INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Modelos de Data Mining para suporte a avaliações sensoriais do vinho

JOSÉ MIGUEL FERNANDES ABREU
Outubro de 2016

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Modelos de Data Mining para suporte a avaliações sensoriais do vinho

José Miguel Fernandes Abreu

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Sistemas Computacionais

Orientadora: Fátima Rodrigues

Júri:

Presidente:

Vogais:

Porto, Outubro 2016

Resumo

Portugal é um país com grande tradição de exportação de vinho, sendo os Vinhos Verdes os primeiros vinhos portugueses conhecidos nos mercados internacionais. Como tal, a certificação dos vinhos e a avaliação da sua qualidade são elementos-chave neste contexto.

Parte integrante do processo de certificação do Vinho Verde são os testes físico-químicos, que são assegurados por análises laboratoriais tais como: valores de pH, açúcar, álcool, entre outros. Adicionalmente, o processo de certificação passa por testes sensoriais (visão, olfato, paladar) realizados por provadores que são peritos certificados para o efeito.

Neste projeto de dissertação, propõe-se desenvolver modelos de *Data Mining* para prever/standardizar a avaliação sensorial efetuada pelos peritos, através da correlação das características sensoriais dos vinhos com as respetivas características físico-químicas. Para tal, serão extraídas da base de dados da entidade certificadora, características físico-químicas e respetivas avaliações sensoriais de diversas amostras de Vinho Verde relativas ao último ano de certificação - 2015, que serão usadas para desenvolver os modelos de *Data Mining*.

Sendo os dados na sua maioria compostos por valores numéricos contínuos e o atributo a prever ordinal (escala de valores compreendida entre 1 e 10) a seleção das técnicas a aplicar são questões críticas. Pelo que várias técnicas serão testadas e avaliadas tais como, árvores modelo*, support vector machine e random forest*. O objetivo principal será desenvolver modelos com uma taxa de previsão aceitável, para apoiar os peritos nas avaliações do vinho e potencialmente, melhorar a qualidade e a velocidade das suas decisões.

Palavras-Chave: Ciência do Vinho Verde, árvores de decisão, máquinas de suporte vectorial, florestas aleatórias

Abstract

Portugal has a remarkable tradition of wine exportation. Vinho Verde was one of the first Portuguese wines to be exported to several countries in the world. Therefore, the certification of wines and their quality evaluation are key elements in this context.

As part of the Vinho Verde wine certification process, physical and chemical tests are performed by the *Comissão de Viticultura da Região dos Vinhos Verdes* (CVRVV) laboratory. Among other tests, they evaluate pH, sugar and alcohol levels. Additionally, the wine certification relies on sensory tests (sight, smell and taste) performed by certified wine experts.

This dissertation aims to develop data mining models for forecasting sensory results performed by wine taster experts through the correlation with both the wine's physical and chemical characteristics. The dataset will be extracted from the certifying organization database, and will consist of several physical and chemical characteristics and their corresponding sensory evaluation of Vinho Verde wine samples. These data will be used to develop data mining models.

The dataset is mostly composed of continuous numeric values, moreover the attribute to predict is an ordinal (scale values between 1 and 10), therefore selecting the applicable techniques is critical. Various techniques will be tested and evaluated such as tree models, support vector machine and random forest. The main goal is to develop models with a reliable forecast rate to support experts in their wine evaluations and potentially improve the quality and speed of their decisions.

Keywords: Vinho Verde Science, decision trees, bagged trees, support vector machine, random forest

Agradecimentos

A todos aqueles que fizeram parte desta etapa, e que contribuíram, com apoio, amizade, sabedoria e conhecimento, o meu muito obrigado.

Um agradecimento especial, a CVRVV, aos amigos, e à família, pois sem os quais a conclusão esta etapa não seria possível de se realizar.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

AE	Agente Económico
BMLP	*Multi layer perceptron*
CART	*Classification And Regression Tree*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CSV	*Comma-separated values*
CV	*Cross Validation*
CVRVV	Comissão de Viticultura da Região dos Vinhos Verdes
DCBD	Descoberta de Conhecimento em Base de Dados
DM	*Data Mining*
DO	Denominação de Origem
GCV	*Generalized Cross Validation*
IG	Indicação Geográfica
KDD	*Knowledge Discovery in Databases*
kNN	*k nearest neighbours*
MAE	*Mean Absolute Error*
MARS	*Multivariate adaptive regression splines*
MLP	*Single layer perceptron*
MR	Regressão Linear/Múltipla
MSE	*Mean Square Error*
R^2	Coeficiente de determinação
RDVV	Região Demarcada dos Vinho Verdes
REC	*Regression Error Characteristic*
RMSE	*Root Mean Square Error*
ROM	Requisitos organoléticos do Vinho Verde
RF	*Random Forest*
SQL	*Structured Query Language*
SVM	*Support Vector Machines*
TAV	Título alcoométrico volúmico total

1 Introdução

A viticultura sempre teve um lugar de destaque na agricultura portuguesa, representando o vinho um importante setor na sua portuguesa. Hoje, a reputação internacional dos vinhos portugueses é inquestionável e prova disso é o dinamismo do setor das exportações de vinho português, ocupando o nosso país o 9º lugar no *ranking* do comércio internacional do vinho [1].

Face a este dinamismo, a indústria do vinho tem vindo a investir em novas tecnologias não só para melhorar os seus processos de produção e de vendas, mas também de certificação, sendo este um elemento chave para a aferição e garantia da qualidade dos vinhos. A certificação permite prevenir a adulteração ilegal, assim como assegurar a qualidade do vinho lançado no mercado, sendo que vinhos de elevada qualidade são vendidos a preços mais elevados.

Atualmente, o processo de certificação do Vinho Verde inclui a realização de testes físico-químicos que são assegurados por análises laboratoriais tais como, valores de pH, açúcar, álcool, entre outros; e testes sensoriais, através da visão, do olfato e do paladar, realizados pelos provadores.

Surge a necessidade de agilizar o processo da análise sensorial (prova organolética), por forma a fazer face ao aumento exponencial de pedidos de certificação dos produtos de Denominação de Origem (DO) e Indicação Geográfica (IG) da Região Demarcada dos Vinhos Verdes, e consequentemente, o elevado número de ensaios físico-químicos e sensoriais necessários para avaliar a qualidade do vinho.

Sendo os dados, na sua maioria, compostos por valores numéricos contínuos e o atributo a prever ordinal (escala de valores compreendida entre 1 e 10) a seleção das técnicas a aplicar são questões críticas, pelo que várias técnicas serão testadas e avaliadas, tais como: árvores modelo*, support vector machine* (SVM) *e random forest*. Assim, o objetivo principal será desenvolver modelos com uma boa taxa de previsão para apoiar os peritos nas avaliações do vinho e potencialmente, melhorar a qualidade e a velocidade das suas decisões.

Em suma, pretende-se que o projeto a desenvolver nesta dissertação possa constituir uma mais valia qualitativa e quantitativa para a Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), organismo que certifica a DO Vinho Verde e IG Minho. Uma vez que os modelos a desenvolver, proporcionarão uma maior eficiência e eficácia no processo de análise e consequentemente potencial redução de custos.

1.1 Contexto

A Comissão de Viticultura da Região dos Vinho Verdes (CVRVV), é um organismo interprofissional que tem por objeto a representação dos interesses das profissões envolvidas na produção e comércio da Denominação de Origem (DO) «Vinho Verde» e da Indicação Geográfica (IG) «Minho» e a defesa do património regional e nacional que constitui [2].

É ao laboratório da CVRVV que compete verificar se os vinhos, aguardentes e vinagres aptos à DO Vinho Verde ou IG Minho obedecem à qualidade e genuinidade dos produtos vitivinícolas da Região Demarcada dos Vinhos Verdes. Esta garantia é dada pelo "Selo de Garantia", sendo a CVRVV a entidade responsável pela sua atribuição. Deste modo, o laboratório analisa as amostras de forma a garantir que as suas características correspondem às estipuladas para produtos vitivinícolas da região. As principais competências são as seguintes:

1. Gerir as condições necessárias para responder a solicitações de análises físico-químicas, sensoriais e microbiológicas;
2. Assegurar as atividades necessárias à certificação (físico-química e sensorial) dos produtos vínicos da região com DO e IG;
3. Dinamizar, apoiar e promover a melhoria da qualidade dos produtos vínicos da região;
4. Apoiar, realizando as atividades necessárias para o esclarecimento enológico aos clientes;
5. Apoiar a área de marketing em ações de promoção;
6. Participar na realização de atividades necessárias para planear e realizar ações de formação.

Atualmente, a CVRVV possui um avançado software de gestão de laboratório, totalmente desenvolvido de acordo com as suas necessidades, denominado *iLab,* para registo, entre outras funcionalidades, das análises físico-químicas, sensoriais e microbiológicas. Este *software* entrou em funcionamento em 2004.

Os produtos vínicos devem revelar determinadas características organoléticas, tais como o seu aspeto visual (limpidez e cor), olfativo (aroma) e gustativo (sabor). Tais características são avaliadas pela Câmara de Provadores que se pronuncia objetivamente em relação às amostras enviadas ao Laboratório da CVRVV, caracterizando-as quanto aos referidos parâmetros organoléticos. O resultado de cada parâmetro organolético avaliado pela Câmara de Provadores deve traduzir a maioria das opiniões formuladas individualmente pelos provadores que a integram.

No caso dos produtos com DO controlada "Vinho Verde", estes devem cumprir os seguintes Requisitos Organoléticos do Vinho Verde (ROM)[3]:

a) Limpidez - O Vinho Verde deve apresentar-se límpido ou ligeiramente opalino. Apenas é admitido que o vinho se apresente opalino quando este já se encontrar engarrafado e certificado, tendo a rotulagem, neste caso, que mencionar a susceptibilidade de originar depósito. Este parâmetro da análise sensorial não é tido em conta na apreciação dos Vinhos Verdes que não se destinem a engarrafamento ou não engarrafados.

b) Cor - O Vinho Verde branco deve apresentar cor entre citrino descorado e ligeiramente dourado. Os Vinhos Verdes tintos devem apresentar cor entre rubi e vermelho retinto. O Vinho Verde tinto "Palhete" ou "Palheto" e o Vinho Verde tinto "Clarete" devem apresentar cor rubi clara ou rubi. O Vinho Verde rosado deve apresentar cor rosada.

c) Aroma e Sabor - Os requisitos mínimos do Vinho Verde em termos de aroma e sabor são: ausência de defeito marcado, qualidade suficiente - notação igual a cinco, e tipicidade -notação igual a cinco, conforme as escalas de qualidade e tipicidade referidas respetivamente na Figura 1 e Figura 2. O Vinho Verde de casta deve cumprir os requisitos de Vinho Verde, isto é, evidenciar a casta e ter uma notação igual ou superior a seis, conforme a escala de qualidade referida na Figura 1. O Vinho Verde com indicação de sub-região ou com designativo de qualidade deve cumprir os requisitos de Vinho Verde e apresentar características organoléticas destacadas, com notação superior ou igual a seis para a sub-região e para os designativos Escolha, Grande Escolha e Reserva, com notação superior ou igual a sete para os designativos Superior e Colheita Selecionada, conforme a escala de qualidade referida na Figura 1.

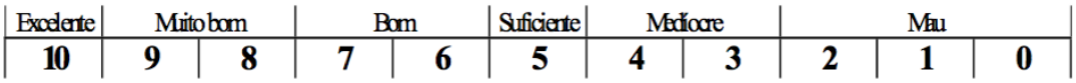

Figura 1 Escala Qualidade

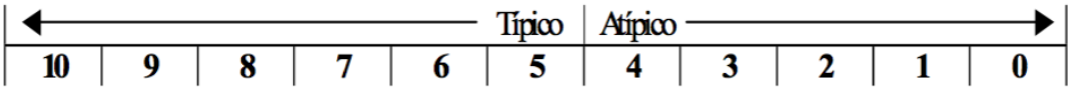

Figura 2 Escala Tipicidade, *Adaptado dos ROM*

1.1.1 *Motivação*

Tal como já foi referido, a CVRVV faz o controlo desde a vinha ao vinho na garrafa e, por isso, neste contexto, qualquer lote de vinho engarrafado necessita de ser previamente certificado pela CVRVV. Esta certificação obriga a que o Agente Económico (AE) disponibilize uma amostra que represente esse lote, de forma a que sejam realizadas as análises físico-químicas e sensoriais (organoléticas) com o intuito de certificar o vinho como apto a engarrafamento.

Cada pedido de certificação de lote exige, consoante o produto a certificar, uma bateria com um número variável de análises físico-químicas e organoléticas regulamentadas pelos estatutos da Região Demarcada dos Vinhos Verdes (RDVV), caso não se enquadre nestes aspetos, o lote não pode ser certificado.

Este processo exige alguns formalismos físicos, nomeadamente:

a) a equipa de provadores que não está alocada exclusivamente ao laboratório, necessita de se reunir no mesmo dia, à mesma hora e local para efetuar a prova;

b) o limite físico do número de provas que cada provador pode efetuar por dia, uma vez que a partir de um determinado número de provas as papilas gustativas já não têm a mesma capacidade de análise.

Assim, de forma a agilizar o processo da prova organolética, este projeto vem contribuir para uma tentativa de extrair resultados sensoriais através dos resultados das análises físico-químicas e com um nível de precisão fidedigno, de modo a reduzir o número de provas efetuadas pelos provadores, mantendo apenas as provas de controle. Consequentemente, obtendo uma maior liberdade para o número de provas a realizar e uma significativa redução dos custos operacionais (custo de cada análise para a CVRVV será menor), que se traduz num aumento do lucro por cada análise de certificação do lote, e finalmente, na redução do tempo de resposta para cada certificação de lote.

Em consequência do exposto, a CVRVV torna o processo mais eficiente e eficaz para com os seus clientes, AE e outras entidades certificadoras que subcontratam ao laboratório da CVRVV as suas análises com intuito de certificarem os seus produtos.

Este trabalho tem como restrições principais a confidencialidade dos dados, bem como, a extração dos dados para a realização desta tese, não permite a identificação do AE.

1.1.2 *Análise de Valor*

Um negócio que seja bem-sucedido e capaz de se posicionar no mercado de forma competitiva necessita de conseguir definir uma proposta de valor, pois a criação de valor é a chave para qualquer negócio, que tem por base a troca de um bem ou serviço tangível e/ou intangível com o intuito de ter o seu valor aceite e recompensado pelos clientes. (Nicola et al. 2012)

A proposta de valor oferece uma visão global do pacote de produtos e serviços que são de valor para o cliente e para a empresa (Osterwalder 2004). Definindo qual o valor que este oferece, e qual o valor percebido dos clientes/consumidores. Zeithmal (1988), sugeriu que o valor percebido é a avaliação global do consumidor sobre a utilidade de um produto, baseado nas perceções do que é recebido e o que é oferecido.

Woodall (2003), revela a existência de potencial na decomposição do valor para o cliente numa perspetiva longitudinal, no sentido em que cada cliente faz uma avaliação diferente do valor ao longo do tempo, tal como, antes da compra *(Ex Ante VC[1])*, no ato da compra *(Transaction VC)*, após a compra *(Ex Post VC)* e depois da utilização/experiência *(Disposition VC)*. Para todos estes casos deve-se relacionar quais os benefícios inerentes e quais os sacrifícios implícitos Tabela 1.

