instituto SAS e refere-se a um processo de um projeto de mineração de dados (Azevedo & Santos, 2008)

Esta metodologia engloba cinco etapas, descritas de seguida:

- Amostragem:
  Esta etapa consiste em criar amostras dos dados, através da extração de uma porção do *dataset*, contendo informação relevante.
- Explorar:
  Esta etapa consiste em explorar os dados, extraídos no passo anterior, procurando por anomalias ou tendências existentes.
- Modificar:
  Nesta etapa, os dados são modificados, criando, selecionando e transformando determinadas variáveis para o processo de escolha do modelo de aprendizagem automática.
- Modelação:
  Esta etapa consiste em aplicar os modelos escolhidos de forma, originando uma previsão de um dado resultado.

- Avaliar:
  Nesta etapa são analisadas as descobertas obtidas e a qualidade das previsões realizadas.

3.2.1 Aprendizagem Automática

Aprendizagem Automática também conhecida como *Machine Learning*, pode ser definida como métodos computacionais que usam informação registada no passado para realizar previsões.(Mohri et al., 2018).

As aplicações da aprendizagem automática vão desde recomendações sobre que filme assistir, que refeição encomendar ou que produtos comprar. Para além de aplicações comerciais, a aprendizagem automática tem tido uma influência elevada em investigações científicas, tais como a análise de sequências de ADN ou disponibilização de tratamento personalizado para determinadas doenças, como por exemplo o cancro (Müller & Guido, 2016).

Na área das telecomunicações, a AT&T [8] utiliza dados de vídeo recolhidos por drones para realizar inspeções mais adequadas às suas torres de comunicações, permitindo realizar alterações em tempo real (Donovan, 2016) enquanto que a Vodafone [9] desenvolveu um *chatbot* responsável por responder a pedidos de serviços dos clientes, permitindo um aumento de 68% na satisfação dos clientes (Shaham, 2020).

Estes são alguns dos exemplos das mais-valias que advêm da implementação de técnicas de aprendizagem automática com o objetivo de melhorar a satisfação e a experiência dos seus clientes.

3.2.2 Categorias de Problemas

Existe um conjunto de problemas que podem ser desenvolvidos utilizando aprendizagem automática. Alguns dos mais comuns incluem:

- Classificação
  Este é um tipo de problema que consiste em atribuir uma categoria a um determinado item. Algumas dos problemas que são resolvidos através de técnicas de classificação são: classificar uma categoria de um documento a um valor, como política, desporto ou economia, ou atribuir uma categoria a uma imagem como sendo um carro, comboio ou um avião, por exemplo.

- Regressão
  Neste tipo de problema o principal objetivo é prever um valor real para cada item. Exemplos de problemas de regressão são a previsão de valores de ações da bolsa de valores.

- Ranking
  Este tipo de problema consiste em aprender a ordenar determinados itens de acordo com um determinado critério. Exemplo de um problema é a visualização de um resultado de uma página web consoante os valores introduzidos.

[8] https://about.att.com/category/all_news.html acedido a 25/02/2021
[9] https://www.vodafoneziggo.nl/en/ acedido a 25/02/2021

- Agrupamento
  Neste tipo de problema o objetivo é tentar agrupar um conjunto de itens em subconjuntos heterogéneos. Este tipo de tarefa pode ter como objetivo direcionar campanhas de marketing mais personalizadas para determinados grupos.

- Redução de dimensionalidade
  Este tipo de problema consiste em transformar uma representação inicial de um determinado item numa representação de baixa dimensão, preservando algumas das suas propriedades. Um exemplo é o pré-processamento de imagens em tarefas relacionadas com a visão computacional.

3.2.3 Terminologia Utilizada

Na área de aprendizagem automática é frequente a utilização de determinados termos. De seguida vão ser apresentados os termos mais comuns:

- Exemplos
  Instâncias de dados utilizadas para a aprendizagem. Um exemplo é uma lista de clientes.

- Características
  Conjunto de atributos correspondentes aos exemplos. No caso de clientes, as suas características podem ser o género, a idade ou número de compras num determinado ano.

- Classes
  Valores ou categorias atribuídas a exemplos. Em problemas de classificação as classes são atribuídas a determinadas categorias, por exemplo, desporto, política ou economia. Em problemas de regressão, as instâncias possuem um valor real.

- Hiper-parâmetros
  Valores introduzidos como parâmetros dos algoritmos de aprendizagem automática

- Dados de treino
  Instâncias utilizadas para treinar um algoritmo. As instâncias são treinadas de acordo com as suas características e as suas classes.

- Dados de validação
  Instâncias utilizadas para otimizar os algoritmos.

- Dados de teste
  Instâncias utilizadas para testar os algoritmos. A amostra de teste é separada da amostra de validação e da de treino. Apenas são utilizadas as características do conjunto de dados.

- Função de erro
  Função que avalia a diferença, entre uma previsão e classes valor real.

3.2.4 Tipos de Aprendizagem

Existem vários tipos de aprendizagem utilizados para desenvolver algoritmos consoante a finalidade pretendida. Os principais tipos de aprendizagem são:

- Aprendizagem supervisionada
  Neste caso o algoritmo recebe um conjunto de dados onde os exemplos estão identificados. Este é o cenário mais frequente para a resolução de problemas de classificação e de regressão.

- Aprendizagem não-supervisionada
  Ao contrário da aprendizagem supervisionada, a aprendizagem não-supervisionada não recebe exemplos identificados. Este é o cenário mais frequente para a resolução de problemas de agrupamento ou de redução de dimensionalidade.

- Aprendizagem semi-supervisionada
  Neste cenário, o algoritmo recebe um pequeno conjunto de exemplos identificados e conjunto grande de exemplos não identificados e tenta prever os conjuntos de exemplos não identificados. Um exemplo da aplicação deste cenário é a classificação de documentos onde é feita a classificação de um pequeno número de documentos para depois ser aplicado, através dos algoritmos, num maior número de documentos.

- Aprendizagem com reforço
  Neste cenário, frequentemente utilizado em jogos e robótica, o algoritmo aprende, através de tentativa/erro, as ações que resultam em melhores resultados.

3.2.5 Métricas de Avaliação

A fase de avaliação do modelo de aprendizagem automática desenvolvido é crucial pois permite perceber a qualidade do modelo desenvolvido, independentemente dos algoritmos escolhidos para a sua construção.

Neste subcapítulo, serão apresentadas algumas das métricas utilizadas para avaliar um modelo.

- Matriz de confusão
  Esta métrica é representada através de uma visualização tabular com duas dimensões. Cada linha da matriz de confusão representa as instâncias de uma classe prevista e cada coluna representa instâncias de uma classe real, ou vice-versa.

Tabela 9 – Matriz de confusão

	Realidade	
Previsão	Verdadeiros Positivos	Falsos Positivos
	Falsos Negativos	Verdadeiros Negativos

Para explicar a Tabela 9, imagine-se duas classes que o modelo tenta prever, P (Positiva) e N (Negativa).

Verdadeiros Positivos (VP) são os valores da classe P classificados como P.

Falsos Positivos (FP) são valores da classe N, mas que foram classificados como P.

Falsos Negativos (FN) são valores da classe P, mas que foram classificados como N.

Verdadeiros Negativos (VN) são os valores da classe N classificados como N.

A partir desta tabela surgem outras métricas, explicadas em seguida.

- Precisão
  Esta métrica representa a taxa de precisão do modelo, isto é, o rácio entre os Verdadeiros Positivos e a soma de Verdadeiros Positivos e Falsos Positivos. Tem como objetivo perceber a taxa de registos positivos que o modelo previu como positivos.

$$Precisão = \frac{VP}{VP + FP} \quad (1)$$

- Exatidão
  Esta métrica tem como objetivo perceber a previsão de acerto de todo o modelo.

$$Exatidão = \frac{VP + VN}{VP + FP + FN + VN} \quad (2)$$

- Erros
  Esta métrica tem como objetivo perceber a taxa de insucesso dos resultados dos modelos.

$$Erros = 1 - Acerto \quad (3)$$

- *Recall*
  Também conhecida como TPR, o *recall* é definido como o rácio de instâncias de uma classe que são corretamente previstas pelo modelo.

$$Recall = \frac{VP}{VP + FN} \quad (4)$$

- F1-*Score*
  Esta métrica é a média harmónica entre a precisão e o *recall.*

$$F1 = \frac{2 * VP}{2 * VP + FP + FN} \quad (5)$$

- Taxa de Acertos de Verdadeiros Negativos
  Esta métrica também conhecida como TNR, Indicam a percentagem de conjuntos negativos corretamente previstos no modelo.

$$TNR = \frac{VN}{VN + FP} \quad (6)$$

- Taxa de Acertos de Falsos Negativos
  Esta métrica também conhecida como FNR, Indicam a percentagem de falsos negativos corretamente previstos no modelo.

$$FNR = \frac{FN}{FN + VN} \quad (7)$$

- Taxa de Falsos Positivos
  Esta métrica também conhecida como FPR, Indicam a percentagem de exemplos negativos previstos como positivos no modelo.

$$FPR = \frac{FP}{FP + VN} \quad (8)$$

- Curva *Receiver Operating Characteristic* (ROC)
  É uma métrica de visualização que mostra o balanço a partir do rácio entre TPR e a FPR.

- Área sob a curva (AUC)
  Estimativa da probabilidade que um modelo irá prever um registo positivo escolhido aleatoriamente com um valor superior a um registo negativo escolhido aleatoriamente.

3.2.6 Algoritmos de Aprendizagem Automática

Neste subcapítulo, são listados e descritos alguns dos algoritmos mais populares de aprendizagem automática. Os algoritmos estudados foram escolhidos devido à versatilidade em serem utilizados em vários tipos de problemas e encontram-se descritos na Tabela 10.

Tabela 10 – Algoritmos de Aprendizagem Automática

Algoritmos	Descrição
Linear Regression	Utiliza-se para estimar valores reais com base em variáveis continuas. Tenta explicar uma relação entre variáveis, ajustando uma equação linear aos dados observados (Yale, 1997).
Logistic Regression	O algoritmo transforma o seu *output* para retornar um valor de probabilidade com a função sigmoide logística e prevê o objetivo pelo conceito de probabilidade (Nabipour et al., 2020).

Algoritmos	Descrição
Decision Tree	O objetivo da técnica é fazer previsão usando regras de decisão fáceis, moldadas a partir do conjunto de dados e recursos relacionados (Nabipour et al., 2020).
Support Vector Machine	Geralmente modelado e resolvido como um problema de otimização quadrática convexa para detetar um hiperplano ótimo. O Support Vector Machine tem vantagens como robustez, boa capacidade de generalização e solução ótima global única no caso do problema convexo (Burges, 1998).
Naive Bayes	Método de classificação probabilístico com base no teorema de Bayes com fortes suposições de independência entre as características dado o valor da variável de classe (Nabipour et al., 2020).
K-Nearest Neighbours	O método segue algumas etapas para encontrar os alvos: Dividir o conjunto de dados em dados de treino e teste, selecionar o valor de K, determinar a função de distância a ser usada, escolher uma amostra dos dados de teste (como uma nova amostra) e calcular a distância para seus n treinando as amostras, classificando distâncias ganhas e guardando amostras de dados *k*-mais próximas e, finalmente, atribuindo a classe de teste à amostra na votação da maioria do seu vizinho *k* (Nabipour et al., 2020).
K-Means	A ideia principal é definir um centróide para cada cluster. Ao definir o centróide, diferentes posições do centróide produzirão diferentes resultados de agrupamento. Portanto, a melhor escolha é mantê-los o mais distantes possível um do outro. Em seguida, cada ponto nos dados é classificado como o centróide mais próximo a ele (Li et al., 2020).
Random Forest	Esta técnica calcula a média do resultado da previsão de árvores, que é chamado de floresta. Além disso, o algoritmo inclui três ideias aleatórias, selecionando, aleatoriamente, dados de treino ao formar árvores, escolhendo aleatoriamente alguns subconjuntos de variáveis ao dividir nós e considerando apenas um subconjunto de todas as variáveis para dividir cada nó em cada árvore de decisão (Nabipour et al., 2020).

Na Figura 6, é esquematizada a informação descrita anteriormente.

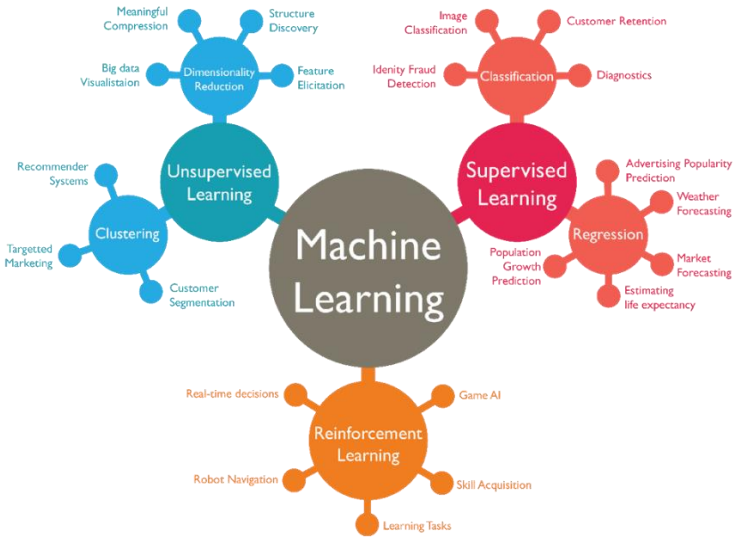


Figura 6 – Esquematização dos conceitos de Aprendizagem Automática [10]

3.2.7 Tecnologias de Aprendizagem Automática

De seguida é apresentado um conjunto de ferramentas e bibliotecas de Aprendizagem Automática existentes:

3.2.7.1 Ferramentas de Aprendizagem Automática

- Accord.net Framework
  Ferramenta de computação desenvolvida em .NET. Normalmente é utilizada em campos da Estatística, Inteligência Artificial, Processamento de Imagens e Álgebra Linear.

- Amazon Machine Learning
  Está presente no conjunto de ferramentas disponibilizadas pela Amazon Web Services [11] e consiste num produto que permite descobrir padrões em dados através de algoritmos, construir modelos matemáticos baseados nesses padrões.

- Apache Mahout
  Ferramenta que permite desenvolver modelos de aprendizagem automática relacionados com álgebra linear.

- Azure ML Studio
  Interface gráfica que permite desenvolver *workflows* de aprendizagem automática em ambientes Azure [12].

- H2O
  Ferramenta *open-source* que permite desenvolver modelos de aprendizagem automática.

- MLlib

[10] Retirado de https://blogs.oracle.com/datascience/types-of-machine-learning-and-top-10-algorithms-everyone-should-know-v2. Acedido a 25/02/2021
[11] https://aws.amazon.com/
[12] https://azure.microsoft.com/

Ferramenta de desenvolvimento de modelos de aprendizagem automática dentro do ambiente Apache Spark [13].

3.2.7.2 Bibliotecas de Aprendizagem Automática

- Mlpack
  Biblioteca de ML para desenvolvimento em C++.
- Scikit-Learn
  Biblioteca *open-source* de aprendizagem automática para desenvolvimento em Python.
- TensorFlow
  Biblioteca desenvolvida pela Google para desenvolvimento de modelos de *Deep Learning.*
- Torch
  Biblioteca *open-source* de aprendizagem automática para desenvolvimento na linguagem Lua.
- Weka
  Ferramenta que contém uma coleção de algoritmos de aprendizagem automática.

Existem também algumas linguagens de programação que com o auxílio de determinadas bibliotecas conseguem resolver problemas de aprendizagem automática. De entre as mais populares destacam-se:

- MatLab
  Disponibiliza bibliotecas para o desenvolvimento de modelos de aprendizagem automática.
- Python
  Normalmente utilizado em conjunto com a biblioteca *scikit-learn* para desenvolvimento de modelos de aprendizagem automática.
- R
  Oferece pacotes R para o desenvolvimento de modelos de aprendizagem automática.

3.3 Trabalhos Relacionados

Como foi analisado no capítulo 2.1, a implementação de soluções que pretendem melhorar os indicadores estratégicos é importante, pois permite perceber, por exemplo, os clientes mais propensos a abandonar um fornecedor ou qual a imagem que o cliente tem de um fornecedor.

Para isso foram estudadas sete abordagens científicas onde são identificadas técnicas para responder a alguns dos principais problemas que estão na base dos indicadores estratégicos existentes, na área das telecomunicações.

Os artigos foram escolhidos por apresentarem algumas das técnicas mais populares de aprendizagem automática na resolução de alguns dos problemas ou oportunidades que permitem melhorar os indicadores estratégicos de uma provedora de serviços de telecomunicações.

[13] https://spark.apache.org/

As três primeiras abordagens tentam mitigar o problema do abandono de clientes e identificação dos fatores de abandono, recorrendo a técnicas de aprendizagem automática distintas.

A quarta abordagem identifica o grau de influência social de um cliente na operadora de telecomunicações, recorrendo a técnicas de *Big Data*, aprendizagem automática e grafos.

Na quinta abordagem, é apresentada uma estratégia para detetar e antecipar problemas de rede em tempo real, apresentando uma arquitetura de *Big Data* e modelos de aprendizagem automática.

De seguida, é realizada uma análise de sentimento a *tweets* de cinco operadoras de telecomunicações, através de ontologias e processamento de linguagem natural.

O último artigo apresenta uma análise Recência, Frequência e Monetária (RFM) agrupando os clientes, de forma a direcionar ofertas mais personalizadas para esses clientes.

3.3.1 Abandono dos Clientes

O artigo de (Ullah et al., 2019) tem como objetivo apresentar um modelo que utiliza técnicas de classificação e de agrupamento para identificar os clientes que estão a abandonar uma operadora de telecomunicações e enumerar os fatores que estão na base dessa decisão.

Depois da classificação como clientes positivos - que presumivelmente não abandonam - e clientes negativos - que presumivelmente abandonam, o modelo propõe ainda segmentar os clientes que poderão abandonar.

Este modelo identificou também os fatores que são essenciais para o abandono dos clientes para que, a partir dos dados dos clientes, as operadoras sejam capazes de recomendar as promoções ou campanhas mais relevantes, para grupos de clientes.

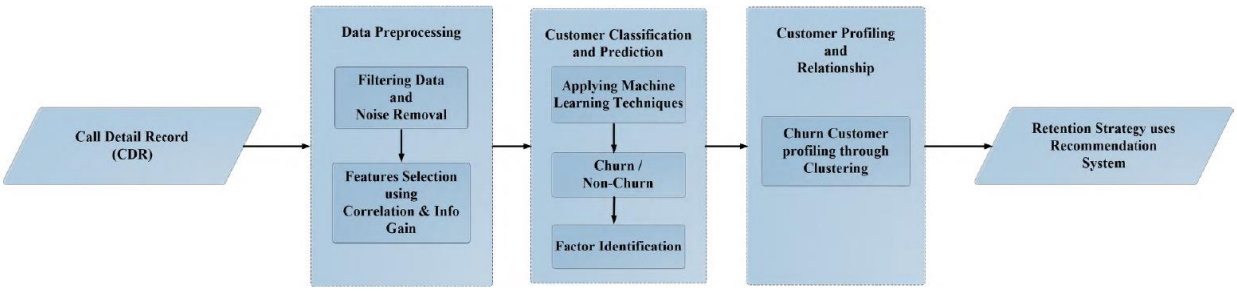


Figura 7 – Modelo proposto para a previsão de abandono dos clientes [14]

O modelo proposto é apresentado na Figura 7. Os dados utilizados para a implementação deste modelo foram recolhidos de uma empresa asiática. Estes dados incluem registos de chamadas telefónicas e mensagens de texto entre clientes.

[14] Retirado de (Ullah et al., 2019)

A primeira atividade desenvolvida foi denominada pré-processamento, e inclui a remoção de valores nulos, remoção de características desequilibradas e normalização dos dados. Nesta atividade foi ainda realizada a seleção das características mais relevantes de acordo uma avaliação de ganho de informação e correlação. Através de técnicas de ganho de informação e correlação de atributos foram selecionadas as 17 características mais relevantes, visíveis na Figura 8, que incluem total de chamadas realizadas e total de minutos realizados em chamadas.

Attributes	Information Gain Ranking Values	Correlation Attributes Ranking values
TOTAL_CALLS	0.010614	0.07856
TOTAL_MINS	0.007962	0.0497
TOTAL_CALLS_REV	0.009111	0.07175
ONNET_CALLS	0.008609	0.06123
ONNET_MINS	0.006335	0.04303
ONNET_REV	0.008882	0.06251
OFFNET_CALLS	0.007919	0.06542
OFFNET_MINS	0.006929	0.06139
OFFNET_REV	0.007164	0.0646
INCOMING_TOTAL_CALLS	0.003773	0.04296
CHRGD_CALLS	0.010331	0.0757
CHRGD_MINS	0.008834	0.05974
CHRGD_REV	0.009111	0.07175
FREE_CALLS	0.005597	0.04683
FREE_MINS	0.006043	0.04066
REVENUE_SMS	0.005483	0.04333
RECHRG_TOTAL_LOAD	0.003697	0.05451

Figura 8 – Ranking dos valores de ganho da informação e de correlação de atributos [14]

Seguidamente, os autores focaram-se em identificar quais os potenciais clientes a abandonar uma operadora. Numa fase inicial recorreu-se a um algoritmo através do qual os dados de teste foram utilizados para criar mais dados. Dado que o conjunto de dados utilizado distingue um cliente que poderá abandonar e um que não, foram, numa segunda fase, aplicados algoritmos de classificação, nomeadamente: *Random Tree*, J48, Random Forest, *Decision Stump*, *AdaboostM1* + *Decision Stump*, *Bagging* + *Random Tree*, *Naïve Bayes*, *Multilayer Perceptron*, *Logistic Regression*, *K-Nearest Neighbours* e *Locally Weighted Learning*.

Após a aplicação dos algoritmos de classificação, concluiu-se que o *Random Forest* obteve melhores resultados, classificando corretamente 88.63% dos dados, visível na Figura 9.

Method used	Incorrectly Classified Instances (%)	Correctly Classified Instances (%)	Time for Building Tree (Sec)
Random Forest	11.37	88.63	108.48
Attribute Selected Classifier	11.66	88.34	4.08
J48	11.42	88.58	7.44
Random Tree	15.66	84.34	2.06
Decision Stump	29.02	70.98	0.97
AdaBoostM1 Classifier + Decision Stump	16.05	83.95	9.24
Bagging + Random Tree	11.39	88.61	13.98
Naïve Bayes	52.37	47.63	0.48
Multilayer Perceptron	17.96	82.04	214.18
Logistic Regression	29.02	70.98	1.87
IBK	19.63	80.37	0.02
LWL	18.41	81.59	0.05

Figura 9 – Resultados obtidos pelos algoritmos de classificação [14]

No entanto, para a identificação dos fatores na base do abandono, o *Random Forest* gera florestas complexas tornando difícil visualizar e criar regras de inferência, segundo os autores.

Assim, na terceira atividade, identificação dos fatores, os autores optaram por recorrer a um classificador de atributos selecionáveis, *Attribute Selected Classifier*. Este classificador permite detetar padrões ou fatores que possam estar na base da saída de um cliente, podendo ser mais tarde utilizado pela operadora para especificar possíveis recomendações

Contrariamente às técnicas de aprendizagem automática abordadas no artigo anterior, as técnicas de *Deep Learning* aprendem, de uma forma intrínseca, as relações entre as várias características de um determinado conjunto de dados, permitindo, em alguns casos, diminuir o tempo de desenvolvimento dos modelos.

Devido a isso, (Umayaparvathi & Iyakutti, 2017) propuseram um modelo em que utilizam dois tipos de redes neuronais, *Feedforward Neural Network* e *Convolutional Neural Network* para detetar o abandono dos clientes.

De forma a treinar as três redes neuronais, utilizaram dois conjuntos de dados, um com 70,831 exemplos e outro com 3,333 exemplos. Para validar a qualidade das redes neuronais, recorreram a uma validação cruzada (k=10) para cada rede neuronal, sendo que a FNN obteve os melhores resultados, classificando corretamente 93,1% e 71,66% dos dados, no *dataset* de menor dimensão e no de maior, respetivamente.

Ainda na problemática de abandono de clientes foi analisado um último artigo (Ahmad et al., 2019) em que os autores desenvolveram um modelo de previsão do abandono, recorrendo a técnicas de *Decision Tree, Random Forest, Gradient Boosted Machine Tree* e *Extreme Gradient Boosting*.

Neste artigo, os autores desenvolveram um modelo utilizando dados de uma provedora de telecomunicações síria. Os dados utilizados abrangiam um período de nove meses com um tamanho de 70 *Terabytes* e continham informação dos clientes, dados de rede e de comunicação (chamadas de voz,

mensagens de texto). Devido ao elevado tamanho dos dados, eles foram introduzidos numa plataforma Hadoop. Na Figura 10, encontra-se representada a arquitetura proposta para a criação do modelo.

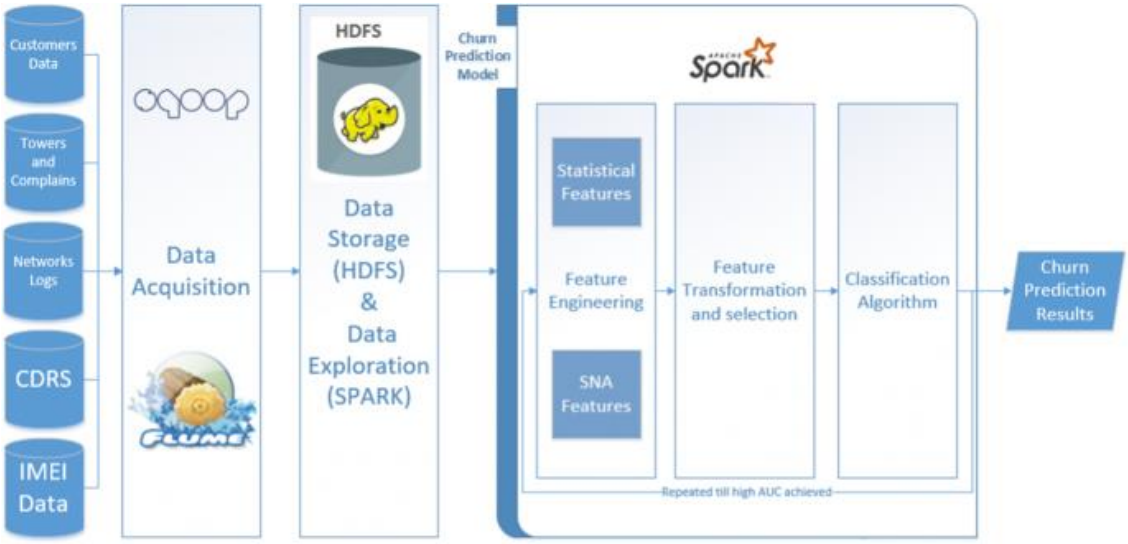


Figura 10 – Arquitetura para a construção do modelo de previsão de abandono [15]

Utilizaram a tecnologia Apache Spark para desenvolver a maior parte das fases de desenvolvimento do modelo: processamento dos dados, engenharia das características (*feature engineering*), treino e teste do modelo. Importa referir que o processo de engenharia das características foi realizado em dois tipos de características: características estatísticas e de análise da rede social da fornecedora, *Social Network Analysis* (SNA).

Para o desenvolvimento do processo de engenharia das características estatísticas, foram agregados dados estatísticos, tais como: média de chamadas realizadas pelo cliente por mês, média dos valores de *upload*/*download*, número de serviços subscritos, rácio entre chamadas e mensagens, etc.

Relativamente ao processo de engenharia das características da rede social da fornecedora, os autores construíram um grafo baseado nos dados de comunicação dos clientes. O grafo consistia em nós, representados pelos números dos clientes, e as arestas bidirecionais, representadas pelas interações entre dois clientes (mensagens, chamadas).

De seguida, o desenvolvimento do modelo de previsão foi realizado dividindo os dados em dois grupos: treino e de teste. O grupo de treino consistia em 70% dos dados e os restantes 30% foram utilizados para testar os modelos. Os algoritmos, referidos anteriormente, utilizados para a criação do modelo foram otimizados e validados recorrendo uma validação cruzada, k=10. Tendo em conta, que os dados estavam desequilibrados, pois tratando-se de um problema de classificação, a classe não-abandono estava em maioria, representando 95% dos dados, os autores resolveram implementar uma técnica de balanceamento denominada subamostragem *(undersampling)*.

[15] Retirado de (Ahmad et al., 2019)

Os autores concluíram que a implementação da análise social permitiu obter excelentes resultados, tendo o *Extreme Gradient Boosting* obtido um valor da área sob a curva, *area under the curve* (AUC), de 93,301%.

3.3.2 Influência Social

Técnicas de deteção dos clientes mais influentes das fornecedoras de serviços são utilizadas para perceber a importância de um cliente na rede. Este tipo de deteção ajuda as fornecedoras a direcionar ofertas mais apropriadas e estudar o comportamento desse tipo de clientes.

Por isso, (Al-Molhem et al., 2019) desenvolveram uma solução que consiste na criação de uma rede social da fornecedora de serviços, através de um grafo bidirecional, representado na Figura 11, com pesos, quantificando a proximidade entre dois clientes (nós).

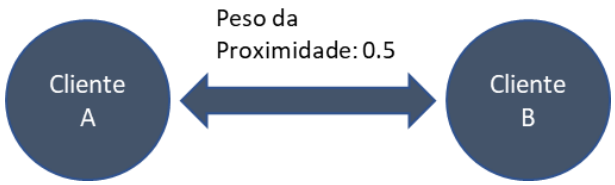


Figura 11 – Representação do grafo

Os autores determinaram a influência através das comunicações entre clientes. No domínio das telecomunicações um cliente é considerado influente pois interage com vários clientes. Estas interações, segundo os autores, podem garantir baixo risco de abandono e potencial elevado para difundir produtos e serviços.

Para além das comunicações entre os clientes, os autores tentaram determinar a influência em clientes que possuem mais do que um cartão *Subscriber Identity Module* (SIM), da mesma fornecedora ou de fornecedoras diferentes, num determinado telemóvel. Este tipo de clientes, segundo os autores tem um potencial mais elevado para abandonarem a operadora, portanto é importante identificar os clientes mais influentes deste tipo.

No desenvolvimento desta rede social, foi utilizado um conjunto de dados, de um período de três meses, que continha informação de chamadas entre clientes, dados dos serviços subscritos pelos clientes, dados dos clientes, como o género ou a idade e dados das torres de telecomunicações.

Os dados foram armazenados na plataforma Hadoop e posteriormente foram desenvolvidos os grafos de relações utilizando a ferramenta Apache Spark. Na Figura 12, encontra-se representada a arquitetura definida pelos autores.

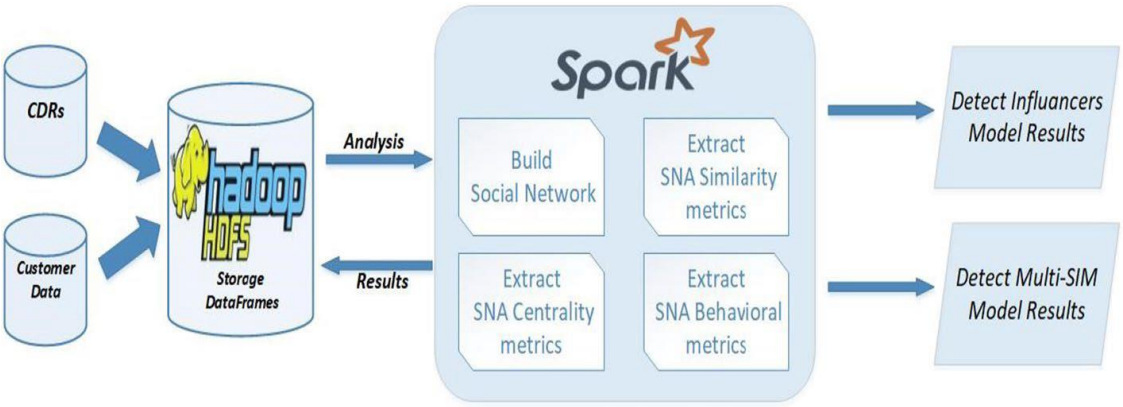


Figura 12 – Arquitetura para desenvolvimento da solução de deteção de influência social [16]

O primeiro passo foi o desenvolvimento da rede, os autores implementaram um grafo com pesos nas arestas. Cada aresta representa uma chamada entre dois clientes. O cálculo do peso dependeu da quantidade e duração das chamadas entre dois clientes. Os valores calculados foram posteriormente normalizados. Na Figura 13, é possível verificar um excerto do grafo.

Source	Destination	Weight
963-9******14	963-9******22	0.0425
963-9******31	963-9******62	0.0496
963-9******94	963-9******11	0.0272
963-9******34	963-9******78	0.0335

Figura 13 – Esquematização dos valores dos grafos

De seguida, os autores calcularam métricas de centralidade para cada cliente, representadas na Figura 14.

Id (GSM)	In-degree	Out-degree	Degree	ND	LCCF
963-9******14	0.039	0.024	0.042	0.02867	0.241417
963-9******31	0.134	0.165	0.1993	0.124	0.022958
963-9******94	0.108	0.085	0.1286	0.092	0.037237
963-9******34	0.021	0.011	0.0213	0.018	0.074074
963-9******89	0.050	0.051	0.0673	0.0453	0.120723

Figura 14 – Representação das métricas de centralidade

Com base nestes valores, os autores calcularam o valor da centralidade autovetor, *Eigenvector centrality* (EV). Este valor mede a influência de um cliente na rede enquanto considera a importância dos clientes com que interage. A ideia principal é que as arestas dos nós mais importantes (conforme medido no passo anterior) são mais valiosas do que as arestas dos nós menos importantes.

[16] Retirado de (Al-Molhem et al., 2019)

Para o cálculo desta centralidade todos os nós começam com um valor de zero e à medida que os nós são analisados, aqueles que têm os maiores valores das métricas de centralidades, calculadas anteriormente, começam a ganhar maior importância. Esta importância propaga-se para os outros nós com que se relacionam. Após um determinado número de iterações, o valor do EV estabiliza e são obtidos os valores definitivos para cada nó. Na Figura 15, é possível verificar o valor do EV para alguns dos clientes.

Id	EV
963-9******14	0.030832
963-9******31	0.094961
963-9******94	0.051983
963-9******34	0.135877
963-9******89	0.046235

Figura 15 – Valores do EV

Para validar a influência desta rede social, os autores selecionaram 50000 clientes da mesma região e dessa amostra escolheram os 15% mais influentes e desenvolveram uma campanha de compra de um determinado serviço, tendo 22255 clientes relacionados com este grupo adquirido o serviço.

3.3.3 Experiência dos Clientes

(Diaz-Aviles et al., 2015) implementaram uma solução que prevê a experiência de um cliente em tempo real, sem qualquer tipo de interação com o cliente, a partir de um conjunto de dados de atividades dos telemóveis e de registos de chamadas para os *call centers*.

O conjunto de dados obtido pelos autores das atividades dos telemóveis consiste em dados anonimizados associados à localização dos dispositivos e dados específicos de determinadas aplicações (taxa de *download,* pacotes retransmitidos). Os autores assumiram que uma má experiência pode ser inferida analisado a *performance* da rede no momento da utilização de determinadas aplicações ou serviços. Tendo em conta a dificuldade em determinar se a experiência é positiva ou negativa consoante a *performance* da rede, os autores testaram a hipótese de os clientes estarem a passar por uma experiência negativa em momentos que antecederam uma chamada ao *call center*. Desta forma, conseguindo provar esta hipótese, as operadoras conseguem antecipar os motivos na base de uma determinada queixa.

Este problema foi classificado como uma tarefa de previsão binária, i.e., irá o cliente ligar para o *call center* ou não, consoante a sua experiência no seu telemóvel. Assim, recorreram a técnicas para a construção do modelo preditivo a partir de dados históricos, de forma antever chamadas para o *call center,* no próprio dia.

Esta solução requer a existência de um sistema com duas capacidades principais:

1. Processar grandes quantidades de dados históricos de rede e de chamadas para o *call center* para conseguir criar modelos preditivos com base no contexto do cliente.
2. Processar grandes quantidades de dados de rede em tempo real.

Para isso recorreram a uma arquitetura, visível na Figura 16, baseada na arquitetura *Lambda* (explicada em detalhe no capítulo **Error! Reference source not found.**)

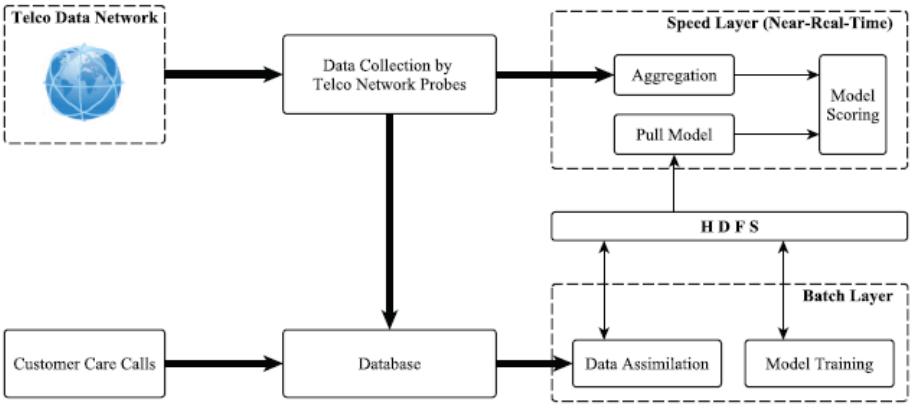


Figura 16 – Arquitetura da solução proposta [17]

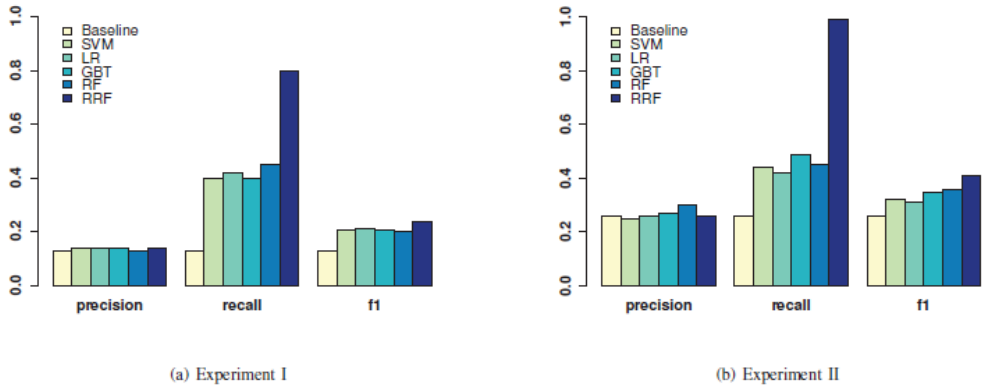
Como foi referido anteriormente, os autores apontaram como principal dificuldade a determinação exata de uma experiência positiva ou negativa num dado momento do dia, sem questionar o cliente.