Tabela 1 Benefícios e Sacrifícios

	BENEFÍCIOS	**SACRIFICÍOS**
EX ANTE VC	Customização	Custos de relacionamento Esforço
TRANSACTION VC	Qualidade Serviço Desempenho Serviço	Custos monetários Tempo
POST VC	Operacionais	Energia Humana
DISPOSITION VC	Económicos	Custos de manutenção

[1] Do inglês, "Value for the Customer"

[2] Mosto - Em vinicultura, o termo é usado para referir-se ao sumo de uvas frescas utilizado

Através do projeto a desenvolver pretende-se trazer valor para o cliente (CVRVV), optimizando o processo das análises sensoriais, beneficiando na futura redução de custos e consequente aumento do lucro. Adequa-se aqui um possível cenário de negociação *Win-Win* (Egyed and Boehm 1998), pois o cliente beneficiará de várias melhorias que superam o custo deste desenvolvimento.

É possível analisar a criação de valor através de modelos conceptuais para a decomposição do valor para o cliente em componentes mais simples. (Nicola et al. 2014). Estes propõem um modelo para a avaliação da proposta de valor, integrando o valor percebido pelo cliente e pelo fornecedor. Este modelo está dividido em 3 passos, sendo que o primeiro demonstra a perspetiva dentro da empresa sobre a relevância dos ativos envolvidos no processo e a relação destes com os benefícios e sacrifícios. A perceção do cliente sobre os mesmos é obtida junto com o cliente no passo seguinte, sendo que no final desta análise são combinadas as duas perspetivas de modo a que a empresa seja capaz de relacionar os ativos com os benefícios e sacríficos ajustando a sua proposta de valor.

1.1.3 *Modelo de negócio*

De forma a esboçar um modelo de negócios que descreve a lógica de criação, entrega e captura de valor por parte de uma organização, utilizou-se o modelo de *Canvas* conforme apresentado na Tabela 2.

Tabela 2 Modelo de negócio

<table>
<tr><td rowspan="2"><u>Parceiros</u>

CVRVV</td><td><u>Atividades Chave</u>

Análise de dados
Descoberta de conhecimento a partir de dados</td><td rowspan="2"><u>Proposta de valor</u>

Redução de Custos
Maior eficiência e produtividade
Captura de conhecimento</td><td><u>Relacionamento com Clientes</u>
Assistência Personalizada
Cocriação de valor</td><td rowspan="2"><u>Segmentos de Clientes</u>

Entidades Certificadoras de Vinho [4]</td></tr>
<tr><td><u>Recursos Chave</u>

Dados</td><td><u>Canais</u>
Reuniões Presenciais ou via (Skype)
E-mail</td></tr>
<tr><td colspan="2"><u>Estrutura de custos</u>
Local de Trabalho
Equipamento Informático</td><td colspan="3"><u>Fontes de receitas</u>
Redução de custos do processo certificação
Aumento receitas com mais processos de certificação</td></tr>
</table>

2 Estado da Arte

2.1 Descoberta de Conhecimento em Bases de Dados

O processo de Descoberta de Conhecimento em Bases de Dados (DCBD), tem como objetivo a extração de conhecimento a partir de grandes bases de dados. O processo global de DCBD, que se desenvolve em várias fases, inclui a utilização dos algoritmos de *DM* e a interpretação dos padrões encontrados pelos mesmos, os quais são posteriormente utilizados no suporte à tomada de decisão.

Data Mining é uma das fases do processo de DCBD que consiste na pesquisa de relacionamentos, padrões ou modelos que estão implícitos nos dados armazenados em grandes bases de dados.

As diversas fases do processo DCBD estão representadas na Figura 3 e incluem:

- Seleção dos dados;
- Tratamento dos dados;
- Pré-processamento dos dados;
- *Data Mining*;
- Interpretação de resultados.

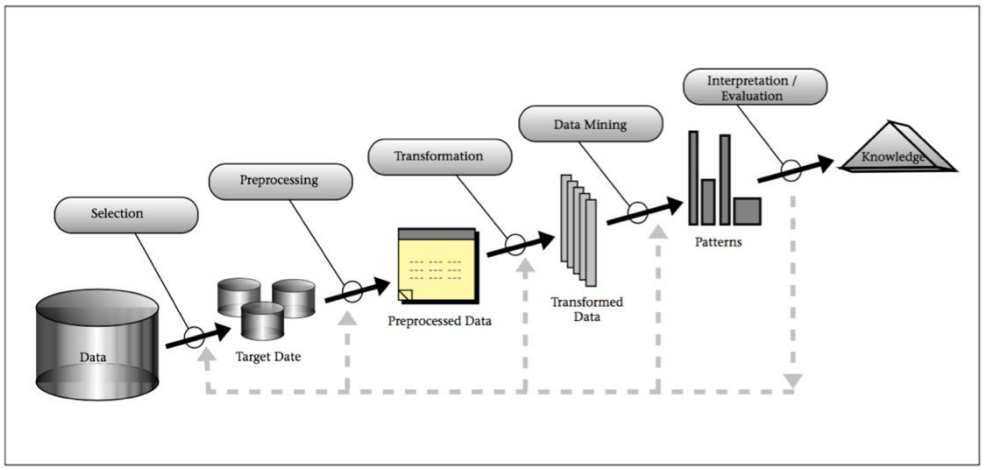

Figura 3 Fases do Processo DCBD, *Adaptado de (Fayyad et al. 1996)*

2.2 Metodologia CRISP-DM

A metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*)[5] de suporte a projetos de *DM*, descreve um modelo de referência que define as fases a seguir, as tarefas a executar e os resultados esperados pela realização de cada uma das tarefas. Esta metodologia encontra-se como o *standard* mais utilizado, devido à flexibilidade da sua implementação em qualquer domínio e à compatibilidade com qualquer ferramenta de *DM*.

Na Figura 4 é possível verificar a organização das etapas definidas pelo CRISP-DM, que consistem em:

- Compreensão do negócio: entender quais os objetivos e requisitos do projeto na perspetiva de negócio, e converter em uma definição de problema *DM*;
- Compreensão dos dados: inicialmente procede-se à aquisição de amostras de dados, seguindo-se a exploração, de forma a tentar entender as relações entre estes e verificação da sua qualidade em conjunto com atividades que possibilitem a familiarização com os dados;
- Preparação dos dados*:* fase que abrange todas as atividades necessárias para construir o conjunto final de dados, transformando e corrigindo qualquer problema de incoerência nos dados para a utilização das ferramentas de modelação;
- Modelação*:* seleção e aplicação das técnicas de modelação e calibração de parâmetros;
- Avaliação*:* construção do(s) modelo(s) que aparentam ter maior qualidade na perspetiva de análise de dados. Antes de proceder ao desenvolvimento do modelo final é importante avaliar e rever todos os passos executados, de forma a ter confiança que o modelo atinge os objetivos impostos;

- Implementação*:* A criação do modelo não é, geralmente o fim do ciclo, apesar de o seu propósito ser aumentar o conhecimento sobre os dados analisados. Contudo, é necessário organizar e apresentar o conhecimento obtido de forma clara e que o cliente a possa reutilizar. Normalmente, este passo envolve a integração dos modelos no processo de decisão da organização, como paginas Web personalizadas, a implementação de um processo repetitivo de *DM*, ou apenas a geração de um relatório.

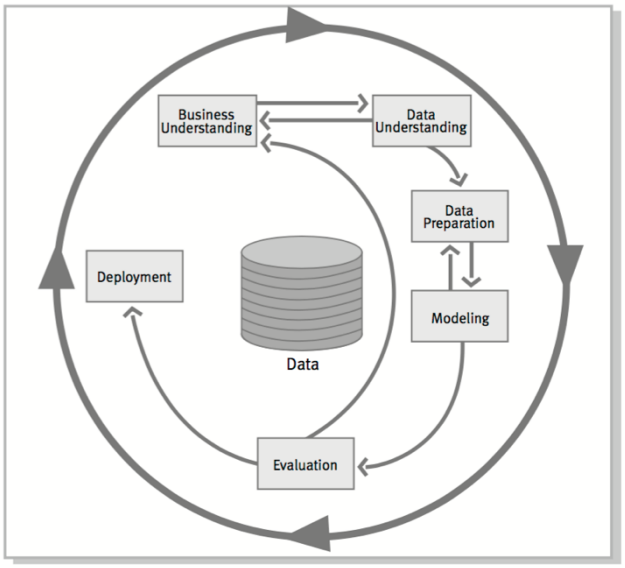

Figura 4 Fases do modelo de referencia CRISP-DM, *Adaptado de (Chapman et al. 2000)*

O modelo CRISP-DM é o processo de *DM* mais robusto, completo e documentado em comparação com DCBD. No entanto estes processos apresentam algumas semelhanças, tais como as que estão representadas na Tabela 3.

KDD	**CRISP-DM**
Pre KDD	Business Understanding
Selection	Data Understanding
Pre processing	
Transformation	Data Preparation
Data mining	Modelling
Interpretation/Evaluation	Evaluation
Post KDD	Deployment

Tabela 3 Comparação KDD e CRISP-DM, *Adaptado de (Azevedo and Santos 2008)*

O processo CRISP-DM implementa todos os passos do DCBD, adicionando a valorização e a relevância do pré conhecimento da aplicação de domínio, os objetivos do cliente e a consolidação do conhecimento obtido.

2.3 Técnicas de *Data Mining*

Existem diversos algoritmos de aprendizagem, sendo que cada um possui vantagens e desvantagens. Dado que não existem fórmulas analíticas que possam determinar se um dado algoritmo terá um desempenho melhor do que o outro, é necessário realizar estudos experimentais.
Quando a modelação dos dados é contínua a Regressão Linear/Múltipla (MR) é a abordagem clássica mais utilizada (Wu et al. 2008). No entanto, como se pretende fazer uma previsão ordinal (valores contínuos) com múltiplas variáveis de previsão e com relações complexas entre as mesmas, nesta situação os modelos de regressão nem sempre são adequados. Assim, outras técnicas serão utilizadas tais como Árvores de Previsão Numérica, *Support Vector Machine* (SVM) e Random Forest.

2.3.1 *Árvores de Previsão Numérica*

Uma árvore de decisão é uma estrutura hierárquica que contém nós, ramos e folhas, representando a decomposição do domínio, tal que:

- Cada nó inclui um teste a um atributo
- Um ramo corresponde ao resultado desse mesmo teste
- Uma folha pertence a uma classe

As árvores de decisão inicialmente foram desenvolvidas para previsão de atributos categóricos – árvores de classificação, mas mais tarde foram também adaptadas para previsão numérica, fazendo apenas pequenos ajustes no algoritmo de construção da árvore. As árvores de previsão numérica são construídas de modo idêntico às árvores de classificação. Começando no nó raiz, os dados são divididos usando uma estratégia dividir-para-conquistar que seleciona o atributo que conduza a uma maior homogeneidade no resultado após a realização da divisão. Em árvores de classificação a homogeneidade é medida pela entropia, que no caso de atributos numéricos não é aplicável, pelo que para árvores de decisão numéricas, a homogeneidade é medida por estatísticas, tais como, variância, desvio padrão, ou o desvio absoluto da média. Dependendo do algoritmo de construção da árvore a medida de homogeneidade pode variar, mas os princípios são basicamente os mesmos das árvores de classificação.

As árvores de previsão numérica dividem-se em duas categorias: árvores de regressão e árvores modelo (*model trees*).

2.3.1.1 Árvores de Regressão

As árvores de regressão foram introduzidas na década de 1980, das quais se destaca o algoritmo*, Classification And Regression Tree* (CART) (Chapman and Hall 1984). Apesar do seu nome as árvores de regressão não fazem previsão usando métodos de regressão linear, mas sim com base no valor médio dos exemplos que atingem as folhas das árvores.

Uma limitação das árvores de regressão simples é que em cada nó terminal a previsão é feita usando a média dos valores das instâncias nesse nó. Como consequência, estes modelos falham na previsão de valores extremos, baixos ou elevados.

2.3.1.2 Árvores Modelo

O segundo tipo de árvores para a previsão numérica é conhecido como árvores modelo. Estas foram introduzidas mais tarde que as árvores de regressão, e são mais poderosas. As árvores modelo são desenvolvidas de modo semelhante às árvores de regressão com exceção dos seguintes parâmetros:

- O critério de divisão é diferente. Utilizando as variáveis independentes, os dados são divididos em vários pontos, onde se verifica o erro entre o valor previsto e o valor real, as variáveis que revelarem menor erro, são as escolhidas para o nó. Este processo ocorre recursivamente.
- Os nós terminais preveem o resultado usando um modelo linear (em oposição à média simples).
- A previsão de uma amostra é muitas vezes uma combinação das previsões de diferentes modelos ao longo do percurso através da árvore da raiz até à folha.

Dependendo do número de nós folha, uma árvore modelo pode construir dezenas ou mesmo centenas de tais modelos. Isto pode fazer com que as árvores modelo sejam mais difíceis de analisar do que as árvores de regressão equivalentes, mas têm a vantagem de serem um modelo mais preciso. A principal implementação deste tipo de árvores é o algoritmo M5, descrito em (Wang and Witten 1997).

2.3.2 *Árvores bagged*

Na década de 1990 surgiram as técnicas de *ensemble* (métodos que combinam previsões de muitos modelos). *Bagging*, abreviação para a agregação de *bootstrap* (secção 2.4.3), originalmente proposta por Leo Breiman, foi uma das primeiras técnicas ensemble desenvolvida (Breiman 1996). *Bagging* é uma abordagem geral que utiliza *bootstrapping* em conjunto com qualquer modelo de regressão ou classificação. O método é simples e é constituído pelos passos do seguinte algoritmo:

```
1- Para i = 1 to m fazer
2- Gerar uma amostra bootstrap a partir do conjunto original
3- Treinar um modelo árvore de decisão com esta amostra
4- fim
```

Cada modelo do conjunto é então usado para gerar uma previsão para uma nova amostra e estas m previsões são combinadas para dar a previsão do modelo *bagged*.

2.3.3 *Support Vector Machines (SVM)*

As SVM constroem um classificador de acordo com um conjunto de padrões, por ele identificados nos exemplos de treino, onde a classificação é conhecida. Considerando o exemplo da Figura 5, nela existe um conjunto de classificadores lineares que separam duas classes, mas apenas um (em destaque) que maximiza a margem de separação (distância da instância mais próxima ao hiperplano de separação das duas classes em questão). O hiperplano com margem máxima é chamado de hiperplano ótimo, que será o objeto de busca do treino do classificador (Gunn 1998).

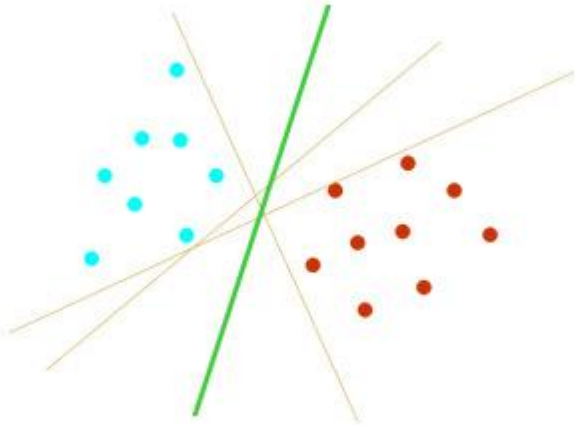

Figura 5 Possíveis hiperplanos de separação e hiperplano ótimo, *Adaptado de (Junior 2010)*

A Figura 6 (a) mostra um dos possíveis hiperplanos de separação com margem pequena e (b) mostra o hiperplano de separação ótimo com a margem maximizada.