Definiram então duas abordagens:

- Cruzaram a informação das ligações ao *call center* com os dados de rede dos momentos que antecederam uma chamada, classificando com uma experiência negativa. Assumiram que o momento após a chamada resultou na resolução do problema, classificando as atividades seguintes como experiências positivas.
- Assumiram que o motivo pelo qual a chamada foi realizada resultou de um período longo de experiência negativa, agregando os dados por utilizador num determinado período de tempo.

A nível de resultados, representados na Figura 17, os autores compararam o modelo *Restricted Random Forest* com outros algoritmos e constataram a baixa precisão do modelo nas duas abordagens. No entanto observaram um aumento do F1 da primeira abordagem para a segunda, o que indica um alto número de falsos positivos, isto é, o modelo classifica erradamente a chamada do cliente. Este detalhe pode ser aproveitado pela operadora para proactivamente identificar motivos de má experiência mesmo que o cliente decida não contactar.

[17] Retirado de (Diaz-Aviles et al., 2015)



(a) Experiment I (b) Experiment II

Figura 17 – Demonstração de resultados [17]

3.3.4 Satisfação dos Clientes

Um sentimento positivo de uma empresa é um indicador de preferência do público por uma determinada marca e um sentimento negativo pode levar um cliente a abandonar uma empresa por outra mais indicada às suas preferências.

No mercado competitivo das telecomunicações, é importante ter valores altos de positividade. Tendo em conta isso, (Ranjan et al., 2018) desenvolveram uma análise de sentimentos a cinco operadoras indianas a partir de *tweets*. Para isso recorreram a um dicionário de sentimentos, que contem palavras, frases, conceitos relacionados e polaridades de sentimentos, utilizando técnicas de processamento de linguagem natural e de ontologias.

Os autores analisaram *tweets* de quatro meses, e compararam a sua previsão de número de crescimento do número de clientes nesses meses com o valor atual de crescimento, obtendo resultados bastantes acertados.

3.3.5 Valor dos Clientes

O artigo de (Win & Bo, 2020) propõe um modelo de prevê a categoria do cliente no ano seguinte baseado no seu *Customer Lifetime Value*, permitindo à fornecedora priorizar quais os clientes em que devem ser desenvolvidos esforços para os manter. Para isso utilizaram o algoritmo *Random Forest* e otimização dos parâmetros deste segundo a técnica *Random Search* para obter o melhor acerto preditivo.

A metodologia adotada está representada na Figura 18.

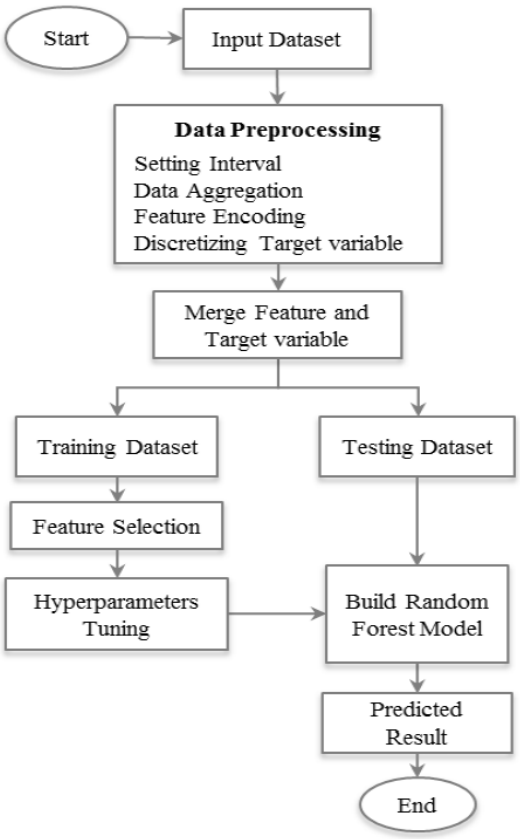


Figura 18 – Metodologia para previsão da categoria do cliente

A partir do conjunto de dados obtidos, os autores adotaram a RFM. A análise RFM é uma técnica de *marketing* utilizada para determinar quantitativamente quais são os melhores clientes, examinando há quanto tempo um cliente comprou (recência), com que frequência ele compra (frequência) e quanto o cliente gasta (monetária).

Depois da segmentação do conjunto de dados numa abordagem RFM, os autores desenvolveram o modelo de aprendizagem automática.

Os resultados permitiram concluir um acréscimo no acerto utilizando a técnica *RandomSearch* em comparação com os parâmetros por defeito do algoritmo *Random Forest*, com 84,27% e 81.46% de acerto respetivamente.

3.4 Mapeamento indicadores estratégicos em Casos de Uso

A análise realizada da literatura e dos indicadores estratégicos permite concluir a existência de várias técnicas para melhorar os indicadores. Assim, foi o criado o mapeamento representado na Figura 19.

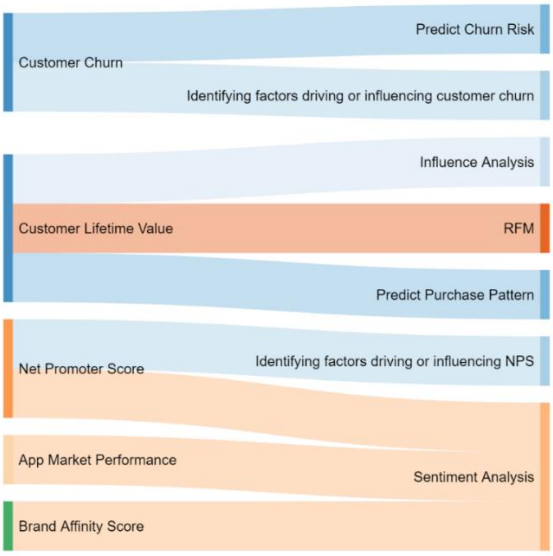


Figura 19 – Mapeamento indicadores estratégicos em Casos de Uso

A escolha dos indicadores, representados à esquerda da figura, é justificada com o facto de incidirem em nos clientes das telecomunicações e na perceção dos clientes para com a empresa.

Começando pelo indicador *Customer Churn*, a análise realizada permite perceber que existem duas técnicas que sendo implementadas ajudarão a melhorar o indicador. A previsão do risco de abandono e a identificação dos fatores que influenciam a saída são pontos de partida para a melhoria deste indicador.

O indicador *Customer Lifetime Value,* que como foi referido no 2.1, mede o valor que um dado cliente tem através das compras que realiza, pode ser acompanhado e melhorado, recorrendo a técnicas de análise de influência, identificando os clientes mais importantes, percebendo os serviços que utilizam mais e tentar acrescentar mais valor a estes. A análise RFM, pode ser utilizada para agrupar os clientes em grupos. E por fim, a previsão do padrão de compra pode ser utilizada para determinar a próxima melhor oferta para um dado cliente.

O indicador *Net Promoter Score*, permite avaliar a satisfação do cliente e a análise realizada mostra que pode ser melhorado através da identificação de fatores que levam à satisfação ou insatisfação. Para além disso, a análise de sentimento pode ser analisada, através das interações que o cliente realiza com a operadora ou em canais disponibilizados pela operadora, como por exemplo o *website*.

Por fim, tanto o indicador *App Market Performance* e o *Brand Affinity Score* estão mais relacionados com a perceção que o cliente tem da operadora. Estes indicadores podem ser atentamente analisados e melhorados, através da análise de sentimento por parte das interações realizadas pelo cliente.

Este mapeamento tem como objetivo servir de guia estratégico interno para a elaboração da melhoria dos indicadores estratégicos, permitindo ser adaptado a qualquer operadora que pretenda melhorar um ou mais indicador estratégicos. A existência de mais do que um caso de uso para cada indicador estratégico é importante pois nem sempre existem dados disponíveis para desenvolver de uma determinada forma.

4 Análise de Valor

Neste capítulo irá ser apresentada a análise de valor para a implementação deste projeto.

O capítulo inicia com a identificação e análise da oportunidade para este projeto. De seguida é apresentada a proposta de valor segundo o modelo Osterwalder.

O capítulo conclui com a apresentação da avaliação e seleção de tecnologias disponíveis para o desenvolvimento do módulo de Aprendizagem Automática.

4.1 Identificação da Oportunidade

O termo *Customer* 360 deriva do conceito foco no cliente, desenvolvido em 1960, por Lester Wunderman. Até essa altura, o foco das empresas estava direcionado em meios de comunicação como a televisão e a rádio. Com o crescimento da internet a partir do final dos anos 90, isto criou diversas oportunidades para explorar o perfil individual dos clientes, melhorando a sua experiência na utilização dos produtos e serviços de uma operadora de telecomunicações.

(Briggs et al., 2020) afirma que "as organizações estão a priorizar a criação de experiências mais centradas nos clientes em vez de utilização de marketing direcionado para a aquisição de clientes". Esta mudança de paradigma é suportada por dois terços dos diretores executivos das empresas pertencentes ao grupo Global 2000, que estão a mudar o seu foco para estratégias digitais mais modernas para melhorar a experiência do cliente (Reinsel et al., 2018).

Em adição, do ponto de vista tecnológico, o crescimento da utilização de técnicas de Aprendizagem Automática e *Big Data* expande as possibilidades para compreender a informação de cada cliente. Na

Figura 20, é possível observar o crescimento de pesquisas no Google sobre o termo *Machine Learning*, no período Janeiro 2004 a Janeiro 2021.

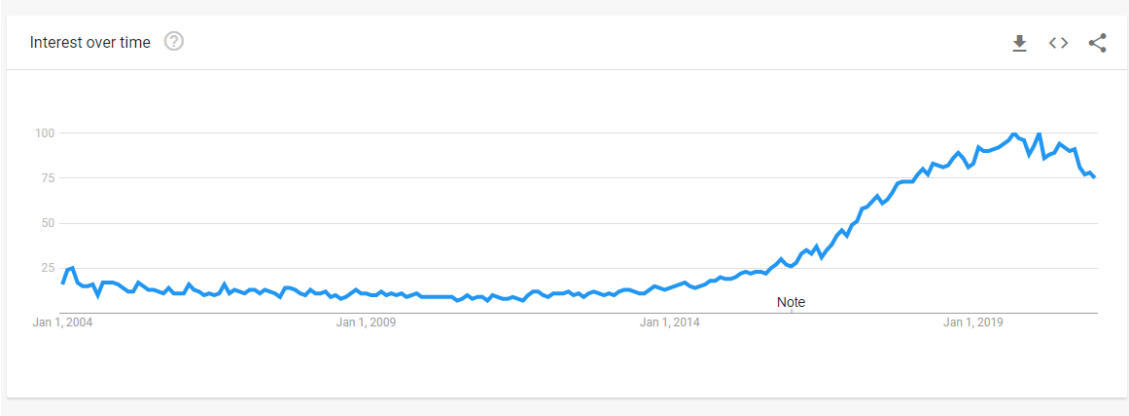


Figura 20 – Tendências de pesquisa *Machine Learning* [18]

4.2 Análise da Oportunidade

Os sistemas que compõem a arquitetura das telecomunicações armazenam grandes quantidades de dados sobre cada cliente, criando uma oportunidade para as provedoras de telecomunicações desenvolverem soluções que permitam: explorar a experiência que um cliente está a sentir, conhecimentos mais alargados do perfil do cliente e que serviços estão a utilizar. Esse conhecimento pode ser aproveitado para se proactivamente envolver com as necessidades dos clientes.

A diversidade de dados armazenados pelas operadoras posiciona-as de forma vantajosa. Isto leva a uma oportunidade que se for bem aproveitada, pode melhorar o índice de satisfação e reduzir o abandono. A operacionalização dos diferentes tipos de dados, tais como dados de faturação ou rede são a chave da transformação para disponibilizar uma melhor experiência para os clientes.

A utilização de técnicas como Aprendizagem Automática e *Big Data* leva à compreensão de determinadas características conduzindo a atividades personalizadas de *marketing*, vendas e atendimento ao cliente para esses clientes. Esta compreensão pode ainda assumir a forma de resultados preditivos que indiquem a possibilidade de um cliente comprar um determinado produto ou uma avaliação do risco de abandono.

Apresenta-se de seguida, na Tabela 11, a análise SWOT para a implementação de uma solução deste género, abstraindo uma operadora especifica de telecomunicações.

Tabela 11 – Análise SWOT

Forças	Quantidade de dados disponíveis. Diversidade de dados dos sistemas.

[18] Retirado de: https://trends.google.com/trends/explore?date=all&q=machine%20learning

Fraquezas	Arquitetura dos sistemas demasiado complexas. Dados nem sempre consistentes.
Ameaças	Regulamento Geral de Proteção de Dados (RGPD). Insatisfação dos clientes com a indústria.
Oportunidades	Crescimento de tecnologias de ML e *Big Data.*

4.3 Proposta de Valor

A proposta de valor segundo o modelo Osterwalder é formada recorrendo a dois blocos específicos – perfil do cliente e proposta de valor da empresa. No caso do desenvolvimento deste projeto, o sistema a ser implementado será utilizado por operadores das telecomunicações e outros intervenientes com responsabilidades e estes pretendem conhecer melhor os clientes que usam os serviços das operadoras.

Relativamente ao perfil do cliente, representado à direita na Figura 21, são especificados os seguintes pontos:

- *Gains*
  Os benefícios pelo qual o cliente espera e deseja.

- *Pains*
  As experiências negativas na obtenção da informação desejável.

- *Customer Jobs*
  Os problemas que tentam ser resolvidos.

Relativamente à proposta de valor da empresa, representado à esquerda na Figura 21, encontra-se segmentada em três pontos:

- *Gain Creators*
  Como o produto adiciona valor para o cliente.

- *Pain Relievers*
  Descreve como o produto alivia as dores dos clientes.

- *Products & Services*
  Os produtos que irão adicionar valor para o cliente.

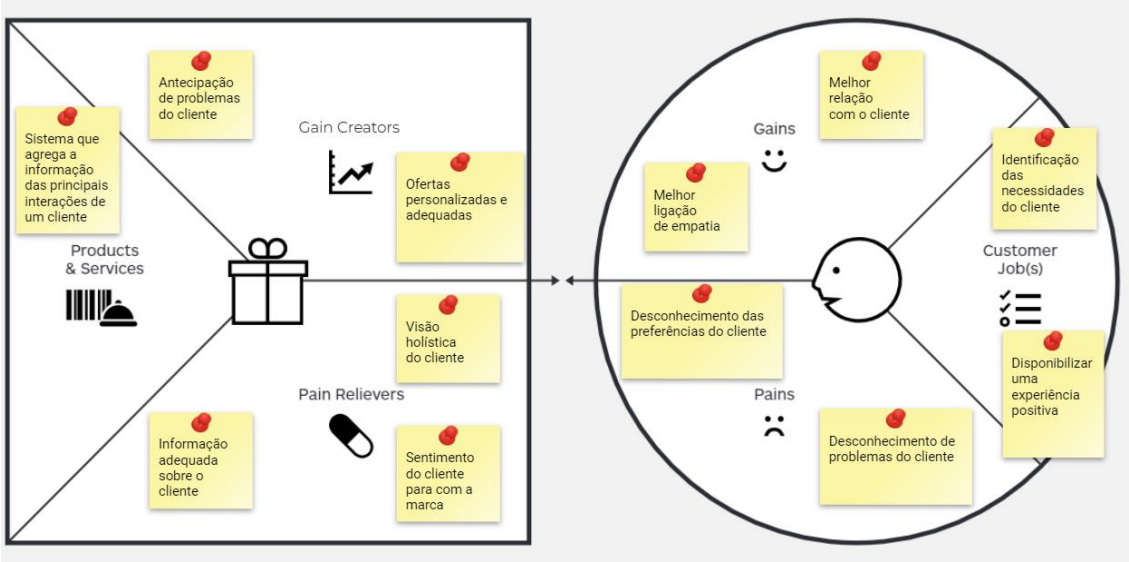


Figura 21 – Proposta de valor baseado no modelo Osterwalder

Esta solução interliga as interações de um cliente das telecomunicações, criando respostas para as necessidades dos clientes. Isto atrai uma variedade de empresas de telecomunicações que estão interessadas em conhecer melhor os seus clientes, promovendo um benefício mútuo entre empresa e cliente.

4.4 Análise Funcional

A *Function Analysis and System Technique* (FAST) é utilizada para definir, analisar e compreender as funcionalidades de um determinado produto e como as funcionalidades se relacionam entre si. É representada através de um diagrama para visualizar funcionalidades numa sequência lógica.

A leitura do diagrama pode iniciar da esquerda para a direita, começando com uma funcionalidade e seguindo essa direção questionando como a funcionalidade é alcançável. O caminho inverso questiona a finalidade da funcionalidade.

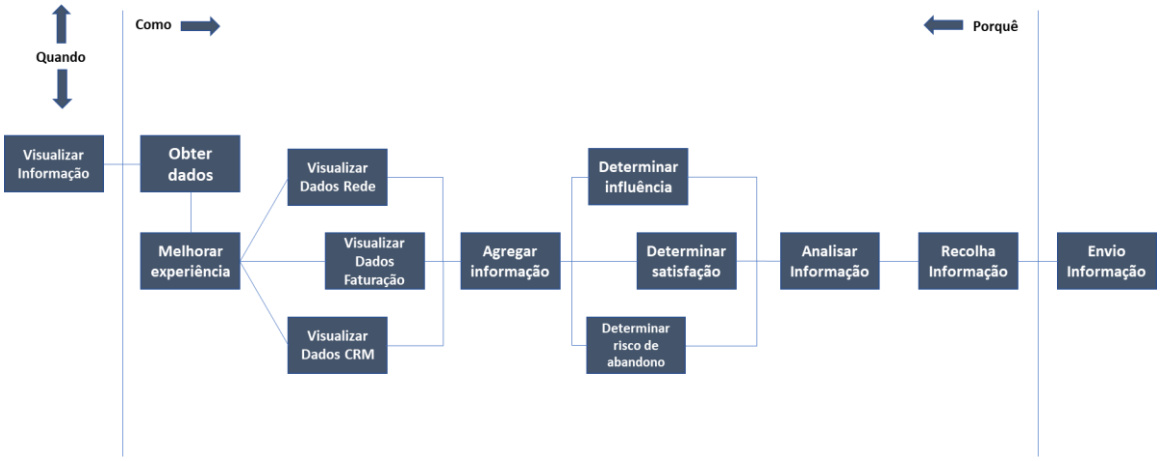


Figura 22 – Diagrama FAST

Podemos concluir a partir da Figura 22, as seguintes funcionalidades:

- Visualizar Dados de Rede
- Visualizar Dados de Faturação
- Visualizar Dados do CRM
- Determinar Influência
- Determinar Satisfação
- Determinar Risco de Abandono

Para além destas funcionalidades, é possível também, identificar duas entidades interessadas:

- Provedora de Serviços
  Empresa que disponibiliza serviços de telecomunicações.

- Clientes
  Clientes da provedora.

Tabela 12 – Análise de comparação dos requisitos

	Visualizar Dados de Rede	Visualizar Dados de Faturação	Visualizar Dados do CRM	Determinar Influência	Determinar Risco de Abandono	Determinar Satisfação	Relevância
Visualizar Dados de Rede		1	1	1	1	1	5
Visualizar Dados de Faturação	0		1	1	1	1	4

	Visualizar Dados de Rede	Visualizar Dados de Faturação	Visualizar Dados do CRM	Determinar Influência	Determinar Risco de Abandono	Determinar Satisfação	Relevância
Visualizar Dados do CRM	0	0		1	1	1	3
Determinar Influência	0	0	0		0	1	1
Determinar Risco de Abandono	0	0	0	1		1	2
Determinar Satisfação	0	0	0	0	0		0

Da Tabela 12, conclui-se que a visualização dos dados de rede é a mais relevante, seguida da visualização dos dados de faturação.

4.5 Avaliação e Seleção de Tecnologias

Neste subcapítulo é apresentado o processo de escolha da tecnologia para o desenvolvimento da solução que permite a melhoria de um dos indicadores estratégicos. Para isso recorreu-se ao método de análise de processo analítico hierárquico, *Analytic Hierarchy Process (AHP).*

Foram selecionadas as seguintes tecnologias:

- Python
- R
- MatLab

Estas tecnologias terão de ser comparadas entre si, através dos seguintes critérios:

- Bibliotecas: quantidade de bibliotecas disponíveis para implementar a solução
- Modularidade: capacidade para substituir mecanismos ou para criar novos
- Documentação: qualidade da documentação e facilidade de acesso

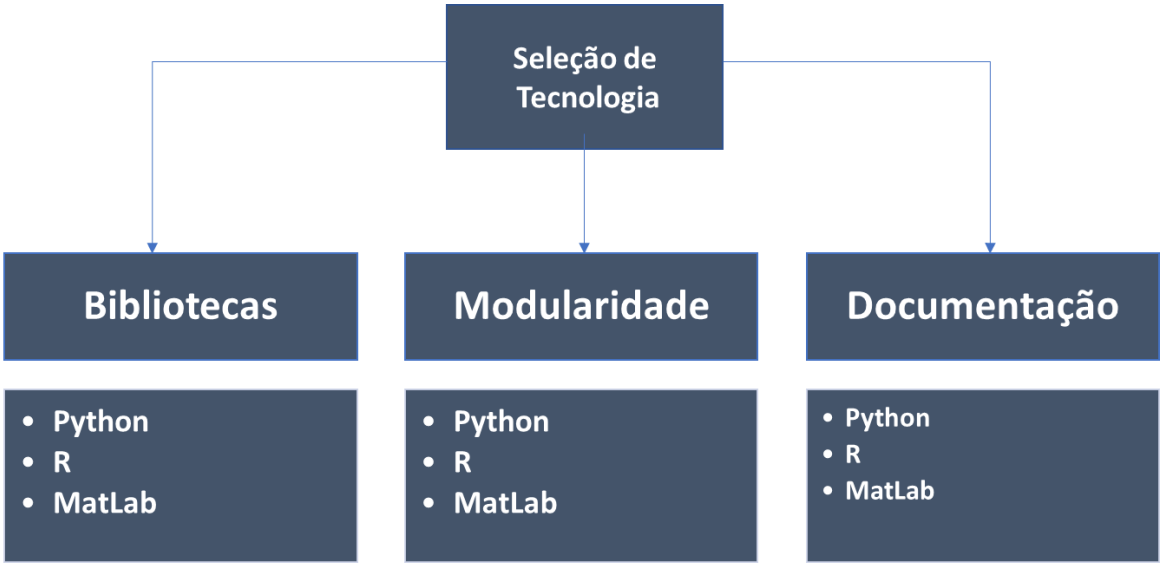


Figura 23 – Árvore hierárquica

4.5.1 Cálculo e Análise da Prioridade Relativa de cada Critério

Tendo em conta a árvore hierárquica definida na Figura 23, foi definido o grau de importância de cada critério. Para determinar o grau de importância foi utilizado a escala fundamental de Saaty, representada na Figura 24.

1	Igual Importância	As duas atividades contribuem igualmente para o objetivo
3	Importância pequena de uma para a outra	A experiência e o julgamento favorecem levemente uma atividade em relação à outra.
5	Importância grande ou essencial	A experiência e o julgamento favorecem fortemente uma atividade em relação à outra.
7	Importância muito grande ou demonstrada	Uma atividade é muito fortemente favorecida em relação à outra.
9	Importância absoluta	A evidência favorece uma atividade em relação à outra com o mais alto grau de certeza.
2,4,6,8	Valores intermediários	Quando se procura uma condição de compromisso entre as duas definições.

Figura 24 – Escala de Saaty

A prioridade para cada critério encontra-se pormenorizada na Tabela 13.

Tabela 13 – Matriz de comparação dos critérios

	Bibliotecas	**Modularidade**	**Documentação**
Bibliotecas	1	4	2
Modularidade	1/4	1	1/3
Documentação	1/2	3	1

De seguida, a matriz é normalizada para se obter a prioridade relativa de cada um dos critérios. Para obter a prioridade deve-se calcular a média aritmética de cada um dos valores dos critérios. Esta normalização está representada na Tabela 14.

Tabela 14 – Matriz normalizada com Pesos

	Bibliotecas	**Modularidade**	**Documentação**	**Prioridade Relativa**
Bibliotecas	4/7	1/2	3/5	0,55
Modularidade	1/7	1/8	1/10	0,13
Documentação	2/7	3/8	2/7	0,32

O passo seguinte passa por calcular o valor da razão de consistência (RC), para medir o quanto dos julgamentos foram consistentes. Para que isto se verifique, o valor do RC terá de ser inferior a 0.1. Caso este valor seja superior, o processo deverá ser reanalisado até que os julgamentos sejam consistentes.

Para determinar o valor do RC, é necessário, calcular o índice de consistência (IC), dado pela fórmula (9), bem como o índice aleatório (IR).

$$IC = \frac{\lambda max - n}{n - 1} \tag{9}$$

Onde λ_{max} é o maior valor próprio da matriz de comparação e n a ordem de matriz. O valor de IR neste caso é 0,58, correspondente à ordem 3 da matriz da Tabela 15.

Tabela 15 – Valores do IR para matrizes quadradas de ordem n

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
IR	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.48	1.56	1.57	1.59

Os dois índices são utilizados para definir a fórmula do RC da seguinte forma:

$$RC = \frac{IC}{IR} \tag{10}$$

Primeiramente é calculado o valor de λ_{max} , de modo a ser utilizado no cálculo do IC. Assim, λ_{max} poderá ser

$$\begin{bmatrix} 1 & 4 & 2 \\ 1/4 & 1 & 1/3 \\ 1/2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 0{,}55 \\ 0{,}13 \\ 0{,}32 \end{bmatrix} = \lambda_{max} \begin{bmatrix} 0{,}55 \\ 0{,}13 \\ 0{,}32 \end{bmatrix} \Leftrightarrow$$

$$\Leftrightarrow \begin{bmatrix} 2/3 \\ 3/8 \\ 1 \end{bmatrix} \begin{bmatrix} 0{,}55 \\ 0{,}13 \\ 0{,}32 \end{bmatrix} = \lambda_{max} \begin{bmatrix} 0{,}55 \\ 0{,}13 \\ 0{,}32 \end{bmatrix} \Leftrightarrow$$

$$\lambda_{max} = \begin{bmatrix} 3{,}02 \\ 3{,}01 \\ 3{,}02 \end{bmatrix}$$

Com esta matriz final, pode-se calcular λ_{max} como sendo a média dos valores obtidos. O valor obtido para o λ_{max} foi aproximadamente 3,02.

Assim, o passo seguinte foi calcular o valor de IC, através da fórmula (11).

$$IC = \frac{\lambda\, max - n}{n-1} \Leftrightarrow \tag{11}$$

$$IC = \frac{3{,}02 - 3}{3-1} \Leftrightarrow IC \simeq 0{,}01$$

Concluindo, através da análise do valor de IR para uma matriz quadrática de ordem 3, ou seja, 0,58, pode-se calcular o RC a partir da fórmula.

$$RC = \frac{IC}{IR} \Leftrightarrow RC = \frac{0{,}01}{0{,}58} \Leftrightarrow RC \simeq 0.02$$

Assim, com um valor RC inferior a 0,1, pode-se concluir que os julgamentos são consistentes. O passo seguinte passa por obter a matriz de comparação para cada alternativa, segundo os critérios definidos.

4.5.2 Construção da Matriz de Comparação de cada Alternativa

A primeira matriz a ser construída diz respeito às bibliotecas disponíveis. A avaliação desta característica é baseada na diversidade e qualidade das bibliotecas disponíveis de aprendizagem automática e estatísticas para a elaboração da melhoria dos indicadores estratégicos.

A tabela com a comparação das alternativas segundo o critério Bibliotecas encontra-se representado na Tabela 16.

Tabela 16 – Matriz de comparação para o critério Bibliotecas

Bibliotecas	Python	R	MatLab	Vetor Próprio
Python	1	2	5	0,55
R	1/2	1	5	0,35
MatLab	1/5	1/5	1	0,09

De seguida, foi construída a matriz segundo o critério Modularidade. Esta característica é importante para perceber a capacidade de as tecnologias em estudo evoluírem de forma a acomodar novas tendências tecnológicas, como por exemplo novos algoritmos.

Na Tabela 17 encontra-se representado a comparação das alternativas segundo o critério Modularidade.

Tabela 17 – Matriz de comparação para o critério Modularidade

Modularidade	Python	R	MatLab	Vetor Próprio
Python	1	2	8	0,57
R	1/2	1	8	0,37
MatLab	1/8	1/8	1	0,06

Por fim, foi construído a matriz segundo o critério Documentação. Este último critério diz respeito à qualidade da documentação existente por parte das equipas que desenvolveram e que mantêm as tecnologias.

A tabela com a comparação das alternativas segundo o critério Bibliotecas encontra-se representado na Tabela 18.

Tabela 18 - Matriz de comparação para o critério Documentação

Documentação	Python	R	MatLab	Vetor Próprio
Python	1	5	2	0,57
R	1/5	1	1/4	0,10
MatLab	1/2	4	1	0,33

Concluídas as matrizes para cada alternativa, foi criado o seguinte resumo, representado na Figura 25.

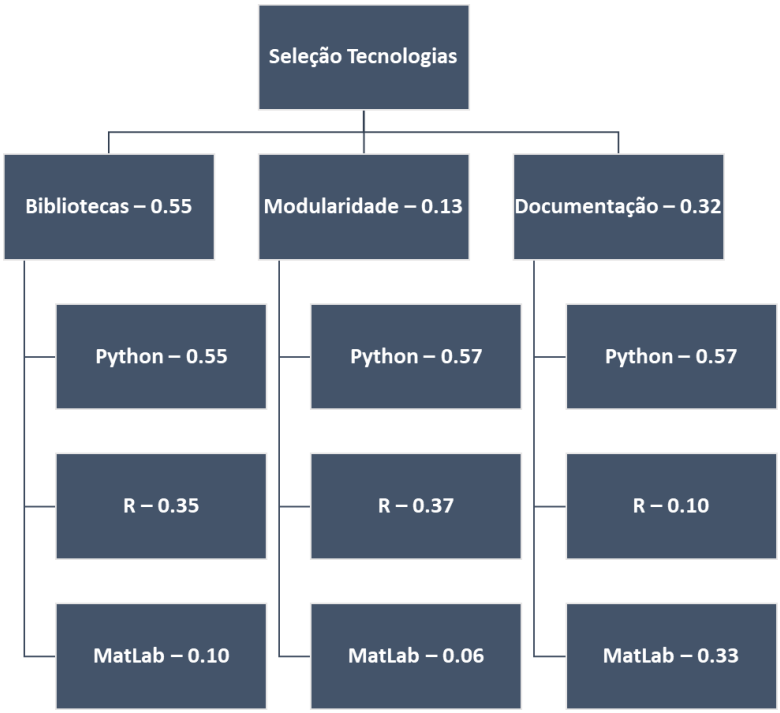


Figura 25 – Resumo dos pesos das alternativas e dos critérios

O passo final passa por escolher a alternativa mais adequada. Para isso será multiplicado a matriz com as prioridades relativas de cada critério com a matriz composta pelos vetores próprios.

$$\begin{bmatrix} 0.55 & 0.57 & 0.57 \\ 0.35 & 0.37 & 0.10 \\ 0.10 & 0.06 & 0.33 \end{bmatrix} \times \begin{bmatrix} 0.55 \\ 0.13 \\ 0.32 \end{bmatrix} = \begin{bmatrix} \mathbf{0.55} \\ 0.12 \\ 0.33 \end{bmatrix}$$

Obtendo os resultados finais, pode-se concluir que a alternativa Python é a que serve melhor para o desenvolvimento da melhoria desta parte da solução.

5 Design

Neste capítulo é apresentada a especificação de *Design* para esta solução, considerando os requisitos e os objetivos deste projeto.

O capítulo lista os requisitos funcionais e não funcionais e a proposta de arquitetura, detalhando os componentes presentes.

5.1 Requisitos Funcionais

Para a definição dos requisitos funcionais, foram realizadas várias reuniões com especialistas na Celfocus, avaliando cenários que são utilizados hoje em dia. Os requisitos encontram-se sintetizados de seguida:

- RF1: O sistema deve ser capaz de processar dados de forma diária
- RF2: O sistema deve ser capaz de processar dados em tempo-real.
- RF3: O sistema deve ser capaz de determinar o risco de abandono de Cliente.
- RF4: O sistema deve ser capaz de consultar dados de Cliente.
- RF5: O sistema deve ser capaz de consultar os atributos que contribuem para o abandono.
- RF6: O sistema deve ser capaz de consultar clientes em risco de abandono.

O requisito 1 e 2, foram identificados como sendo complementares. O requisito 1, tem como objetivo processar todos os dados relevantes presentes nos vários sistemas externos, no entanto poderá haver falhas dado tratar-se de um grande volume de dados, surge então o requisito 2 que pretende complementar eventuais falhas, mas também apresentar a informação mais atualizada possível.

O requisito 3 foi identificado, pois corresponde à melhoria do indicador estratégico *Customer Churn*.

O requisito 4 foi capturado, com o objetivo de serem consultado os dados, de um dado cliente, que existem nos sistemas externos, como por exemplo sistemas de CRM e *Billing*.

O requisito 5 pretende identificar os atributos/características que contribuem para o abandono dos clientes, percebendo assim as características que devem ser analisadas pela operadora com o objetivo de melhorar o serviço prestado aos clientes.

Por fim, o requisito 6, tem como objetivo consultar os clientes que apresentem as características identificadas no requisito 5.

5.2 Requisitos Não Funcionais

Os requisitos não funcionais pretendem especificar o comportamento que o sistema deve ter. Tal como os requisitos funcionais, os não funcionais foram definidos com os especialistas na Celfocus.

- Interface gráfica deverá ser apelativa, intuitiva e amigável para o utilizador.
- Utilização de REST (*Representational State Transfer)* para comunicação entre APIs.
- Boas práticas de design, como por exemplo, utilização do padrão MVC (Modelo, Vista, Controlador; *Model, View, Controller)*.
- Utilização de ReactJS [19], preferencialmente a versão mais recente, para o desenvolvimento da interface gráfica.
- Os dados devem ser persistidos em base de dados MongoDB [20].
- Utilizar tecnologias de código aberto, preferencialmente, com licença Apache.

5.3 Arquitetura

Na Figura 26, é apresentada a proposta de arquitetura para esta solução, através de um diagrama de componentes. A arquitetura resultou do estudo das duas arquiteturas mais conhecidas de *Big Data*, tendo sido optado por desenhar uma arquitetura baseada na arquitetura *Lambda*.

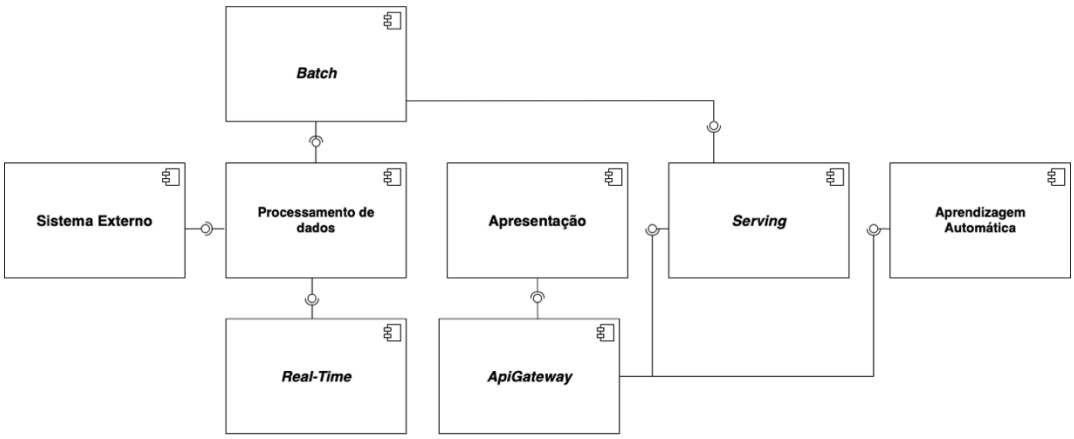


Figura 26 – Arquitetura escolhida

[19] https://reactjs.org
[20] https://www.mongodb.com/

Como foi referido anteriormente no capítulo 3.1.3, a arquitetura *Lambda* difere da arquitetura *Kappa*, pois apresenta uma camada que é responsável por armazenar dados históricos e processá-los de forma periódica, ao contrário da arquitetura *Kappa* onde os dados são apenas extraídos e processados das fontes. Também a existência de duas camadas de processamento de dados cria mais robustez ao sistema permitindo que haja duas componentes de onde a informação é extraída, eliminando pontos únicos de falha de processamento.

Assim, tendo em conta estes argumentos, a escolha recaiu na arquitetura *Lambda*.

A arquitetura é composta pelas seguintes componentes:

- Aprendizagem Automática
  Componente responsável por gerir os modelos de Aprendizagem Automática.

- API *Gateway*
  Componente com a responsabilidade de gerir as ligações a partir do componente de Apresentação para os componentes *Serving* e Aprendizagem Automática.

- Apresentação
  Componente que tem como objetivo disponibilizar uma interface gráfica.

- *Batch*
  Componente responsável por processar e armazenar os dados históricos.

- Processamento de Dados
  Componente responsável por extrair e processar a informação dos vários sistemas para o componente *Batch* ou *Real-Time.*

- *Real-Time*
  Componente responsável por processar e armazenar os dados em tempo-real.

- *Serving*
  Componente responsável por disponibilizar a informação ao componente de apresentação.

- Sistema Externo
  Representação de um sistema pertencente às telecomunicações, e.g. faturação, onde serão retirados os dados.

Esta escolha permite definir fronteiras entre cada um dos componentes, distribuindo as responsabilidades de cada componente.

5.3.1 Alternativa considerada

Tendo em conta que foram apenas estudadas duas arquiteturas, a alternativa considerada é baseada na arquitetura *Kappa* e encontra-se representada na Figura 27.