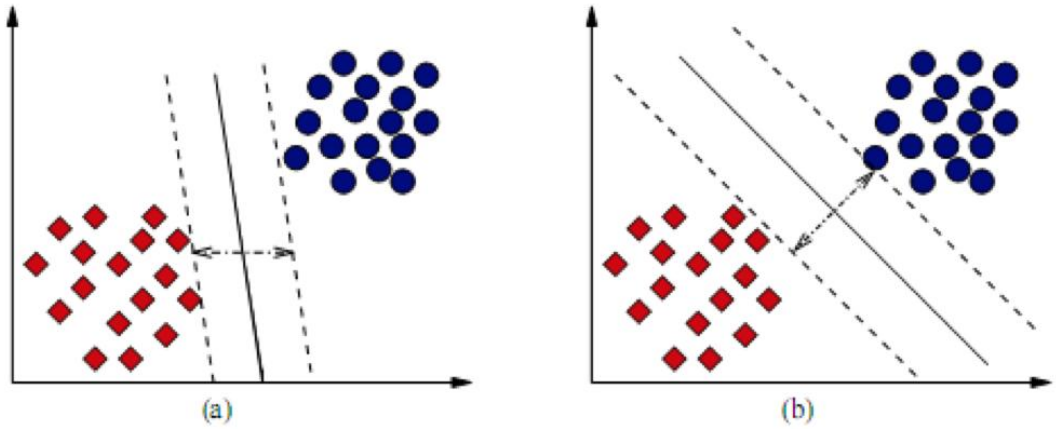

Figura 6 (a) Hiperplano com margem pequena. (b) Hiperplano com margem máxima, *Adaptado de (Junior, 2010)*

Porém, as amostras de dados nem sempre são linearmente separáveis. As funções de *Kernel* tem a finalidade de projetar os vetores de características de entrada num espaço de características de alta dimensão, para classificação de problemas que se encontram em espaços não linearmente separáveis. À medida que se aumenta o espaço da dimensão do problema, aumenta também a probabilidade desse problema se tornar linearmente separável em relação a um espaço de baixa dimensão. No entanto, para obter uma boa distribuição para esse tipo de problema é necessário um conjunto de treino com um elevado número de instâncias (Gonçalves 2010).
A Figura 7 representa esquematicamente o processo de transformação de um domínio não linearmente separável, num problema linearmente separável através do aumento da dimensão, onde é feito um mapeamento por uma função de *Kernel* $F(x)$.

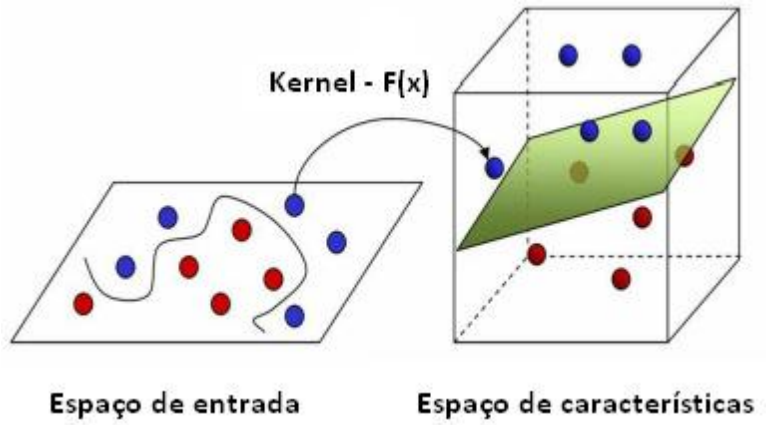

Figura 7 Transformação: problema não linearmente separável em um problema linearmente separável, *Adaptado de (Rebelo 2008)*.

Existem vários tipos de *kernels* que podem ser usados, porém as funções de *Kernel* mais usadas são polinomiais, gaussiano e sigmoidal.
As *SVMs* (Lorena and Carvalho 2007) podem ser usadas de várias formas. É possível utilizar diferentes tipos de funções *Kernel* e diferentes parâmetros a variar de acordo com o *Kernel* escolhido. Além disso, um melhor parâmetro para uma base, não é necessariamente, e muito provavelmente não será, o melhor parâmetro para outra.

2.3.4 *Random Forest*

Random Forest consiste numa coleção de classificadores baseados em árvores de decisão em que é cada árvore dá origem a um voto para a previsão da saída. O *Random Forest* produz um modelo constituído por *n* árvores de decisão (ensemble), onde cada árvore é baseada num certo número de instâncias do conjunto de treino, escolhidas aleatoriamente.

Cada nó de cada árvore é construído a partir de um subconjunto aleatório dos atributos. Ao receber uma instância de teste cada árvore irá decidir (votar) sobre qual a classe a que pertence. A classe mais votada será a classe prevista pelo modelo. O algoritmo *Random Forest* mais divulgado é o algoritmo de Breiman (Breiman et al. 1984).

2.3.5 *Redes neuronais*

Uma rede neuronal é uma máquina de processamento paralelo e distribuído, composto por unidades de cálculo simples que têm a capacidade inata de aprender, guardar e utilizar conhecimento. As redes neuronais são caracterizadas por dois aspetos, a arquitetura, que se encontra relacionada com o tipo e número de unidades de processamento e a forma como os neurónios (nós da rede) estão conectados. O segundo aspeto é a aprendizagem, que diz respeito às regras utilizadas para o ajuste dos pesos da rede e a informação utilizada pela rede.

A rede neuronal mais usada, denomina-se por rede neuronal *feed-forward.* Esta tipologia de rede, é composta por vários nós interligados por pontes e organizados por camadas. A informação analisada pela rede apenas transita em um sentido. A existência, ou não, de camadas escondidas na rede neuronal, define o tipo de rede como *multi layer perceptron-(BMLP)* ou *single layer perceptron-(MLP)*. (Wilamowski 2003)

O tipo de rede BMLP, é considerada mais precisa, devido a utilização da técnica *back-propagation*. Esta técnica baseia-se na comparação dos valores de saída da rede, com os valores de treino, para calcular o valor do erro. Por consequência é propagado para trás e utilizado para o ajustamento dos pesos de cada conexão. A repetição deste processo converge a rede neuronal para um ponto em que o erro tende a diminuir. (Gershenson 2003)

2.3.6 *k-nearest neighbours (kNN)*

A técnica kNN, também conhecida como k-vizinhos mais próximos, procura classificar grupos de objetos - *k*, no conjunto de treino, que se encontram mais próximos do objeto de teste. Existem três elementos-chave nesta abordagem, tais como, a identificação dos conjuntos marcados, a distância ou a semelhança métrica para calcular a distância entre dois objetos e o valor de *k*, que indica o número de vizinhos mais próximos. A classificação de um objeto não rotulado, é calculada a partir da distância deste com os seus *k*-vizinhos mais próximos. (Cover et al. 1967)

A escolha do valor de *k*, é um aspeto importante, pois afeta diretamente o desempenho desta técnica. Se o valor de *k* for demasiado pequeno, o resultado pode ser afetado por ruído nos dados, tornando-se bastante sensível, caso contrário, os dados vizinhos a analisar, podem conter muitos pontos de outras classes, obtendo assim, resultados incorretos.

2.3.7 *Multivariate adaptive regression splines (MARS)*

Multivariate adaptive regression splines (MARS) (Friedman 1991), é uma metodologia para aproximar funções com múltiplas variáveis de entrada. É um método não paramétrico, baseado numa estratégia dividir-para-conquistar, que particiona os conjuntos de dados de treino em segmentos lineares (*splines*) interligados, resultando num modelo flexível adaptado a contextos lineares e não lineares (Zhang and Goh 2016). Adequa-se a problemas multidimensionais, e embora os tempos de formação para este método tendem a ser mais rápidos do que as redes neurais *feed-forward* usando *back-propagation*, problemas mais complexos com múltiplas aproximações, tendem a tornar este método bastante lento.

2.4 Avaliação de modelos

2.4.1 *Técnicas para estimativa de erro dos modelos*

A estimativa da qualidade de um modelo, inferido através de uma amostra de dados usando um algoritmo de DM, é sem dúvida muito importante, pois permite avaliar a capacidade de generalização do modelo, a partir de um novo conjunto de dados. Busca-se estimar o quão preciso é o modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

Para conhecermos o valor do erro real de um modelo com exatidão teríamos de testar o modelo com todos os exemplos possíveis. Este processo não é, no entanto, realístico pois em casos normais é impraticável a obtenção de todos os valores do universo em causa, tendo-se geralmente disponíveis apenas um reduzido número de exemplos. É necessário pois estimar este valor da forma mais rigorosa possível, tendo o cuidado de não efetuar um cálculo viciado, isto é, demasiado otimista ou demasiado pessimista. Existem várias técnicas de estimação de erro. Seguidamente serão apresentadas apenas aquelas que irão ser usadas nesta dissertação discutindo-se as suas vantagens e desvantagens.

2.4.2 *Validação cruzada*

O conjunto inicial de dados é dividido em k-subconjuntos de igual tamanho (Breiman et al. 1984). Em cada iteração são usados $(k-1)$ subconjuntos para treino e um subconjunto para teste (Figura 8).

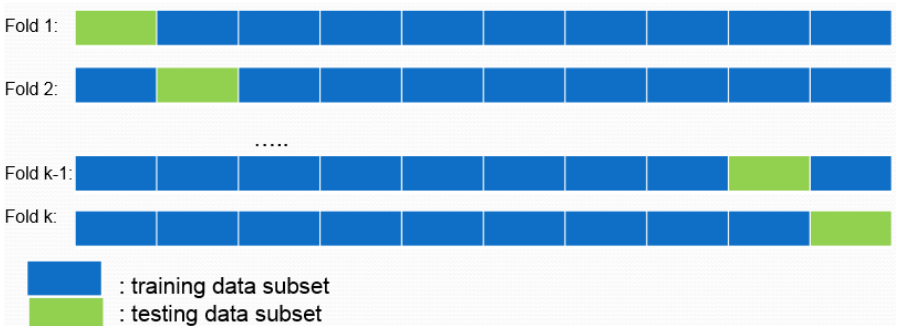

Figura 8 K-Subconjuntos Validação cruzada [6]

Ao final de *k* iterações cada registo é usado o mesmo número de vezes para treino e uma só vez para teste. A avaliação é realizada k vezes (*k*=10 é o número mais usado). Através de experiências intensivas demonstrou-se que, *k*=10 é a melhor escolha para se obter uma estimativa fiável e um compromisso entre precisão e complexidade (Tan et al. 2005).
O valor estimado para o erro do classificador final será então a média dos erros estimados para cada um dos k classificadores parciais assim calculados.
As vantagens deste método é que utiliza o máximo de dados possível para treino, os conjuntos de teste são mutuamente exclusivos, uma vez que cobrem todo o conjunto de dados, e a estratificação reduz a variância das estimativas. A maior desvantagem deste método é essencialmente notória, uma vez que envolve grande complexidade analítica. O cálculo dos diversos classificadores bem com o seu teste pode implicar um volume de processamento incomportável para conjuntos de treino de dimensões razoáveis.

2.4.3 *Estimativa por bootstrapping*

A técnica de estimação por *bootstrapping* é especialmente indicada para os casos em que o conjunto de treino é bastante limitado. Para a estimação por *bootstrapping* é formado um novo conjunto de treino a partir do conjunto de treino original, amostrando *n* vezes com reposição os *n* exemplos do conjunto original. Os exemplos repetidos são eliminados, podendo ser demonstrado que a percentagem esperada de exemplos no novo conjunto de treino assim obtido é de 63.2% do número de exemplos do conjunto original. Os casos que não se encontrarem no novo conjunto de treino constituirão o conjunto de teste.

O erro estimado para o classificador será a média de várias iterações deste algoritmo, sendo normalmente cerca de 200 iterações um número considerado apropriado para uma boa estimação (Efron and Tibshirani 1986). É, portanto, evidente a grande complexidade deste método, devido ao grande número de iterações necessárias.

Todas as técnicas indicadas nestas heurísticas, são técnicas de reamostragem de modo a garantir que todos os exemplos do conjunto de treino inicial são usados tanto para o treino do classificador como para o seu teste. Este facto elimina a possibilidade de escolha de conjuntos de treino/teste incaracterísticos e consequente obtenção de resultados excessivamente otimistas ou pessimistas.

2.4.4 *Medidas para Avaliação de Modelos*

2.4.4.1 MSE - Erro médio quadrado

O erro médio quadrado (*Mean Square Error* – MSE) é calculado pela soma do quadrado dos resíduos. Cada valor Y_i é comparado com o valor previsto $\hat{Y}_i$ para ver o quão longe se encontram. A expressão $\hat{Y}_t - Y_t$ define o erro, numa previsão perfeita $\hat{Y}_t$ será igual a Y_t e o erro será zero.

$$MSE = \frac{1}{n} \Sigma_{t=1}^{n} (\hat{Y}_t - Y_t)^2$$

2.4.4.2 RMSE - Raiz quadrada do Erro médio quadrado

A medida mais comum para avaliar modelos preditivos numéricos é obtendo a raiz quadrada dos erros (RMSE). Esta métrica é uma função dos resíduos do modelo, que são os valores observados menos as previsões do modelo. O RMSE é, então, calculado através da raiz quadrada do MSE, sendo este erro na mesma unidade dos dados originais.

$$RMSE = \sqrt{MSE}$$

Este valor é geralmente interpretado como a distância média entre as previsões e os valores reais.

2.4.4.3 R^2 - Coeficiente de Determinação

Outra métrica comum é o Coeficiente de Determinação, comummente escrito como R^2. Este valor pode ser interpretado como a proporção de informação nos dados que é explicada pelo modelo. Assim, um valor de R^2 de 0.75 implica que o modelo consegue explicar três quartos da variação do resultado. Existem várias fórmulas para o cálculo desta métrica (Kvalseth 1985), embora a versão mais simples encontra o coeficiente de correlação entre os valores observados e os previstos (geralmente designado por R) e calcula o quadrado desta medida R^2.

2.4.4.4 Curva REC - *Regression Error Characteristic*

Outra técnica de avaliação e comparação de modelos de regressão é chamada de *Regression Error Characteristic (REC)* (Breiman et al. 1984). As curvas REC mostram a taxa de acerto global (eixo do *y*) para diversos valores de tolerância (*T*) de erro absoluto (eixo do *x*). Usando curvas REC os não peritos podem avaliar facilmente a capacidade de previsão de um modelo.

A precisão, ou taxa de acertos, é definida com a percentagem de pontos que se encaixam dentro da tolerância. Se a tolerância for zero apenas os pontos de ajuste são considerados. Escolhendo uma tolerância que exceda o erro máximo observado para o modelo, então todos os pontos considerados serão corretos. Assim existe um *trade-off* claro entre a tolerância de erro e a precisão da função de regressão. O conceito de tolerância de erro é atraente porque muitas vezes os dados de regressão são imprecisos devidos por exemplo a erros de medição.

2.5 Abordagens

A utilização de técnicas de DM neste domínio é ainda bastante escassa e os trabalhos publicados baseiam-se no geral em conjuntos de dados de pequena dimensão. Na Tabela 4 encontra-se em destaque uma revisão bibliográfica de abordagens que utilizam técnicas de DM para prever a qualidade do vinho.