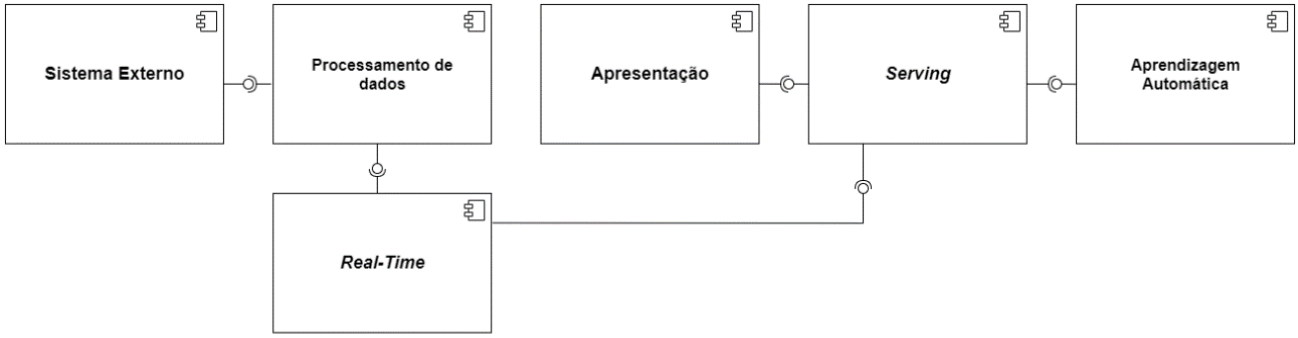


Figura 27 – Arquitetura alternativa

5.4 Arquitetura Detalhada

Neste capítulo será pormenorizado as tecnologias presentes em cada componente, através de um diagrama de componentes de granularidade mais baixa, visível na **Error! Reference source not found.**. Posteriormente, serão apresentados os motivos para as escolhas das tecnologias que pertencem a cada componente.

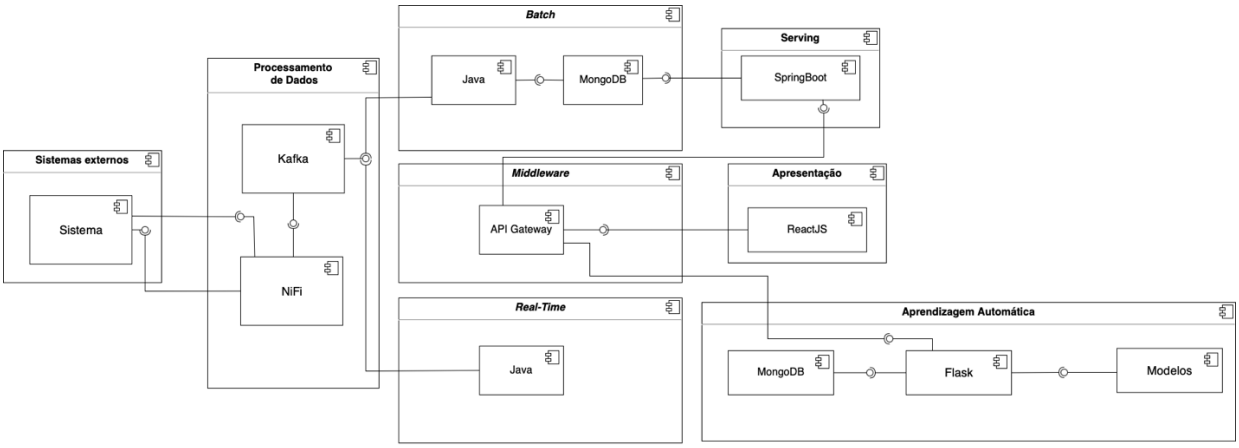


Figura 28 - Diagrama de Componentes

5.4.1 Sistemas Externos

O componente de Sistemas Externos representa os sistemas que compõem uma operadora de telecomunicações.

5.4.2 Processamento de Dados

O componente Processamento de Dados tem como objetivo gerir o processamento dos dados provenientes dos sistemas externos e o componente de dados históricos (*Batch*), bem como o componente de dados em tempo real (*Real-Time*).

A tecnologia Apache Nifi foi escolhida pela capacidade de processar, extrair e transformar os dados de várias fontes distintas, como por exemplo, base de dados ou ficheiros de texto, podendo ser definida a frequência com que a extração é efetuada. No contexto deste projeto, esta tecnologia será utilizada para processar os dados que foram atualizados em tempo-real e processá-los de uma forma diária.

Ainda neste componente, selecionou-se a tecnologia Apache Kafka com a responsabilidade de gerir os dados de forma temporária, mas também pela integração entre consumidores e produtores, sem a existência de bloqueios para os produtores e sem deixar que os produtores saibam quem são os consumidores finais.

Tendo em conta a exigência em processar rapidamente os dados que são criados ou atualizados nos vários sistemas, a combinação de tecnologias do Apache Nifi para extrair e processar e o Apache Kafka como um mediador de mensagens permite cumprir a exigência de um processamento rápido e sem bloqueios.

5.4.3 Dados em Tempo-Real

Neste componente os dados são consumidos de um tópico Kafka presente no componente de Processamento de Dados e reencaminhados para o componente de dados históricos para serem realizadas operações de limpeza dos dados e armazenamento

5.4.4 Dados Históricos

Este componente assenta no desenvolvimento de uma aplicação Java para transformar e agregar os dados e armazenar numa base de dados NoSQL, MongoDB.

A aplicação Java é essencial pois tem como objetivo não só a transformação dos dados que serão processados diariamente, mas também dos dados que são atualizados em tempo-real. A escolha por uma base de dados MongoDB deve-se à capacidade de processar grandes volumes de dados e tráfego.

5.4.5 Serving

O componente *Serving* tem como objetivo realizar pesquisas nos dados armazenados na Base de Dados MongoDB. Este componente será desenvolvido através da tecnologia SpringBoot [21], devido à experiência do autor com a tecnologia, mas também pelo facto de ser a *framework* Java mais utilizada pela Celfocus.

[21] https://spring.io/projects/spring-boot

5.4.6 Aprendizagem Automática

O componente de Aprendizagem Automática corresponde ao componente onde será desenvolvido o modelo, utilizando um ambiente Jupyter Notebook [22], recorrendo à linguagem Python, utilizando bibliotecas sklearn que disponibilizam os algoritmos de aprendizagem automática necessários para prever o abandono dos clientes.

Neste componente é também disponibilizada uma API desenvolvida recorrendo à ferramenta Flask, onde será realizada a previsão para um ou mais clientes. Neste componente, também estará presente uma base de dados MongoDB para guardar os resultados das previsões efetuadas, permitindo, no futuro, aprimorar o modelo desenvolvido.

5.4.7 Apresentação

O componente de Apresentação tem como objetivo servir de interface gráfica do sistema Customer 360. Para o desenvolvimento deste componente será utilizada a tecnologia ReactJS. Esta tecnologia, desenvolvida em *Javascript*, foi implementada e é assegurada de uma forma *open-source* e foi escolhida com base em reuniões entre o autor e os especialistas na Celfocus.

5.4.8 API Gateway

Este componente tem como objetivo desacoplar as ligações internas entre componentes de Apresentação e a lógica implementada nos componentes de *Serving* e de Aprendizagem Automática. Permite redirecionar os pedidos para os componentes de uma forma mais direta, disponibilizando uma interface mais simples para os clientes interagirem.

Escolheu-se a tecnologia NodeJS [23] devido à facilidade em implementar pedidos REST a APIs externas, bem como a simples integração com o componente de Apresentação dado a existência de bibliotecas em comum.

5.5 Diagramas de Sequência

O presente capítulo expõe, através de diagramas de sequência, os requisitos funcionais capturados para este projeto.

5.5.1 Processar Registos em Tempo Real

A funcionalidade processar registos em tempo real inicia com o envio dos registos por parte dos sistemas externos para o componente de processamento de dados, que por sua vez, insere os registos num tópico Kafka. Os registos são mais tarde consumidos pelo componente de processamento de dados em tempo

[22] https://jupyter.org
[23] https://nodejs.org

real e enviados para o componente de processamento de dados históricos, através de um outro tópico Kafka.

Este fluxo encontra-se representado na Figura 29.

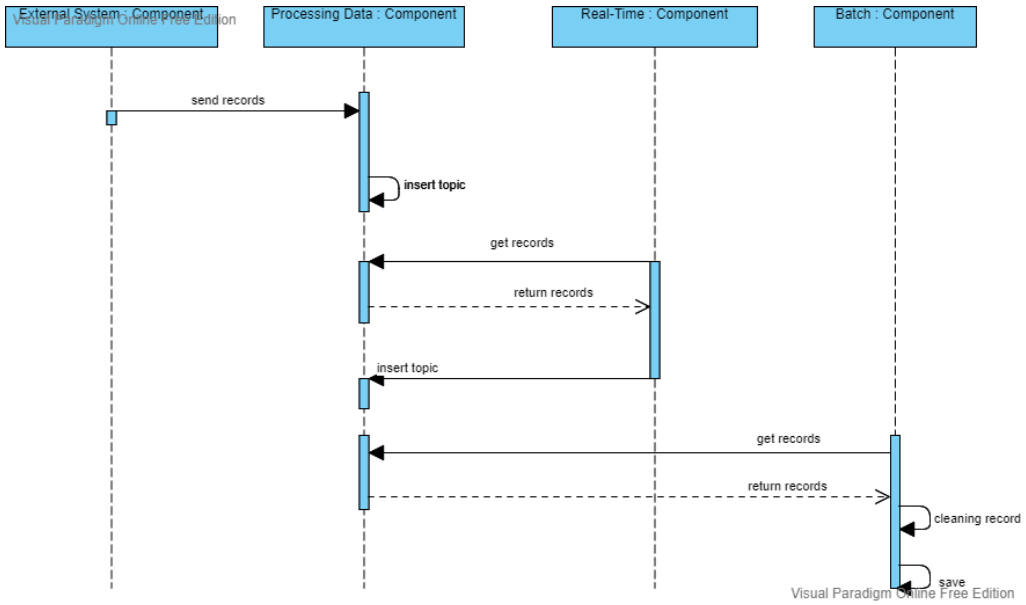


Figura 29 - Diagrama de Sequência da Funcionalidade Processar Registos em Tempo Real

5.5.2 Processar Registos de Forma Diária

Na Figura 30, encontra-se representado o desenho da funcionalidade processar registos de forma diária. O processo inicia pelo componente de processamento de dados solicitando os registos modificados no último dia, recebendo esses registos do sistema externo.

No momento em que obtém os registos, os mesmos são inseridos num tópico Kafka de forma a serem consumidos pelo componente de processamento de dados históricos, onde é realizado a limpeza e transformação dos dados e a inserção numa coleção MongoDB.

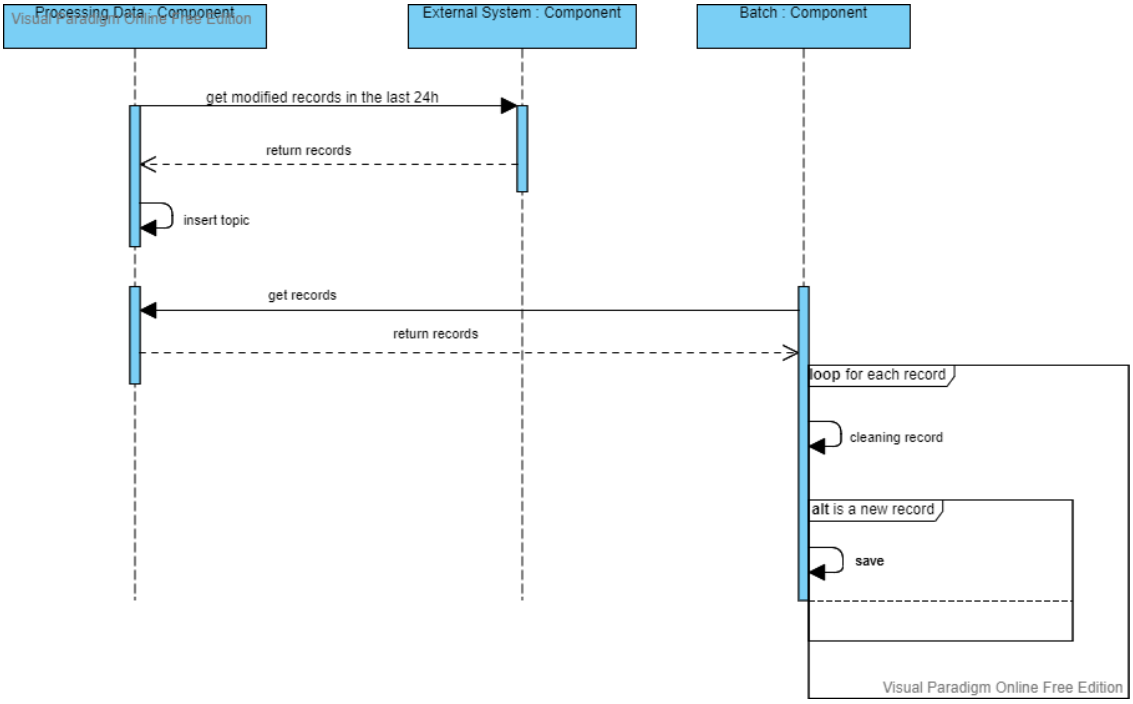


Figura 30 – Diagrama de Sequência da Funcionalidade Processar Registos de Forma Diária

5.5.3 Consultar Atributos que Contribuem para o Abandono

A funcionalidade consultar atributos que contribuem para o abandono inicia-se pelo Utilizador a solicitar a informação. O pedido é enviado para a API Gateway, que por sua vez remete para a aplicação Flask. Os atributos são enviados pelos mesmos componentes até ao Utilizador.

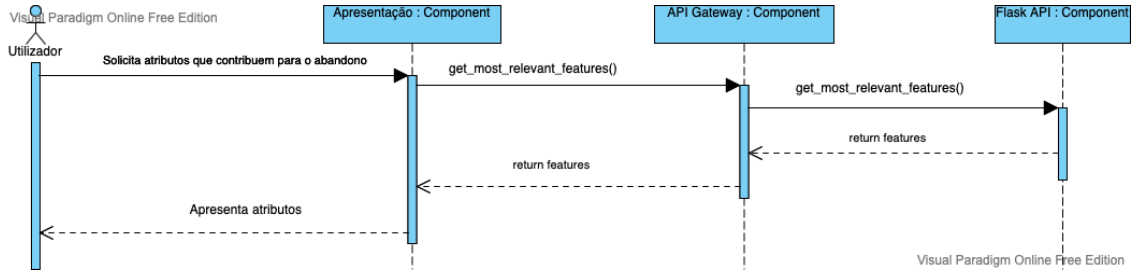


Figura 31 – Diagrama de sequência do requisito consultar atributos que contribuem para o abandono.

5.5.4 Consultar Dados dos Clientes

A funcionalidade consultar dados dos clientes é solicitada pelo utilizador, pesquisando pelas informações, através do ID do cliente. O pedido é enviado pela API *Gateway* até ao componente *Serving*, e este devolve a informação até chegar ao componente API *Gateway*. Aqui é validado se o cliente se encontra ativo e se for o caso, então, é solicitado à API Flask a previsão para o cliente. O pedido é recebido, a previsão realizada e guardada na base de dados. Depois, os dados completos são enviados até ao cliente.

Caso o cliente não esteja ativo, então não é feita a solicitação de previsão à API Flask.

O processo completo encontra-se ilustrado na Figura 32.

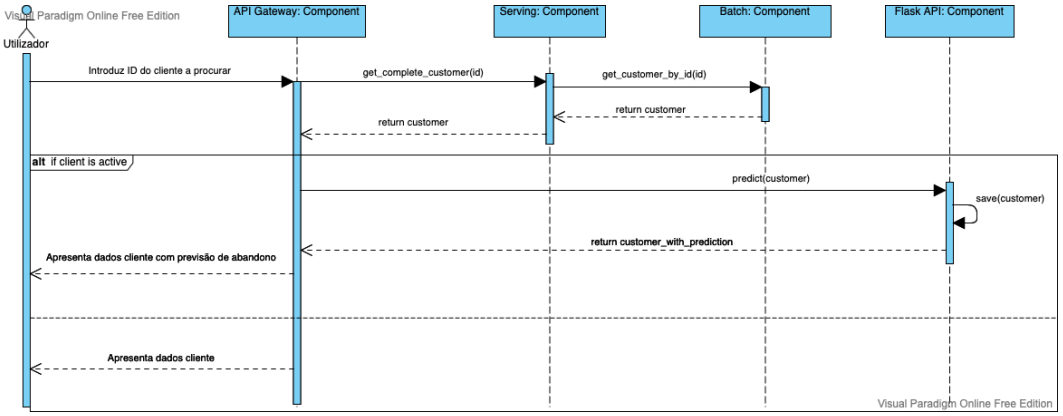


Figura 32 - Diagrama de Sequência da Funcionalidade Consultar Dados do Cliente

5.5.5 Consultar Clientes em Risco de Abandono

Na Figura 33, encontra-se representada a funcionalidade consultar clientes em risco de abandono. O processo inicia com a solicitação da informação pelo utilizador. O pedido é enviado para a API *Gateway* que envia para o componente *Serving*, e este solicita os clientes ativos ao componente de processamento de dados históricos. A informação é recebida e por cada cliente em risco, é adicionado à lista de clientes. No final, a lista é devolvida ao cliente.

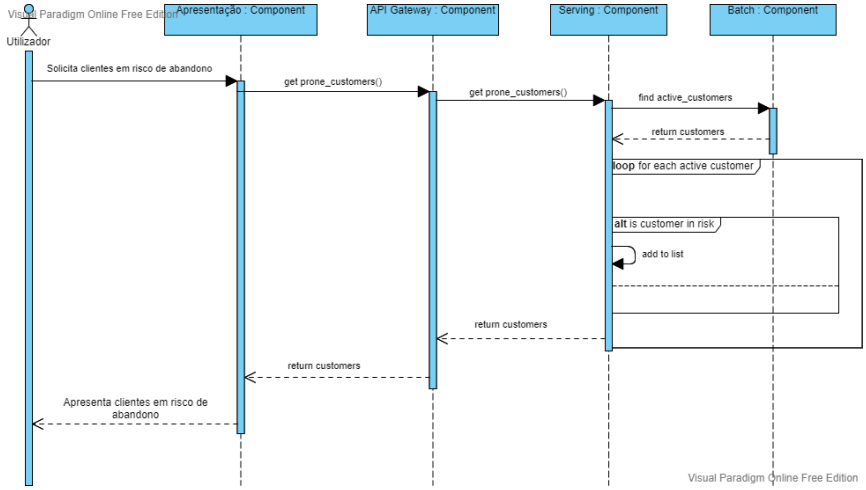


Figura 33 - Diagrama de Sequência da Funcionalidade Consultar Clientes em Risco de Abandono

6 Implementação da Solução

Neste capítulo são apresentados os dados e os passos utilizados para a implementação da solução proposta. Como foi referido anteriormente para o desenvolvimento deste projeto utilizaram-se técnicas de *Big Data* e de Aprendizagem Automática.

O primeiro subcapítulo introduz a descrição dos dados utilizados.

O subcapítulo seguinte descreve a implementação dos componentes que fazem parte da arquitetura escolhida.

O terceiro subcapítulo demonstra o processo elaborado para a criação de um modelo de aprendizagem automática com o objetivo de prever se um dado cliente abandonará a operadora.

Por fim, o último subcapítulo contextualiza sobre a implementação do componente de Apresentação.

6.1 Apresentação dos Dados Utilizados

Para o desenvolvimento deste projeto pretendeu-se utilizar dados reais pertencentes a clientes das telecomunicações. Esses dados seriam utilizados para o processamento e para a criação do modelo.

Tendo em conta a sensibilidade que estes tipos de dados têm para uma dada operadora e apesar das tentativas em obter dados reais, tal não foi possível.

A alternativa passou pela utilização de dados fictícios criados pela IBM. Os dados foram criados com o objetivo de prever o abandono de clientes, mas após uma análise mais aprofundada verificou-se que seriam também utilizados para a componente de processamento.

Os dados encontram-se divididos em três ficheiros, mais concretamente: dados demográficos, dados de serviços e dados de abandono dos clientes. Os ficheiros representam dados de 7043 clientes de uma operadora fictícia.

O objetivo da arquitetura é ser capaz de processar dados modificados de fontes externas em tempo real, em tempo agendado e armazenar os dados para, caso seja necessário, mais tarde serem consultados através de uma interface web.

Tendo em conta que não foi possível obter dados reais, nem efetuada nenhuma ligação a fontes de dados, foi utilizada uma base de dados PostgreSQL [24] para simular um local de armazenamento dos dados, semelhante a um existente numa infraestrutura tecnológica de uma operadora.

Partindo dos três ficheiros existentes, os registos de cada ficheiro foram introduzidos, manualmente, em tabelas SQL, visíveis através da representação do modelo relacional da Figura 34.

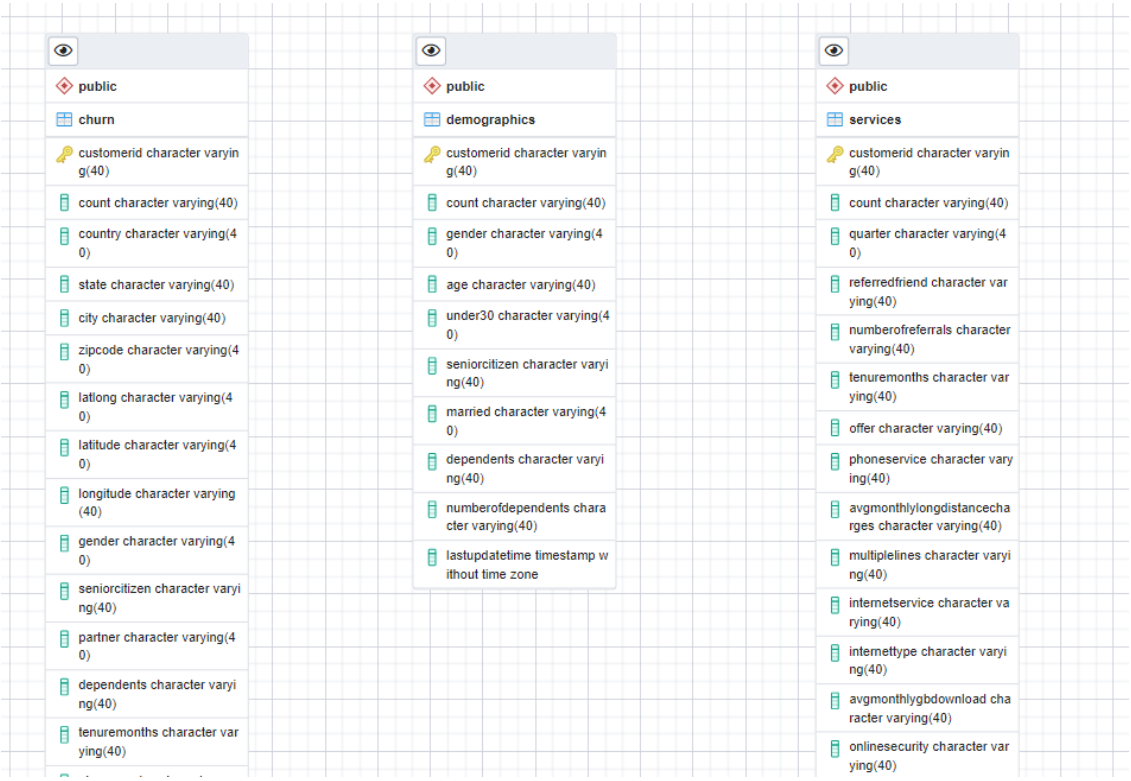


Figura 34 – Modelo Relacional

6.2 Implementação do Componente de Processamento de Dados em Tempo Real

Neste subcapítulo irá ser apresentado a implementação do componente de processamento dos dados em tempo real, através do processo descrito nos diagramas de sequência da Figura 29.

[24] https://www.postgresql.org/

Para esta implementação pretende-se que seja processado um determinado registo após a alteração de um campo no sistema a que pertence. Um exemplo para este projeto é a alteração do tarifário de um cliente no sistema de CRM, que deve ser refletida na base de dados MongoDB deste projeto.

Para isso, como foi referido anteriormente, estão a ser utilizadas três tabelas SQL de uma base de dados PostgreSQL que simula uma infraestrutura, de dimensão reduzida, de uma operadora.

Para que se cumprir esta funcionalidade é necessário a existência de um mecanismo para detetar o registo alterado e ser processado de seguida.

Após alguma pesquisa, foi possível adotar o comando NOTIFY, presente em base de dados PostgreSQL. Este comando foi utilizado num *trigger* para cada umas três tabelas existentes, para notificar, através de um canal estabelecido, a alteração de um registo juntamente com a respetiva data de modificação.

No Apache Nifi, utilizou-se um processador, ExecuteSQLRecord, onde no espaço para query é definido o comando "LISTEN <canal>" e a periocidade em que a consulta é realizada, neste caso imediata.

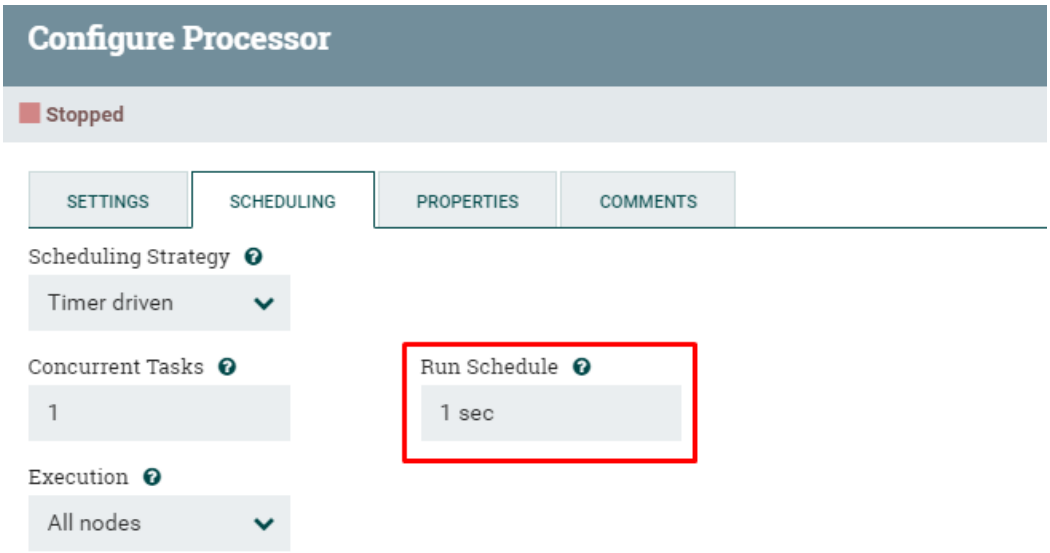


Figura 35 – Periocidade de Consulta dos Dados em Tempo-Real

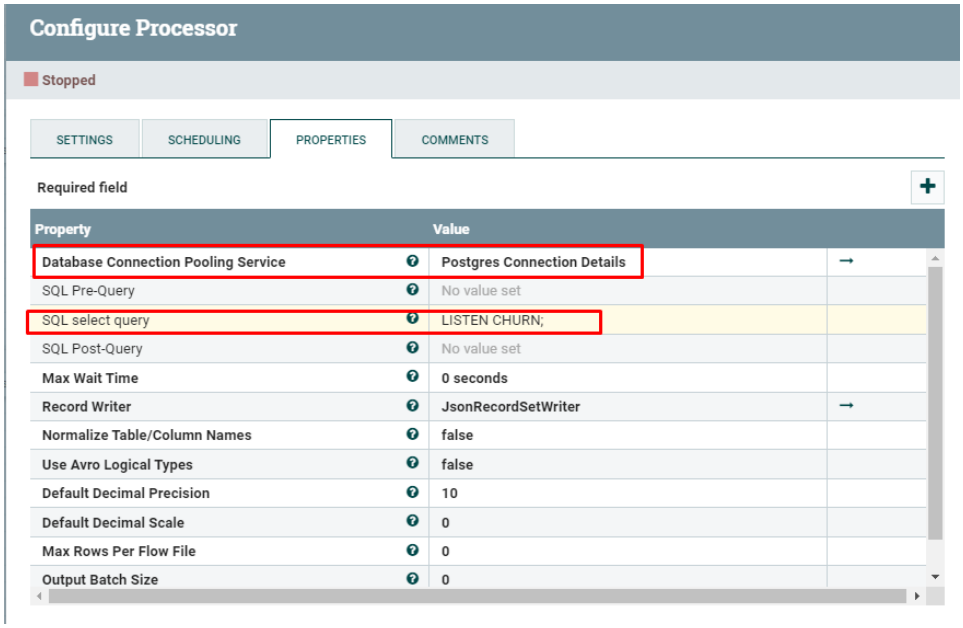


Figura 36 – Definição do mecanismo de leitura dos dados

Os dados processados são colocados, através de um produtor Kafka, num tópico para serem processados por uma aplicação Java e enviados para o componente de dados históricos. Após o registo ser consumido, compara-se a data de modificação do registo existente com o registo alterado e caso seja mais recente, o registo é atualizado.

6.3 Implementação do Componente de Processamento de Dados Históricos

Neste subcapítulo irá ser apresentado a implementação do componente de processamento dos dados históricos, através do processo descrito nos diagramas de sequência da Figura 30.

Para implementar este componente foi utilizada a tecnologia Apache Nifi. Como referido no capítulo 3.1.4, esta tecnologia permite fazer a ligação a ficheiros, base de dados para extrair, processar e transformar registos, através de processadores desenvolvidos pela comunidade de desenvolvedores. Adicionalmente permite inserir os ficheiros noutras fontes de armazenamento.

Havendo como requisito o processamento dos registos de forma diária, foi necessário utilizar o processador ExecuteSQLRecord, para realizar a ligação à base de dados PostgreSQL e extrair os registos alterados num espaço de um dia. Tendo em conta, a existência de três tabelas às quais queremos obter informação, utilizou-se três processadores onde cada um consulta os dados de uma tabela.

O processador ExecuteSQLRecord permite especificar o período – em segundos - pelo qual serão extraídos os dados e utilizar uma ligação a uma base de dados com uma query SQL especifica.

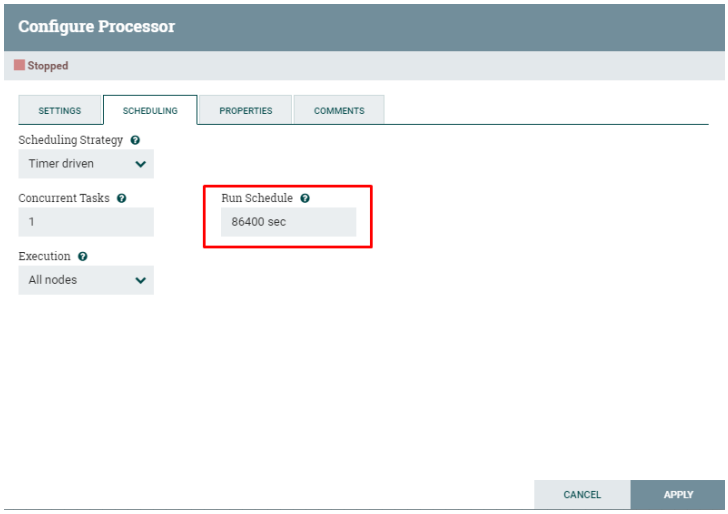


Figura 37 – Definição do período de *polling*

Na Figura 37 é possível verificar a especificação de 86400 segundos, o que corresponde a um dia. Este processador recolhe dados de uma forma diária, através de uma query pré-definida.

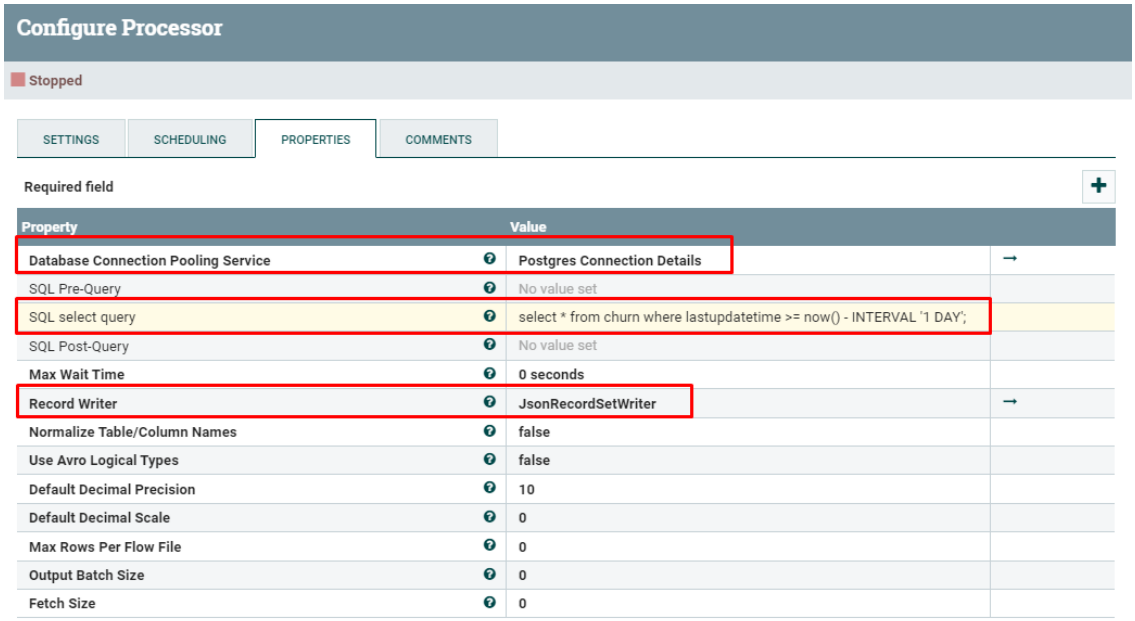


Figura 38 – Configurações do processador de polling dos dados

Na Figura 38, encontra-se a definição do serviço de base de dados utilizado, a *query* utilizada para extrair os dados e o tipo de escritor utilizado para recolher os dados.

Depois, de forma iterativa, por cada registo recolhido são convertidos num formato JSON (*JavaScript Object Notation*). Este passo torna mais simples a leitura que será realizada posteriormente.

O passo final passa pelo envio dos dados para um produtor Kafka que inseriu num tópico. A Figura 39 ilustra as configurações necessárias para a escrita dos dados num tópico de *Kafka*.

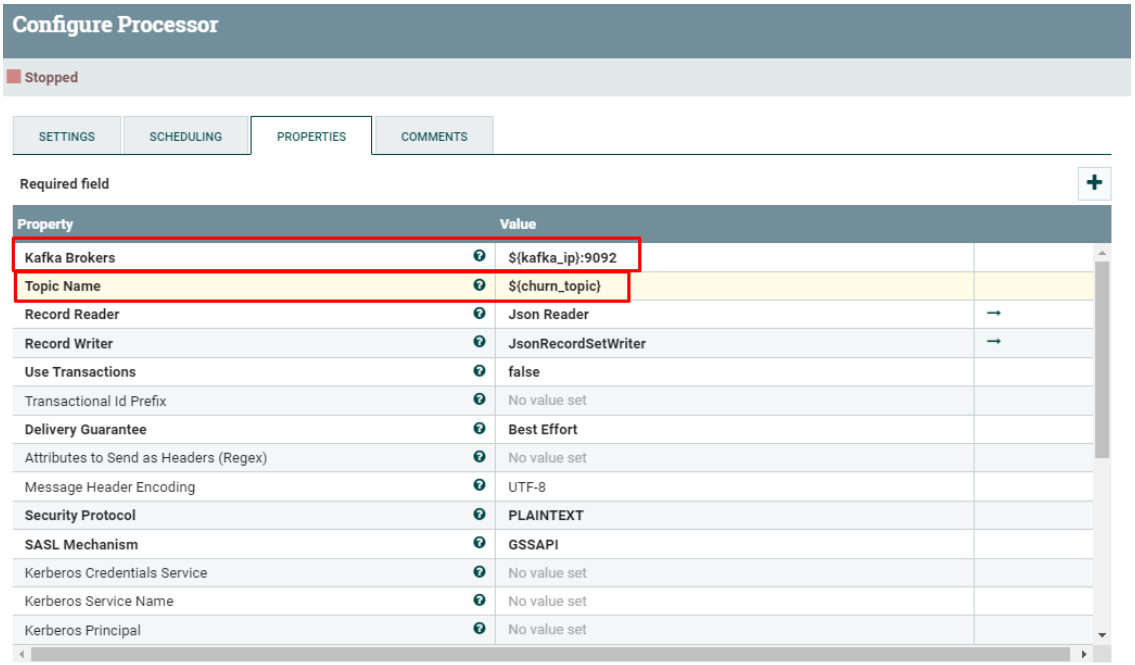


Figura 39 – Configuração da escrita dos dados num tópico de *Kafka*

Para este passo, especifica-se a instância do Kafka e em que tópico queremos inserir os dados.

O processo completo do Apache Nifi encontra-se representado na Figura 40.

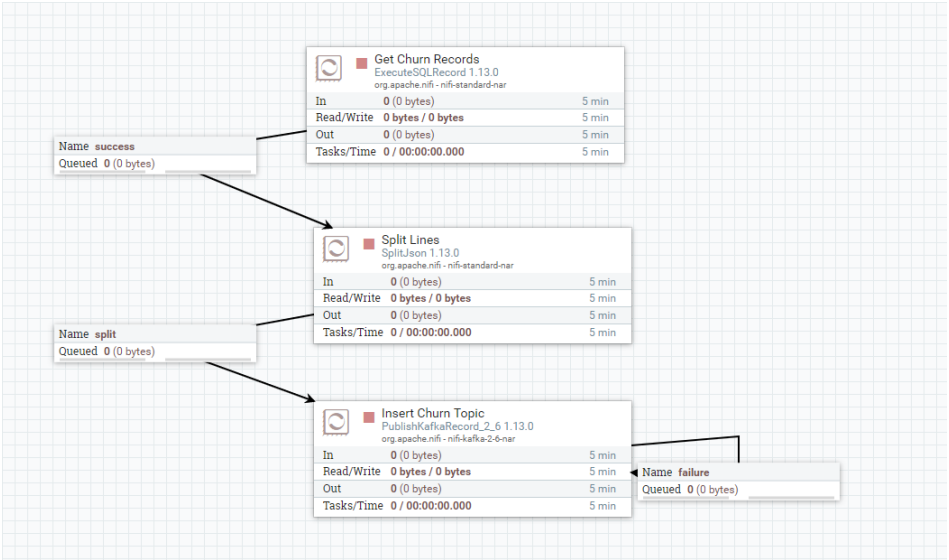


Figura 40 – Fluxo de processamento dos dados históricos

Após os dados estarem inseridos no tópico, a aplicação Java utiliza um consumidor Kafka, retirando os registos em formato JSON, convertendo para um objeto Java, realizando operações de limpeza e transformação dos dados, para serem inseridos numa coleção MongoDB.

6.4 Implementação do Componente *Serving*

Como indicado, o componente *Serving* utiliza a tecnologia SpringBoot para consultar os dados dos clientes armazenados no componente dos dados históricos, mais precisamente na base de dados MongoDB.