Tabela 4 Abordagens existentes

Referencia	**Objetivo**	**Dados**	**Tecnologia/Métodos**	**Resultados**
[Cortez et al. 2009]	Prever as preferências gustativas do ser humano com base nos testes analíticos disponíveis na etapa de certificação do vinho.	Dois conjuntos de dados distintos. 1599 registos vinho tinto, com 6 classes e 4898 registos vinho branco com 7 classes.	Modelos de Regressão SVM *Neural Network*	Apresentam um método que executa a seleção simultânea de modelos e classes para as técnicas *Neural Network* e SVN. Obtêm uma precisão global de 89.0% para Vinhos Verdes Tintos e 86,8% para Vinhos Verdes Brancos.
[Buratti et al. 2007]	Testar a capacidade de extrair a partir de técnicas analíticas, como a língua eletrónica e nariz eletrónico informações sobre as propriedades sensoriais do vinho e qualidade, por exemplo, amargor, acidez, corpo, aroma e cor.	As medições foram realizadas em 15 vinhos italianos. Obtendo um conjunto de dados com 15 registos (amostras de vinho) e 56 classes (variáveis)	Aplicaram algoritmos genéticos para selecionar os atributos e construir modelos de regressão preditivos	Os resultados obtidos demonstram a possibilidade de utilizar estas técnicas inovadoras, com a finalidade de descrever e prever uma grande parte da informação.
[Tian and Pang 2010]	Desenvolver modelo preditivo capaz de classificar e identificar a qualidade do vinho.	Conjunto de dados com 3655 registos e 12 classes (analises físico-químicas)	Linear/Múltipla Regressão Árvores de decisão *Neural Network*	*Neural Network* é a tecnologia mais indicada para desenvolver o modelo, obtendo 84% de precisão (matriz *forecast*)
[Hosu et al 2014]	Criação de instrumentos preditivos, capazes de identificar as propriedades mais relevantes dos vinhos e capazes de caracterizar os diferentes tipos de qualidade.	Conjunto limitado de 28 vinhos Romanos.	*Neural Network*	*Neural Networks* revelam-se confiáveis para a avaliação ou validação das características essenciais do vinho, capazes de classificar o ano de colheita e a origem de cada vinho sem qualquer erro.
[Omatu et al. 2015]	Desenvolver um novo sistema de cariz eletrónico baseado no algoritmo *learning vector quantization*	Foram utilizadas 24 amostras, com dados de odor, de dois tipos de vinho (branco e tinto), produzidos por duas empresas.	Algoritmo especifico de *Neural Network* – *learning vector quantization*	Classificado dois tipos de vinho produzido em empresas distintas. Obtendo uma taxa de acerto de 97,5% para vinhos brancos e 83,4% para vinhos tintos.

3 Design da solução

3.1 Ferramentas

A aplicação *DM* do presente trabalho será desenvolvida em R [7], sendo uma ferramenta *Open Source* (gratuita) e multiplataforma (disponível para Windows, Linux e Mac) de computação estatística. Embora não sendo uma ferramenta especificamente desenvolvida para a área de *DM*, esta conta com uma comunidade muito ativa, que disponibiliza periodicamente novos pacotes capazes de alargar as capacidades desta ferramenta, com novas técnicas e algoritmos. Contendo mais de 7800 pacotes disponíveis no *The Comprehensive R Archive Network* (CRAN) [8].

RStudio [9] é um *IDE* que torna o R mais fácil de usar e mais produtivo. *RStudio* combina um conjunto de ferramentas de produtividade em um só ambiente, incluindo:

- Editor de código - destaque de sintaxe, *auto-complete* de código, indentação e definições
- Depuração - depuração na consola, *breakpoints* e *tracebacks*
- Visualização - exibição de dados, visualização gráfica e manipulação de dados

3.2 Pacotes

Os algoritmos usados neste projeto, estão implementados em vários pacotes disponíveis no CRAN, sendo os principais, RPart, Caret, iPred, randomForest, RWeka, nnet e earth.

Rpart (Therneau et al. 2015), disponibiliza a construção de modelos de classificação ou regressão, através de particionamento recursivo. Estes modelos são apresentados como árvores binárias onde a subdivisão dos dados tem por base, várias variáveis independentes.

O processo é denominado recursivo, porque cada subconjunto pode ser dividido num número indeterminado de vezes, terminado aquando um critério em particular é alcançado.

Caret (Kuhn et al. 2016), oferece um conjunto de funções que visam simplificar o processo de criação de modelos preditivos. Tais como, separação dos dados, pré-processamento e otimização de modelos utilizando reamostragem.

IPred (Peters et al. 2015), permite a implementação de *bagging*, para árvores de regressão e classificação.

RandomForest (Liaw and Wienere 2002), permite a implementação do algoritmo, e a criação de modelos de classificação ou regressão com base em um conjunto de árvores, utilizando variáveis de entrada aleatórias.

RWeka (Hornik et al. 2009), interface em R para *Weka*, sendo um conjunto de algoritmos de *Machine Learning* para tarefas de *DM* desenvolvidos em Java, este pacote permite assim o uso de várias ferramentas, pré-processamento de dados, classificação, regressão, *clustering*, regras de associação e visualização.

Nnet (Venables and Ripley 2002), necessário para a formação de redes neuronais, permite a configuração das mesmas através de configurações flexíveis em função do erro.

Earth (Millborrow 2016), permite a construção de modelos de regressão usando as técnicas de *Multivariate Adaptive Regression Splines* e *Fast MARS*.

3.3 Arquitetura

Neste trabalho não existe propriamente um design conceptual da solução do problema, mas sim um conjunto de passos a seguir de acordo com a metodologia CRISP-DM (secção 2.2).
O desenvolvimento de um procedimento para a escolha do período de datas, dos dados a extrair da base de dados, o tratamento de limpeza e preparação dos conjuntos de dados de treino e teste, de forma a aplicar os algoritmos do processo de *DM*. Finalizando com a avaliação e documentação dos resultados obtidos.

4 Previsão da Qualidade do Vinho Verde através de Técnicas de *Data Mining*

4.1 Compreensão do Negócio

A análise sensorial da CVRVV tem como principal objetivo detetar a presença de defeitos e avaliar o cumprimento de critérios inerentes ao produto. Como tal, a previsão destes resultados a partir das análises físico-químicas, suscitam interesse, no sentido de melhorar a qualidade e a velocidade das suas decisões.

4.2 Compreensão e Preparação dos Dados

4.2.1 *Aquisição dos dados – processo de aquisição (procedimento de SQL, tratamento prévio)*

Em primeira instância e com vista à aquisição do conjunto de dados, desenvolveu-se um procedimento em *Structured Query Language (SQL)*. Para tal, foi necessário efetuar uma análise da base de dados, bem como a relação entre as tabelas onde se encontram os dados a extrair para análise. Este procedimento envolveu a junção de seis tabelas pela realização de *joins* para a consequente obtenção dos dados que se encontram dispersos pela base de dados, bem como a relação com as regras definidas pelo negócio. Preparou-se também o procedimento para futuras extrações, sendo possível a indicação do intervalo de tempo e do tipo de produto a extrair.
No presente trabalho, os dados recolhidos, referem-se a todas as análises realizadas no ano 2015.

4.2.2 *Descrição dos dados*

De forma a importar os dados para a ferramenta de análise, estes foram inicialmente exportados para um ficheiro *CSV*, sendo que o conjunto de dados extraído contém 35590 registos, composto por 3 atributos: referência da amostra, a descrição do tipo de ensaio efetuado (29 ensaios distintos) e o resultado obtido.

Todos os produtos certificados pela CVRVV, regem-se por uma legislação específica aos quais são aplicadas regras de conformidade de acordo com cada tipo de produto. Estas regras definem que tipo de análises físico-químicas e sensoriais são necessárias efetuar e balizam a aceitação dos resultados.

No excerto de dados demonstrado na Tabela 5, é exemplificada a informação obtida de um boletim de análise, o qual dará origem à atribuição de selos de garantia. De salientar que todos os dados extraídos correspondem a boletins de análise válidos, ou seja, todos os resultados estão conforme os parâmetros definidos nas regras de conformidade.

Tabela 5 Excerto do conjunto de dados

AMOSTRA	DESCRIÇÃO DO ENSAIO	VALOR
2015000000	Acidez fixa	6.1
2015000000	Acidez total	6.31
2015000000	Acidez volátil	0.13
2015000000	Ácido cítrico	0.34
2015000000	Açúcares totais	3.2
2015000000	Cloretos	0.021
2015000000	Dióxido de enxofre livre	38
2015000000	Dióxido de enxofre total	112
2015000000	Extrato não redutor	21.0
2015000000	Extrato seco total	24.2
2015000000	Massa volúmica	0.9926
2015000000	pH	3.50
2015000000	Sulfatos	0.95
2015000000	Título alcoométrico volúmico adquirido	11.3
2015000000	Ácido ascórbico	19.0
2015000000	Ácido sorbico	62.0
2015000000	Cobre	0.03
2015000000	Dióxido Carbono	12.0
2015000000	Metanol	67.0
2015000000	Sobrepressão	0.80
2015000000	Título alcoométrico volúmico total	11.5
2015000000	Aroma – Defeito marcado	2
2015000000	Aroma – Qualidade	7
2015000000	Aroma – Tipicidade	7
2015000000	Aspeto – Cor	2
2015000000	Aspeto – Limpidez	1
2015000000	Sabor – Defeito marcado	2
2015000000	Sabor – Qualidade	7
2015000000	Sabor – Tipicidade	7

4.2.2.1 Descrição dos atributos previsores

Acidez total – Representa a soma das componentes acidez fixa e volátil dos vinhos. Importante índice caracterizador das potenciais propriedades ácidas do vinho. O perfil ácido do vinho depende da concentração dos ácidos presentes e da sua força química, sendo que também determina o equilíbrio físico-químico e proteção do produto.

Acidez volátil – O ácido acético é o ácido orgânico volátil mais comum no vinho. Estes componentes conferem características organoléticas negativas ao vinho e estão relacionados com a atividade de leveduras e bactérias. Relacionada com a qualidade sanitária do vinho.

Acidez fixa – Diferença entre Acidez total e Acidez Volátil; representa o teor em ácidos fixos como ácido tartárico, málico, cítrico, lático entre outros menos representativos. Estes ácidos orgânicos fixos são responsáveis pelas propriedades organoléticas ácidas do vinho.

Acido cítrico – Já referido como sendo ácido orgânico fixo; confere ao vinho sabor citrino/frescura, atuando por isso no melhoramento gustativo; pratica enológica com aumento significativo.

pH – Está relacionado com a componente ácida real/disponível do vinho.

Açúcares totais – Os açúcares do mosto[2] são transformados em álcool no decurso da fermentação alcoólica por ação das leveduras. Os açúcares residuais mais comuns no vinho são frutose e glucose, sendo responsáveis pelo gosto "doce", embora com importâncias organoléticas diferenciadas, sendo a frutose a que confere uma intensidade mais acentuada.

Cloretos – Importante elemento da constituição mineral do vinho. Um elevado teor pode ser associado à prática enológica fraudulenta.

Sulfatos – Importante elemento da constituição mineral do vinho. O mineral é quantificado sob a forma do sal mineral "sulfato de potássio". Elevado teor pode estar associado à prática enológica fraudulenta.

Cobre – Importante elemento da constituição mineral do vinho, menos abundante. Valores elevados originam turvação do vinho (instabilidade). Pode ter origem endógena e/ou exógena por meio de tratamento enológicos de eliminação de aromas desagradáveis.

[2] Mosto - Em vinicultura, o termo é usado para referir-se ao sumo de uvas frescas utilizado antes do processo de fermentação.

Dióxido de enxofre livre e total – Tem uma função fisiológica protetora atuando quer como agente antimicrobiano e antioxidante. Está presente no vinho sob duas formas: livre (responsável pela proteção contra a oxidação) e combinado com outros compostos. O total refere-se à soma das componentes, a sua adição constitui uma prática enológica autorizada.

Acido ascórbico/sórbico – Tem uma função fisiológica protetora atuando como agente antioxidante. Existe naturalmente no vinho, no entanto, a sua adição é uma prática enológica autorizada.

Extrato seco total – Representa a soma dos componentes que não se volatizam; constitui uma ferramenta para avaliar a quantidade do vinho pois está relacionado com o corpo/estrutura do produto.

Extrato não redutor – Refere-se ao extrato seco total diminuído dos açúcares.

Massa volúmica – Relação entre a massa e o volume de uma determinada amostra a 20ºC.

Titulo alcoométrico volúmico adquirido – Representa o teor de álcool do produto. Resulta da fermentação alcoólica dos açúcares e tem diferentes especificações dependendo do produto, casta e/ou região.

Título alcoométrico volúmico total (TAV) – Contempla, para além do atributo anterior o TAV potencial calculado com base no teor de açúcares presentes no vinho que poderiam fermentar.

Dióxido carbono – Resulta das fermentações alcoólica e malolática e daí os vinhos novos apresentarem um teor mais elevado, que vai diminuindo com a estabilização do produto. Afeta significativamente o aspeto do vinho enquanto característica organolética e permite realçar o seu aroma e frescura.

Sobrepressão – Constitui um cálculo efetuado com base no teor de CO_2, TAV adquirido e açúcares.

Metanol – Produto resultante da atividade de enzimas e em elevadas concentrações pode ser tóxico.

4.2.2.2 Descrição dos atributos a prever

Sabor – Defeito marcado / Aroma – Defeito marcado
Refere-se a possíveis defeitos presentes no vinho e que conferem notação negativa ao produto.

Sabor – Qualidade / Aroma – Qualidade
Parâmetro que expressa o equilíbrio do produto numa escala definida internamente. Cada produto possui nota mínima e máxima, definidas internamente nos Requisitos Organoléticos do Vinho Verde (ROM).

Sabor – Prova descritiva / Aroma – Prova descritiva
Características que justificam o resultado na escala de qualidade.

Sabor – Tipicidade / Aroma – Tipicidade
Reflete a origem do produto: típico da região ou não.

Aspeto – Cor / Aspeto – Limpidez
Parâmetros que traduzem as respetivas características do produto, em escalas definidas internamente no ROM.

4.2.2.3 Formato da extração

A referência da amostra, isto é, identifica um determinado vinho em específico – boletim de análise, sem revelar qualquer informação sobre o operador económico, bem como a marca correspondente. Para cada boletim de análise, existem 22 atributos referentes a análises físico-químicas e 8 atributos referentes à prova organolética.
O tipo de análise específica é dado pela "Descrição do Ensaio", sendo que se encontram discriminadas as análises físico-químicas e organoléticas, com o respetivo resultado obtido.

4.2.2.4 Qualidade dos dados

Tal como foi anteriormente referido, todos os dados extraídos encontram-se em conformidade com a legislação. Apenas são encontrados valores numéricos ou com ausência de valor, sendo uma exceção para aos atributos "Aroma – Prova Descritiva" e "Sabor – Prova descritiva", que correspondem à avaliação do provador.

4.2.2.5 Tratamento dos dados

Após uma análise ao conjunto de dados verificou-se a existência de atributos com mais de 50% de ausência de valores, estas situações correspondem às análises descritas na Tabela 6 por serem análises efetuadas apenas quando o produto se destina a exportação e estarem obrigados às regras do país a que se destina.
Devido ao grande número de ausências de valor, estes atributos não trazem qualquer benefício para a construção dos modelos, pelo que foram excluídos do conjunto de dados.

Tabela 6 Atributos com > 50% de valores em falta

Ensaio	Ácido Ascórbico	Ácido Sórbico	Cobre	Dióxido Carbono	Metanol	Sobrepressão
% de Valores em falta	0.642	0.704	0.738	0.986	0.730	0.621

Ao importar o conjunto de dados para R, decorreu a necessidade de reformular a apresentação do atributo "Descrição do Ensaio", devido a acentuação utilizada na descrição das análises não ser adequada para o desenvolvimento do trabalho no ambiente do *RStudio*.

Durante o processo de importação do conjunto de dados para o ambiente R, foi realizado a transposição dos registos que se encontravam descriminados por varias linhas, para um só registo, por referência da amostra. Obtendo assim o *conjunto de dados* a utilizar com 1339 registos e 26 atributos.

No decorrer deste processo, verificou-se a existência de ensaios em duplicado, a ocorrência destes registos corresponde a pedidos específicos do operador económico, sendo que foram consideradas apenas as análises obrigatórias pela legislação em vigor.

4.2.3 *Exploração dos dados*

De forma a verificar a distribuição dos dados, procedeu-se a visualização gráfica dos atributos a prever. Estes gráficos demonstram a frequência dos valores existentes para cada atributo objetivo em específico.
Na análise aos gráficos obtidos (Figura 9) pode-se concluir pela dispersão dos dados, que estes encontram-se muito próximos em comparação com a escala definida na ROM. Os atributos Sabor e Aroma correspondentemente à sua Qualidade e Tipicidade, têm na sua maioria valores entre 5 e 7, que pela escala na ROM, são classificados como suficientes e bons.