Para a construção deste componente, foi utilizado o padrão de design arquitetural MVC. Este padrão delimita as responsabilidades da aplicação em camadas de dados (Modelo), de disponibilização da informação (Vista) e de controlo da informação (Controlador).

Para o acesso à consulta dos dados na base de dados, utilizou-se o padrão Repositório (*Repository*). Este padrão abstrai a lógica de negócio presente na camada Modelo e o acesso a fontes externos, neste caso apenas a base de dados.

A camada de disponibilização da informação utiliza objetos DTO (Objeto de Transferência de Dados, *Data Transfer Object*) para serem enviados na resposta aos pedidos efetuados.

Este componente disponibiliza dois tipos de acessíveis através de pedidos REST, descritas na Tabela 19.

Tabela 19 - Funcionalidades disponíveis no Componente *Serving*

Descrição	**Método**	**URL**
Consultar dados de um cliente	GET	{URL_BASE}/*customers*/{id}
Consultar clientes mais propensos a abandonar	GET	{URL_BASE}/*prone_customers*

6.5 Implementação da API *Gateway*

A API *Gateway* tem como objetivo servir a camada de apresentação das consultas realizadas à camada *Serving* e de Aprendizagem Automática. Esta camada, desenvolvida em NodeJS, disponibiliza três métodos, enumerados na Tabela 20.

Tabela 20 - Métodos disponibilizados pela API *Gateway*

Descrição	**Método**	**URL**
Consultar dados de um cliente com previsão de abandono	GET	{URL_BASE}/*customer_with_prediction*/{id}
Consultar clientes mais propensos a abandonar	GET	{URL_BASE}/*prone_customers*
Consultar atributos que contribuem para abandono	GET	{URL_BASE}/*features_importance*

6.6 Desenvolvimento do Modelo

Neste subcapítulo irá ser apresentado o processo de desenvolvimento do modelo de aprendizagem automática, nomeadamente as etapas que levaram à construção do modelo e concluindo com os resultados obtidos.

6.6.1 Entendimento do Problema

Através da análise efetuada no capítulo 2 e da experiência do autor, no contexto das telecomunicações, concluiu-se que a componente de aprendizagem automática deveria incidir no desenvolvimento de um modelo que permitisse prever o abandono dos clientes, através das informações dos mesmos.

O principal objetivo passa por construir uma API que disponibilize, através de uma interface REST, um serviço de previsão de abandono de um ou mais clientes a partir de um modelo de aprendizagem automática.

6.6.2 Análise Exploratória dos Dados

De seguida, na fase de entendimento dos dados, começou-se por recolher os dados disponibilizados pela IBM.

Como foi referido no subcapítulo 5.1, foram recolhidos três ficheiros. O primeiro passo foi analisar os dados provenientes de cada ficheiro.

Na Figura 41 estão representados os dados demográficos.

	Customer ID	Count	Gender	Age	Under 30	Senior Citizen	Married	Dependents	Number of Dependents
0	8779-QRDMV	1	Male	78	No	Yes	No	No	0
1	7495-OOKFY	1	Female	74	No	Yes	Yes	Yes	1
2	1658-BYGOY	1	Male	71	No	Yes	No	Yes	3
3	4598-XLKNJ	1	Female	78	No	Yes	Yes	Yes	1
4	4846-WHAFZ	1	Female	80	No	Yes	Yes	Yes	1

Figura 41 – Dados demográficos

Na Figura 42, estão representados os dados dos serviços que os clientes possuem.

Avg Monthly Long Distance Charges	Multiple Lines	Internet Service	Internet Type	Avg Monthly GB Download	...	Contract	Paperless Billing	Payment Method	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue	CustomerID
0.00	No	Yes	DSL	8	...	Month-to-Month	Yes	Bank Withdrawal	39.65	39.65	0.00	20	0.00	59.65	8779-QRDMV
48.85	Yes	Yes	Fiber Optic	17	...	Month-to-Month	Yes	Credit Card	80.65	633.30	0.00	0	390.80	1024.10	7495-OOKFY
11.33	Yes	Yes	Fiber Optic	52	...	Month-to-Month	Yes	Bank Withdrawal	95.45	1752.55	45.61	0	203.94	1910.88	1658-BYGOY
19.76	No	Yes	Fiber Optic	12	...	Month-to-Month	Yes	Bank Withdrawal	98.50	2514.50	13.43	0	494.00	2995.07	4598-XLKNJ
6.33	Yes	Yes	Fiber Optic	14	...	Month-to-Month	Yes	Bank Withdrawal	76.50	2868.15	0.00	0	234.21	3102.36	4846-WHAFZ

Figura 42 - Dados de Serviços

E por fim, na Figura 43, uma representação dos dados que apresentam os clientes que abandonaram, que permanecem e outras informações adicionais.

City	Zip Code	Lat Long	Latitude	Longitude	Gender	...	Contract	Paperless Billing	Payment Method	Monthly Charges	Total Charges	Churn Label	Churn Value
Los Angeles	90003	33.964131, -118.272783	33.964131	-118.272783	Male	...	Month-to-month	Yes	Mailed check	53.85	108.15	Yes	1
Los Angeles	90005	34.059281, -118.30742	34.059281	-118.307420	Female	...	Month-to-month	Yes	Electronic check	70.70	151.65	Yes	1
Los Angeles	90006	34.048013, -118.293953	34.048013	-118.293953	Female	...	Month-to-month	Yes	Electronic check	99.65	820.5	Yes	1
Los Angeles	90010	34.062125, -118.315709	34.062125	-118.315709	Female	...	Month-to-month	Yes	Electronic check	104.80	3046.05	Yes	1
Los Angeles	90015	34.039224, -118.266293	34.039224	-118.266293	Male	...	Month-to-month	Yes	Bank transfer (automatic)	103.70	5036.3	Yes	1

Figura 43 - Dados de Abandono

Através desta análise, percebeu-se que os três ficheiros podiam ser integrados num só de forma a facilitar o processo, através do atributo-chave *CustomerID,* removendo colunas duplicadas e preservando os valores.

Depois de criado este ficheiro e através da explicação de cada coluna, disponibilizada pela IBM, foram removidas algumas colunas que não apresentavam valor e foram convertidos atributos com tipo de dados texto para numérico.

Iniciou-se o processo de análise gráfica aos valores numéricos e categóricos, através de histogramas, *boxplots* e gráficos de barras, começando pela distribuição de clientes que abandonaram e que se mantêm na operadora, representados na Figura 44.

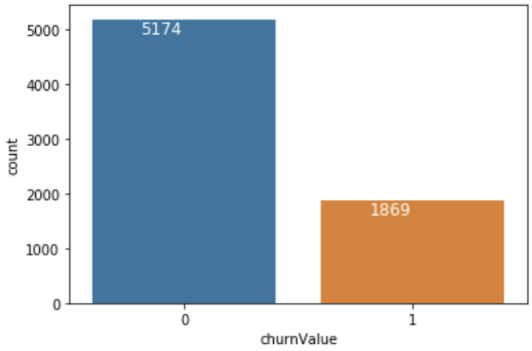


Figura 44 - Distribuição de Abandono

Na Figura 44, está representada a quantidade de clientes que permanecem na operadora, representados por '0' e clientes que abandonaram, representados por '1'.

Verifica-se, ainda, que não existe um desequilíbrio significativo nos valores de abandono e não abandono, não sendo necessário aplicar técnicas para equilibrar os dados, como por exemplo sobreamostragem *(oversampling),* sobamostragem *(undersampling)* ou SMOTE (*Synthetic Minority Over-Sampling Technique)*.

De seguida, a análise prosseguiu com os restantes atributos centrando a análise na distribuição de abandono.

Uma das questões mais relevantes nas operadoras prende-se com os clientes que possuem contrato mês-a-mês sem fidelização. Pela ausência de fidelização, torna-se mais fácil para os clientes abandonarem ao contrário dos clientes que possuem contratos de um ou de dois anos.

Os dados que foram utilizados comprovam uma maior taxa de abandono em contratos mês-a-mês, na Figura 45.

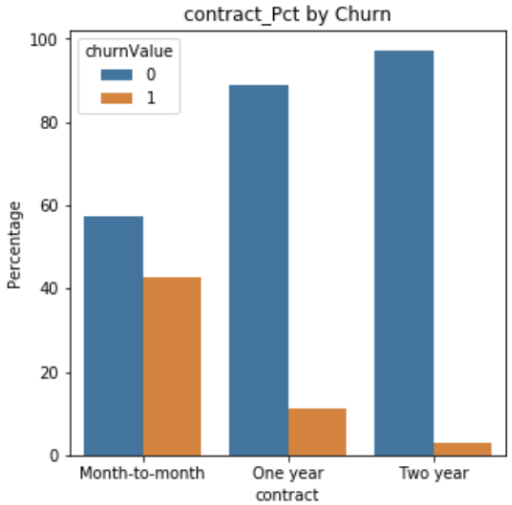


Figura 45 - Distribuição de Abandono por Tipo de Contrato

Altos valores de pagamento mensal, estão por norma, também associados a taxas de abandono mais elevadas comparativamente com valores mais baixos.

Na Figura 46, é possível verificar que os clientes que abandonaram apresentavam valores da mediana de cerca de 83 dólares. Por outro lado, os clientes que permanecem na operadora, apresentam um valor da mediana, de pouco mais de 60 dólares.

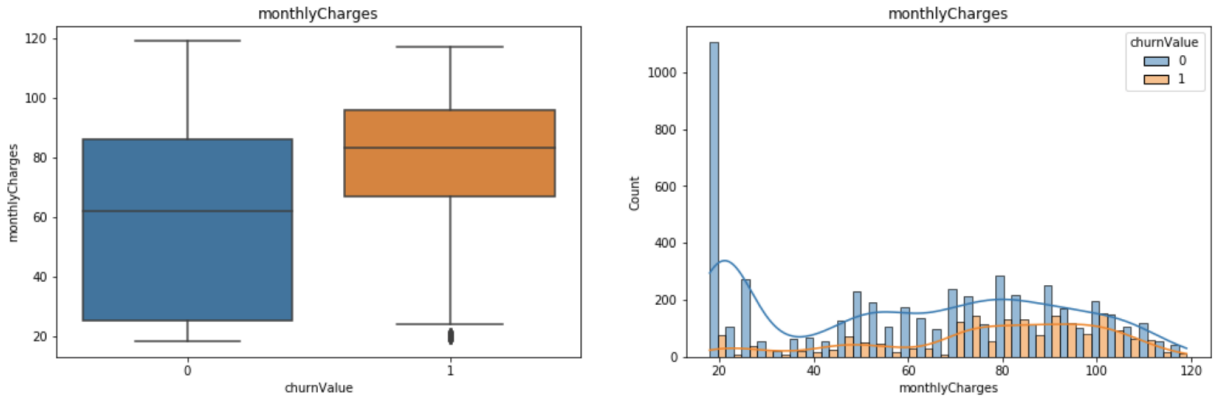


Figura 46 – Distribuição dos Valores de Pagamento Mensal

Estes valores explicam-se pela alta competitividade existente no sector e na procura dos clientes por uma alternativa mais barata, sem perda de qualidade, à que possuem.

Uma das descobertas mais interessantes nesta análise foi a relação da idade com o abandono.

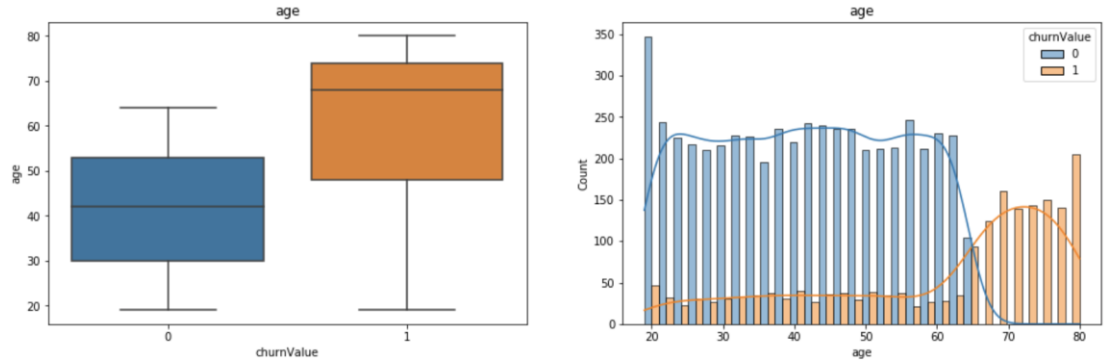


Figura 47 - Relação da Idade com o Abandono

Pela análise da Figura 47, verifica-se que não existem clientes com mais de 65 anos que tenham abandonado a operadora e entre os clientes que abandonam, a grande maioria encontra-se entre os 50 e 75 anos.

6.6.3 Preparação dos Dados

A análise prosseguiu com a preparação dos dados para mais tarde serem utilizados pelos algoritmos de aprendizagem automática.

Numa fase inicial, foram convertidos os valores categóricos em valores numéricos, e para isso utilizou-se a função *get_dummies* da biblioteca Pandas. Esta função cria novos atributos numéricos em função dos

valores categóricos existentes e atribui um valor de ‘0’ caso o atributo não esteja presente e ‘1’ caso o atributo esteja presente.

Na Figura 48, é possível verificar o exemplo para o atributo contrato, antes da aplicação da função *get_dummies*.

	contract
0	Month-to-month
1	Month-to-month
2	Month-to-month
3	Month-to-month
4	Month-to-month

Figura 48 - Valores do Contrato Antes da Aplicação da Função *get_dummies*

E na Figura 49, após a aplicação da função *get_dummies*.

	contract_Month-to-month	contract_One year	contract_Two year
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

Figura 49 - Valores do Contrato Após Aplicação da Função *get_dummies*

Este procedimento foi aplicado a todos os atributos categóricos, mas também ao atributo da idade que inicialmente caracterizava-se como numérico. No caso da idade, foram transformados os valores numéricos em valores categóricos, foram segmentados em intervalos de valores e foi aplicada, novamente, a função *get_dummies*, obtendo o resultado representado na Figura 50.

	Age_Label_33-46	Age_Label_47-60	Age_Label_61-80	Age_Label_Under32
0	0	0	1	0
1	0	0	1	0
2	0	0	1	0
3	0	0	1	0
4	0	0	1	0

Figura 50 - Aplicação da Função *get_dummies* no Atributo Idade

Após esta transformação foram selecionados dos atributos mais relevantes através da correlação de Pearson. Para isso, foram selecionados os atributos com valores absolutos superiores a 0.25, em relação à variável de abandono ou de permanência.

A força da correlação de cada atributo encontra-se representada pelo mapa de calor da Figura 51.

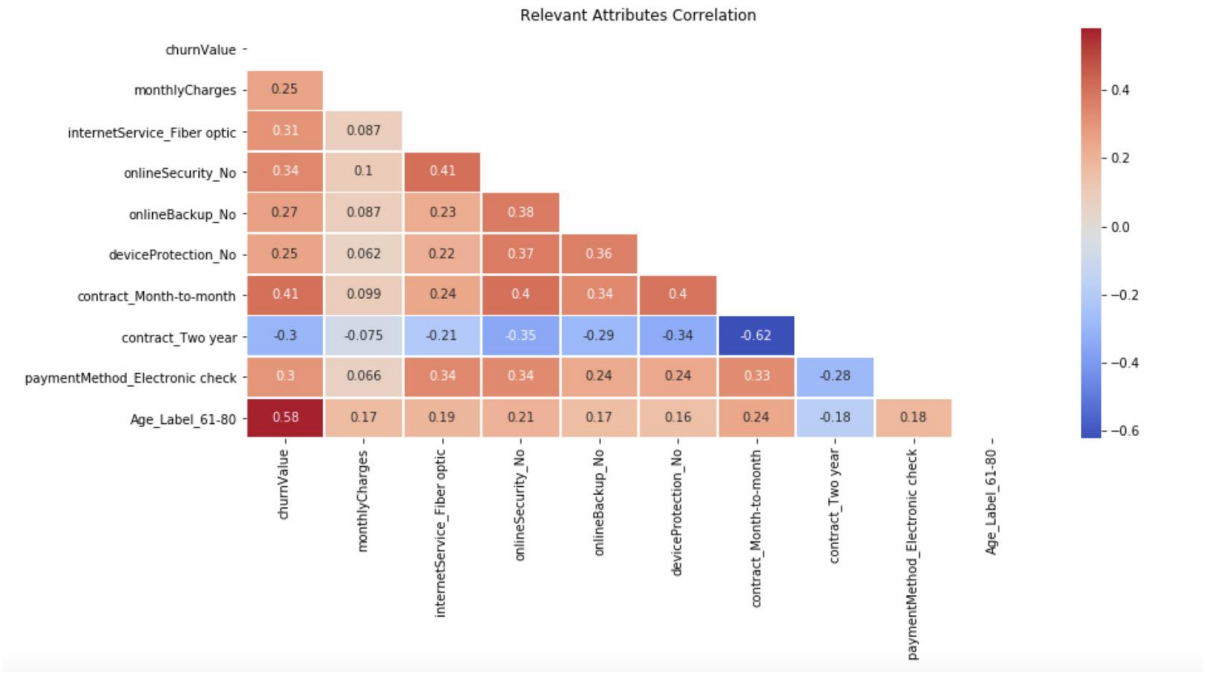


Figura 51 - Mapa de Calor dos Atributos Mais Relevantes

Na Tabela 21, encontra-se uma descrição dos nove campos utilizados.

Tabela 21 - Descrição dos atributos selecionados

Atributo	**Descrição**
Age Label 61-80	Cliente possui idade entre 61 e 80, inclusive
Payment Method: Electronic Check	Utiliza cheque eletrónico como método de pagamento
Contract: Two Year	Possui um contrato de dois anos
Contract: Month-to-Month	Possui um contrato mês a mês
Device Protection: No	Não possui proteção adicional para os dispositivos
Online Backup: No	Não possui *backups* adicionais
Online Security: No	Não possui serviços de segurança online adicionais
Internet Service: Fiber Optic	Serviço de internet é disponibilizado por fibra ótica
Monthly Charges	Valor a pagar mensalmente pelos serviços subscritos

Foram estes atributos utilizados no desenvolvimento de um modelo de aprendizagem automática. Com base na literatura estudada no capítulo 3.3 e tratando-se de um problema de classificação, foram selecionados os algoritmos *Random Forest, Support Vector Machine* e *Logistic Regression*.

6.6.4 Modelação e Avaliação

Neste passo, foram realizadas três etapas para escolher o algoritmo que apresentava melhores resultados em função dos dados existentes. Um passo partilhado por todas as etapas descritas de seguidas foi a normalização dos dados.

A primeira etapa foi dividir os dados em conjuntos de treino e de teste, representando 70% e 30% dos dados, respetivamente.

Esta técnica foi aplicada em cada algoritmo, utilizando os parâmetros por defeito, e de seguida são apresentados os resultados para a exatidão:

Tabela 22 - Exatidão dos Algoritmos com Parâmetros por Defeito

Algoritmo	Exatidão
Random Forest	83,2
Support Vector Machine	85,9
Logistic Regression	85,2

Apesar dos bons resultados obtidos pelos três algoritmos, esta técnica apresenta algumas desvantagens. Dado que o conjunto de treino e de teste é aleatoriamente selecionado, existe a possibilidade de determinados *clusters* dentro do *dataset* fiquem exclusivamente no conjunto de treino e outros *clusters* fiquem no conjunto de teste, traduzindo-se num viés inevitável (Gunasegaran & Cheah, 2017).