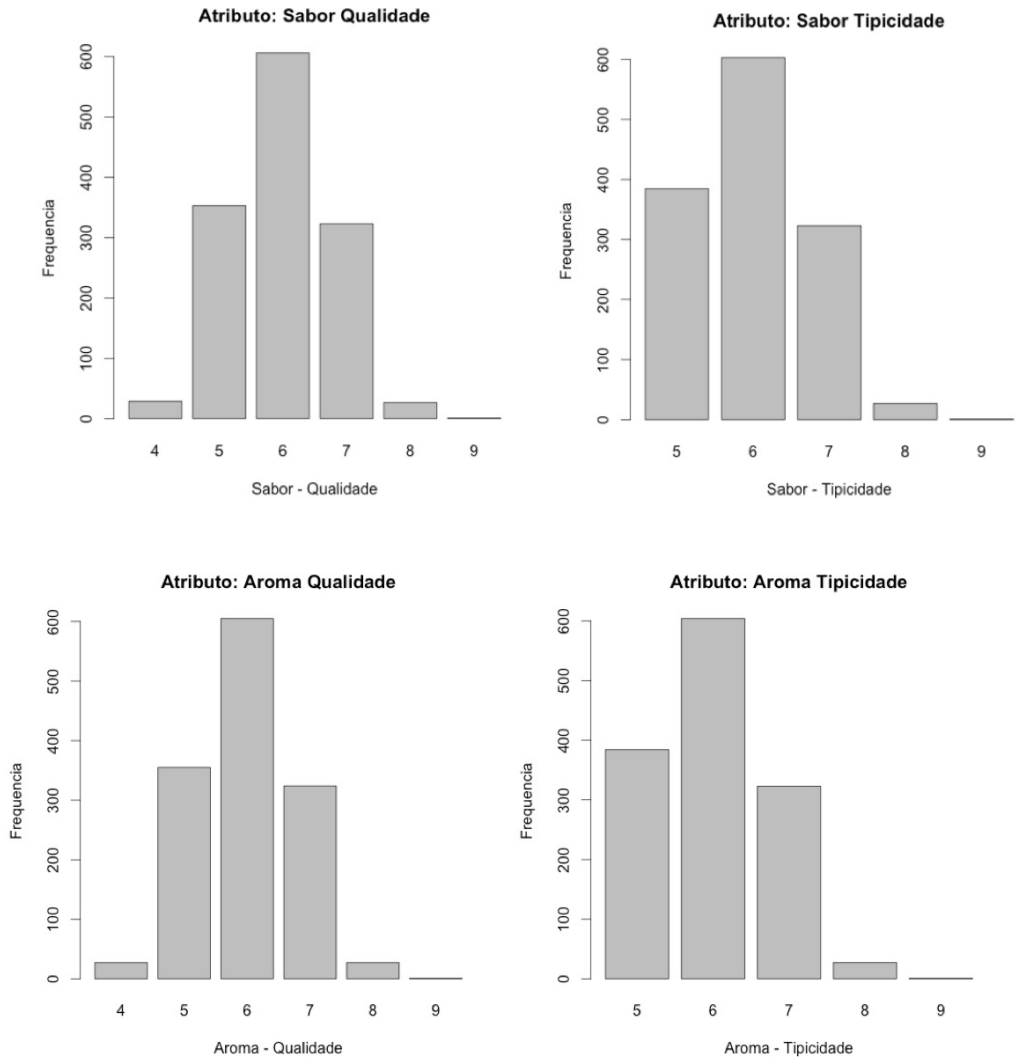

Figura 9 Frequência de valores dos atributos objetivo

É possível também verificar, a partir dos gráficos apresentados anteriormente, que não existem incongruências nos dados, quanto à sua escala definida na ROM, para valores máximos e mínimos atribuídos para cada amostra.

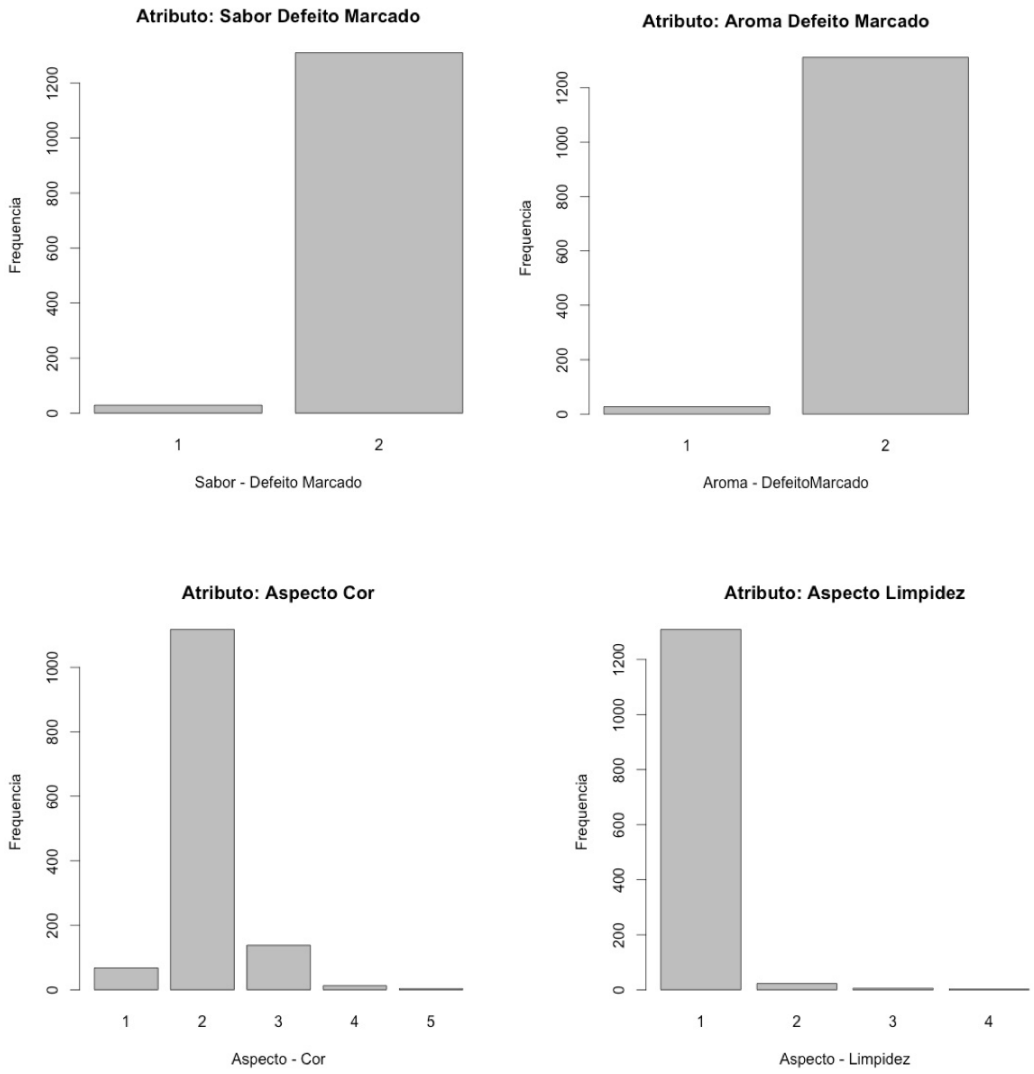

Figura 10 Frequência de valores dos atributos objetivo

Na Figura 10, podemos observar que o atributo Defeito Marcado quanto ao Sabor e Aroma, apresenta para todos os registos o valor 2, que corresponde à ausência de defeito no vinho, podendo já concluir atempadamente que as técnicas de previsão, irão se sobre-ajustar aos dados, uma vez que as classes se encontram desajustadas.
O mesmo se pode esperar quanto ao atributo Aspeto Cor e Aspeto Limpidez.

4.3 Modelação

4.3.1 *Técnicas de modelação*

Para a aplicação dos algoritmos de previsão, é necessário a preparação do conjunto de dados de treino. Esta preparação, incide sobre a metodologia *holdout* estratificado e validação cruzada (CV). No primeiro caso, o conjunto de dados é dividido em dois subconjuntos de treino e teste. O conjunto de treino corresponde 70% dos dados iniciais, e o conjunto de teste compõe os restantes 30%. Durante este processo da criação dos subconjuntos, foi tida em consideração a proporção dos valores possíveis dos atributos objetivo para que estes tenham igual representação em ambos os conjuntos de treino e teste, eliminando a possibilidade de modelos sobre-ajustados.

A utilização da metodologia de validação cruzada em conjunto com *holdout* estratificado, está presente na maioria das técnicas utilizadas, com exceção das Árvores *Bagged* e *randomForest,* nos quais apenas foi utilizado o *holdout* estratificado.

4.3.2 *Construção do modelo*

Este projeto envolveu a utilização de nove algoritmos diferentes de modelação preditiva, cada um aplicado aos oito atributos objetivo. Para a construção dos modelos utilizou-se as funções *train* e *prediction,* disponíveis no pacote *caret.* Sendo que os métodos utilizados para a implementação dos algoritmos encontram-se na seguinte lista:

- Regressão linear – método 'lm' do R
- Árvores de regressão – método 'rpart2' do pacote rpart
- Árvores *Bagged* – método 'bagging' do pacote iPred
- Árvores modelo – método 'M5' do pacote RWeka
- RandomForest – método 'randomForest'
- Redes Neuronais – método 'avNNet' do pacote nnet
- SVM – método 'svRadial'
- MARS – método 'earth'
- kNN – método 'kNN'

4.4 Avaliação

A avaliação dos modelos implementados debruçou-se na análise dos valores obtidos nas diferentes medidas de quantificação do erro. Tais como, a RMSE, MAE, o coeficiente de determinação R^2 e a correlação entre os valores previstos pelos modelos e o valor real do conjunto de teste.

4.4.1 *Atributo Sabor Qualidade*

Como se pode verificar na Tabela 7, o modelo de Árvores de Regressão apresenta o menor valor de RMSE, 50%. Em comparação com os restantes modelos a diferença é superior a 10%. Para a criação deste modelo os atributos que mais contribuíram para a previsão do atributo sabor qualidade foram: Título Alcoométrico Volúmico Adquirido e Total, Massa Volúmica, Açúcares Totais e Cloretos.

Tabela 7 Resultados Sabor Qualidade

Sabor Qualidade

	RMSE	MAE	R^2	Correlação
Regressão Linear	0.6299202		0.3929998	
Árvores de Regressão	**0.5013695**	0.5607443		0.5348615
Árvores Modelo	0.6438199	0.5141645		0.6042422
Redes Neuronais	0.6246007	0.4966302		0.6339346
SVM	0.6304296	0.4825686		
Árvores Bagged	0.6243867	0.5135276		0.6358718
Random Forest	0.6075926	0.4833023		0.6599462
MARS	0.6840666		0.3258042	
kNN	0.6450438		0.3644590	

A árvore resultante pode ser visualizada na Figura 11. Esta, possui uma profundidade de 5 níveis, e é composta por 10 nós interiores e 11 folhas.
A árvore começa por dividir as 940 observações de treino, entre as quais 330 possuem Título Alcoométrico Volúmico Adquirido inferior a 10.41, e as restantes 610 superior a 10.41. Pela observação da árvore, os vinhos com maior valor de Sabor Qualidade, são os vinhos que tem um teor alcoólico superior a 10.41.

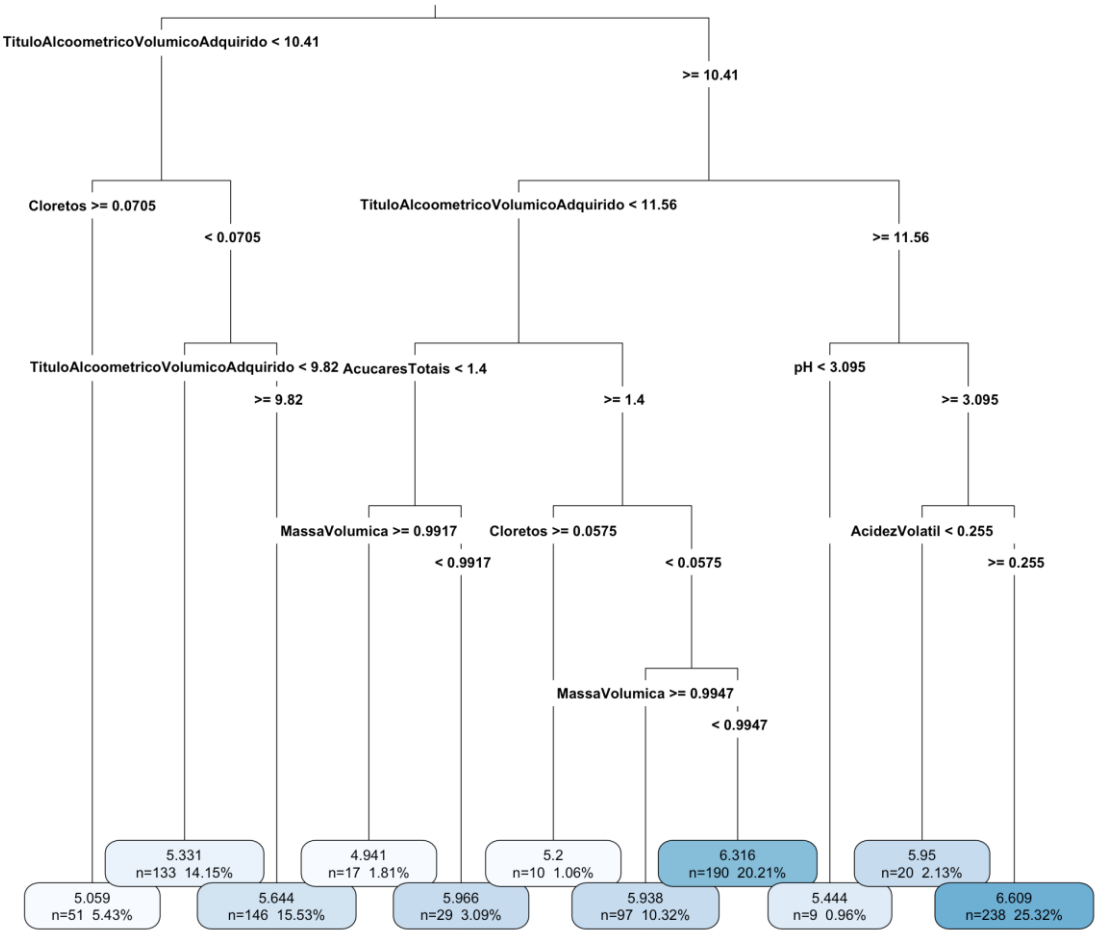

Figura 11 Modelo Árvore de Regressão para Sabor Qualidade

Comparando com os valores do atributo Sabor Qualidade do conjunto original na Figura 9 que demonstra a frequência das classes, verifica-se que o modelo é incapaz de prever os extremos desta, relativamente aos vinhos com Sabor Qualidade inferior a 5 e superior a 7, pelo facto de não haver dados suficientes para caracterizar estas classes.

4.4.2 *Atributo Sabor Tipicidade*

De acordo com os resultados descritos na Tabela 8 de resultados, para a previsão do atributo objetivo Sabor Tipicidade, o menor valor de RMSE é conseguido através do modelo baseado em *Random Forest*. No entanto, neste caso, a diferença entre os diferentes modelos não é muito significativa.

Tabela 8 Resultados Sabor Tipicidade

Sabor Tipicidade

	RMSE	MAE	R^2	Correlação
Regressão Linear	0.6437907		0.3201003	
Árvores de Regressão	0.6705284	0.5537735		0.5183856
Árvores Modelo	0.6562794	0.5381255		0.5428442
Redes Neuronais	0.6520619	0.5311756		0.5619822
SVM	0.637639	0.5112792		
Árvores Bagged	0.6308276	0.5205473		0.5861559
Random Forest	**0.6122503**	0.4969267		0.619049
MARS	0.6232310		0.3837928	
kNN	0.6604754		0.2917532	

O modelo de *random forest* foi construído utilizando 500 árvores, com um valor mínimo de 5 folhas por árvore e a cada *split* são avaliadas 5 das 15 variáveis previsoras. A partir da Figura 12, é possível observar os valores das variáveis no que diz respeito à sua importância e ao decréscimo da impureza nas decisões. As variáveis Título Alcoométrico Volúmico Adquirido, Título Alcoométrico Volúmico Total e Massa Volúmica, são as três variáveis mais importantes e que mais influenciam a pureza dos nós.

Verifica-se também que %IncMSE indica a importância das variáveis para o modelo. Neste caso, as variações no valor de Titulo Alcoométrico Adquirido têm um maior impacto no valor a ser previsto do que o atributo Massa Volúmica.

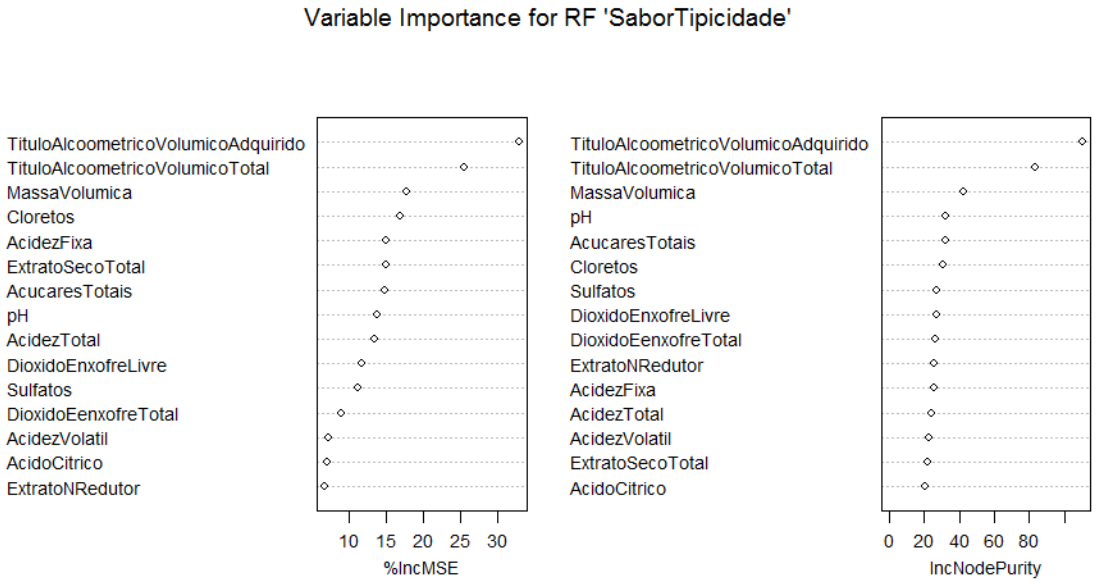

Figura 12 Atributos previsores mais contributivos RF Sabor Tipicidade

Random Forests são tipicamente tratadas como caixas negras, tornando difícil a sua interpretação no que consta à visualização das decisões tomadas. De forma a se tentar

compreender melhor o modelo e a influência das variáveis previsoras no atributo objetivo, pode-se recorrer a gráficos de dependências parciais. A Figura 13 apresenta os gráficos de dependências parciais das três variáveis mais importantes para o modelo. Ambas as variáveis Título Alcoométrico Volúmico Adquirido e Título Alcoométrico Volúmico Total, causam subidas acentuadas no valor de Sabor Tipicidade, quando o seu valor se aproxima do valor 10. Finalmente, a massa volúmica diminui o valor de Sabor Tipicidade à medida que o seu valor aumenta, no entanto, sem causar grandes variações no valor do atributo objetivo.

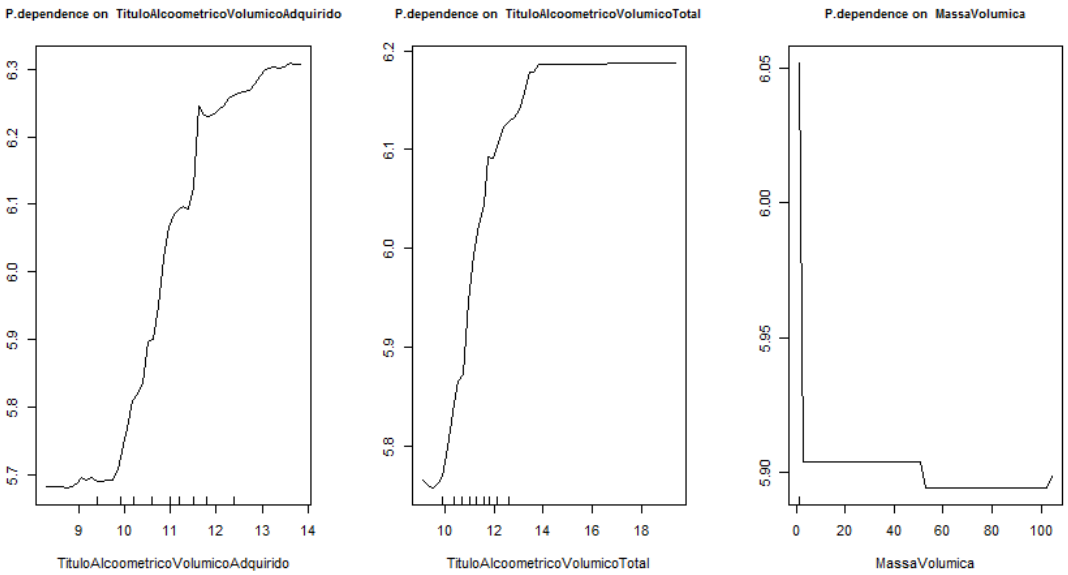

Figura 13 Dependência parcial de Sabor tipicidade das 3 variáveis mais importantes

4.4.3 *Atributo Sabor Defeito Marcado*

Na previsão do atributo Sabor Defeito Marcado, verifica-se que o erro gerado pelos modelos é praticamente idêntico, independentemente do modelo. O motivo pelo qual o valor de RMSE ser bastante baixo, deve-se ao facto dos modelos se sobre-ajustarem aos dados, pois as classes do atributo objetivo estão proporcionalmente desajustadas, para uma boa previsão. Na Tabela 9, o modelo gerado pelo algoritmo MARS é o que apresenta melhores resultados, com os atributos de maior importância, o Dióxido Enxofre Livre, Cloretos, Açúcares Totais, Extrato Seco Total, Extrato Não Redutor e pH, apesar da limitação das classes, conforme se verifica, na Figura 10 apresentada anteriormente. De notar que a diferença entre o algoritmo MARS e KNN não é significativa.

Tabela 9 Resultados Sabor Defeito Marcado

Sabor Defeito

	RMSE	MAE	R^2	Correlação
Regressão Linear	0.15222741		0.0001189218	
Árvores de Regressão	0.1475135	0.03294884		0.05636321
Árvores Modelo	0.1506643	0.03862613		0.02446201
Redes Neuronais	0.1493333	0.04329745		0.03100045
SVM	0.1402535	0.02682177		
Árvores Bagged	0	0		1
Random Forest	0.1464052	0.03950083		0.09141597
MARS	**0.1383963**		0.0007365382	
kNN	0.1393244		0.01616418	

Para a construção do modelo final, foram avaliados quais os melhores valores para os parâmetros: grau de interação e o número máximo de termos após a poda. O modelo resultante, representado na Figura 14, com menor valor RMSE foi o modelo com os valores 2 e 3 respetivamente aos parâmetros mencionados. Salienta-se a presença do previsor Açúcares Totais nos dois termos usados para a previsão, evidenciando a sua importância.

```
                                                    coefficients
(Intercept)                                           1.98374250
h(AcucaresTotais-2.5) * h(Cloretos-0.111)            -0.16687658
h(AcucaresTotais-2.5) * h(TituloAlcoometricoVolumicoTotal-13)  -0.00921279

Selected 3 of 24 terms, and 3 of 15 predictors
Termination condition: RSq changed by less than 0.001 at 24 terms
Importance: AcucaresTotais, Cloretos, TituloAlcoometricoVolumicoTotal, AcidezFixa-unused, ...
Number of terms at each degree of interaction: 1 0 2
GCV 0.01978831    RSS 18.32474    GRSq 0.09780734    RSq 0.1074102
```

Figura 14 Resultado do modelo para Sabor Defeito Marcado.

O modelo removeu 12 das variáveis previsoras, por não considerar que as variáveis tenham efeito na previsão da classe. A importância de todas as variáveis escolhidas pelo modelo, segundo o critério *Generalized Cross Validation* (GCV), pode ser vista na Figura 15. Confirma-se a contribuição de Açúcares Totais e também de Cloretos com os valores GCV.

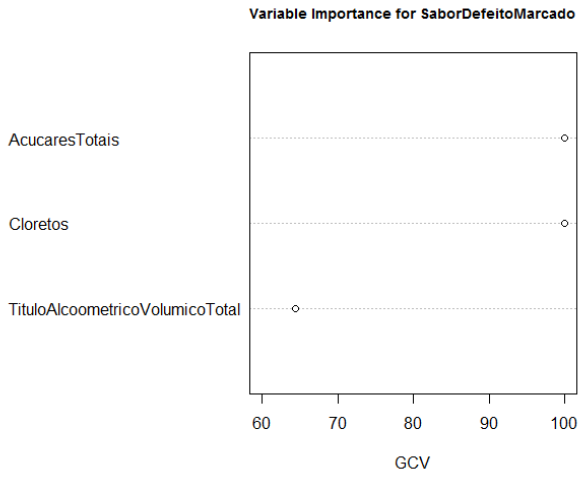

Figura 15 Importância variáveis previsoras do modelo para Saber Defeito Marcado

4.4.4 *Atributo Aspeto Limpidez*

Os resultados obtidos pelos modelos para a previsão do atributo Aspeto Limpidez, apresentam uma boa previsão, com erros na ordem dos 15%, representados na Tabela 10.
De entre os modelos aplicados, o modelo SVM é o que apresenta um valor de RMSE mais baixo, sendo que para a previsão do atributo objetivo as variáveis que mais contribuíram estão representadas na Figura 16, sendo estas, os Açúcares Totais, o Extrato não Redutor e o Extrato Seco total.

Tabela 10 Resultados Aspeto Limpidez

Aspeto Limpidez

	RMSE	MAE	R^2	Correlação
Regressão Linear	0.147930826		0.002077349	
Árvores de Regressão	0.1541431	0.03746807		0.04608008
Árvores Modelo	0.1496333	0.04935857		0.07703472
Redes Neuronais	0.1504612	0.06543165		0.1720487
SVM	**0.1446004**	0.03593114		
Árvores Bagged	0.1457088	0.0402709		0.1365202
Random Forest	0.151232	0.0417946		0.1396065
MARS	0.2145908		0.0660987976	
kNN	0.148622100		0.001253057	

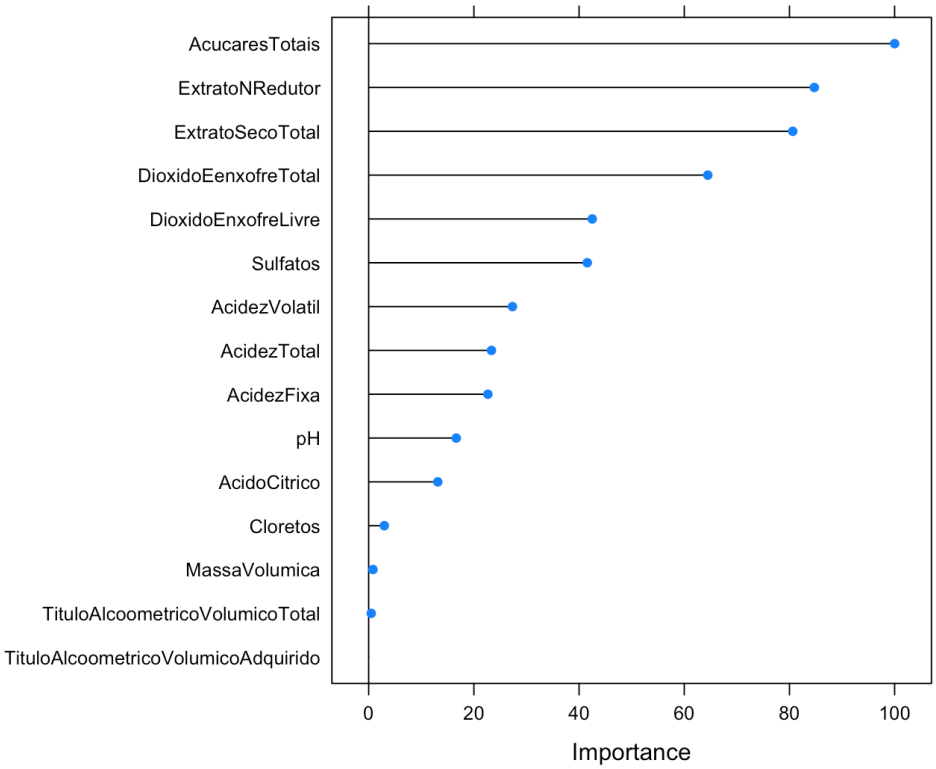

Figura 16 Atributos Previsores mais contributivos para o modelo de regressão linear

4.4.5 *Atributo Aspeto Cor*

Na previsão do atributo objetivo Aspeto Cor, o modelo com melhor prestação foi o de árvores modelo, com um valor RMSE abaixo dos 40% que se pode confirmar através da Tabela 11.

Tabela 11 Resultados Aspeto Cor

Aspeto Cor

	RMSE	MAE	R^2	Correlação
Regressão Linear	1.069324e-16		1	
Árvores de Regressão	0.4486863	0.2385		0.3132267
Árvores Modelo	**0.3986972**	0.2401998		0.4632833
Redes Neuronais	0.4354993	0.2598485		0.3862617
SVM	0.4091933	0.2113831		
Árvores Bagged	0.4049802	0.2338479		0.4354536
Random Forest	0.4038452	0.2352536		0.4405616
MARS	0.4155580		0.20571972	
kNN	0.4239360		0.1233836	

A árvore modelo foi construída utilizando o algoritmo M5P, resultando numa árvore podada de 11 de altura, composta por 87 nós totais, entre os quais 44 são os modelos lineares finais.

Neste modelo, as variáveis previsoras mais relevantes para o atributo objetivo são Dióxido Enxofre Livre e Título Alcoométrico Volúmico Total, uma vez que se encontram nas primeiras divisões da árvore. A Figura 17 retrata uma representação gráfica parcial da árvore modelo resultante, na qual se destaca o último modelo linear a ser construído, o modelo LM44.

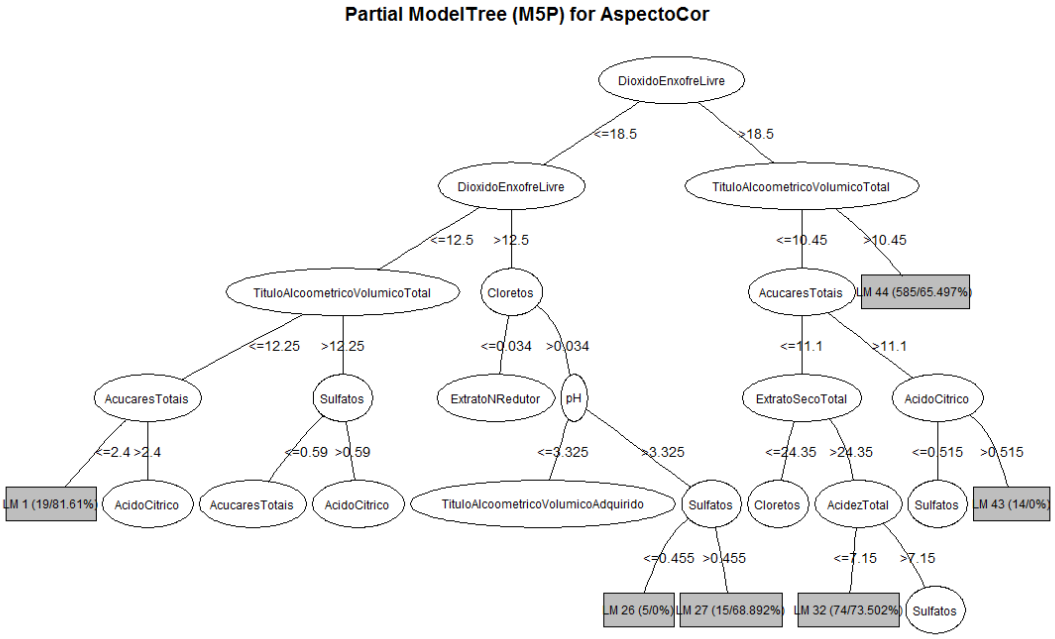

Figura 17 Árvore modelo parcial do atributo objetivo Aspeto Cor

Recaíram no modelo LM44, 585 observações das 938 totais do conjunto de treino, representando 62% do total. Para tal, o modelo avalia em primeiro lugar: se o valor de Dióxido Enxofre Livre é superior a 18.5 e se o Título Alcoométrico Volúmico Total é superior a 10.45. Por exemplo, pode-se observar pela equação do modelo LM44, que uma unidade de Acidez Fixa causará um aumento em 0.2286 no valor do Aspeto Cor. Em comparação, o modelo LM32 é o segundo modelo com mais observações, ao albergar 74 instâncias que obedecem a um conjunto de regras mais extenso.

$$AspectoCor = 0.2286 \times AcidezFixa - 0.1693 \times AcidezTotal + 0.5219 \times AcidezVolatil - 0.3534 \times AcidoCitrico + 0.0173 \times AcucaresTotais + 0.0675 \times Cloretos - 0.0021 \times DioxidoEnxofreLivre + 0.0014 \times DioxidoEnxforeTotal + 0.0228 \times ExtratoNRedutor - 0.0105 \times ExtratosecoTotal + 0.0003 \times MassaVolumica + 0.2865 \times TituloAlcoometricoVolumicoAdquirido - 0.2966 \times TituloAlcoometricoVolumicoTotal + 0.2518 \times pH + 0.6865$$

Comparando algumas estatísticas dos valores de Aspeto Cor previstos com os de conjunto de teste, representados na Tabela 12, pode-se observar que o modelo apresenta valores semelhantes, exceto a incapacidade de prever *outliers* de vinhos com valores de Aspeto Cor acima de 3.

Tabela 12 Comparação entre os valores previsto e o conjunto de teste - Árvores Modelo

	Min.	*1º Quartil*	*Mediana*	*Média*	*3º Quartil*	*Máx.*
Previsão	1.206	1.980	2.069	2.076	2.144	3.020
Conj. Teste	1.000	2.000	2.000	2.085	2.000	5.000

4.4.6 *Atributo Aroma Qualidade*

Analisando a Tabela 13 de resultados, para a previsão do atributo objetivo Sabor Tipicidade, o menor valor de RMSE é conseguido através do modelo baseado em *Random Forest*, pelo que, não existem diferencias muito significativas para os restantes modelos.

Tabela 13 Resultados Aroma Qualidade

Aroma Qualidade

	RMSE	MAE	R^2	Correlação
Regressão Linear	0.6444851		0.3492621	
Árvores de Regressão	0.6792078	0.5439495		0.5345993
Árvores Modelo	0.6526028	0.5171783		0.5763986
Redes Neuronais	0.6300256	0.5113381		0.6148286
SVM	0.6432871	0.5046999		
Árvores Bagged	0.6405109	0.5243484		0.6080558
Random Forest	**0.6235602**	0.4927455		0.634143
MARS	0.6790788		0.3351021	
kNN	0.6531778		0.3539443	

O modelo de *random forest* construído é bastante idêntico ao já apresentado para o atributo Sabor Tipicidade, utilizando 500 árvores, com um valor mínimo de 5 folhas por árvore e a cada *split* são avaliadas 5 das 15 variáveis previsoras. As variáveis Título Alcoométrico Volúmico Adquirido, Titulo Alcoométrico Volúmico Total e Açúcares totais, são as mais importantes e que mais influenciam a pureza dos nós, como se demonstra na Figura 18.

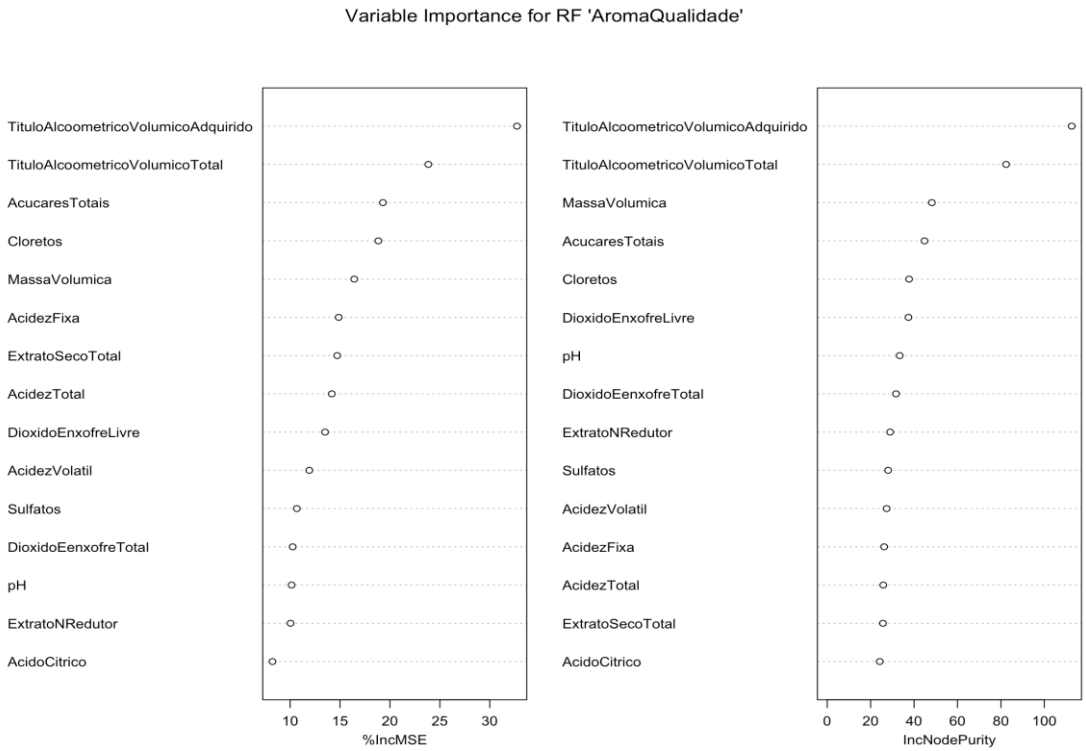

Figura 18 Atributos previsores mais contributivos RF Aroma Qualidade

Relativamente às dependências parciais das variáveis mais importantes para o modelo, conforme a Figura 19, ambas as variáveis Título alcoométrico Volúmico Adquirido e Título Alcoométrico Volúmico Total, causam subidas acentuadas no valor de Aroma Qualidade, à medida que este se aproxima do valor 10. Quanto aos Açúcares totais, apesar de aparentar uma curva acentuada, não têm grande influência na alteração do valor a prever.

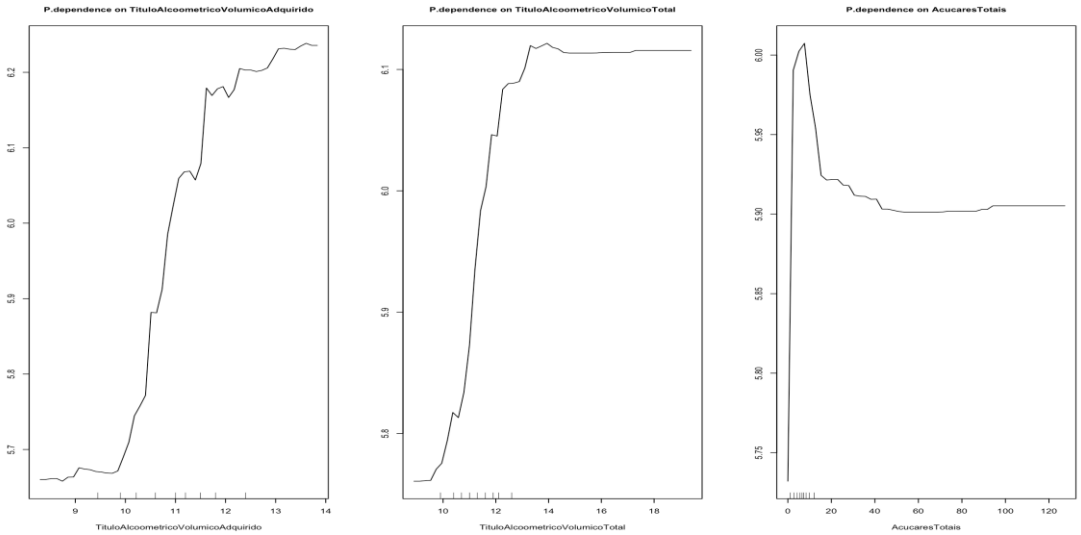

Figura 19 Dependência parcial de Aroma Qualidade das 3 variáveis mais importantes

4.4.7 *Atributo Aroma Tipicidade*

Para a classe Aroma Tipicidade, à semelhança de algumas previsões anteriores, a diferença entre o valor de RMSE dos diferentes modelos não é relevante. Neste caso, o modelo *random forest,* produziu um erro de 61%, conforme apresentado na Tabela 14.

Tabela 14 Resultados Aroma Tipicidade

Aroma Tipicidade

	RMSE	MAE	R^2	Correlação
Regressão Linear	0.6356014		0.3405919	
Árvores de Regressão	0.6591571	0.5466123		0.5425682
Árvores Modelo	0.6478962	0.5352023		0.5615644
Redes Neuronais	0.6338559	0.5196156		0.5876241
SVM	0.6265802	0.5046999		
Árvores Bagged	0.6299917	0.5237287		0.5939924
Random Forest	**0.6127388**	0.5026176		0.6221556
MARS	0.6281474		0.3638398	
kNN	0.6427720		0.3321877	

A *random forest* desenvolvida, é constituída por 500 árvores, com um valor mínimo de 5 folhas por árvore e a cada *split* são avaliadas 5 das 15 variáveis previsoras. Para este modelo, as variáveis que mais influência têm na previsão da classe objetivo são: Título Alcoométrico Volúmico Adquirido, Título Alcoométrico Volúmico Total e Cloretos. No entanto, é relevante realçar que a importância dos Cloretos é relativa, uma vez que os subsequentes 5 atributos possuem valores aproximados. A importância destas variáveis encontra-se representada na Figura 20.

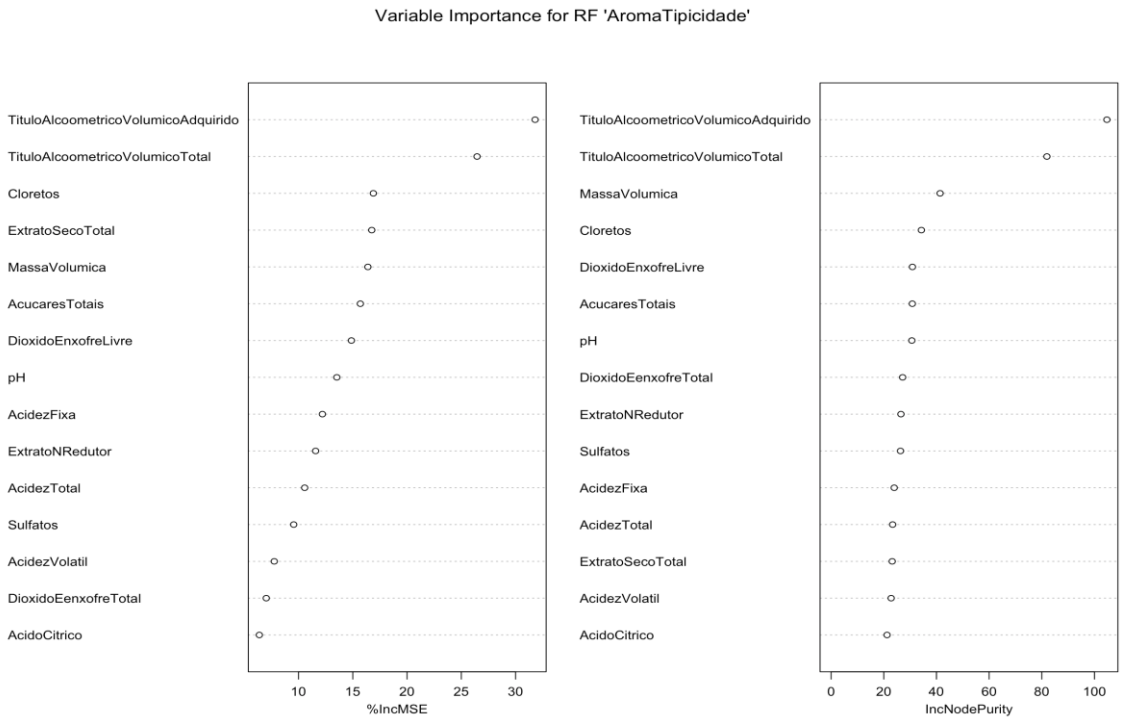

Figura 20 Atributos previsores mais contributivos RF Aroma Tipicidade

As duas variáveis mais importantes, Título Alcoométrico Volúmico Adquirido e Total, são, as mesmas consideradas pelo modelo usado na previsão do Sabor Tipicidade. O mesmo pode ser constatado nos gráficos da dependência parcial da classe objetivo com estes atributos, representados na Figura 21. As duas primeiras variáveis apresentam curvas semelhantes às apresentadas no modelo Sabor Tipicidade, o que possivelmente quererá dizer que as mesmas características que influenciem o aroma de um vinho estarão relacionadas com o sabor do mesmo.

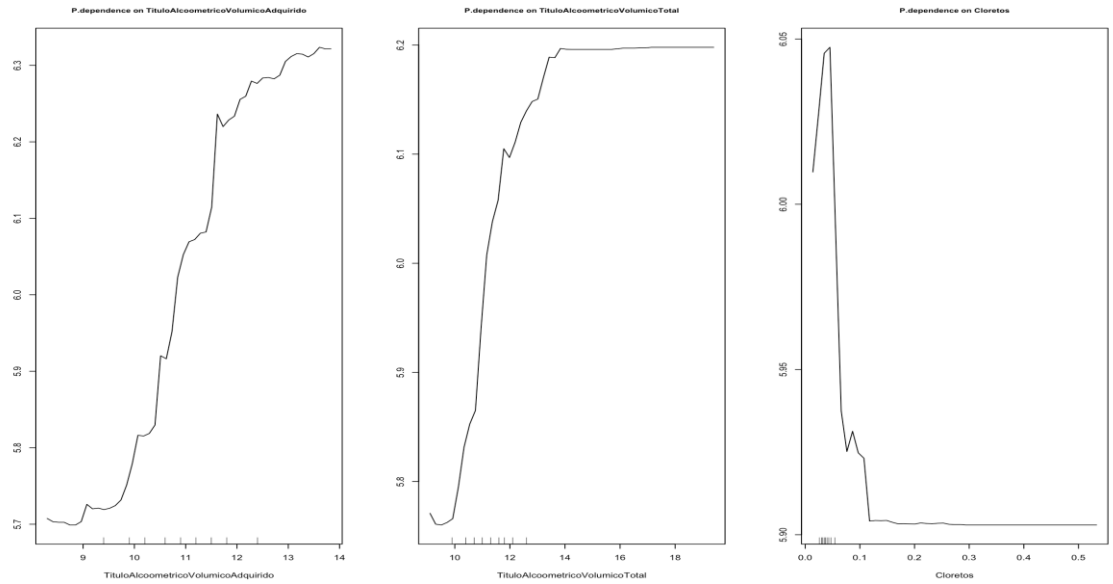

Figura 21 Dependência parcial de Aroma Tipicidade das 3 variáveis mais importantes

4.4.8 *Atributo Aroma Defeito Marcado*

Para a previsão do atributo Aroma defeito marcado, o modelo kNN, Tabela 15, foi o que devolveu melhores resultados, utilizando o atributo Cloretos como a variável previsora mais importante, conforme Figura 22.

Tabela 15 Resultados Aroma Defeito

Aroma Defeito

	RMSE	MAE	R^2	Correlação
Regressão Linear	1.440414e-01		6.475471e-05	
Árvores de Regressão	0.150302	0.03515765		-0.03276465
Árvores Modelo	0.1407283	0.0350264		0.03076226
Redes Neuronais	0.1428984	0.04014126		-0.00533830
SVM	0.1311012	0.02696283		
Árvores Bagged	0	0		1
Random Forest	0.1402371	0.03753084		0.05819439
MARS	0.1354069		0.0009050399	
kNN	**0.13072964**		0.01400278	

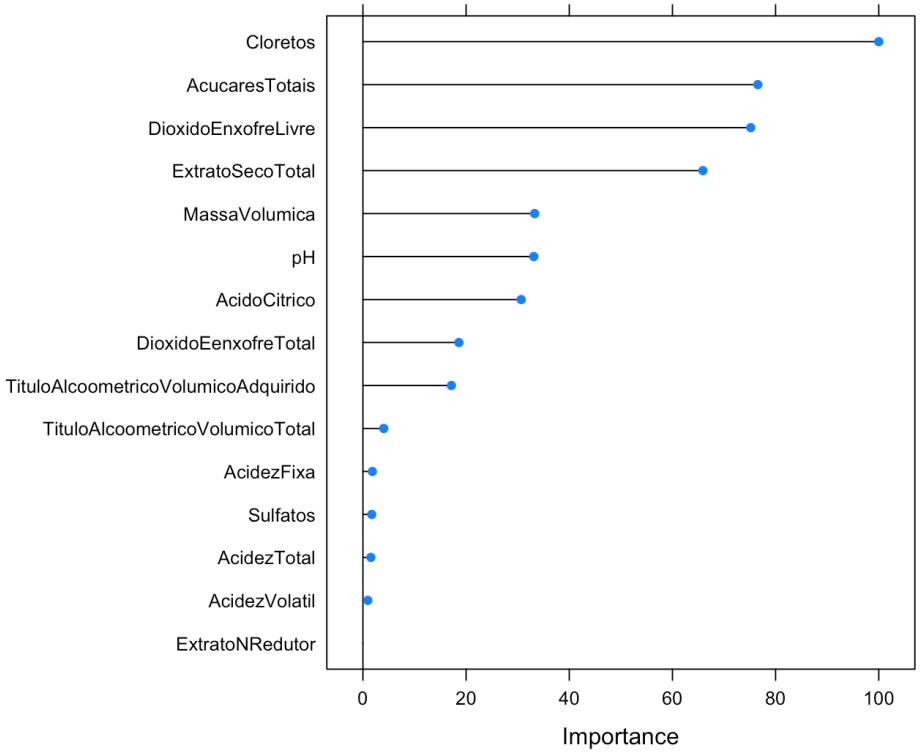

Figura 22 Atributos previsores mais contributivos kNN Aroma Defeito Marcado

Demonstra-se pela Figura 23, o comportamento do modelo criado, que através do atributo Cloretos quantos mais vizinhos são considerados para a previsão do Aroma Defeito Marcado, o erro tende a baixar para valores abaixo dos 10%. A ocorrência deste comportamento deve-se ao facto dos valores do atributo Cloretos se encontrarem dispersos no mesmo intervalo de valores, e o atributo objetivo ter 98% dos registos na mesma classe.

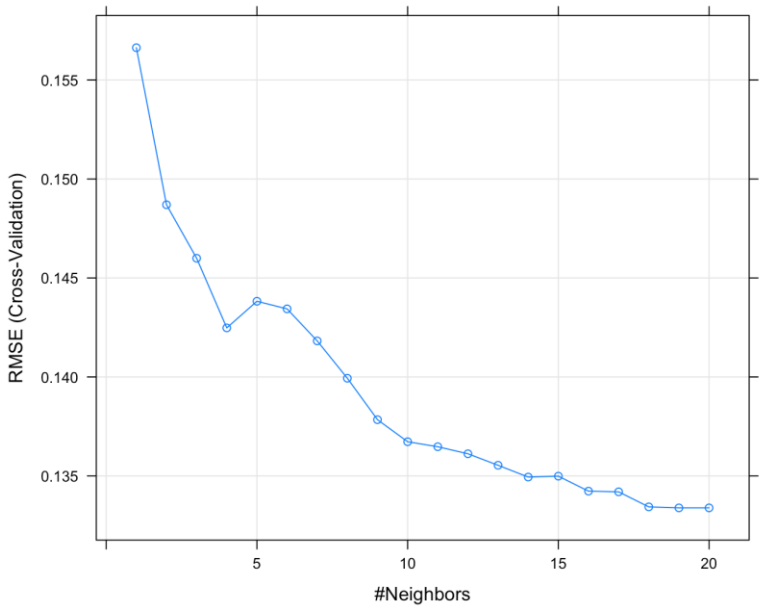

Figura 23 Relação entre RMSE e o número de vizinhos mais próximos

4.5 Discussão dos Resultados

Dos vários modelos desenvolvidos para caracterização dos diferentes atributos objectivo os que apresentaram resultados mais fiáveis dizem respeito à previsão dos atributos Aspecto Limpidez (modelo SVM, RMSE=14.5%), Sabor defeito Marcado (modelo MARS, RMSE=13%), e Aroma Defeito Marcado (modelo KNN, RMSE=13%). Para os restantes atributos, Sabor Qualidade/Tipicidade, Aroma Qualidade/Tipicidade o RMSE é cerca de 60% e para o atributo Aspecto Cor cerca de 40%. Estes resultados devem-se sobretudo à falta de dados para caracterização das classes extremas (classes 4, 8 e 9), como pode ser observado nos histogramas da figura 9.

A importância dos atributos previsores expressa nas várias técnicas usadas apontam como os atributos mais informativos o álcool, cloretos, pH e açucares. Estes resultados confirmam a teoria enológica de que um aumento do teor álcool (considerada a variável mais relevante) tende a resultar num vinho de qualidade superior. Além disso os níveis de açúcar são importantes para o vinho branco.

5 Conclusões e Trabalho Futuro

O Vinho Verde é um produto de excelência do nosso país, que muito tem contribuído pela imagem de Portugal internacionalmente. Para além do relevante contributo na balança agro-comercial é um bem cujo valor é, em regra, 100% nacional. Todo e qualquer trabalho que possa ajudar e contribuir para o sucesso e crescimento deste bem, revela por si só um interesse especial.

O problema deste projeto enquadra-se no tempo e nos custos despendidos do processo de certificação do Vinho Verde, devido à necessidade de realização de análises físico-químicas e organoléticas. O objetivo deste projeto consistiu em explorar uma possível relação teórica entre os atributos resultantes das análises físico-químicas, de forma a prever as características organoléticas (sabor, aroma e aspeto) do vinho. Para tal, foram aplicadas várias técnicas de *DM* que conduzem à descoberta de conhecimento, tendo sido desenvolvido um projeto na linguagem R, em que foram aplicados nove algoritmos distintos de previsão. A escolha destes algoritmos, teve um cuidado especial, pois na sua maioria os dados são compostos por valores numéricos contínuos e o atributo a prever ordinal.

A implementação destes algoritmos, resultou na impossibilidade de considerar um só modelo como uma solução fiável para este projeto. No sentido em que os atributos como Sabor Qualidade/Tipicidade e Aroma Qualidade/Tipicidade com mais interesse na sua previsão, o valor do erro encontra-se em cerca de 60% para todos os modelos. Destacando-se os modelos criados com base em *random forests,* que obtiveram na sua maioria o valor mais baixo.

Como perspetiva de trabalho futuro, é necessário definir e implementar um processo que verifique, se os resultados dos algoritmos de aceitação estão em conformidade ou não, e definir os métodos de auditoria.

A procura por outras análises físico-químicas, para além das obrigatórias, e que sejam potencialmente mais contributivas para os algoritmos, também é um ponto a ser analisado. Pois, verificou-se que no decorrer deste projeto, as análises como Título Alcoométrico, Açúcares e Cloretos, revelaram-se como os atributos com mais peso na previsão dos modelos, contudo, não são suficientemente discriminatórios para uma previsão fiável. Sobre este prisma, é necessário ponderar os custos associados e a valorização que a empresa dá na otimização e aumento de produtividade.

Referências

(Nicola et al. 2012) Nicola, S., Ferreira, E. P., & Ferreira, J. J. P.(2012), Int. J. Info. Tech. Dec. Mak. 11, 661-703.

(Osterwalder 2004) Osterwalder, A., & Pigneur, Y. (2004). An ontology for e-business models. Value creation from e-business models, 65-97.

(Zeithmal 1988) Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: a means-end model and synthesis of evidence. The Journal of marketing, 2-22.

(Egyed and Boehm 1998) Boehm, B., & Egyed, A. (1998, August). WinWin requirements negotiation processes: A multi-project analysis. In 5th International Conference on Software Processes (pp. 125-136).

(Woodall 2003) Woodall, T. (2003). Conceptualising'value for the customer': an attributional, structural and dispositional analysis. Academy of marketing science review, 2003, 1.

[Nicola et al. 2014] Nicola, S., Ferreira, E. P., & Ferreira, J. J. P. (2014). A Quantitative Model for Decomposing & Assessing the Value for the Customer. Journal of Innovation Management, 2(1), 104-138.

(Fayyad et al. 1996) Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

(Chapman et al. 2000) Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

(Cover et al. 1967) Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

(Azevedo 2008) Azevedo, A. I. R. L. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.

(Chapman and Hall 1984) Classification and Regression Trees by L. Breiman, J.H. Friedman, C.J. Stone, and R.A. Olshen.

(Wang and Witten 1997) Induction of model trees for predicting continuous classes, Proceedings of the Poster Papers of the European Conference on Machine Learning by Y. Wang and I.H.Witten (1997).

(Breiman et al. 1984) Classification and Regression Trees by L. Breiman, J.H. Friedman, C.J. Stone, and R.A. Olshen (Chapman & Hall, 1984)

(Gunn 1998) Gunn S. R. Support Vector Machine for Classification and Regression. Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science.

(Junior, 2010) Junior, G. M. O. Máquina de vetores suporte: estudo e análise de parâmetros para otimização de resultado, 2010. Trabalho de Graduação em Ciência da Computação, Univ. Federal de Pernambuco

(Gonçalves 2010) Gonçalves A. R. Maquinas de Vetores suporte.

(Rebelo 2008) Rebelo, L. D. T., Avaliação automática do resultado estético do tratamento conservador do cancro de mama. Faculdade de Engenharia da Universidade do Porto

(Lorena and Carvalho 2007) Lorena, A. C., & Carvalho, A. C. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, *14*(2), 43-67.

(Efron and Tibshirani 1986) Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, 54-75.

(Tan et al. 2005) Tan, W. P. N., Steinbach, M., & Kumar, V. (2005). General approach to solving a classification problem. Introduction to Data Mining.

(Kvalseth 1985) Kvalseth T (1985). "Cautionary Note About R2." American Statistician, 39(4), 279–285.

(Cortez et al. 2009) Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), pp. 547-553.

(Buratti et al. 2007) Buratti, S., Ballabio, D., Benedetti, S., & Cosio, M. S. (2007). Prediction of Italian red wine sensorial descriptors from electronic nose, electronic tongue and spectrophotometric measurements by means of Genetic Algorithm regression models. *Food Chemistry*, *100*(1), 211-218.

(Tian and Pang 2010) Tian, H., & Pang, Q. (2010). Data mining application for upgrading quality of wine production. In The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding pp. 109-111.

(Hosu et al. 2014) Hosu, A., Cristea, V. M., & Cimpoiu, C. (2014). Analysis of total phenolic, flavonoids, anthocyanins and tannins content in Romanian red wines: prediction of antioxidant activities and classification of wines using artificial neural networks. *Food chemistry*, *150*, 113-118.

(Omatu et al. 2015) Omatu, S., Yano, M., & Ikeda, Y. (2015, April). Smell classification of wines by the learning vector quantization method. In Proceedings of the 30th Annual ACM Symposium on Applied Computing (pp. 195-200). ACM.

(Therneau et al. 2015) Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. https://CRAN.R-project.org/package=rpart

(Kuhn et al. 2016) Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z.,
Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y. & Candan, C. (2016). caret: Classification and Regression Training. R package version 6.0-71. https://CRAN.R-

project.org/package=caret

(Peters et al. 2015) Peters, A. & Hothorn, T. (2015). ipred: Improved Predictors. R package version 0.9-5. https://CRAN.R-project.org/package=ipred

(Liaw and Wienere 2002) Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

(Venables and Ripley 2002) Venables, W. & Ripley, B. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

(Millborrow 2016) Milborrow, S. (2016). earth: Multivariate Adaptive Regression Splines. R package version 4.4.6. https://CRAN.R-project.org/package=earth

(Hornik et al. 2009) Hornik, K., Buchta, C. & Zeileis A. (2009). "Open-Source Machine Learning: R Meets Weka." *Computational Statistics*, **24**(2), pp. 225–232. doi: 10.1007/s00180-008-0119-7

(Friedman 1991) Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1-67.

(Wilamowski 2003) Wilamowski, B. M. (2003, December). Neural network architectures and learning. In Industrial Technology, 2003 IEEE International Conference on (Vol. 1, pp. TU1-T12). IEEE.

(Gershenson 2003) Gershenson, C. (2003). Artificial neural networks for beginners. *arXiv preprint cs/0308031*.

(WU et al. 2008) Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.

(Zhang and Goh 2016) Zhang, W., & Goh, A. T. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, *7*(1), 45-52.

Referências Web

[1] O setor do vinho, http://www.viniportugal.pt/OSector, Fevereiro 2016

[2] CVRVV, http://portal.vinhoverde.pt, Fevereiro 2016

[3] ROM, http://portal.vinhoverde.pt/pt/documentacao, Janeiro 2016

[4] Entidades Certificadoras, http://www.ivv.min-agricultura.pt/np4/212.html, Fevereiro 2016

[5] CRISP-DM, https://the-modeling-agency.com/crisp-dm.pdf, Janeiro 2016

[6] *Performance Evaluation of Learning Algorithms*, www.icmla-conference.org/icmla11/PE_Tutorial.pdf, Fevereiro 2016

[7] *R Core Team, R: A Language and Environment for Statistical Computing*, http://www.R-project.org, Fevereiro 2016

[8] *The Comprehensive R Archive Network*, https://cran.r-project.org, Fevereiro 2016

[9] *RStudio*, https://www.rstudio.com, Fevereiro 2016