Devido a esse ponto negativo, na próxima etapa utilizou-se a técnica de validação cruzada. Nesta técnica os dados de entrada são fragmentados em vários subconjuntos e são realizadas várias iterações (*folds*). Em cada iteração um dos conjuntos é utilizado para teste e os restantes como treino.

O processo é concluído até todos os conjuntos terem sido utilizados como dados de teste. Na Figura 52, encontra-se representado este processo.

Figura 52 - Validação Cruzada

Para cada um dos três algoritmos, com parâmetros por defeito, foi utilizado a técnica de validação cruzada com 10 *folds*, tendo-se obtido, na Tabela 23, os resultados relativos às métricas estudadas anteriormente.

Tabela 23 - Resultados com Validação Cruzada

Algoritmo	Exatidão	Precisão	*Recall*	F1	ROC-AUC
Random Forest	83,4	80,5	58,7	67,9	77,6
Support Vector Machine	86,5	82,2	60,6	70,6	77,9
Logistic Regression	86,2	80,3	60,0	71,0	78,0

Estes resultados, na Tabela 23, comprovam uma melhoria dos resultados comparativamente à técnica de seleção dos conjuntos de teste e de treino aleatória.

Apesar das melhorias nos resultados, foi utilizada uma outra técnica denominada *GridSearchCV*. Esta técnica tem como objetivo determinar os parâmetros otimizados para cada algoritmo em função dos dados existentes e encontra esses parâmetros utilizando validação cruzada.

Na Tabela 24, encontra-se o resultado da otimização para o algoritmo *Support Vector Machine*.

Tabela 24 – Parâmetros otimizados do algoritmo *Support Vector Machine*

Parâmetro	Descrição	Valor
Kernel	Tipo de Kernel a ser utilizado (linear, polinomial, sigmoide, base radial)	rbf
C	Parâmetro de penalização. O objetivo é regular o sobre ajustamento dos dados.	100
Gamma	Coeficiente do *Kernel*. Um valor baixo coloca os limites da 'curva' de decisão baixo levando que a região de decisão muito abrangente. Um valor alto coloca os limites da 'curva' de decisão altos, criando 'ilhas' de decisão à volta dos pontos a serem avaliados.	0.01

Na Tabela 25, encontra-se o resultado da otimização para o algoritmo *Logistic Regression*.

Tabela 25 - Parâmetros otimizados do algoritmo *Logistic Regression*

Parâmetro	Descrição	Valor
Solver	Tipo de algoritmo a ser utilizado para otimizar (*liblinear, newton-cg, lbfgs, saga, sag*).	*liblinear*
C	Parâmetro de penalização. O objetivo é regular o sobre ajustamento dos dados.	0.033
Penalty	Parâmetro utilizado para especificar a norma a ser utilizada de penalização.	L1

Na Tabela 26, encontra-se o resultado da otimização para o algoritmo *Random Forest*.

Tabela 26 - Parâmetros otimizados do algoritmo *Random Forest*

Parâmetro	Descrição	Valor
Criterion	Função utilizada para medir a qualidade de uma divisão	*entropy*
MaxDepth	Número máximo de níveis em cada árvore de decisão	0.033
MaxFeatures	Número de características a serem utilização em divisão de cada nó	auto
Estimators	Número de árvores a serem utilizadas	200

Foi assim avaliado o resultado de cada algoritmo utilizando os parâmetros definidos anteriormente, verificando-se, pela análise da Tabela 27, que os três algoritmos apresentam melhorias consecutivas nos resultados.

Tabela 27 - Resultados com *GridSearchCV*

Algoritmo	Exatidão	Precisão	*Recall*	F1	ROC-AUC
Random Forest	86,6	86,0	65,9	70,1	91,1
Support Vector Machine	86,6	83,1	60,0	70,0	86,4
Logistic Regression	86,3	85,1	64,4	70,7	87,7

O modelo que utiliza o algoritmo *Random Forest* destaca-se pelo melhor resultado em todas as métricas avaliadas, sendo por isso escolhido como modelo final.

Para que o modelo final seja utilizado pelos utilizadores, foi necessário exportar externamente para ficheiro. E para isso recorreu-se à biblioteca *pickle*.

Como foi referido anteriormente, os dados que foram utilizados para o desenvolvimento do modelo passaram por um processo de seleção de atributos, sendo por isso também necessária a exportação dos atributos selecionados no desenvolvimento.

6.6.5 Implantação

Após a conclusão do desenvolvimento do modelo, a etapa seguinte passa por disponibilizar o modelo aos utilizadores. Para isso, foi desenvolvida uma aplicação Flask.

Esta aplicação tem como principal funcionalidade a exposição de métodos, desenvolvidos em linguagem Python, através de pedidos REST. Escolheu-se esta tecnologia tendo em conta a sua facilidade de integração com o modelo desenvolvido em linguagem Python.

Esta aplicação disponibiliza dois métodos, listados na Tabela 28.

Tabela 28 – Métodos disponibilizados pela aplicação Flask

Descrição	Método	URL
Realizar previsão	POST	{URL_BASE}/*predict*
Consultar atributos que contribuem para o abandono	GET	{URL_BASE}/*features_importance*

6.7 Implementação do Componente de Apresentação

Tal como referenciado no capítulo 5.4.7, o componente de Apresentação foi desenvolvido utilizando ReactJS.

O ReactJS permite a construção da interface gráfica de duas formas, utilizando *functions* ou *React.Components*.

No caso deste projeto foram utilizados *functions* para criar três módulos principais:

- *Navbar*
- *ProfilePage*
- *Dashboard*

O módulo *Navbar*, visível na Figura 53, tem como objetivo disponibilizar uma barra de navegação comum às duas páginas principais da aplicação. Neste componente está presente uma caixa de texto que permite pesquisar por um cliente e uma hiperligação que permite redirecionar para a página de *Dashboard*.

Figura 53 – Módulo *Navbar*

ProfilePage é o módulo que disponibiliza a informação dos dados do cliente após a pesquisa do utilizador.

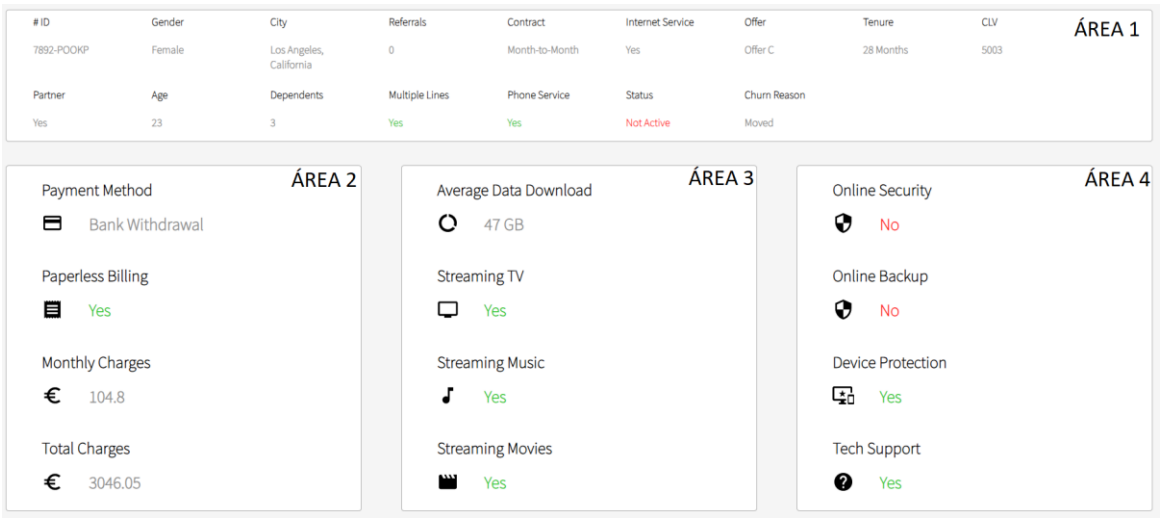


Figura 54 – Módulo ProfilePage

Este módulo encontra-se dividido em quatro áreas distintas, exibidas na Figura 54. A área 1 representa os dados demográficos, serviços e ofertas subscritas pelo cliente. Na área 2 encontra-se a informação relativa a dados financeiros do cliente. Na área 3 encontram-se os consumos móveis utilizados pelo cliente consoante o tipo de serviços que tenham subscrito. E por fim, na área 4 encontram-se a informação dos tipos de serviços de proteção.

O módulo de *Dashboard* contêm os dez atributos que contribuem mais para o abandono de clientes e os clientes em risco de abandonarem em função de cada atributo, representado na Figura 55.

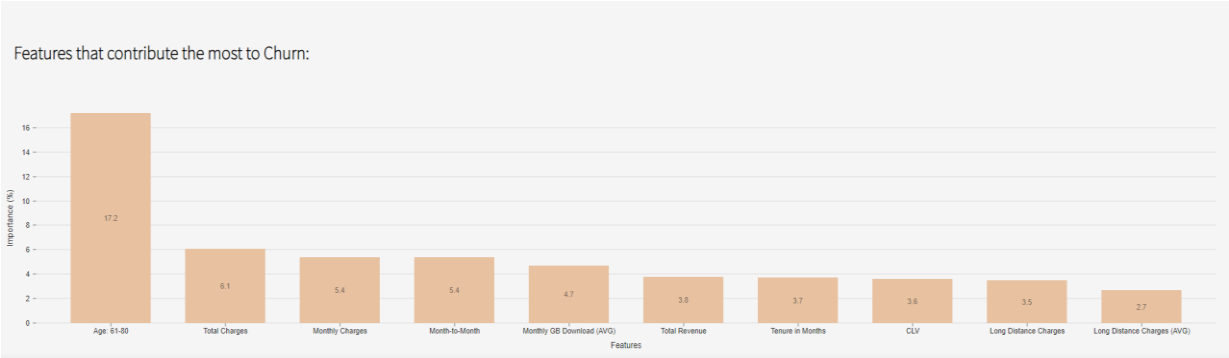


Figura 55 – Atributos que contribuem para o abandono relativo ao módulo de *Dashboard*

Ainda neste módulo e para cada atributo que contribui para o abandono, foram determinados quais os clientes que em função de um atributo selecionável pelo utilizador estão em risco de abandonar.

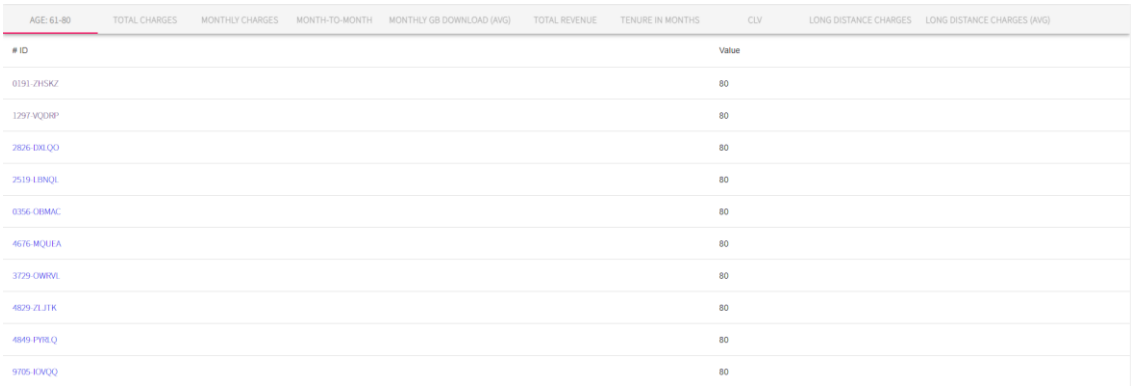


Figura 56 – Clientes mais propensos a abandonar

Neste caso, na Figura 56, são apresentados os dez clientes mais velhos no intervalo [60,81], tendo em conta que este intervalo foi identificado como contribuindo, mais do que qualquer um, para o abandono.

7 Conclusões

Esta tese apresenta uma proposta de solução centrada no cliente, denominada *Customer* 360, com o objetivo de disponibilizar às operadoras um sistema que agrega os dados presentes nos vários sistemas bem como uma proposta para melhorar um dos indicadores mais importantes para uma operadora de telecomunicações, o abandono de clientes.

No desenvolvimento desta tese, foram estudados os sistemas que compõem a infraestrutura tecnológica de uma operadora de telecomunicações, com o objetivo de recolher a informação mais relevante de cada um desses.

Para além disso, foram estudados os indicadores estratégicos, propostos pelas várias operadoras, com o objetivo de a partir dos dados dos sistemas, definir estratégias para os melhorar.

Foram recolhidas, através de artigos científicos, as metodologias e técnicas de Aprendizagem Automática e de *Big Data* utilizadas para melhorar alguns dos problemas presentes nas operadoras.

Após a análise da literatura foi mapeado os indicadores estratégicos, mais centrados no cliente e nas interações que o cliente realiza para com a operadora, e as técnicas que foram analisadas na literatura. Isto permite determinar, perante os principais problemas de uma operadora, as técnicas a adotar para mitigar o problema.

Tendo em conta a dimensão dos sistemas e a potencial informação existente, foram estudadas arquiteturas de *Big Data* existentes (*Lambda* e *Kappa*) que suportassem os objetivos para este projeto. Após levantamento dos requisitos, foi criada uma proposta de arquitetura baseada na arquitetura *Lambda*.

De forma a validar uma solução *Customer 360*, foram implementados os componentes propostos para esta solução e através de dados criados com o objetivo de prever o abandono de clientes, conseguiu-se implementar uma solução baseada na arquitetura considerada.

Para além disso, foi implementado o módulo que prevê o abandono de um cliente, através da metodologia CRISP-DM, obtendo 86,65% de exatidão, superiorizando-se a alguns resultados da literatura estudada.

Por fim, foi construído, também, uma interface gráfica que disponibiliza a informação que foi agregada dos sistemas externas, permitindo a operadores e a outros interessados, visualizar a informação completa correspondente a um dado cliente.

Em suma, considera-se que os objetivos propostos foram cumpridos na totalidade e com resultados satisfatórios. Foi compreendido melhor o negócio das telecomunicações e o potencial que as tecnologias de *Big Data* e de Aprendizagem Automática tem na melhoria no negócio das operadoras de telecomunicações.

7.1 Limitações e Trabalho Futuro

Ao longo do desenvolvimento desta tese, foi-se percebendo a complexidade e antiguidade existente dos sistemas externos, dificultando a extração dos dados existentes nesses sistemas. Apesar dos bons resultados obtidos com o modelo de Aprendizagem Automática não é possível extrapolar conclusões mais aprofundadas dada a impossibilidade de utilização de dados reais. Isto, sem dúvida, que daria uma qualidade superior ao trabalho realizado, pois aumentaria a qualidade de um modelo com dados reais e compostos por mais informações do que as presentes nos dados que foram utilizados.

No entanto, no desenvolvimento do componente de Aprendizagem Automática recorreu-se a uma das metodologias mais reconhecidas para projetos deste tipo, podendo, no futuro, utilizar as mesmas etapas para uma implementação com dados reais.

Referências

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6

Al-Molhem, N. R., Rahal, Y., & Dakkak, M. (2019). Social network analysis in Telecom data. *Journal of Big Data*, *6*(1), 99. https://doi.org/10.1186/s40537-019-0264-6

ASCI. (2020). *Benchmarks by Sector*. https://www.theacsi.org/acsi-benchmarks/benchmarks-by-sector

Azevedo, A., & Santos, M. F. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. 6.

Bahga, A., & Madisetti, V. (2016). *Big data science & analytics: A hands-on approach*. Verlag nicht ermittelbar.

Briggs, B., Scott Buchholz, & Sandeep Kumar Sharma. (2020). Macro technology forces. *Deloitte Insights*. https://www2.deloitte.com/us/en/insights/focus/tech-trends/2020/macro-technology-trends.html

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Chen, T. F. (2012). Applying Artificial Intelligence in CRM: Case Studies of Intelligent Virtual Agents and Pegasystems. *Applied Mechanics and Materials*, *182–183*, 878–882. https://doi.org/10.4028/www.scientific.net/AMM.182-183.878

Diaz-Aviles, E., Pinelli, F., Lynch, K., Nabi, Z., Gkoufas, Y., Bouillet, E., Calabrese, F., Coughlan, E., Holland, P., & Salzwedel, J. (2015). Towards real-time customer experience prediction for telecommunication operators. *2015 IEEE International Conference on Big Data (Big Data)*, 1063–1072. https://doi.org/10.1109/BigData.2015.7363860

Donovan, J. (2016). Drones Taking Our Network to New Heights. *TECHNOLOGY BLOG*. https://about.att.com/innovationblog/drones_new_heights

Dumbill, E. (2013). Making Sense of Big Data. *Big Data*, *1*(1), 1–2. https://doi.org/10.1089/big.2012.1503

Idris, A., & Khan, A. (2012). Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers. *2012 15th International Multitopic Conference (INMIC)*, 23–27. https://doi.org/10.1109/INMIC.2012.6511498

ITU. (n.d.-a). *Fixed broadband subscriptions (per 100 people)—United States, Euro area, World*. Retrieved April 3, 2021, from https://data.worldbank.org/indicator/IT.NET.BBND.P2?end=2019&locations=US-XC-1W&name_desc=false&start=2001&view=chart

ITU. (n.d.-b). *Mobile cellular subscriptions (per 100 people)—United States, Euro area, World*. Retrieved April 3, 2021, from https://data.worldbank.org/indicator/IT.CEL.SETS.P2?end=2019&locations=US-XC-1W&name_desc=false&start=2001&view=chart

Li, L., Wang, J., & Li, X. (2020). Efficiency Analysis of Machine Learning Intelligent Investment Based on K-Means Algorithm. *IEEE Access*, *8*, 147463–147470. https://doi.org/10.1109/ACCESS.2020.3011366

Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. *Information Systems*, *34*(1), 87–107. https://doi.org/10.1016/j.is.2008.04.003

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning* (Second edition). The MIT Press.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists* (First edition). O'Reilly Media, Inc.

Nabipour, M., Nayyeri, P., Jabani, H., S., S., & Mosavi, A. (2020). Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. *IEEE Access*, *8*, 150199–150212. https://doi.org/10.1109/ACCESS.2020.3015966

PEGA. (n.d.). *Building powerful, yet transparent omni-channel AI*. Retrieved October 5, 2021, from https://www.pega.com/topic/artificial-intelligence#p-3200bf31-f24d-4259-b5d2-6dd7d51715a8

Ranjan, S., Sood, S., & Verma, V. (2018). Twitter Sentiment Analysis of Real-Time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies. *2018 4th International Conference on Computing Sciences (ICCS)*, 166–174. https://doi.org/10.1109/ICCS.2018.00035

Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World From Edge to Core*. DATA AGE 2025. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Shaham, H. (2020). Vodafone – Global Innovator Chooses AR Remote Assistance. *Tech See*. https://techsee.me/blog/vodafone-innovation-augmented-reality/

TM FORUM. (2018). *Inspire loyalty with customer lifecycle management.*

TM FORUM. (2019). *The CSPs Guide to AI-Driven 360 Degree Customer Profiles*.

TM FORUM. (2020). *30 Strategic KPIs for Digital Transformation*.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

Umayaparvathi, V., & Iyakutti, K. (2017). Automated Feature Selection and Churn Prediction using Deep Learning Models. *International Research Journal of Engineering and Technology (IRJET)*, *4*(3), 1846–1857.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

Win, T. T., & Bo, K. S. (2020). Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm. *2020 International Conference on Advanced Information Technologies (ICAIT)*, 236–241. https://doi.org/10.1109/ICAIT51105.2020.9261792

Yale. (1997). *Linear Regression*. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm