INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Sistema de previsão de resultados de jogos de futebol através de técnicas de Data Mining

ÂNGELO JOÃO LOUREIRO PINTO
Setembro de 2018

POLITÉCNICO
DO PORTO

isep Instituto Superior de Engenharia do Porto

Instituto Superior de Engenharia do Porto ⊙ Rua Dr. António Bernardino de Almeida, 431 ⊙ 4249-015 Porto
Tel. +351 228 340 500 ⊙ Fax +351 228 321 159 ⊙ mail@isep.ipp.pt ⊙ www.isep.ipp.pt

INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

isep

Sistema de previsão de resultados de jogos de futebol através de técnicas de Data Mining

ÂNGELO JOÃO LOUREIRO PINTO
Setembro de 2018

INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA

isep




Sistema de previsão de resultados de jogos de futebol através de técnicas de Data Mining

ÂNGELO JOÃO LOUREIRO PINTO
Setembro de 2018

POLITÉCNICO DO PORTO

Instituto Superior de Engenharia do Porto

Sistema de previsão de resultados de jogos de futebol através de técnicas de Data Mining

Ângelo João Loureiro Pinto

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

Orientadora: Professora Doutora Fátima Rodrigues

Júri:

Presidente:

Vogais:

Porto, Setembro de 2018

Resumo

O futebol é um dos desportos com mais impacto em todo o mundo e tem muitos adeptos em Portugal. Os negócios há volta deste desporto têm aumentado ao longo do tempo e nos últimos anos as apostas em jogos de futebol têm crescido. Desde 2015 foram licenciadas em Portugal várias casas de apostas, tendo estas milhares de apostadores, e a publicidade a estas casas tem também crescido, prevendo-se um crescimento deste mercado. No entanto é difícil obter lucros ao fazer apostas desportivas, devido à dificuldade em prever o resultado final dos jogos. É por isso necessário criar ferramentas para prever os resultados de jogos e ajudar os apostadores a não terem prejuízos.

Uma das formas de prever resultados de jogos de futebol é usar os dados estatísticos de jogos anteriores, como o número de golos, remates das equipas, entre outros. Utilizando esses dados em conjunto com técnicas de *Data Mining* é possível fazer a previsão dos jogos.

Esta tese utilizou técnicas de *Data Mining*, testando vários algoritmos, para criar um modelo de previsão capaz de prever os jogos da Liga Inglesa. O modelo desenvolvido foi integrado num Sistema de Apoio à Decisão que indica ao utilizador as apostas que deve fazer, bem como se as apostas têm um risco baixo ou elevado.

Palavras-chave: Data Mining, Apostas desportivas, Sistema de Apoio à Decisão, Classificação, Futebol

Abstract

Football is one of the sports with most impact around the world, including in Portugal. Businesses related to football have been growing and over the past years bets on football matches have increased. Since 2015 many bookmakers have been licensed in Portugal, serving thousands of gamblers. Marketing on these bookmakers has been rising and predictions say that this market will grow. However, it is difficult to get profits when betting on sports, due to the difficulty in predicting the final result of matches. So it is necessary to create tools to predict results and help gamblers not lose money.

One of the ways to predict the result of football matches is using statistics from previous matches, such as the number of goals or the shots made by the teams. Using these data, together with Data Mining techniques, it is possible to predict results.

In this thesis Data Mining was used, testing several algorithms to create a prediction model of results of the English Premier League. This model was integrated in a Decision Support System that tells the user which bets should be made, as well as if the bets have a low or high risk.

Keywords: Data Mining, Sports betting, Decision Support System, Classification, Football

Agradecimentos

Agradeço a todas as pessoas que me ajudaram neste projeto e que de alguma forma contribuíram para o melhorar.

Agradeço à minha orientadora, a professora Fátima Rodrigues, pela orientação dada e por toda a ajuda dada ao longo do projeto.

Quero agradecer também ao Tiago Coelho por me ter dado a ideia para esta tese.

Índice

Lista de Figuras

Lista de Tabelas

Glossário

Época Período de tempo entre Agosto e Maio onde são disputados os jogos de um campeonato de futebol.

KNN K-nearest Neighbor

Odd Cotação dada a um jogo, representa o retorno monetário quando se faz uma aposta e reflete a probabilidade de ocorrência do evento. Quanto menor for a *odd* maior é a probabilidade de ocorrência do evento.

Jornada Conjunto de jogos disputados numa semana para um campeonato. Num campeonato com 20 equipas, em cada semana há 10 jogos, esses 10 jogos correspondem a uma jornada. Uma equipa tem de jogar duas vezes com cada equipa ao longo de uma época, num campeonato com 20 equipas tem de fazer um total de 38 jogos, por isso o número total de jornadas numa época é de 38.

RLM Regressão Logística Multinomial

RNA Redes Neuronais Artificiais

RF Random Forest

SAD Sistema de Apoio à Decisão

SVM Support Vector Machines

Xgboost Extreme Gradient Boosting

1 Introdução

Neste capítulo é feita uma abordagem ao contexto e problema do projeto, são explicados os objetivos da tese e é descrita a análise de valor, resultados esperados e abordagem a ser feita, terminando com a descrição da organização do documento.

1.1. Contexto

O futebol é um dos desportos mais importantes em todo o mundo, sendo muito importante a nível europeu. Nas últimas décadas o desporto cresceu bastante, tendo-se tornado um negócio que movimenta milhões de euros. Tal como já tinha acontecido com outros desportos, surgiram as apostas em jogos de futebol. Quando se iniciaram as apostas no futebol, em Portugal não existiam casas de apostas e por isso a maioria dos apostadores eram sobretudo jovens, não havendo um número relevante de apostadores. A partir de 2015 foi legalizada em Portugal a primeira casa de apostas física, o Placard. Este fator veio alavancar muito as apostas em eventos desportivos, muito por culpa da publicidade feita. De acordo com os dados disponíveis [1], no primeiro ano do Placard mais de 9% dos portugueses apostaram 300 milhões de euros, um valor que demonstra bem o grande valor comercial deste negócio. Dentro das apostas feitas 78% foram em jogos de futebol, sendo este o desporto que mais dinheiro movimenta. Atualmente, para além do Placard, existem mais 6 casas de apostas online legalizadas. O *marketing* feito por estas casas de apostas tem vindo a crescer e espera-se que o número de apostas continue a subir.

1.2. Problema

Apostar em eventos desportivos e obter lucros é muito difícil, especialmente quando se trata de futebol. Para cada jogo de futebol há vários tipos de apostas, mas as que são feitas mais frequentemente são as “1x2”, onde se aposta no resultado de um jogo, indicando se uma dada equipa vai ganhar, empatar ou perder o jogo. A antecipação dos resultados é uma tarefa muito complexa devido à grande quantidade de fatores que podem influenciar os jogos. Devido às características do próprio jogo, é possível uma equipa claramente superior a outra perder o jogo, o que dificulta ainda mais qualquer tipo de aposta. A imprevisibilidade do jogo torna difícil apostar sem fazer qualquer tipo de análise aos jogos. Por isso muitos apostadores têm a necessidade de ter algum apoio quando fazem apostas para não terem prejuízos. Uma forma de os ajudar seria com uma ferramenta que fizesse a previsão dos resultados dos jogos e lhes indicasse qual o resultado mais provável.

Atualmente existem sites que oferecem de forma gratuita dicas de apostas de futebol. As previsões desses sites baseiam-se na força das equipas, golos marcados ou cálculos matemáticos [2], [3], [4], [5]. Há também aplicações para Android como a "Betting Tips" [6] e para iOS como a "Bet Predictor" [7].

Apesar de estes sites fornecerem informação variada, não são 100% confiáveis e apostar cegamente com base nessas previsões pode levar a grandes prejuízos. Um exemplo concreto é o do site "forebet" [5]. Por análise das previsões da jornada 19 da liga inglesa da época 2017/2018 foi possível ver que este site apenas previu corretamente 2 dos 10 jogos da jornada, o que demonstra o risco de confiar nestas previsões. Há portanto uma falta de boas ferramentas de apoio nesta área e é necessário melhorar a forma como se tenta prever jogos de futebol.

1.3. Objetivos

Uma forma de tentar prever resultados futuros de jogos de futebol é analisando dados históricos de jogos anteriores. Dados como a forma recente de uma equipa podem ajudar a prever se essa equipa irá ter um bom ou mau resultado num jogo. A disponibilização de dados relativos aos jogos nas várias ligas de futebol é cada vez mais detalhada o que torna viável a recolha desta informação para análise. Estes dados podem depois ser usados para aplicar técnicas de *Data Mining.* As ferramentas de *Data Mining* permitem fazer previsões e já foram usadas em diferentes áreas, incluindo no desporto, com resultados muito positivos. Como a liga inglesa é uma das ligas mais importantes do mundo e é uma das que tem mais dados disponíveis, esta é a competição mais indicada para testar este tipo de previsão, razão pela qual será adotada nesta tese.

O objetivo do projeto é aplicar várias técnicas de *Data Mining* para previsão de resultados de jogos de futebol da Liga Inglesa usando dados de jogos relativos a vários anos anteriores. Para fazer a previsão serão extraídos dados de diferentes fontes. Para além dos modelos de previsão é também objetivo a criação de um Sistema de Apoio à Decisão que indique aos apostadores quais os resultados mais prováveis para uma dada jornada. O sistema indicará aos utilizadores quais as apostas com maior ou menor risco, permitindo-lhes apostar mais facilmente e com menor risco, ajudando os apostadores menos experientes a obter lucros e evitando a necessidade de fazer a análise dos jogos.

1.4. Análise de Valor

Nos últimos anos as apostas desportivas têm aumentado e surgiram novas casas de apostas. Em Portugal, numa só casa de apostas, foram apostados 300 milhões de euros em 1 ano. Atualmente existe também um aumento da publicidade a casas de apostas o que dá mais visibilidade a este mercado e as previsões são de que este mercado continue a crescer. Devido

à dificuldade de acertar nos resultados dos jogos é necessário desenvolver soluções para ajudar os apostadores a não obterem prejuízos. Como o futebol é o desporto em Portugal onde se aposta mais dinheiro, um sistema que seja capaz de dizer a um apostador as melhores apostas a fazer para obter mais dinheiro tem um grande mercado.

1.5. Resultados Esperados

Os resultados esperados incluem o desenvolvimento de modelos de previsão capazes de acertar em resultados de jogos de futebol. Esses modelos devem ser consistentes, ou seja, não deve haver semanas em que o número de jogos corretamente previstos seja demasiado baixo para que os utilizadores não percam dinheiro. Os modelos devem também ter uma boa percentagem de acerto nos 3 resultados possíveis de um jogo. Os modelos serão utilizados para desenvolvimento de uma aplicação para indicar aos apostadores qual a melhor aposta a fazer para um determinado jogo, permitindo aos utilizadores deixar de ter prejuízos e obter lucros.

1.6. Abordagem Preconizada

A abordagem a fazer inclui o desenvolvimento de modelos de previsão usando *Data Mining*. Os modelos terão como base dados de jogos anteriores das equipas. Para assegurar a qualidade do desenvolvimento dos modelos de previsão, será utilizada a metodologia padrão CRISP-DM (Wirth e Hipp, 2000). Como o problema envolve uma tomada de decisão será desenvolvido um Sistema de Apoio à Decisão. Esse sistema utilizará o melhor modelo de previsão para indicar previsões de resultados dos jogos, ajudando assim os apostadores a fazer apostas. O sistema indicará aos utilizadores quais as apostas que têm menor e maior risco de modo a que estes possam fazer apostas mais ponderadas.

1.7. Organização do Documento

Este documento encontra-se organizado em 9 capítulos. O primeiro descreve de uma forma geral o problema e os objetivos a atingir. O segundo capítulo apresenta o estado da arte relativo ao problema e descreve a análise de valor do projeto. No terceiro capítulo é feita a avaliação das abordagens para resolver o problema, tendo em conta o contexto apresentado no capítulo 2. O quarto capítulo apresenta o *design* da solução para o problema. O capítulo 5 descreve a forma de avaliação do projeto. O sexto capítulo apresenta o modo como foi feita a análise e tratamento dos dados obtidos. O capítulo 7 descreve o modo como foram desenvolvidos os modelos de previsão e os resultados obtidos. O capítulo 8 descreve o sistema de apoio à decisão desenvolvido. Por fim, o nono capítulo apresenta as conclusões.

2 Contexto e Estado da arte

2.1 Área de Negócio

As apostas desportivas consistem em apostar uma determinada quantidade de dinheiro no resultado de um evento desportivo. O objetivo é fazer um prognóstico correto que permita acertar no resultado e obter lucros. Por este motivo estas apostas não devem ser feitas com base na sorte e devem ser ponderadas. Há vários tipos de apostas desportivas, sendo um exemplo as apostas 1X2, onde se aposta no resultado de um evento desportivo, no caso do futebol há 3 resultados possíveis, 1 – vitória da equipa da casa, X- empate, 2 – vitória da equipa visitante. Há também outros tipos de apostas como apostar no número de cartões amarelos ou vermelhos que vão ser mostrados a uma das equipas, quem vai marcar golos ou quantos cantos vai haver num jogo. Consoante os diferentes desportos há tipos de apostas bastante variadas. Há ainda a probabilidade de apostar em vários eventos em simultâneo, aumentando o lucro que se pode obter, mas reduzindo as chances de obter retorno, visto ser necessário acertar em mais do que um resultado. [8]

O lucro que se pode obter numa aposta é obtido multiplicando o valor da aposta pela *odd* da aposta. Uma *odd* é um valor, por exemplo "1.20", que reflete a probabilidade de um determinado evento acontecer. Se a probabilidade for muito alta a *odd* é menor e caso a probabilidade seja baixa a *odd* é maior [9]. Uma vez que a probabilidade de acertar ao acaso num jogo de futebol é de 33,3% [10] há a necessidade de oferecer algum apoio quando se aposta neste tipo de evento.

2.2 Análise de valor

Para realizar a análise de valor deste projeto foi aplicado o modelo NCD (*New Concept Development Model*). Este modelo foi desenvolvido por Peter Kohen e é usado para descrever as etapas de análise, criação e inovação de um produto (Koen *et al.*, 2014). O modelo NCD tem 5 elementos-chave: "1) Identificação de oportunidade, 2) Análise de oportunidade, 3) Geração de ideias, 4) Seleção de ideias, 5) Definição de conceito" (Koen *et al.*, 2014, p. 2). De seguida são descritas as 5 etapas do modelo NCD no âmbito deste projeto e os métodos envolvidos em cada uma dessas etapas.

- **Identificação de oportunidade**
  As apostas em eventos desportivos têm aumentado ao longo dos últimos anos e em 2015 foram licenciadas em Portugal 2 casas de apostas desportivas online , a "Bet.pt" [11] e a Betclic [12], e também o jogo Placard da Santa Casa [13]. Isto fez com que fosse mais fácil apostar em Portugal, visto que antes só era possível apostar em sites

estrangeiros e o número de apostadores era reduzido. A análise do mercado demonstrou que no 1º ano do Placard o número de apostadores foi de mais de 900 mil e o valor apostado situou-se nos 300 milhões de euros [1].

O futebol representou 93% das apostas feitas no Placard [14] e 78% das apostas online [15]. O retrato dos apostadores do Placard permitiu concluir que em muitos casos há prejuízos [16], o que é compreensível dada a imprevisibilidade dos jogos.

O número de casas de apostas desportivas em Portugal tem vindo a aumentar, existindo atualmente 7 casas de apostas. Em 2017 foram licenciadas 2 casas de apostas desportivas, o Estoril Sol Casinos [17] e o Casino Portugal [18]. No presente ano de 2018 houve mais 2 licenças atribuídas para a "Nossa Aposta" [19] e para o "Placard.pt" [20]. Nos próximos anos prevê-se a atribuição de mais licenças a casas de apostas desportivas [21]. Os valores apostados estão também a aumentar, os dados disponíveis de Janeiro de 2018 indicam que o valor total de apostas no campeonato português situou-se em média nos 288 milhões de euros por jornada [22], sendo portanto apostados por época perto de 10 mil milhões de euros.

Assim identificou-se a necessidade de dar algum tipo de apoio a quem faz apostas em jogos de futebol, indicando a um apostador qual o resultado final de um jogo de futebol.

- **Análise de oportunidade**

  Nesta etapa foi feita a análise do segmento de mercado em Portugal e identificou-se que os apostadores vivem nos grandes centros de Porto, Lisboa e Braga e a maioria são homens [15]. Existem 2 segmentos de mercado distintos, o 1º inclui apostadores que não se importam de arriscar e pretendem obter lucros altos e o 2º inclui apostadores mais conservadores que preferem fazer apostas seguras de baixo risco.

  A oportunidade existe não só em Portugal, mas também a nível mundial. Em 2016 foram apostados mais de 53 mil milhões de euros numa das maiores casas de apostas desportivas em todo o mundo [23]. As expectativas são de que o mercado de apostas mundial cresça e atinja o valor de 635 mil milhões de dólares em 2022 [24], um valor quase 3 vezes superior ao PIB de Portugal [25]. A prova de que a oportunidade existe é o facto de já existirem entidades que ajudam os apostadores a apostar melhor. Atualmente existem potenciais concorrentes como sites de dicas de apostas ou apostadores profissionais que vendem dicas de apostas. O principal concorrente é o site SokkerPro [26], visto já ter um negócio consolidado e apresentar bons resultados nas suas dicas de apostas.

- **Geração de ideias**

  Nesta fase foi feito um *brainstorming* com diferentes pessoas como engenheiros informáticos e pessoas que apostavam regularmente. Isto permitiu identificar mais facilmente que tipo de tecnologias poderiam ser usadas para solucionar o problema e quais os requisitos que uma aplicação deste tipo deveria cumprir. Foi também feita uma pesquisa sobre o modo como se podia tentar prever o resultado final de um jogo de futebol, tentando identificar casos de sucesso neste contexto.

- **Seleção de ideias**

  Tendo em conta que a aplicação necessita de prever resultados futuros torna-se imperativo o uso de técnicas de *Data Mining*. As previsões serão baseadas em estatísticas de jogos passados das equipas, como a média de golos, média de remates, entre outras estatísticas a selecionar. Esta solução prevaleceu sobre outras porque já tinha demonstrado algum sucesso como no caso de (Gomes *et al.*, 2015) e é o método mais usual para tentar resolver problemas que requerem decisões baseadas em conhecimento, decorrem em ambientes de mudança, têm métodos subotimizados, têm dados acessíveis, suficientes e relevantes e têm grande retorno com a tomada de decisões certas.

- **Definição de conceito**

  O sistema a desenvolver deve basear-se num sistema de apoio à decisão capaz de auxiliar os apostadores, permitindo-lhes deixar de ter prejuízos e passar a ter lucros. Para fazer a previsão devem ser usadas técnicas de *Data Mining* e o modelo de previsão deve usar dados disponíveis livremente na internet. O sistema irá fornecer apoio para apostar nos jogos da liga inglesa, indicando para cada jogo de uma jornada 1 de 3 resultados possíveis, 1- vitória da equipa da casa, 2 – empate, 3 – vitória da equipa de fora. Os objetivos passam por implementar um sistema que indique aos utilizadores quais as apostas que são menos arriscadas para que possa apostar com segurança.

Para que um negócio tenha algum sucesso é necessário que se faça a sua análise de valor. O valor pode ser definido "em diferentes contextos teóricos como necessidade, desejo, interesse, crenças, atitudes e preferências. O valor depende portanto da percepção." (Nicola *et al.*, 2012, p. 661). O valor deve ainda ser definido como valor para o cliente e valor percebido. O valor para o cliente é definido como a sensação de vantagem pessoal que um cliente obtém de uma oferta feita por uma organização e pode ocorrer quando há redução de sacrifícios para o cliente ou quando há benefícios (Woodall, 2003, p. 2). O valor percebido, em inglês "*perceived value"*, tem a seguinte definição: "avaliação global feita pelo consumidor da utilidade de um produto baseado na percepção sobre o que é recebido e dado." (Zeithaml, 1988, p. 14). Em relação a este projeto os benefícios para o cliente são maximizar o lucro obtido com as apostas, sem precisar de considerar a mais variada informação associada aos jogos de futebol, obtendo assim mais tempo disponível. Assim é eliminado o sacrifício de ter de analisar em detalhe vários jogos antes de apostar.

O sacrifício é o preço a pagar pelo uso da aplicação. O valor do projeto é portanto o de dar apoio ao fazer apostas desportivas. O valor para os clientes será o facto de estes poderem apostar de forma mais segura e sem prejuízos, o que naturalmente dependerá da percepção de cada cliente. Para um cliente o valor percebido irá depender dos lucros que obtiver ao apostar baseando-se na aplicação, esta é a melhor maneira de a avaliar. Na Tabela 1 estão enquadrados os benefícios e sacrifícios numa perspetiva longitudinal de valor com 4 fases (Woodall, 2003, p. 10).

Tabela 1 – Benefícios e sacrifícios do projeto numa perspetiva longitudinal de valor.

Fase	Benefícios	Sacrifícios
Pré-compra		Valor pago para usar a aplicação
Transação	Ganho de tempo ao deixar de analisar jogos antes de apostar	
Pós Compra	Obtenção de lucros	
Após uso/experiência	Possibilidade de investir em apostas de futebol com baixo risco	

Para melhor definir este projeto numa perspetiva de negócio foi usado o modelo de negócios Canvas (Osterwalder *et al.*, 2010) . Na Figura 1 é demonstrado o modelo de Canvas, o modelo foi feito através da ferramenta "Strategyzer" [27].

A proposta de valor deste serviço é definida como:

- Sistema que apoia na tomada de decisão de apostas de futebol indicando as apostas que se devem fazer, permitindo obter lucros ao apostar. É útil tanto para apostadores experientes porque não precisam de analisar a informação relativa aos jogos antes de apostar e também para apostadores inexperientes porque lhes permite apostar com baixo risco.

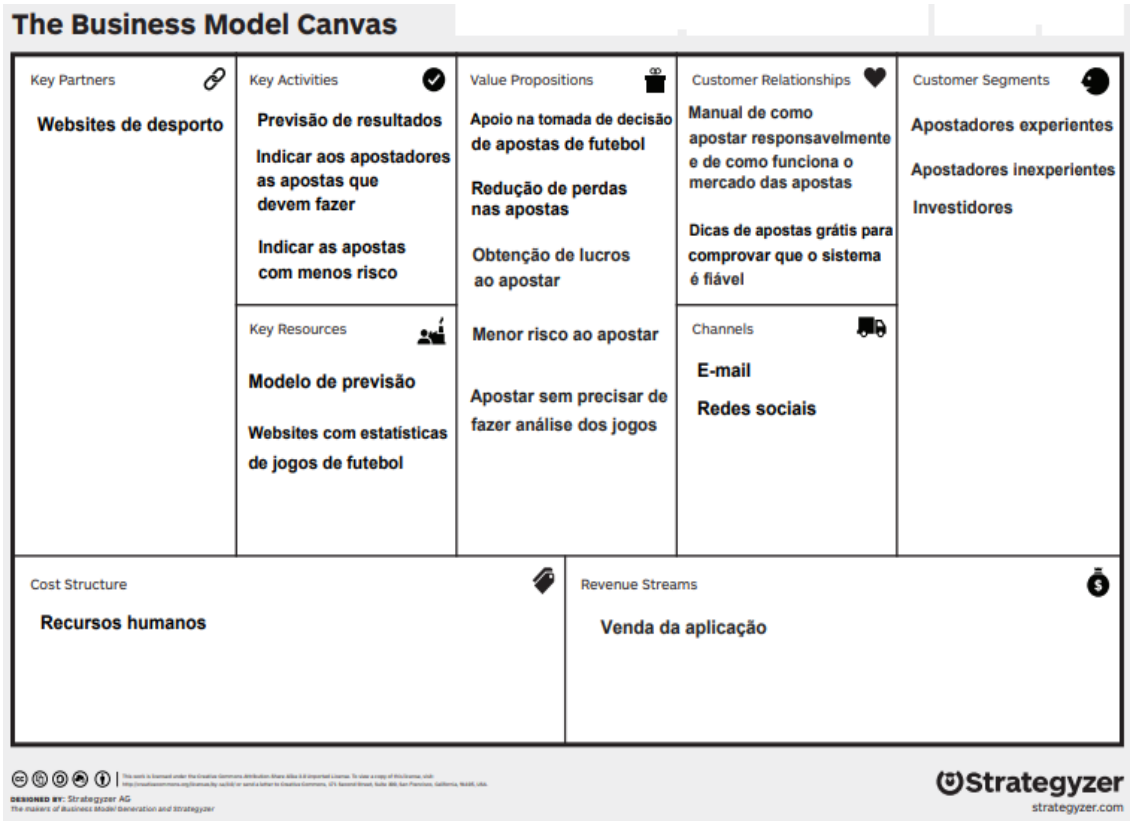

Figura 1 – Modelo de Canvas.

2.3 Descoberta de Conhecimento em Bases de Dados

Nos últimos anos o crescimento das tecnologias de informação levou à recolha e armazenamento de grandes quantidades de dados em várias áreas (Vercellis, 2011, p. 77). Isto faz com que atualmente haja uma grande necessidade de ferramentas para análise de dados. Simultaneamente no mundo atual há uma grande competição entre organizações. O uso de técnicas de extração de conhecimento a partir dos dados pode auxiliar muitas empresas na tomada de decisões podendo vir a tornar-se essencial para organizações e empresas no futuro (Turban, 2011, p. 194). Segundo Turban, a descoberta de conhecimento a partir de dados proporciona às empresas um conhecimento mais rigoroso sobre os seus negócios e clientes de modo que as pode ajudar a fortalecer os seus negócios e a ultrapassar os seus concorrentes.

A Descoberta de Conhecimento em Bases de Dados (em inglês *Knowledge Discovery in Databases*) pode ser descrita como o processo global que leva à descoberta de conhecimento útil em bases de dados. É um processo planeado que envolve vários passos como não só o uso de algoritmos de *Data Mining,* mas também a preparação dos dados e avaliação dos resultados, podendo ser necessário a repetição destes passos em diferentes iterações. (Fayyad *et al.*, 1996)

A Descoberta de Conhecimento em Bases de Dados é constituída pelas 9 etapas (Maimon e Rokach, 2010) expressas a seguir e ilustradas pela Figura 2:

1. Compreensão do domínio de aplicação e do que deve ser feito para atingir os objetivos.
2. Seleção e criação do conjunto de dados onde vai ser feita a descoberta de conhecimento.
3. Pré-processamento e limpeza dos dados.
4. Transformação dos dados de modo a que possam ser usados pelos algoritmos de *Data Mining*.
5. Escolha do tipo de tarefa (*clustering,* classificação).
6. Escolha do algoritmo.
7. Implementação/aplicação do algoritmo.
8. Avaliação dos resultados.
9. Uso do conhecimento adquirido.

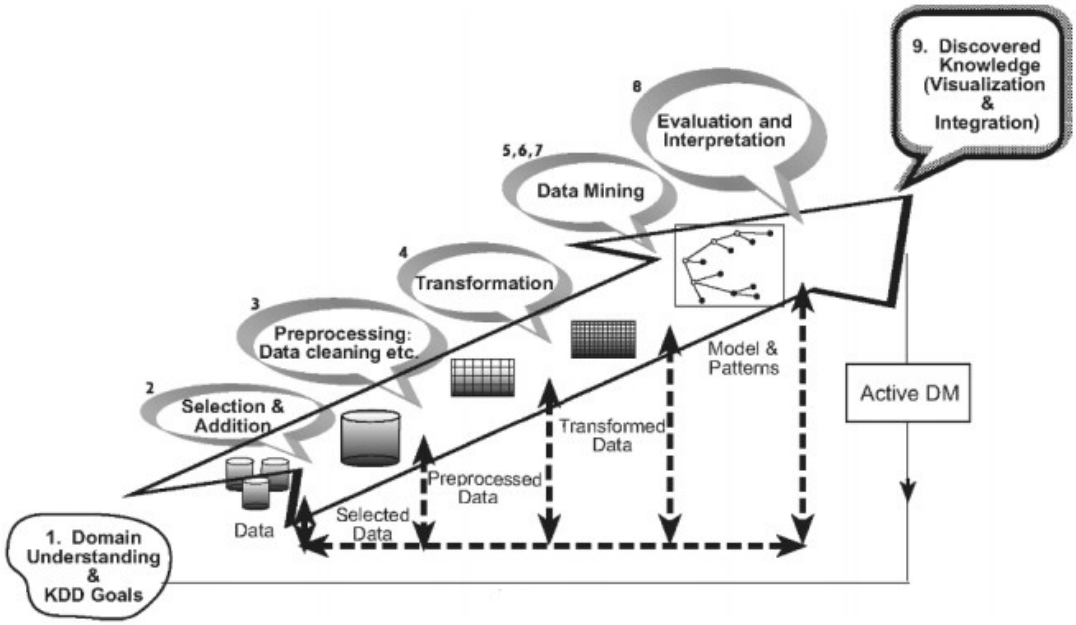

Figura 2 – Fases do processo de Descoberta de Conhecimento em Bases de Dados, imagem retirada de (Maimon e Rokach, 2010, p. 3)

O processo de Descoberta de Conhecimento em Bases de Dados dá grande importância às etapas antes e depois do uso de algoritmos de *Data Mining* de modo a que se consiga extrair informação compreensível dos dados e assim adquirir conhecimento. Isto é essencial pois se o *Data Mining* for aplicado sem a correta preparação dos dados podem ser obtidos maus resultados e descoberta de conhecimento não útil, tornando-se contraproducente (Fayyad *et al.*, 1996).

A metodologia CRISP-DM (Wirth e Hipp, 2000) fornece um guia do ciclo de vida de um projeto de exploração de dados. Contém as fases do projeto e suas respetivas tarefas. O ciclo de vida de um projeto de mineração de dados é baseado nas fases apresentadas na Figura 2 (Chapman *et al.*, 2000). A sequência das fases é indicada pelos números e cada fase depende do resultado das fases anteriores, podendo ser necessária a repetição de fases.

2.4 Data Mining

Data Mining é uma fase no processo de Descoberta de Conhecimento e pode ser descrito como um processo que tem o objetivo de encontrar informação interessante e útil, que se encontra sob a forma de padrões implícitos num conjunto de dados (Fayyad *et al.*, 1996).

O *Data Mining* foi inicialmente definido como um processo de descoberta de padrões. Turban (Turban, 2011) amplia este conceito, considerando que pode ser um processo de análise de dados com o objetivo de aumentar a eficiência e eficácia de uma organização.

O *Data Mining* tem dois tipos de processos, processos de previsão e de descrição (Tan *et al.*, 2005, p. 7). Dentro dos processos de previsão há ainda a distinção entre aprendizagem supervisionada ou não-supervisionada (Maimon e Rokach, 2010). A aprendizagem supervisionada lida com atributos, tentando relacioná-los com um atributo objetivo. Já na aprendizagem não-supervisionada não há um objetivo de aprendizagem predefinido, são os algoritmos que autonomamente procuram padrões nos dados, um exemplo é o agrupamento de dados (Maimon e Rokach, 2010).

Um dos processos de previsão é a classificação. De acordo com (Tan *et al.*, 2005), a classificação é um processo de análise de dados e de reconhecimento de padrões que necessita da construção de um modelo de classificação. Esse modelo é uma função que atribui uma classe a uma dada instância descrita por um conjunto de atributos. Alguns exemplos são a classificação de um e-mail como sendo spam ou não ou a previsão se uma célula é benigna ou maligna.

Para utilizar um algoritmo de classificação é necessário compreender alguns conceitos. Esses conceitos são explicados de seguida, considerando o problema de classificar um e-mail como sendo spam ou não (Mohri *et al.*, 2012, p. 2):

- Exemplo – Instância de um conjunto de dados do problema, neste caso os e-mails a serem verificados;
- Características – conjunto de atributos que caracterizam as instâncias de dados, um e-mail pode ser caracterizado por características como o nome do remetente ou certas palavras presentes na mensagem (Mohri *et al.*, 2012, p. 2);
- Classe – Valor ou categoria dado aos exemplos, neste caso são SPAM ou não-SPAM;
- Exemplos de treino – Exemplos/dados usados para treinar o modelo de classificação. Neste exemplo são e-mails onde já se sabe se são ou não spam;
- Exemplos de teste – Exemplos usados para testar o desempenho do modelo de classificação. Estes exemplos são classificados pelo modelo numa das classes definidas;
- Função avaliação – avalia a taxa de acerto do modelo. “Uma função que mede a diferença, ou perda, entre uma classe prevista e uma classe verdadeira” (Mohri *et al.*, 2012, p. 2).

Nas subsecções seguintes são descritos os algoritmos de *Data Mining* utilizados neste trabalho.

2.4.1 Regressão Logística Multinomial

A Regressão Logística Multinomial (RLM) é um método de Regressão Logística utilizado quando se pretende prever mais do que 2 classes, ou seja quando a classificação não é binária [28]. A RLM faz uma análise que verifica as relações entre a variável que se quer prever e as restantes variáveis, permitindo fazer previsões. Assim este método pode ser usado como um método de classificação [28]. A Regressão Logística é uma função que indica a probabilidade

de um determinado *input* pertencer a uma classe [29]. Quando a classificação é binária este modelo pode ser descrito pela equação (1) [30]. Na equação (1), quando $g(x)$ tende para $+\infty$ então $P(Y = 1) = 1$, quando $g(x)$ tende para $-\infty$ então $P(Y = 1) = 0$.

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}} \quad (1)$$

A classificação pode ser feita da seguinte forma, se $P(Y = 1) > 0.5$ então Y está na classe 1, caso $P(Y = 1) < 0.5$ então Y está na classe 0 [30].

Quando o problema de classificação tem mais de 2 classes é necessário avaliar a razão entre a classe base e as restantes classes (Bittencourt, 2003). A classe base escolhida é normalmente a classe mais comum, a que tem um maior número de registos [31]. Tomando como exemplo um caso com 3 classes possíveis, onde a classe Y pode tomar os valores 1, 2 e 3, é necessário avaliar duas razões. Se a classe base for $Y = 1$, é necessário avaliar a razão entre $Y = 1$ e $Y = 2$, dada pela equação (2), e a razão entre $Y = 1$ e $Y = 3$, dada pela equação (3) (Bittencourt, 2003, p. 79).

$$g_1(x) = \ln\frac{P(Y = 2)}{P(Y = 1)} \quad (2)$$

$$g_2(x) = \ln\frac{P(Y = 3)}{P(Y = 1)} \quad (3)$$

Depois é possível calcular, para um dado conjunto de observações x, a probabilidade de x pertencer a cada uma das 3 classes, através das equações (4), (5) e (6) (Bittencourt, 2003). Assim calcula-se a probabilidade de $Y = 1$, $Y = 2$ e $Y = 3$. A observação x é classificada na classe onde a probabilidade é maior (Bittencourt, 2003, p. 80).

$$P(Y = 1|\, x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}} \quad (4)$$

$$P(Y = 2|\, x) = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}} \quad (5)$$

$$P(Y = 3|\, x) = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}} \quad (6)$$

2.4.2 Árvores de decisão

As árvores de decisão são uma das técnicas de classificação mais conhecidas e é fácil compreender o conceito desta técnica (Tan *et al.*, 2005, pp. 150-152). Esta técnica pode ser explicada considerando um problema de classificação onde se classifica um animal como sendo mamífero ou não-mamífero. Numa árvore de decisão há um nó raiz que verifica o primeiro atributo para poder classificar o animal. Depois existem nós internos que verificam mais atributos, até chegar aos nós folha que classificam o animal tendo em conta os testes

feitos nos nós anteriores (Tan *et al.*, 2005, pp. 150-152). A Figura 3 exemplifica o processo de funcionamento de uma árvore de decisão.

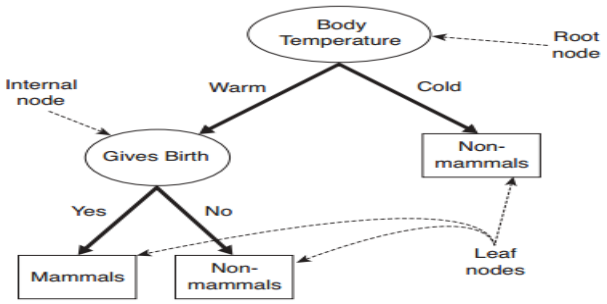

Figura 3 - Classificação de um animal como sendo mamífero ou não, usando uma árvore de decisão. (Tan *et al.*, 2005, p. 151)

Segundo Tan (Tan *et al.*, 2005, pp. 155-163) há que ter vários aspetos em conta quando se constrói uma árvore de decisão. Um dos aspetos é que dependendo do problema de classificação é possível alterar a ordem dos atributos numa árvore. Outro aspeto a considerar é o tipo de atributos usados no problema. Numa árvore de decisão é ainda considerado o grau de pureza dos dados o que permite dividir melhor os ramos na árvore. Esta medida avalia o modo como os dados estão distribuídos. Considerando o exemplo anterior, se houvesse 4 animais e esses animais fossem todos de uma única classe (mamífero ou não-mamífero) a impureza dos dados seria 0. Se 2 dos animais fossem mamíferos e os outros 2 não, havendo um número igual de elementos em cada classe a impureza seria 1. Uma divisão com menos impureza é melhor porque é mais fácil determinar a classe em que uma instância deve ser classificada. Para medir a impureza de um conjunto de dados são usadas medidas como a Entropia, o *Gini Index,* e o Erro de classificação (Tan *et al.*, 2005, pp. 155-163). Um dos algoritmos mais utilizados para construir árvores de decisão é o C5.0 [32].

2.4.3 Xgboost

Há ainda outros métodos baseados em árvores de decisão que oferecem bons resultados de previsão. Um exemplo é a biblioteca Xgboost. O Xgboost, que significa *Extreme Gradient Boosting* [33], baseia-se no método *Gradient Boosting* [34], este método consiste na criação de vários modelos, onde os novos modelos criados permitem corrigir os erros dos modelos iniciais de modo a criar um modelo final mais correto [35]. Um exemplo do funcionamento do *Gradient Boosting* é mostrado na Figura 4. Nesse exemplo o objetivo é prever se uma pessoa irá gostar ou não de jogos de computador e a previsão é feita com base em 2 árvores diferentes, uma avalia a idade do utilizador e a outra avalia se o utilizador usa o computador diariamente [36]. A previsão é dada pela soma das previsões de cada árvore. O Xgboost não explora todas as árvores de decisão possíveis [37], em vez disso utiliza uma função, que pode ser a soma do quadrado dos erros [33], para determinar qual das possibilidades leva a um menor erro e explora essas possibilidades. Uma das vantagens deste algoritmo é o facto de poder ser usado em diferentes linguagens de programação como Python, C++, R, Julia, Scala e

Java [38], além disso é um algoritmo que consome poucos recursos computacionais (Chen e Guestrin, 2016).

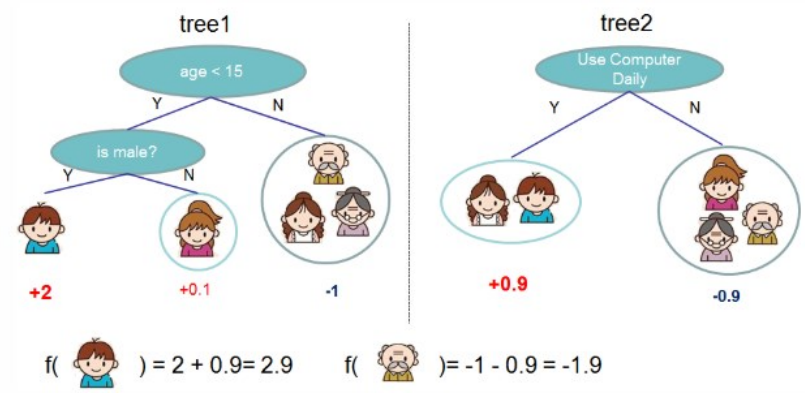

Figura 4 – Exemplo de funcionamento de *gradient boosting*. [33]

2.4.4 Random Forest

Outro algoritmo baseado em árvores de decisão é o RF (*random forest)*, este algoritmo constrói várias árvores de decisão diferentes e combina-as de modo a obter um modelo final que dê o melhor resultado possível [39]. O funcionamento do algoritmo *random forest* é demonstrado na Figura 5. O algoritmo RF começa por gerar diferentes árvores de decisão. As árvores de decisão são criadas de forma aleatória de modo a terem características diferentes (Tan *et al.*, 2013, p. 293). Tomando como exemplo as árvores D1 e D2 da Figura 5, se ambas tiverem um nó "equipa marca mais de 5 golos por jogo" a árvore D1 pode indicar que a equipa vai ganhar o jogo e a D2 pode indicar que a equipa vai empatar o jogo. No final as previsões são combinadas e as decisões são feitas com base na maioria dos votos (Tan *et al.*, 2013, p. 293). Por exemplo se houver 4 árvores de decisão e 3 delas indicarem que uma equipa que marca mais de 5 golos por jogo ganha esse jogo, então a árvore final também vai considerar que a equipa ganha o jogo.

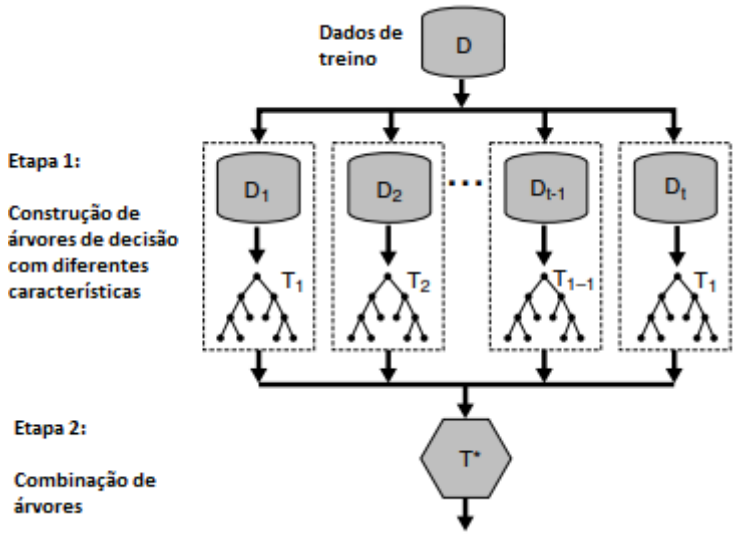

Figura 5 – Funcionamento do algoritmo *random forest*. Adaptado de (Tan *et al.*, 2013, p. 293)

Foi provado teoricamente que, quando o número de árvores de uma RF é suficiente, o erro de uma RF é dado pela equação (7) (Tan *et al.*, 2013, p. 291), onde $\bar{p}$ representa a média da correlação das árvores da RF e s representa a força dos modelos de classificação das árvores.

$$Erro\ RF = \frac{\bar{p}(1 - s^2)}{s^2} \tag{7}$$

Através da equação (7) é possível verificar que quanto maior for a correlação entre as árvores de decisão ou menor for o desempenho dos modelos maior será o erro da RF. Por este motivo quanto maior for a aleatoriedade da RF melhor será o seu desempenho, visto a correlação entre as árvores ser menor.

2.4.5 Naive Bayes

O algoritmo *Naive Bayes* tem por base a teoria de Bayes. A teoria de Bayes é uma teoria probabilística que tem por base o teorema (8) (Tan *et al.*, 2013, p. 230), onde $P(X)$ é a probabilidade de ocorrência de X, $P(Y)$ é a probabilidade de ocorrência de Y e $P(X|Y)$ é a probabilidade de ocorrência de X dada a ocorrência de Y. Este teorema permite calcular a probabilidade de ocorrência de Y dada a ocorrência de X.

$$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)} \tag{8}$$

Para fazer a classificação de um registo, o algoritmo *Naive Bayes* calcula a probabilidade desse registo pertencer a uma dada classe [40]. Este algoritmo considera que as variáveis são independentes o que permite classificar um registo a partir dos exemplos do conjunto de treino, um registo é classificado na classe onde a sua probabilidade é maior (Vercellis, 2011, p. 253).

2.4.6 Support Vector Machines

As *Support Vector Machines* (SVM) são uma técnica de aprendizagem supervisionada. As SVM são definidas como "uma família de métodos de separação para classificação e regressão desenvolvidos no contexto da teoria de aprendizagem estatística" (Vercellis, 2011, p. 262). Esta técnica funciona com base em planos de decisão que separam objetos de diferentes classes. O objetivo desta técnica é criar planos de decisão de modo a que os objetos de classes diferentes fiquem o melhor separados possível [41]. A Figura 6 ilustra o funcionamento de uma SVM. É possível ver que os objetos de 2 classes diferentes (azuis e vermelhos) estão separados por uma linha. Esta técnica atinge bons resultados quando comparada com outras técnicas, sendo aplicável a problemas de várias áreas e problemas de grande dimensão (Vercellis, 2011).

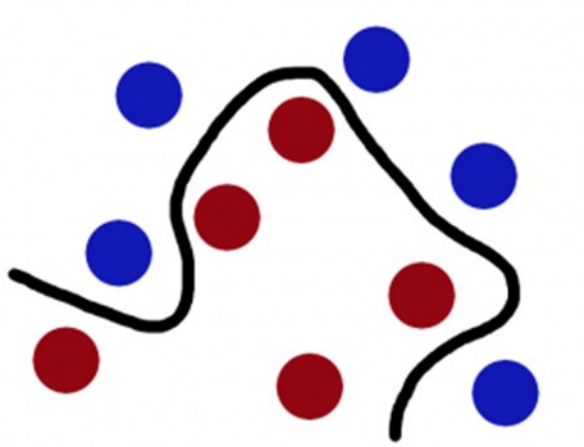

Figura 6 – Exemplo de SVM [42]

2.4.7 Redes Neuronais Artificiais

As RNA (Redes Neuronais Artificiais), em inglês ANN (Artificial Neural Network), foram desenvolvidas para simular o comportamento do cérebro humano (Vercellis, 2011, p. 259). Tal como o cérebro humano é constituído por neurónios, as RNA são compostas por vários nós que se conectam entre si. A Figura 7 mostra uma RNA com 3 conjuntos de nós (Vercellis, 2011, p. 261). Uma rede neuronal é composta por um mínimo de 3 conjuntos (em inglês denominados por *layers*) de nós (Vercellis, 2011, p. 261):

- Conjunto de *input* – é o conjunto de entrada onde são introduzidos os dados para análise.
- Conjunto *hidden* – é o conjunto intermédio onde é feito o processamento dos dados pelos vários neurónios, é possível haver mais do que um conjunto intermédio.
- Conjunto de *output* – é o conjunto de saída onde se podem ver os resultados do processamento.

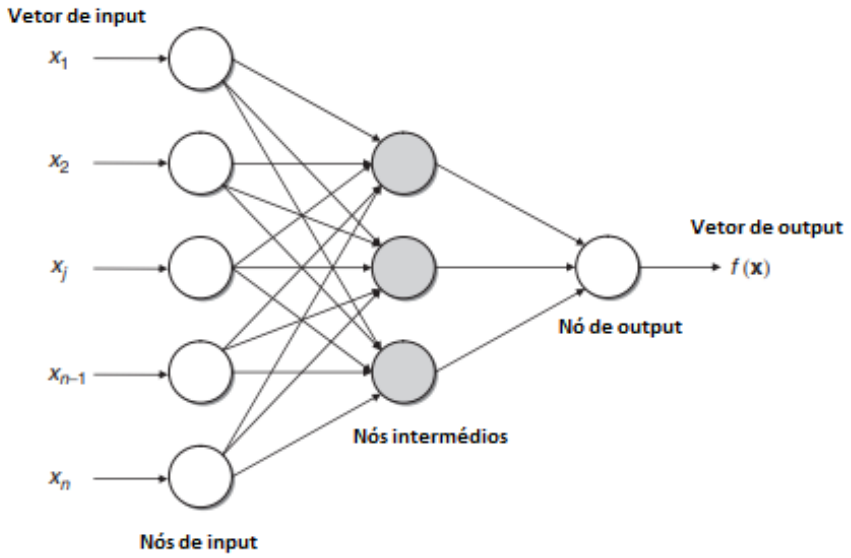

Figura 7 – Funcionamento de uma rede neuronal artificial. Adaptado de (Vercellis, 2011, p. 261)

2.4.8 KNN

O algoritmo KNN (*k-nearest neighbour*) é um algoritmo que pode ser usado, entre outras tarefas, para classificação. O KNN classifica um novo registo comparando-o com os registos do conjunto de treino. Ao novo registo é atribuída a classe do registo mais próximo desse novo registo (Larose, 2014). O novo registo é comparado com um número *k* de registos mais próximos e é-lhe atribuída a classe maioritária desses *k* registos. Para determinar os registos mais próximos é calculada a distância entre os registos. A distância pode ser calculada com base em funções de distância como a Euclidiana, Manhattan, Minkowski ou Hamming, a melhor função varia consoante o caso [43].

2.5 Sistemas de Apoio à Decisão

Um Sistema de Apoio à Decisão (SAD) pode ser descrito como um sistema feito para auxiliar a tomada de decisão, ajudando a encontrar a solução para um dado problema (Turban, 2011, p. 75). Pode ser usado como um termo genérico para "descrever qualquer sistema informatizado que apoie a tomada de decisão numa organização" (Turban, 2011). Estes sistemas analisam várias soluções possíveis para um problema e têm como principais capacidades:

- Apoio a gestores em vários níveis (apoiam indivíduos e organizações);
- Capacidade de ajudar em decisões interdependentes ou sequenciais;
- Usam o processo de tomada de decisão;
- Modelação e análise;
- Acesso a dados;
- Funcionam em modo *standalone*, *web-based* ou integrados num sistema.

Os SAD podem ser usados em múltiplas situações e o uso de sistemas informatizados deste tipo tem vindo a crescer (Turban, 2011). Um sistema deste tipo permite que uma decisão seja tomada de forma (Holsapple, 2008):

- Mais produtiva – mais rápida, com menos custos, com menos esforço;
- Com maior agilidade – permite lidar com aspetos inesperados;
- Inovadora – com mais criatividade;
- Maior qualidade – maior acerto, qualidade e confiança;
- Mais satisfação por parte dos *stakeholders*.

Segundo (Holsapple, 2008, p. 164), a constante evolução do conceito de SAD faz com que haja uma grande variação nas características de um SAD e o que define estes sistemas não é a sua arquitetura, mas sim as suas funções. Um SAD pode fornecer apoio em três níveis: (Hackathorn e Keen, 1981); (Turban, 2011)

- Individual – O apoio é dado a uma pessoa que realiza uma atividade que é independente de outras tarefas;
- Grupo – Apoio dado a um grupo para fazer atividades dependentes entre si;
- Organizacional – Envolve apoio a operações de diferentes áreas.

Os SAD podem ser usados nas mais variadas áreas. Na saúde um exemplo é um SAD que ajuda a detetar glaucomas através da análise de imagens de olhos de pacientes, classificando as imagens para saber se um paciente tem ou não a doença (Karkuzhali e Manimegalai, 2017). Outros exemplos de áreas onde os SAD são usados são (Sauter, 2011):

- Política - Em 2008 a campanha do presidente Obama usou um SAD que recolhia informações de sites e outras fontes para encontrar votantes indecisos e aconselhava o pessoal da campanha a melhor maneira para tentar convencer os indecisos a votar no presidente Obama;
- Retalho – Os SAD são usados por empresas como a "Kroeger" para os ajudar a perceber o comportamento dos consumidores e prever consumos futuros, ajudando-os a organizar as lojas e fazer ofertas que permitam vender mais;
- Desporto – Na liga de basebol norte-americana é usado um SAD para definir o calendário dos jogos. Os jogos são marcados em dias em que haja muitos telespectadores, aumentando o retorno financeiro, ao mesmo tempo são tidos em conta os dias de descanso mínimo para cada equipa.

Há vários tipos de SAD, normalmente um SAD pode ser classificado num dos seguintes tipos (Wienclaw, 2013):

- *Model-Driven* – Utilizam informações financeiras, otimização e simulações para apoiar a decisão;
- *Data-Driven* – Recolhem dados e apresentam informação essencial, importante e atualizada para uma organização;
- *Knowledge-Driven* – "sistemas pessoa/computador com capacidades especializadas de resolução de problemas que podem fazer sugestões ou recomendações a um utilizador" (Wienclaw, 2013);
- *Group Decision* – "permitem que grupos de trabalho processem e interpretem informação em conjunto, mesmo quando não estão juntos fisicamente" (Wienclaw, 2013).

Tendo em conta o âmbito deste projeto, um SAD adequado seria um SAD do tipo *Knowledge-Driven* que fornecesse apoio a nível individual, permitindo apostas mais produtivas.

2.6 Estado da arte em abordagens existentes

A utilização de *Data Mining* no desporto aumentou bastante nos últimos anos, tendo permitido ajudar no processo de tomada de decisão nesta área (Schumaker *et al.*, 2010). Foram feitos estudos comparativos em diferentes desportos que confirmaram que o *Data Mining* é uma boa ferramenta para ajudar a prever resultados de eventos desportivos (Haghighat *et al.*, 2013).

Um trabalho recente (Bunker e Thabtah, 2017) com foque na análise de *Machine Learning* para previsão de resultados desportivos propôs uma nova *framework* para lidar com o problema da previsão de resultados no desporto. Este trabalho realçou a necessidade de criar modelos de previsão mais precisos nesta área. Isto é imperioso devido ao aumento do número de apostas desportivas e também pela necessidade que os treinadores têm hoje em dia de informação útil que os ajude a delinear estratégias de jogo.

Um dos casos que melhor demonstra a valência do *Data Mining* na tomada de decisão no desporto é o da equipa de basebol norte-americana, os *Oakland Athletics*. Inclusive, este caso foi retrato em 2011 no filme de Hollywood "*Moneyball*" [44]. Esta equipa de basebol tinha um orçamento muito inferior às outras equipas do campeonato e usou técnicas de *Data Mining* para analisar estatísticas de jogadores da liga de basebol americana. Assim conseguiu descobrir bons jogadores que estavam subavaliados e contratá-los a um baixo preço conseguindo bater o recorde de vitórias seguidas na liga.

No basquetebol também há um bom exemplo do uso de *software* baseado em *Data Mining.* O "*Advanced Scout*" foi desenvolvido pela IBM e está disponível para todas as equipas da NBA, ajudando a preparar os jogos e analisar os adversários. Há também a "APBRmetrics", uma ferramenta que permite avaliar o desempenho individual dos jogadores de basquetebol. (Schumaker *et al.*, 2010)

No futebol o *Data Mining* também já foi aplicado por clubes como o AC Milan para ajudar a prever futuras lesões de jogadores (Schumaker *et al.*, 2010). No mundial de futebol de 2014 a campeã do mundo Alemanha utilizou um *software* para os ajudar a estudar as equipas e oponentes e monitorizar os seus jogadores [45]. Os dados recolhidos eram analisados para ajudar o treinador da equipa alemã a tomar decisões. Estes factos ilustram bem a potencialidade do *Data Mining* oferecer bons resultados na área do desporto.

Há ainda outros casos de aplicação de estatísticas no futebol, mas que não usam *Data Mining.* Num desses casos [46] era feita a previsão dos resultados dos jogos da Liga dos Campeões e era também determinada qual a probabilidade de uma equipa ganhar essa competição. As previsões eram baseadas no *ranking* de cada uma das equipas participantes e nas probabilidades de resultados de cada jogo. Depois os jogos eram simulados milhares de vezes, levando à correta previsão de mais de metade dos jogos de uma jornada.

Relativamente à previsão de resultados de eventos desportivos houve um caso de estudo feito no âmbito de uma tese de mestrado (Cao, 2012) para prever os resultados de jogos da liga de basquetebol americana, a NBA. A NBA é uma liga com mais de 60 anos de história e é a mais importante do mundo neste desporto. Neste caso de estudo o objetivo era prever 1 de 2 resultados possíveis, vitória da equipa da casa ou vitória da equipa visitante. Os dados eram relativos a 6 anos de estatísticas da NBA da época 2005/06 à época 2010/11 e foram recolhidos e preparados por analistas da NBA. Os dados incluíam inúmeras estatísticas como as equipas iniciais, estatísticas individuais de jogadores, estatísticas globais de cada equipa e dados sobre lesões. Para extrair os dados foi usada a linguagem Python com a biblioteca

Beautiful Soup [47]. Devido à complexidade dos dados foi criado um *Data Mart* o que permitiu organizá-los melhor. Para utilizar apenas dados relevantes no modelo de previsão o autor realizou um processo de extração de *features* (características). As *features* obtidas diziam respeito às estatísticas dos jogos anteriores entre as equipas, estatísticas de confronto direto, número de jogos nos últimos 5 dias, dias de descanso desde o último encontro e desempenho na época anterior. O modelo foi testado com vários algoritmos como Redes Neuronais, *Naive Bayes*, *Support Vector Machines* e *SimpleLogistics*. Os dados de treino usados eram das primeiras 5 épocas recolhidas sendo os de teste os da 6ª época, a de 2010/11. O algoritmo *SimpleLogistics* foi o que obteve melhores resultados com uma taxa de acerto de 69,67%. Uma das conclusões tiradas foi que os dados históricos de confronto entre 2 equipas ajudam a prever corretamente os resultados dos jogos.

Dentro de todos os casos analisados, houve um que se destacou pelo detalhe e coerência do caso de estudo. Esse caso (Gomes *et al.*, 2015), tal como esta tese, tinha como propósito a previsão de resultados de jogos de futebol, sendo o objetivo prever a vitória, empate ou derrota das equipas. Para obter resultados foram analisadas estatísticas de jogos de futebol de modo a identificar padrões que permitissem sugerir qual o resultado de um jogo. Nesse projeto foram seguidos os seguintes passos:

- Recolha de dados de jogos de futebol (número de golos, remates…);
- Tratamento dos dados;
- Criação de modelos de previsão;
- Avaliação dos modelos;
- Escolha do melhor modelo de previsão.

Este caso de estudo seguiu a metodologia CRISP-DM, descrita na secção 2.3. Na fase inicial o problema foi classificado como semiestruturado. Um problema semiestruturado é um problema onde o seu contexto, bem como os objetivos a atingir, estão bem definidos, mas não se sabe qual a melhor solução para resolver o problema (Turban, 2011, p. 12). Nesse caso de estudo foi implementado um sistema de apoio à decisão com a capacidade de ajudar a tomar a melhor decisão ao fazer apostas em jogos de futebol. O sistema foi desenvolvido com o software WEKA (Frank *et al.*, 2016) em conjunto com o *Exsys Corvid Expert System* [48], uma ferramenta que ajuda a desenvolver sistemas inteligentes para auxiliar a tomada de decisão. O *Exsys Corvid Expert System* implementa o sistema desejado com base num conjunto de regras e permite criar facilmente interfaces para o utilizador [48]. Neste caso de estudo os dados incluíam estatísticas de mais de 4900 jogos de 13 épocas, da época 2000/01 à 2012/13. Uma fase importante neste caso de estudo foi a criação de variáveis que pudessem existir antes do começo de um jogo, como as médias de golos de uma equipa. As médias para a equipa que jogava em casa tinham apenas em conta os jogos em casa dessa equipa e para as equipas visitantes apenas tinham em conta os jogos jogados fora.

Foram testados vários modelos de previsão e o que apresentou melhores resultados usava o algoritmo *Support Vector Machines*. O modelo foi testado em 7 jornadas da época 2013/14, perfazendo um total de 70 jogos, e foi obtida uma taxa de acerto de 54,29%, com uma margem de lucro de 20%.

Apesar do futebol ser um desporto imprevisível onde tudo é possível, quando o *Leicester City* ganhou a liga inglesa na época 2015/16 a comunidade futebolística ficou perplexa. Uma investigação aprofundada (Ruiz *et al.*, 2017) foi feita para tentar perceber o que levou a essa vitória surpreendente e tentar perceber como futuras previsões podem ser feitas. A investigação demonstrou que a conquista foi obtida devido ao excelente desempenho do guarda-redes do *Leicester* e ao facto de terem sido muito eficazes a marcar em contra-ataques. Outro fator importante foi haver vários jogadores do *Leicester* que fizeram um grande número de intercepções de passes que à partida tinham uma probabilidade de mais de 80% de terem sucesso. Neste caso de estudo foi ainda criado um modelo para prever o número de remates e golos que uma equipa iria marcar durante um jogo. Chegou-se à conclusão que um modelo que incluísse informação sobre os tipos de remates feitos, por exemplo remates feitos numa jogada de contra-ataque ou feitos depois de um cruzamento, obtinha melhores resultados de previsão.

Há também um caso onde a previsão de jogos de futebol foi feita usando um sistema multiagente (Cañizares *et al.*, 2017). O método de aprendizagem usado foi o *Multilayer Perceptron* e foi testado com dados da época 2015/2016 da liga espanhola. Os dados de treino correspondiam a 80% dos jogos dessa época e os de teste a 20% dos jogos dessa mesma época. A taxa de acerto obtida foi de 61%. Noutro caso de estudo (Prasetio e Harlili, 2016), foi usada regressão logística para prever os jogos da época 2015/16 da liga inglesa. O modelo de previsão apenas previa a vitória ou derrota da equipa da casa, excluindo o resultado de empate. A taxa de acerto rondou os 69,5% e concluiu-se que as variáveis que mais influenciavam a previsão eram as da defesa da equipa da casa e visitante.

Foram feitas ainda duas teses de mestrado com o mesmo intuito desta tese. Na primeira tese (Duarte, 2015) foi feita a previsão de jogos do campeonato português, usando para treino dados da época 2012/13. Quando efetuados testes de previsão na época 2012/13 a taxa de acerto situou-se nos 59%, mas quando o modelo foi aplicado em novos dados da época 2013/14 a taxa de acerto caiu para 45%. O autor referiu que a queda podia ser devida a um problema de *overfitting* e que portanto o modelo estaria "*demasiado ajustado aos dados de treino*" (Duarte, 2015). Na segunda tese (Zan, 2017) uma das conclusões foi que as ligas onde mais facilmente se podem obter lucros são a inglesa, a espanhola, a sueca e a holandesa. Há ainda mais duas teses com objetivos semelhantes ao desta tese. Em (Kumar, 2013) o objetivo da tese era identificar o modo como os especialistas determinam os *ratings* dos jogadores. Nessa tese também foi testada a previsão de resultados na Liga Inglesa, utilizando dados da época 2011/2012, e a taxa de acerto obtida foi de 53,4%. A outra tese (Pettersson e Nyquist, 2017) focou-se mais na previsão de resultados durante o período de jogo, tentando prever os resultados com os dados disponíveis a cada 15 minutos de jogo. Foram utilizados dados de ligas de todo o mundo com jogos disputados de 2015 a 2017. Também foi testada a previsão de resultados antes da ocorrência dos jogos, mas a melhor taxa de acerto foi de apenas 43,96%.

Um dos objetivos da presente tese é desenvolver modelos de previsão melhores do que os das teses anteriormente descritas e obter maiores margens de lucro. Outro objetivo é fazer

um caso de estudo maior, onde os testes fossem feitos não em apenas alguns jogos, mas numa época inteira.

Há também casos de estudo em que foram feitas previsões através de métodos matemáticos, em vez de usar *Data Mining*. Um desses exemplos (Boldrin, 2017) fez as previsões com base na distribuição de Poisson. O modelo de previsão conseguiu uma taxa de acerto de 51,8% considerando empates e uma taxa de acerto de 69% não considerando possível o resultado de empate.

Há ainda estudos feitos nesta área que tentaram descobrir quais as melhores variáveis para prever o resultado de um jogo de futebol. Um desses estudos (Zuccolotto *et al.*, 2014) utilizou o *software* R para essa análise, incluindo exemplos do código utilizado e explicações sobre esse código. Foram testados diferentes algoritmos e o que obteve melhores resultados foi o algoritmo de Regressão Logística Multinomial. Neste caso foram utilizados dados da época 2010/11 da liga italiana, tendo sido usados 300 jogos para treino e 80 para teste. Uma das conclusões tiradas neste estudo foi que uma equipa que faça muitas jogadas com lances aéreos aumenta a probabilidade de empatar ou perder o jogo.

Houve dois casos de estudo encontrados que relataram terem atingido resultados muito acima do que é habitual na previsão de jogos de futebol. Um desses casos (Shin e Gasparyan, 2016) fez previsões de jogos da liga espanhola e acertou em 70% dos jogos. O outro caso (Igiri e Nwachukwu, 2014) fez previsões de jogos da liga inglesa e acertou em 85% dos resultados considerando vitórias, empates e derrotas e 93% considerando apenas vitórias e derrotas.

2.7 Estado da arte em tecnologia relevante

As linguagens de programação mais utilizadas atualmente para *Data Mining* são a linguagem R e Python [49]. A linguagem R é uma linguagem e ambiente de programação *open-source* para computação estatística e gráfica [50]. O R fornece algoritmos para análise de dados, *Data Mining*, construção de sistemas de recomendação, extração de dados web, otimização matemática, entre outros. Uma das principais características desta linguagem é a facilidade para desenhar gráficos que ajudem a analisar e resolver melhor os problemas. O R está em constante melhoramento visto que é um ambiente de programação livre. Há constantes atualizações de bibliotecas do R e semanalmente são adicionadas novas bibliotecas que permitem lidar com mais problemas. O R tem disponíveis inúmeros algoritmos de *Data Mining* e foi o *software* utilizado num caso de estudo anteriormente apresentado (Zuccolotto *et al.*, 2014) para tentar prever resultados de jogos de futebol. O software R foi também utilizado num caso para prever eventos sísmicos perigosos com base na atividade sísmica na região (Dusza *et al.*, 2016).

O Python é uma linguagem *open-source* orientada a objetos [51]. Tal como a linguagem R, o Python permite fazer análise de dados e inclui bibliotecas para usar algoritmos de *Data Mining*. Pode ser facilmente integrado com outras linguagens de programação e a sua

utilização para desenvolver aplicações de *Data Mining* tem aumentado nos últimos anos [49]. Um exemplo do uso de Python com *Data Mining* é o de um caso de estudo em que foi usado para treinar um modelo de árvores de decisão capaz de analisar o comportamento de condutores de automóveis, com o objetivo de classificar os condutores como sendo mais ou menos agressivos (Hwang *et al.*, 2018). Outro exemplo é um estudo (Awoyemi *et al.*, 2017) onde foram testados vários algoritmos, implementados em Python, para detetar fraudes em cartões de crédito através da análise de transações financeiras. Os modelos determinavam se havia fraude classificando as transações financeiras como sendo fraudulentas ou não.

3 Avaliação de soluções e abordagens existentes

Tal como foi referido no capítulo 2, as linguagens mais utilizadas para *Data Mining* são o R e o Python. No entanto, a linguagem R é melhor para análise de dados [52]. Para além disso nos casos de estudo encontrados a linguagem R era mais utilizada para resolver o tipo de problema desta tese. Por estas razões e pelo facto de já ter experiência com a linguagem R, optou-se por utilizar a linguagem R para desenvolver a componente de *Data Mining* do projeto.

A análise dos casos de estudo feitos anteriormente nesta área permitiu encontrar uma causa comum para os problemas que podiam ter afetado a previsão de resultados. A maioria dos autores reportaram que o mau desempenho das previsões poderia estar relacionado com erros nos dados e a necessidade de fazer tratamento dos dados de entrada. Na maioria dos trabalhos que não atingiu bons resultados não tinha sido feita uma fase de preparação antes de começar a criar o modelo de treino. Os melhores resultados foram conseguidos quando foi feito um planeamento estruturado de todo o processo, sendo este um ponto a ter em conta quando se desenvolve uma aplicação deste tipo.

Dentro dos casos de estudo encontrados, as melhores taxas de acerto foram de 75% e de 85%. No entanto estes casos não estavam tão bem detalhados como os restantes. No caso de (Shin e Gasparyan, 2016), que obteve 75% de acerto, não foi indicada a época em que os jogos ocorreram. Para além disso não é indicada a percentagem de jogos que foi usada para treino e para teste, podendo ter sido usados os mesmos dados para treino e para teste. Para uma aplicação real não podem ser usados os mesmos dados para treinar e para testar. Estes fatores diminuem a credibilidade da taxa de acerto atingida. No caso de (Igiri e Nwachukwu, 2014), apesar de melhor documentado, só são usados 110 jogos de uma época para treinar o modelo. Para além disso também não foram indicados quais os jogos usados para treino e para teste, podendo o modelo estar ajustado aos dados de treino. Um modelo de previsão de jogos de futebol deve ser experimentado com jogos de várias épocas porque de época para época pode haver grandes variações nas equipas. Como foram usados poucos jogos para treino, e de uma só época, a taxa de acerto pode estar inflacionada. Seria provável que com mais jogos e de diferentes épocas a taxa de acerto diminuísse. Apesar disso estes dois casos indicam variáveis interessantes para explorar, como dados relativos aos jogadores das equipas. O outro caso que obteve melhores resultados utilizou novas variáveis como a vitória nos últimos 5 jogos das equipas. Esta é também uma variável relevante para testar num modelo de previsão de jogos. A Tabela 2 mostra para os casos de estudo detalhados na secção 2.6, que utilizaram *Data Mining* para prever jogos de futebol, as variáveis usadas para prever os resultados dos jogos e a taxa de acerto obtida.

Tabela 2 – Variáveis usadas para prever jogos e respetivas taxas de acerto.

Caso de estudo	Variáveis	Taxa de acerto
(Duarte, 2015)	Golos marcados e sofridos; Desempenho nos últimos 5 jogos; Confronto direto e fator casa;	45%
(Cañizares *et al.*, 2017)	Resultados e diferença de golos nos últimos 5 jogos; Resultados e diferença de golos nos últimos 5 confrontos diretos; Valor financeiro das equipas; *Ratings* da defesa, meio-campo e ataque das equipas; Posse de bola, passes, remates por jogo; Minutos jogados, quilómetros percorridos, lesões; Número de *penalties*, cartões amarelos e vermelhos;	61%
(Zuccolotto *et al.*, 2014)	Ataque em lances aéreos da equipa da casa; Remates da equipa da casa; Defesa da equipa da casa; Defesa da equipa de fora; Remates da equipa de fora; Capacidade para fazer contra-ataque da equipa de fora;	65%
(Prasetio e Harlili, 2016)* *Apenas considerou vitórias e derrotas	*Rating* ofensivo e defensivo da equipa da casa; *Rating* ofensivo e defensivo da equipa de fora;	69,5*
(Shin e Gasparyan, 2016)	Remates à baliza; Golos marcados; Cartões amarelos e vermelhos; Atributos dos jogadores retirados do jogo;	75%
(Igiri e Nwachukwu, 2014)	Golos marcados e remates; Cantos; Odds dos jogos; Força da equipa da casa; Desempenho dos jogadores e treinadores; Vitórias dos treinadores; Série de vitórias;	85%
(Gomes *et al.*, 2015)	Equipa de casa e equipa visitante; Média de golos da equipa de casa e visitante; Média de remates da equipa de casa e visitante; Média de remates à baliza da equipa de casa e visitante; Vitórias nos últimos 5 jogos das equipas de casa e visitante; Vitórias nos últimos 5 jogos entre as 2 equipas;	54,3%

3.1 Abordagem adotada

Para desenvolver o projeto é primeiro necessário extrair dados de jogos da liga inglesa. Depois serão testados modelos de previsão. Posteriormente serão adicionadas ao modelo novas variáveis como as vitórias nos últimos 5 jogos e estatísticas dos jogadores. Deste modo poder-se-á verificar se esses dados melhoram a previsão.

De acordo com uma perspetiva baseada em Descoberta de Conhecimento em Bases de Dados e a metodologia de trabalho CRISP-DM este projeto enquadra-se da seguinte forma:

1. Compreensão do domínio de aplicação – A análise do estado da arte permitiu concluir esta tarefa.
2. Seleção e criação do conjunto de dados – Os dados são a peça principal num projeto de *Data Mining.* Recolher dados suficientes e que contenham informação relevante é fundamental para atingir sucesso e obter bons resultados. Na fase inicial vão ser extraídos dados de sites como http://www.football-data.co.uk/.
3. Pré-processamento e limpeza dos dados – Limpar os dados de jogos que sejam incoerentes.
4. Transformação dos dados – Criação de variáveis como a vitória nos últimos 5 jogos.
5. Escolha do tipo de tarefa de *Data Mining* – Para este problema a tarefa é a classificação.
6. Escolha dos algoritmos – Foram identificados vários algoritmos para criação dos modelos de classificação de modo a que se pudesse descobrir qual o que oferecia melhores resultados. Os algoritmos utilizados são descritos no capítulo 7 e o seu funcionamento é explicado na secção 2.4.
7. Execução dos algoritmos para implementação dos modelos de classificação – Nesta etapa são implementados modelos de previsão com os algoritmos escolhidos anteriormente.
8. Avaliação dos resultados – Os modelos foram avaliados usando a taxa de acerto e também considerando o lucro que se obteria ao apostar com base nas previsões feitas pelos modelos.
9. Uso do conhecimento adquirido – Utilização do conhecimento no sistema de apoio à decisão para os apostadores de jogos de futebol.

De acordo com a definição de SAD explicada no capítulo 2, o sistema proposto será *Knowledge-Driven* uma vez que se foca em resolver um problema e fazer recomendações. O problema em questão é qual a aposta a fazer e o SAD indica qual deve ser feita. O sistema atua a nível individual, uma vez que se espera que cada pessoa utilize o sistema isoladamente. O SAD vai permitir que a decisão seja mais produtiva, visto que um apostador terá menos esforço ao apostar pois não tem que analisar os jogos antes de cada aposta. Para além disso, como perde menos dinheiro nas apostas terá menos custos.

4 Design

Para desenvolver a solução proposta é necessário cumprir determinados requisitos. De seguida são apresentados os requisitos funcionais e não funcionais para o sistema.
Funcionais:

- Extração de dados de jogos de futebol;
- Desenvolvimento de modelos de previsão de resultados de jogos de futebol através de *Data Mining;*
- Escolha do melhor modelo para previsão de resultados de jogos.

Não funcionais:

- Apresentar de uma forma clara as apostas a fazer, explicitando as que têm maior e menor risco. Para isso vão ser recolhidos dados das probabilidades de ocorrência do resultado previsto pelo modelo de previsão. Com base na maior ou menor probabilidade de ocorrência do evento as apostas vão ser classificadas como sendo de baixo, médio ou alto risco.

A Figura 8 demonstra o diagrama UML de casos de uso de funcionalidades geral do sistema e a Tabela 3 descreve esses casos de uso.

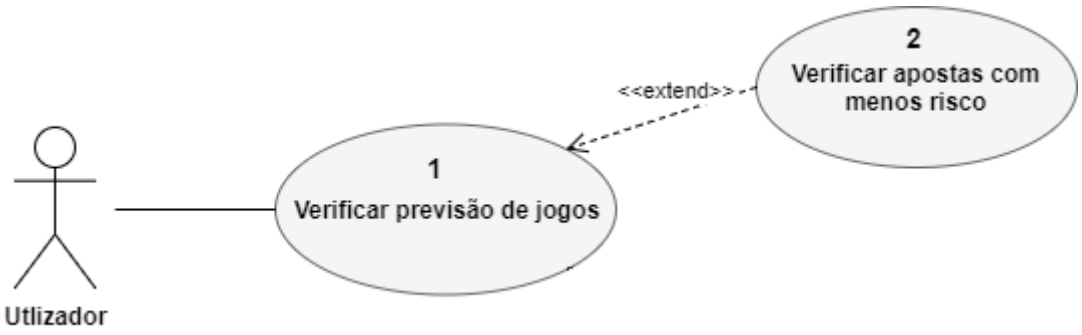

Figura 8 – Diagrama de casos de uso de funcionalidades geral.

Tabela 3 – Descrição dos casos de uso da aplicação.

Caso de uso	Descrição
1	O utilizador verifica para a jornada seguinte do campeonato a previsão feita pelo algoritmo de *Data Mining*.
2	O utilizador verifica as apostas identificadas como tendo menor risco, havendo uma maior probabilidade desse evento acontecer.

A principal componente do sistema é o módulo Data Mining de previsão de resultados de jogos que foi desenvolvido em linguagem R. Para o utilizador poder ver facilmente essas previsões existe uma interface em Java. A linguagem R funciona como um interpretador de código e não como um compilador. Quando se usa um interpretador o código está menos seguro, pois fica visível para os utilizadores ao contrário do que acontece com código compilado [53]. Assim o código desenvolvido em linguagem R não pode ficar na mesma máquina do SAD para diminuir os riscos de segurança. Por este motivo a componente de *Data Mining* ficará num servidor e a aplicação utilizada pelo utilizador apenas terá acesso aos resultados da previsão e não ao código-fonte. Para detalhar melhor o *design* do sistema é apresentado na Figura 9 o diagrama de componentes do sistema em linguagem UML.

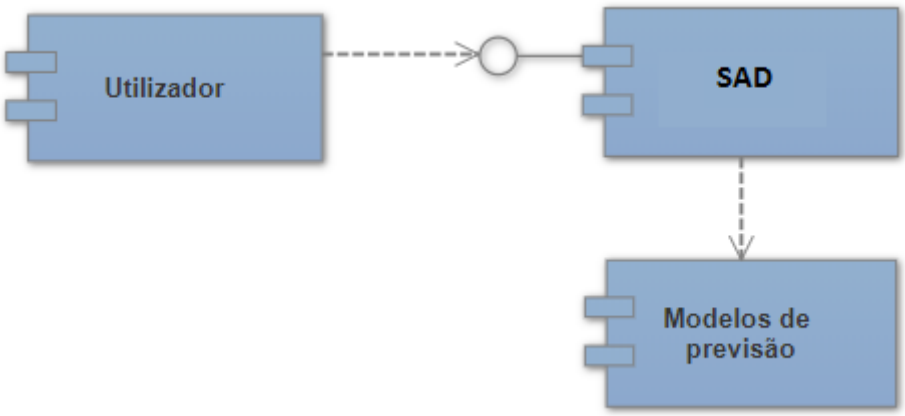

Figura 9 – Diagrama UML de componentes do Sistema.

Foram consideradas alternativas de design como desenvolver uma aplicação web ou criar um *data mart* para armazenar os dados. Tendo em conta que a componente principal do projeto é a componente de Data Mining, não se justifica fazer uma aplicação web no âmbito desta tese de mestrado. O *data mart* também não é essencial para o projeto, uma vez que para uma época de dados de jogos de futebol são apenas precisos 380 registos. Este número é reduzido quando comparado com aplicações que têm milhões de registos. Assim não se justifica a criação de um *data mart*.

5 Avaliação

Este capítulo descreve os métodos utilizados para avaliar o projeto, nomeadamente a componente de previsão de resultados e a solução final.

5.1 Avaliação de modelos de classificação

Para se fazer uma correta estimativa do erro dos modelos de previsão podem ser usadas diferentes técnicas. Dentro das diferentes técnicas as principais são [54]:

- Holdout – Os dados são divididos em 2 conjuntos independentes, um conjunto para treino e outro para teste. Normalmente são utilizados 2/3 dos dados para treino e 1/3 para teste. A vantagem deste método é o facto de os dados de teste não serem usados para treino, evitando assim a sobreposição dos dados (Refaeilzadeh *et al.*, 2009).
- Validação Cruzada – Os dados são divididos em *k* conjuntos de igual tamanho. Se, por exemplo, *k*=10 vão ser feitas um total de 10 iterações. Em cada iteração vão ser usados 9 subconjuntos de dados para treino e 1 subconjunto para teste. Em cada iteração o subconjunto de teste é diferente. A vantagem deste método é que todos os k subconjuntos são usados uma vez para teste.
- Método *Bootstrap* – Neste método, a partir de um conjunto inicial de dados com *N* registos são retiradas amostras com reposição e criados conjuntos de treino com *M* (M < N) registos mas com alguns registos repetidos. Os registos não selecionados para treino são utilizados para teste. Se, por exemplo, um conjunto inicial tiver 3 registos (10, 20, 30), os conjuntos de treino podiam ser (10,10,20), (30,20,20) ou (20,20,20) (Rodrigues, 2016).

Há ainda diferentes métricas para avaliação dos modelos de classificação. Uma dessas métricas é a matriz de confusão, exemplificada na Tabela 4. Esta matriz permite identificar 4 fatores [54]:

- TP (*True Positive*) – Registos positivos corretamente previstos.
- FP (*False Positive*) – Registos negativos indevidamente classificados como positivos.
- TN (True Negative) – Registos negativos classificados como negativos.
- FN (*False Negative*) – Registos positivos indevidamente classificados como negativos.

Outras métricas que podem ser derivadas da matriz de confusão e utilizadas são a *accuracy* (exatidão ou taxa de acerto), a precisão e o *recall*. A taxa de acerto, em inglês *accuracy*, representa a razão entre o número de acertos e o número total de registos. É uma das técnicas mais utilizadas para avaliar o desempenho de um algoritmo de classificação. A taxa

de acerto é calculada dividindo o número de exemplos corretamente classificados pelo número total de exemplos e pode ser usada tanto em problemas binários, como em problemas multi-classe [54].

Tabela 4 – Matriz de confusão, adaptado de [54]

	Classe positiva prevista	**Classe negativa prevista**
Classe positiva	TP	FN
Classe negativa	FP	TN

A precisão permite avaliar a quantidade de previsões positivas que são realmente positivas. O *recall* avalia a proporção de elementos positivos corretamente classificados. Estas métricas são calculadas, respetivamente, a partir das equações (9), (10) e (11) mostradas a seguir [54].

$$Accuracy\ (Taxa\ de\ acerto) = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Precisão = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

5.2 Avaliação da solução final

Para avaliar o projeto vai ser usada principalmente a medida *accuracy* (taxa de acerto) que indica a percentagem de jogos corretamente classificados. Esta é uma medida adequada visto tratar-se de um problema de classificação com mais de 2 classes, neste caso são 3. O modelo de previsão vai ainda ser avaliado tendo em conta o lucro previsto. Este lucro será calculado tendo em conta as *odds* dos jogos acertados pelo modelo e descontando o montante apostado nos jogos que não foram corretamente previstos. O método para calcular o lucro é detalhado na secção 7.2. A hipótese a testar é a previsão de quem ganha, empata ou perde um jogo de futebol com base em variáveis estatísticas de jogos anteriores como o número de golos ou número de remates. Para testar esta hipótese foram usados os algoritmos de classificação descritos na secção 2.4.

A metodologia de avaliação terá 2 partes. A primeira parte é a de análise dos dados, nessa fase, entre outras abordagens, é feita uma matriz de correlação de todas as variáveis para entender os relacionamentos entre elas. A segunda parte envolve a avaliação dos modelos de previsão. Nessa fase o modelo de previsão será testado em jogos de uma época completa. Como as equipas vão mudando ao longo de uma época é essencial testar os modelos em jogos de uma época inteira para o número de testes ser suficiente. No capítulo 7 é explicada com maior detalhe a avaliação feita.

6 Análise e processamento de dados

6.1 Descrição dos dados

Os dados utilizados correspondem a jogos de futebol de 5 épocas, as épocas 2011/2012, 2012/2013, 2013/2014, 2014/2015, 2015/2016 e 2016/2017, perfazendo um total de 1900 jogos. Os jogos são relativos à 1ª divisão de futebol de Inglaterra, cujo nome oficial é *Premier League* [55]. Os dados foram obtidos em formato "csv" do site "football-data" [56] que disponibiliza dados estatísticos de jogos de futebol de forma gratuita. Para cada jogo existem várias informações como o nome das equipas, golos marcados, remates efetuados e *odds* relativas a cada jogo. A lista completa de atributos disponíveis para cada jogo é a seguinte:

- Date – data de realização do jogo (dd/mm/aa);
- HomeTeam – equipa da casa;
- AwayTeam – equipa visitante;
- FTHG – número de golos marcados pela equipa da casa no final do jogo;
- FTAG – número de golos marcados pela equipa visitante no final do jogo;
- FTR – resultado final do jogo (H - vitória da equipa da casa, D – empate, A – vitória da equipa visitante);
- Referee – árbitro do jogo;
- HS – remates da equipa da casa;
- AS – remates da equipa visitante;
- HST – remates à baliza feitos pela equipa da casa;
- AST – remates à baliza feitos pela equipa visitante;
- HC – número de cantos para a equipa da casa;
- AC – número de cantos para a equipa visitante;
- HF – número de faltas cometidas pela equipa da casa;
- AF – número de faltas cometidas pela equipa visitante;
- HY – número de cartões amarelos mostrados a jogadores da equipa da casa;
- AY – número de cartões amarelos mostrados a jogadores da equipa visitante;
- HR – número de cartões vermelhos mostrados a jogadores da equipa da casa;
- AR – número de cartões vermelhos mostrados a jogadores da equipa visitante;
- B365H – *odd* para a vitória da equipa da casa na casa de apostas bet365 [57];
- B365D – *odd* para a ocorrência de empate na casa de apostas bet365;
- B365A – *odd* para a vitória da equipa visitante na casa de apostas bet365;

Nestes 1900 jogos a equipa da casa ganhou em 861 jogos (45,3%), houve 470 empates (24,7%) e a equipa visitante ganhou em 569 jogos, o que representa 29,9% do total de jogos. Para além destes dados, foram utilizados dados extraídos do site "sofifa.com" [58], que contêm informação estatística relativa à qualidade das equipas de futebol.

As equipas são avaliadas numa escala de 0 a 100 e os dados correspondem aos dados do jogo FIFA [59]. Os dados extraídos foram:

- OVA – classificação global da equipa numa escala de 0 a 100;
- ATT – classificação do ataque de uma equipa numa escala de 0 a 100;
- MID – classificação do meio-campo de uma equipa numa escala de 0 a 100;
- DEF – classificação da defesa de uma equipa numa escala de 0 a 100.

Para cada época foram extraídos estes 4 registos para cada equipa, por isso na mesma época estes valores são constantes para cada equipa.

6.2 Limpeza de dados

Nesta fase foi feita a limpeza do conjunto inicial de dados para as 5 épocas extraídas. Foi verificada a existência de atributos em falta para cada jogo e caso houvesse valores em falta para um dado jogo o registo desse jogo era eliminado para não diminuir a eficácia do modelo de classificação. Optou-se por utilizar este método visto existir um número inicial de dados elevado, o que permitia eliminar alguns registos e ainda assim continuar a ter um número de dados suficiente para fazer uma previsão correta. No entanto, como não foram encontrados valores em falta em nenhuma época, não foi necessário eliminar registos. Também não foram encontrados registos repetidos por isso o número de registos depois da fase de limpeza continuou nos 1900.

6.3 Criação de novos dados

De forma a obter melhores resultados, procurou-se criar novas variáveis que permitissem prever melhor o resultado final dos jogos. Por este motivo, tal como foi feito no caso de estudo (Gomes *et al.*, 2015), foram criadas as seguintes variáveis:

- HWINLAST5 – número de vitórias em casa da equipa que joga em casa;
- AWINLAST5 – número de vitórias fora da equipa visitante.

Nos casos de estudo analisados apenas se utilizavam dados relativos aos golos marcados pelas equipas e nunca aos golos sofridos. Por este motivo decidiu-se calcular os golos sofridos pelas equipas uma vez que estas estatísticas poderiam melhorar os resultados das previsões. As variáveis criadas foram:

- HGOLSOFR – média de golos sofridos em casa pela equipa da casa;
- AGOLSOFR – média de golos sofridos em jogos realizados fora pela equipa visitante.

6.4 Transformação de dados

Para que se possam fazer previsões de jogos de futebol que funcionem antes da ocorrência dos jogos é necessário fornecer aos modelos de previsão *Data Mining* dados que estejam disponíveis antes do início de cada jogo. Uma vez que os dados extraídos eram relativos ao final de cada jogo, como o número de golos e remates de cada equipa, estes dados não podiam ser usados diretamente para treinar os modelos de previsão. Por esta razão foi necessário transformar estes dados. A solução adotada neste caso, tal como em (Gomes *et al.*, 2015), foi considerar as médias dos dados disponíveis como a média de golos ou a média de remates de uma equipa antes de um dado jogo. Para cada jogo foram calculadas as médias dos seguintes atributos (descritos na secção 6.1):

- FTHG, FTAG, HS, AS, HST, AST, HC, AC, HF, AF, HY, AY, HR, AR.

Assim, para um determinado jogo, a média de golos para a equipa da casa é calculada com base nos jogos anteriores feitos em casa por essa equipa durante essa época. Para as equipas visitantes as médias são também calculadas com base nos jogos anteriores feitos como visitantes por parte dessas equipas durante a época em que é realizado o jogo. Para as restantes variáveis as médias são calculadas da mesma maneira. As novas variáveis criadas foram:

- AVGH – média de golos marcados da equipa da casa;
- AVGA – média de golos marcados da equipa visitante;
- AVGSH – média de remates da equipa da casa;
- AVGSA – média de remates da equipa visitante;
- AVGSTH – média de remates à baliza da equipa da casa;
- AVGSTA – média de remates à baliza da equipa visitante;
- HFOULAVG – média de faltas cometidas pela equipa da casa;
- AFOULAVG – média de faltas cometidas pela equipa visitante;
- HCORAVG – média de cantos para a equipa da casa;
- ACORAVG – média de cantos para a equipa visitante;
- HYELAVG – média de cartões amarelos para a equipa da casa;
- AYELAVG – média de cartões amarelos para a equipa visitante:
- HREDAVG – média de cartões vermelhos para a equipa da casa:
- AREDAVG – média de cartões vermelhos para a equipa visitante:

6.5 Exploração de dados

Depois das etapas anteriores e antes da etapa de previsão foi feita uma exploração dos dados para verificar no total de 31 variáveis disponíveis quais as que mais se relacionavam entre si. Verificou-se também quais as variáveis que poderiam prever de forma mais clara o atributo objetivo e variáveis que podiam ser excluídas.

6.5.1 Correlação entre variáveis

Para verificar o relacionamento entre variáveis foi feita uma matriz de correlação entre todas as variáveis numéricas. A matriz gerada é mostrada na Figura 10, a matriz apresenta em cada linha cada uma das variáveis, mostrando em cada coluna a correlação com todas as outras variáveis. Os pontos azuis representam uma correlação positiva e os vermelhos uma correlação negativa. Os pontos maiores e mais escuros representam uma correlação maior entre variáveis.

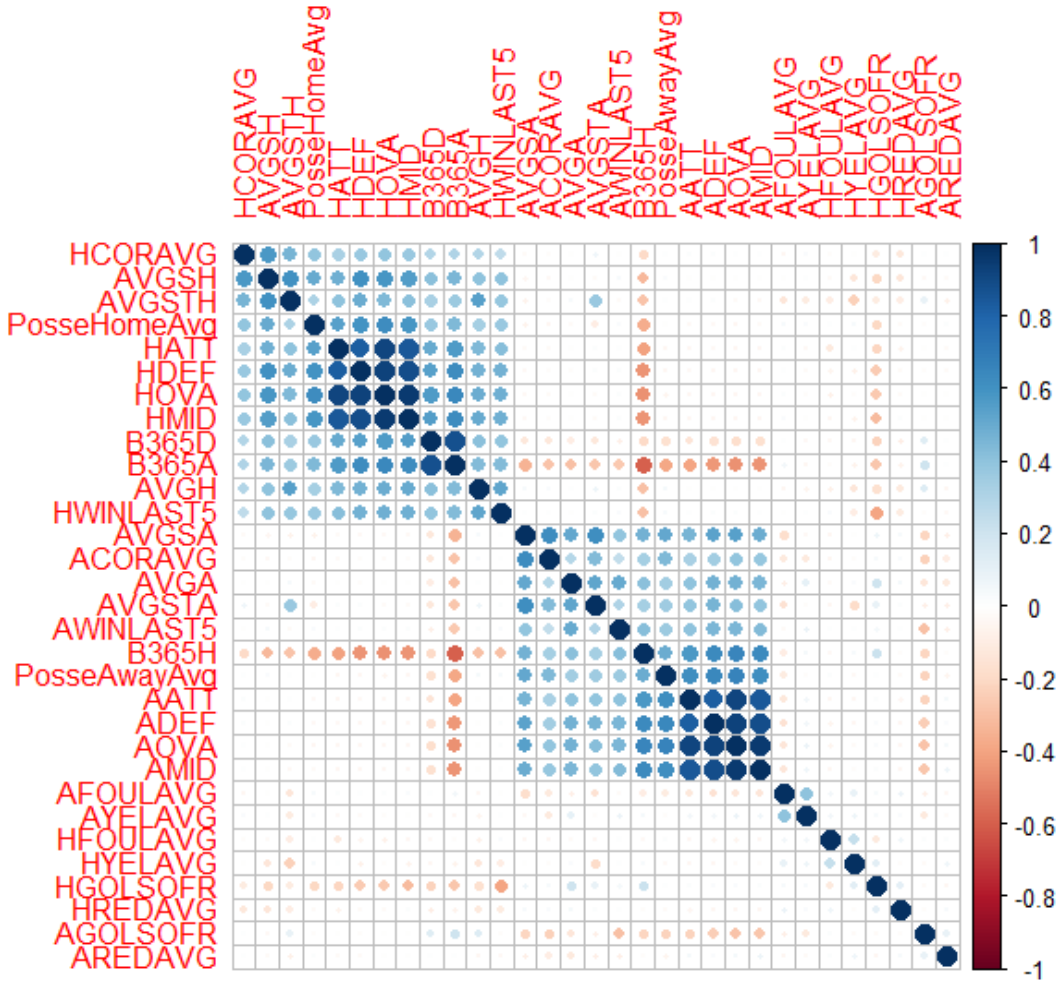

Figura 10 – Correlação entre variáveis.

A correlação entre 2 variáveis nunca é maior que 1, visto que a correlação de uma variável consigo mesma é 1. A análise da Figura 10 permitiu tirar algumas conclusões. As variáveis que têm maior correlação positiva entre si são HATT, HDEF, HOVA e HMID e AATT, ADEF, AOVA, e AMID. Isto era de esperar visto serem todas variáveis que avaliam a qualidade de uma equipa de futebol. Há também uma forte correlação entre B365D e B365A que representam as *odds* para a ocorrência de empate e vitória da equipa visitante. Para além disso consegue-se verificar que as variáveis relacionadas com a equipa da casa têm maior correlação entre si, tal como as variáveis relacionadas com a equipa visitante também têm maior correlação positiva entre si. As variáveis relativas ao número de faltas, cartões amarelos e vermelhos e número de golos sofridos não têm relação positiva com outras variáveis. No entanto verifica-se que HGOLSOFR tem uma correlação negativa com HWINLAST5, o que seria de esperar visto as equipas que sofrem mais golos terem mais dificuldades para ganhar jogos. Pôde-se ainda concluir que a variável B365H está relacionada negativamente com variáveis como B365A, HATT ou HMID. Isto é normal visto que quanto maior for a *odd* para a equipa da casa, menor é a *odd* para a equipa visitante. Da mesma forma, quanto melhor for a equipa da casa ou seja, quanto maiores forem os valores de HATT e HMID, menor é a *odd* para a vitória da equipa da casa.

A matriz de correlação foi também importante para identificar variáveis a remover do conjunto inicial de dados. Num processo de classificação devem ser removidas as variáveis com grande correlação com outras variáveis (Dasgupta, 2018). Isto deve ser feito para que a importância dessas variáveis não seja sobrevalorizada, prejudicando a previsão dos resultados. No caso de haver 2 variáveis idênticas, uma delas passa a ser redundante, não acrescentado informação relevante para o modelo de treino. Por estas razões foram removidas as variáveis com correlação maior ou igual a 0.9. Para remover as variáveis foi utilizada a função “findCorrelation” do *package* “caret” do software R [60]. Esta função encontra as variáveis que têm correlação maior que 0.9 entre si e remove a variável com maior correlação média absoluta. As variáveis removidas neste processo foram as variáveis HATT e AATT, correspondentes à força de ataque da equipa da casa e visitante.

6.5.2 Relação das variáveis com atributo objetivo

Uma etapa importante antes da fase de previsão era tentar identificar as variáveis que ajudassem a prever mais facilmente o atributo/variável objetivo, neste caso a variável FTR. Assim foram feitos gráficos *boxplot* para verificar a relação entre todas as variáveis e os valores de FTR, 1 - para a vitória da equipa da casa, 2- para empate e 3 - para a vitória da equipa visitante. Os gráficos permitiram verificar que há 4 variáveis que melhor ajudam a prever o resultado final de um jogo. Essas variáveis são a B365H, B365A, AOVA, HWINLAST5. A Figura 11 mostra o gráfico relativo à variável B365H.

B365H

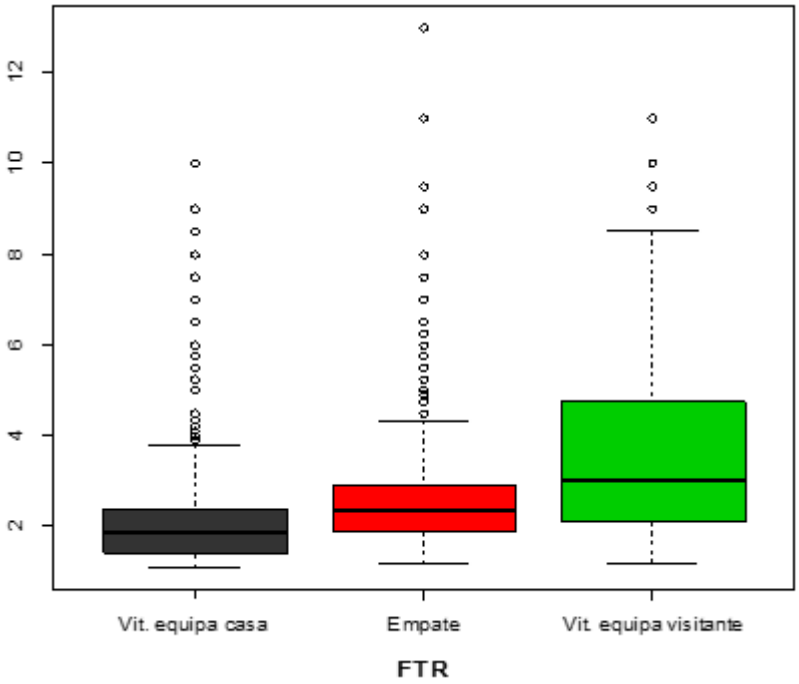

Figura 11 – Boxplot de B365H – *odd* da equipa da casa

Através da análise da Figura 11 é possível verificar que numa grande parte dos jogos ganhos pela equipa visitante, correspondendo aos valores entre o 2º e o 3º quartil, a *odd* é maior que 3, para além disso, acima destes valores ocorrem poucas vitórias da equipa da casa e empates. A Figura 12 mostra o gráfico relativo à variável B365A.

B365A

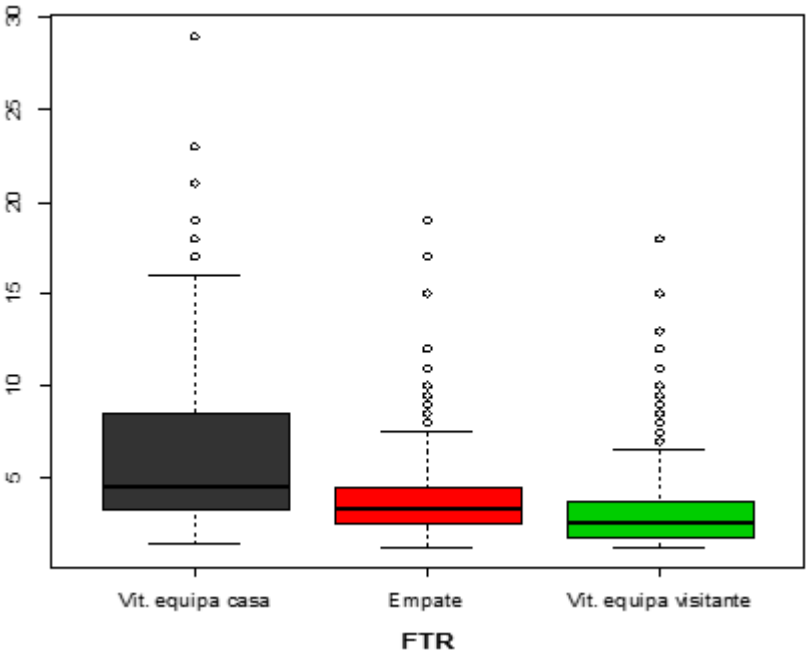

Figura 12 – Boxplot de B365A – *odd* para a vitória da equipa visitante

Na Figura 12 é possível verificar que quando a *odd* para a vitória da equipa visitante é maior que 5 ocorrem mais de 50% das vitórias da equipa caseira. Ao mesmo tempo ocorrem menos de 25% dos empates e vitórias dos visitantes, correspondentes aos valores acima do 3º

quartil. A Figura 13 mostra o gráfico relativo à variável AOVA. Neste gráfico pode-se constatar que quando as equipas visitantes são mais fortes, ou seja, quando o valor de AOVA é maior, há mais vitórias das equipas visitantes. Quando o valor de AOVA é menor ocorrem mais vitórias da equipa da casa. A Figura 14 mostra o gráfico relativo à variável HWINLAST5.

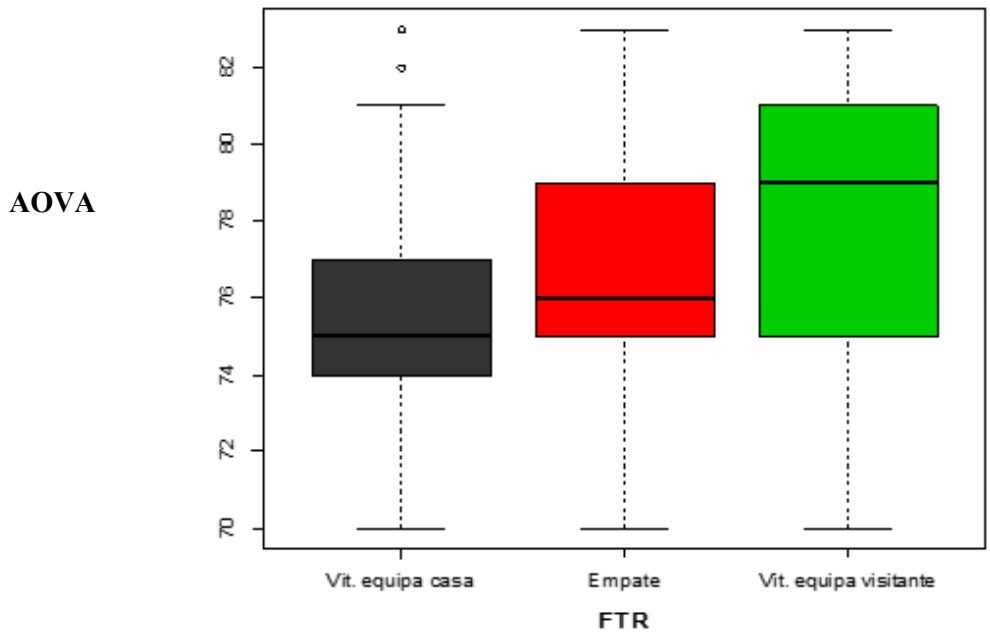

Figura 13 – Boxplot de AOVA – classificação média da equipa visitante

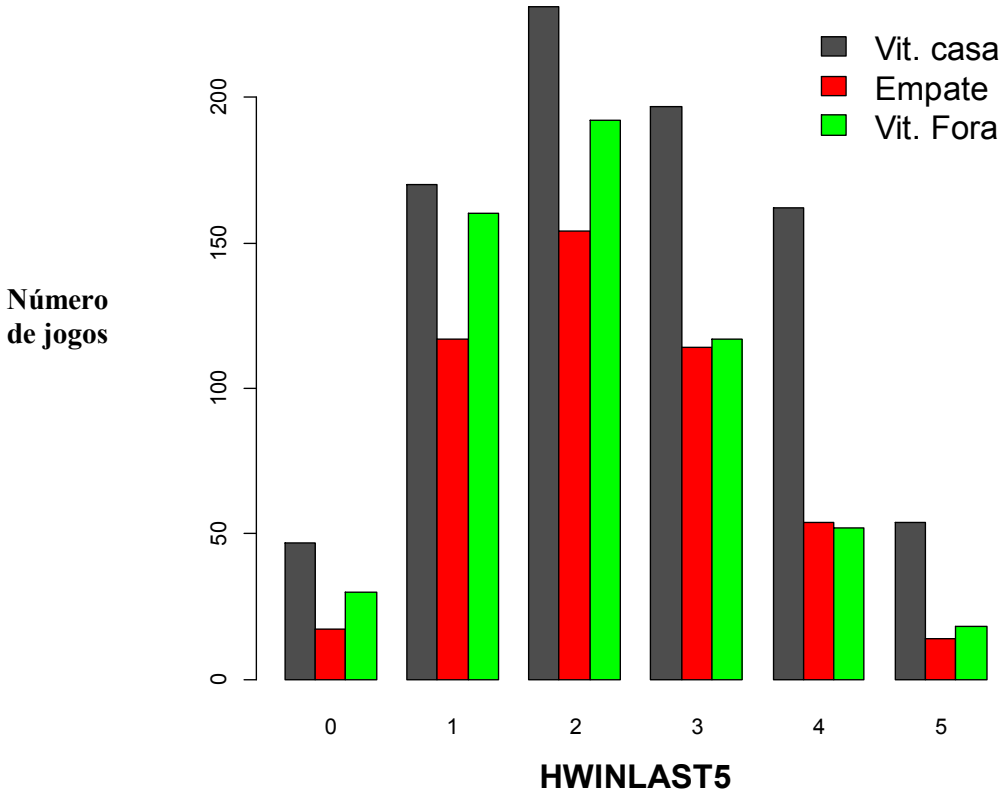

Figura 14 - Gráfico de HWINLAST5 com FTR – vitórias nos últimos 5 jogos da equipa da casa

A análise do gráfico de barras da Figura 14 permite constatar que há diferenças claras no resultado final do jogo quando há 3 ou 4 vitórias nos últimos 5 jogos por parte da equipa da casa. Nestes 2 casos, correspondentes ao 4º e 5º conjunto de barras do gráfico, o número de jogos que acabam empatados ou com a vitória da equipa visitante é idêntico. Já o número de jogos em que a equipa da casa vence é bastante superior, no caso em que há 4 vitórias nos últimos 5 jogos é até superior ao total de empates e vitórias da equipa visitante. Assim esta variável permite prever mais facilmente se a equipa da casa ganha ou não o jogo. No entanto, esta variável não ajuda a prever mais facilmente empates ou vitórias da equipa visitante, visto o número de vitórias ser idêntico nestes 2 casos.

Para além de verificar as variáveis que poderiam oferecer uma melhor previsão, foram também analisadas as que teriam pior capacidade para prever o atributo objetivo. Verificou-se que as variáveis HREDAVG, AREDAVG, HCORAVG e ACORAVG seriam pouco relevantes para treinar o modelo de previsão. Por este motivo estas 4 variáveis não foram utilizadas para treinar os modelos de previsão. A Figura 15 e a Figura 16 mostram os gráficos para a variáveis HREDAVG e AREDAVG. Como é possível verificar na Figura 15 a alteração do número de cartões vermelhos não tem influência no resultado final de um jogo. Na Figura 16 verifica-se que há diferenças muito reduzidas.

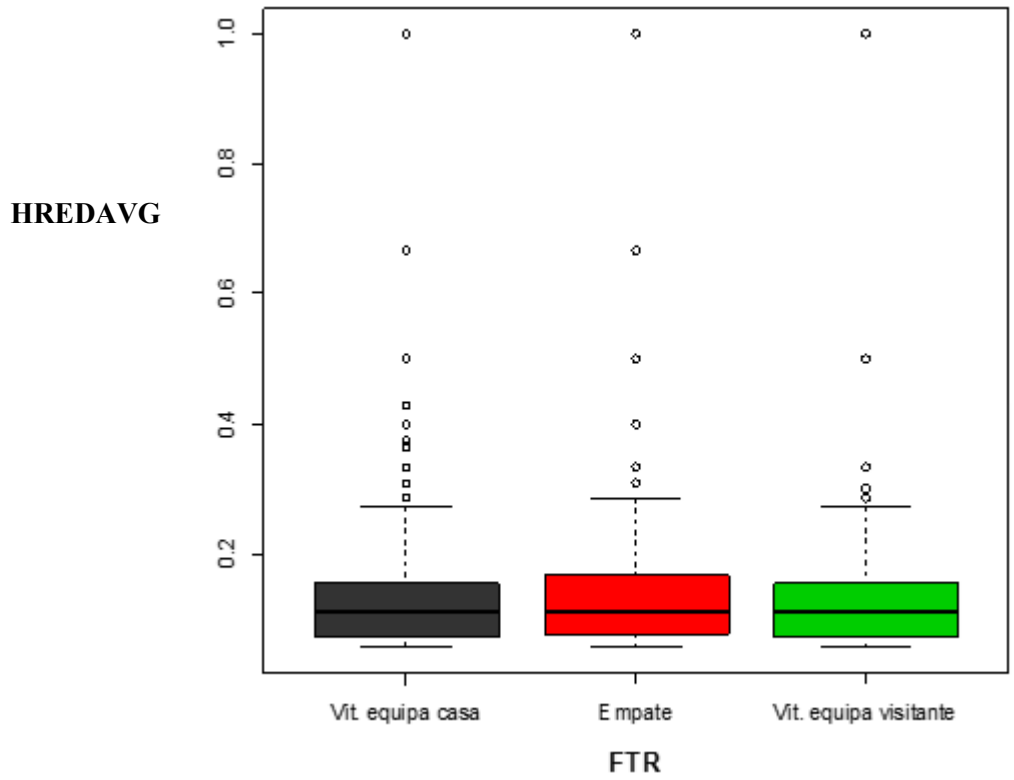

Figura 15 – Boxplot de HREDAVG – média de cartões vermelhos da equipa da casa.

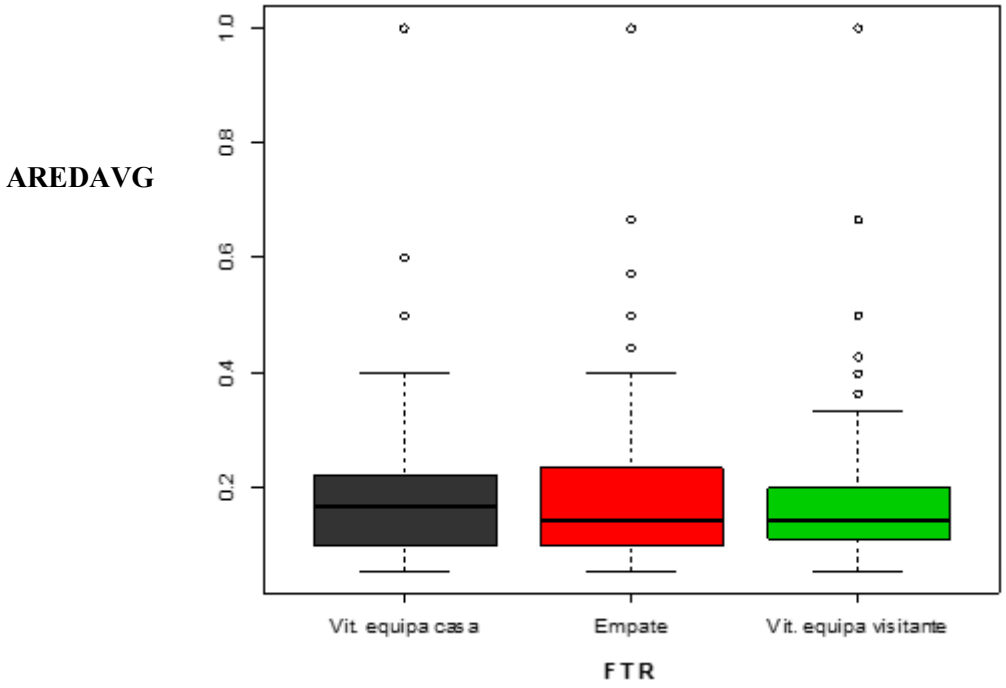

Figura 16 – Boxplot de AREDAVG – média de cartões vermelhos da equipa visitante.

A Figura 17 e a Figura 18 mostram os gráficos *boxplot* para as variáveis HCORAVG e ACORAVG. Os gráficos permitem verificar que não há diferença significativa no número de cantos quando os resultados dos jogos são diferentes. Dentro das variáveis disponíveis as variáveis HCORAVG e ACORAVG eram das que tinham menos capacidade para prever o atributo objetivo, por isso estas variáveis não foram utilizadas para treinar os modelos de previsão.

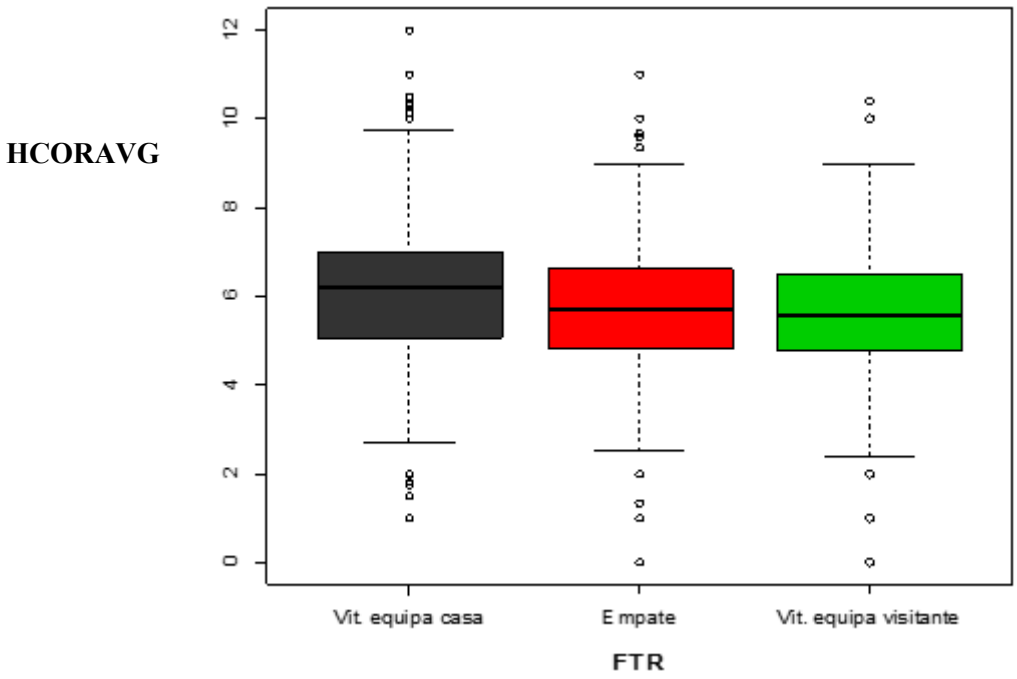

Figura 17 – Boxplot da variável HCORAVG – média de cantos para a equipa da casa.

ACORAVG

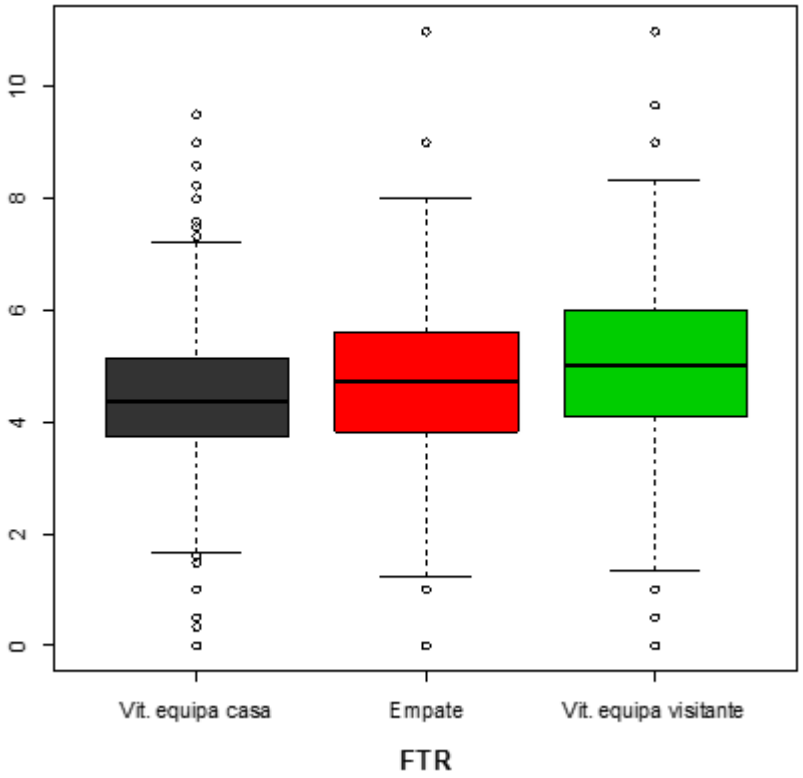

Figura 18 – Boxplot da variável ACORAVG – média de cantos para a equipa visitante.

7 Previsão de resultados

Após a etapa de processamento e análise dos dados foi feita a previsão de resultados de jogos utilizando *Data Mining*. Esta fase começou pela escolha dos dados para treino e para teste. Nos casos de estudo analisados, verificou-se que o número de jogos usados para teste era reduzido, nunca sendo maior que 110. Para este trabalho considerou-se que o conjunto de teste deveria incluir todos os jogos de uma época, um total de 380 jogos. Há vários motivos para se ter escolhido esta metodologia. Um desses motivos é que ao longo de uma época o desempenho das equipas vai variando. No início da época as equipas podem ainda não estar no seu nível normal e no final da época podem estar desgastadas, ou já ter atingido os seus objetivos, levando algumas equipas a obter um rendimento abaixo do normal. Estes fatores podem levar à ocorrência de resultados inesperados, por isso, para que um modelo de classificação possa ser considerado credível, deve ser testado durante uma época inteira. Há ainda outro fator importante que foi considerado. Caso um modelo seja treinado com poucos exemplos pode haver risco de *overfitting*, ou seja, de o modelo de treino se ajustar demasiado aos dados de treino. Por estes motivos, uma vez que existiam dados de 5 épocas, foram utilizadas 4 épocas para treino e 1 época para teste. As 4 épocas de treino correspondem a um total de 1520 jogos das épocas 2011/2012, 2012/2013, 2013/2014 e 2014/2015. A época de teste foi a 2016/2017 com 380 jogos.

7.1 Metodologia

Para conseguir encontrar o melhor modelo de classificação possível foram testados 8 algoritmos. Foram testados vários algoritmos porque todos têm características diferentes e assim era possível verificar qual dos algoritmos se adaptava melhor ao problema em questão. Todos os algoritmos foram executados com o software R. Em seguida são descritos os algoritmos utilizados, bem como as bibliotecas usadas para os executar com o software R:

- *Naive Bayes* – método "naiveBayes" do *package* e1071;
- K-vizinhos mais próximos (KNN) – *packag*e kknn;
- *Random Forest* (RF) – *package* randomForest;
- Support Vector Machines (SVM) - método "svm" do *package* e1071;
- C5.0 (árvores de decisão) – *package* C50;
- Xgboost – *package* xgboost;
- Regressão Logística Multinomial (RLM) – método "multinom" do *package* nnet;
- Redes Neuronais Artificiais (RNA) – método "nnet" do *package* nnet.

Antes de testar os diferentes algoritmos foi feita a normalização dos dados. Normalizar os dados faz com que estes tenham uma escala comum e permite eliminar o efeito de grandes variações de valores (Dasgupta, 2018). A normalização é útil para melhorar o desempenho dos algoritmos de previsão (Shalabi *et al.*, 2006). A normalização foi feita através do método

"z-score" [61]. O "z-score" foi utilizado porque permite incluir valores isolados na análise [62]. Por exemplo, um valor isolado como uma equipa ter uma média de 5 golos por jogo pode ajudar a prever mais facilmente a vitória dessa equipa.

Depois de normalizar os dados foram identificadas as variáveis mais importantes para a previsão dos resultados. Este processo foi feito para que os modelos de previsão utilizassem apenas as variáveis mais relevantes e que garantissem melhores previsões. Para identificar as melhores variáveis foi utilizado o algoritmo Boruta [63]. O Boruta é um algoritmo heurístico de seleção de variáveis baseado em *random forests* que tem o objetivo de encontrar as variáveis mais relevantes num conjunto de dados (Kursa e Rudnicki, 2010). Os resultados da execução deste algoritmo mostram as variáveis relevantes e não relevantes. Este algoritmo foi utilizado porque não procura uma solução subótima [63]. Ao invés, tenta encontrar todas as variáveis com informação relevante, permitindo assim eliminar variáveis que afetariam negativamente os modelos de previsão. Os resultados de execução do Boruta mostraram que as variáveis não relevantes eram as variáveis HomeTeam, AwayTeam, HYELAVG, AYELAVG, HFOULAVG, AFOULAVG e AWINLAST5. Por este motivo estas variáveis não foram utilizadas para treinar os modelos de classificação. As 18 variáveis consideradas importantes pelo algoritmo Boruta e utilizadas na construção dos modelos de classificação foram:

- B365H, B365D, B365A – *odds* para a vitória da equipa da casa, empate e vitória da equipa visitante.
- AVGH, AVGA – média de golos marcados pela equipa da casa e visitante.
- AVGSH, AVGSA – média de remates feitos pela equipa da casa e visitante.
- AVGSTH, AVGSTA – média de remates à baliza feitos pela equipa da casa e visitante.
- HWINLAST5 – número de vitórias nos últimos 5 jogos da equipa da casa.
- HGOLSOFR, AGOLSOFR – média de golos sofridos pela equipa da casa e visitante.
- HOVA, AOVA – força global da equipa da casa e visitante.
- HMID, AMID – força dos jogadores do meio-campo da equipa da casa e visitante.
- HDEF, ADEF – força dos jogadores da defesa da equipa da casa e visitante.

7.2 Resultados Previsões – Fase 1

Para poder verificar de forma mais correta as diferenças entre os modelos de classificação foram utilizadas diferentes medidas como a *accuracy* do modelo e a percentagem de jogos corretamente previstos para os empates e para as vitórias da equipa da casa e visitante. Considerou-se ainda o lucro que se obteria caso se acertasse ou errasse em cada aposta. Uma vez que o objetivo de construção dos modelos é o desenvolvimento de um sistema de recomendação de apostas, calcular o lucro obtido é essencial para verificar se o modelo é bem-sucedido. O lucro foi calculado considerando um valor de 2 euros por aposta. Tendo em conta que há 380 jogos de teste o valor total apostado seria de 760 euros, que seriam apostados ao longo de 9 meses. Caso se falhasse uma aposta o lucro diminuía 2 euros. Se se acertar o lucro é calculado de acordo com a equação (12).

$$Lucro = valor\ apostado * odd\ da\ aposta - valor\ apostado \quad (12)$$

Por exemplo, num jogo em que se apostasse num empate e a *odd* do empate fosse 1.5 o lucro seria de 1 euro. Como o valor apostado é sempre de 2 euros o lucro seria igual a 2*1.5-2=1. A Tabela 5 apresenta os resultados obtidos das previsões feitas com os 8 modelos que utilizaram os 8 algoritmos descritos na secção 7.1.

Tabela 5 – Resultados das previsões com 18 variáveis.

Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
Bayes	53,42%	17,40€	51,87%	30,95%	73,79%
KNN	57,63%	78,02€	78,07%	15,48%	55,05%
RF	59,21%	85,20€	75,40%	21,43%	60,55%
SVM	61,32%	95,06€	88,77%	3,57%	58,72%
C5.0	55,26%	42,52€	72,73%	23,81%	49,54%
Xgboost	59,47%	72,80€	77,54%	10,71%	66,06%
RLM	57,63%	32,56€	78,07%	5,95%	62,34%
RNA	50,00%	18,28€	58,29%	30,95%	50,46%

Os resultados das previsões foram satisfatórios. O melhor algoritmo foi o SVM, conseguindo uma percentagem de acerto acima de 61%. O lucro de 95,06 euros, apesar de não ser elevado, corresponde a uma margem de lucro de 12,51%, que é uma margem de lucro razoável. De realçar ainda que todos os algoritmos permitem obter lucro. No entanto, a melhor taxa de acerto obtida não superou as melhores taxas de acerto encontradas nos casos de estudo referenciados. O melhor modelo acertou apenas em 3,57% dos empates, sendo este valor baixo. Uma vez que os resultados ainda não estavam de acordo com o esperado decidiu-se proceder a uma segunda fase de construção de modelos de previsão.

7.3 Resultados Previsões – Fase 2

Nesta segunda fase de construção de modelos decidiu-se testar todas as combinações possíveis com as 18 variáveis pré-selecionadas, de modo a atingir a melhor taxa de acerto possível. Todavia, com 18 variáveis, o número total de combinações a testar seria superior a 260 mil. Este número é elevado porque seria necessário testar todas as combinações de 18, 17, 16, 15… variáveis até 2 variáveis. O tempo necessário para testar tantas combinações seria demasiado elevado e não seria razoável optar por esta abordagem. Por isso decidiu-se reduzir o número de combinações testadas e combinar apenas 12 variáveis. O número total de combinações necessárias neste caso é de 4094, sendo esta abordagem exequível.

Para testar esta nova abordagem foram identificadas as 6 variáveis mais importantes das 18 iniciais. Todas as combinações testadas utilizaram estes 6 atributos mais importantes, variando apenas os restantes 12. As variáveis mais importantes foram obtidas utilizando o método *Backwards Feature Sel*ection "rfe" do *package* "caret" [60] do software R. Este algoritmo começa por verificar a importância das variáveis usando todas as variáveis fornecidas, depois faz sucessivas iterações onde vai retirando algumas variáveis, ficando apenas com as mais importantes em cada iteração [64]. No final as variáveis mais importantes são as que foram utilizadas no teste que obteve melhor resultado. Os 6 atributos identificados como mais importantes foram: B365H, B365D, B365A, AVGH, HOVA e AGOLSOFR.

7.3.1 Combinações com 1 variável

Combinando os 6 atributos mais importantes com um dos restantes 12 atributos a combinação que obteve melhor taxa de acerto foi com o atributo AVGA e o algoritmo SVM. Os resultados são apresentados na Tabela 6. Utilizando apenas 7 variáveis para prever os resultados a taxa de acerto subiu para 62,62%, um aumento de mais de 1%. O lucro também aumentou, tal como o acerto nos empates e vitórias da equipa de fora. Apenas diminuiu a percentagem de apostas corretas em vitórias da equipa da casa. Confirmou-se assim que utilizando esta abordagem é possível obter melhores resultados do que os iniciais.

Tabela 6 – Melhor previsão com 7 variáveis.

Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
SVM	62,62%	134,50€	51,87%	30,95%	73,79%

7.3.2 Combinações com 2 variáveis

Ao combinar as 6 variáveis mais importantes com outras 2 variáveis foi possível aumentar a taxa de acerto, subindo esta para 63,68%. Neste caso houve 2 combinações que atingiram os mesmos resultados, a primeira com as variáveis AVGA e AOVA e a segunda com as variáveis AVGA e ADEF. A Tabela 7 mostra os melhores resultados obtidos. Os resultados obtidos são muito semelhantes, sendo que no primeiro o lucro é maior tal como o acerto em vitórias caseiras. No segundo caso a percentagem de acerto em empates é superior em mais de 2%. Comparando com os resultados obtidos anteriormente, estes resultados são melhores, visto que há um maior número de apostas corretas e há um maior lucro.

Tabela 7 – Melhores previsões com 8 variáveis.

Nº Teste	Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
1	SVM	63,68%	157,22€	90,91%	4,76%	62,39%
2	SVM	63,68%	154,22€	89,84%	7,14%	62,39%

7.3.3 Combinações com 3 variáveis

As combinações feitas com 3 variáveis permitiram aumentar mais uma vez a *accuracy* atingida para 63,95%. O melhor algoritmo foi mais uma vez o SVM e a melhor combinação foi feita com as variáveis AVGA, AOVA e AMID. Os resultados são apresentados na Tabela 8.

Tabela 8 – Melhor previsão com 9 variáveis.

Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
SVM	63,95%	160,12€	91,44%	3,57%	63,30%

Este modelo de previsão permitiu mais uma vez aumentar o lucro obtido. É possível verificar que a percentagem de empates acertados é baixa, sendo que o principal fator para obter bons resultados é o facto de se acertar em mais de 91% das apostas em vitórias da equipa da casa.

7.3.4 Combinações com 4 variáveis

Neste caso a taxa de acerto foi igual à obtida combinando 3 variáveis. De realçar que o melhor algoritmo foi o Xgboost, enquanto que em todos os casos anteriores o melhor tinha sido o SVM. A Tabela 9 apresenta os resultados obtidos com a melhor combinação de variáveis. Esta combinação foi obtida com as variáveis AVGA, AVGSA, AVGSH e ADEF. Em comparação com os resultados da Tabela 8 o lucro e o acerto em vitórias da equipa da casa diminuíram, mas aumentou a taxa de acerto em empates e vitórias da equipa visitante.

Tabela 9 – Melhor previsão com 10 variáveis.

Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%

7.3.5 Combinações com 5 variáveis

A taxa de acerto obtida situou-se nos 63,95%, igual à dos resultados obtidos nas seções 7.3.3 e 7.3.4. Houve 3 combinações que obtiveram esta taxa de acerto, tal como é mostrado na Tabela 10. As 3 combinações foram obtidas com as seguintes variáveis:

- Teste 1 – AVGA, AVGSA, AVGSH, ADEF e HMID;
- Teste 2 – AVGA, AVGSA, AVGSH, ADEF e AOVA;
- Teste 3 – AVGA, AVGSA, AVGSH, ADEF e AMID.

Tabela 10 – Melhores previsões com 11 variáveis

Nº Teste	Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
1	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
2	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
3	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%

Uma vez que os resultados dos 3 testes são iguais, isto pode significar que as variáveis HMID, AOVA e AMID têm uma influência para a previsão dos resultados semelhante entre si. Os resultados são iguais aos obtidos com o melhor modelo que combinou 4 variáveis da secção 7.3.4. A única diferença desse modelo é a não utilização de nenhuma das variáveis HMID, AOVA e AMID. Por esta comparação pode-se concluir que a influência destas 3 variáveis é baixa, quando comparada com as restantes, visto os resultados serem iguais incluindo ou não essas variáveis.

7.3.6 Combinações com 6 variáveis

Neste conjunto de teste voltou a haver 3 combinações a atingir uma *accuracy* de 63,95%. Novamente o melhor algoritmo nas três foi o Xgboost, tal como é mostrado na Tabela 11.

Tabela 11 – Melhores previsões com 12 variáveis.

Nº Teste	Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
1	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
2	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
3	Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%

As 3 combinações foram obtidas com as seguintes variáveis:

- Teste 1 – AVGA, AVGSA, AVGSH, ADEF, HMID e AOVA;
- Teste 2 – AVGA, AVGSA, AVGSH, ADEF, HMID e AMID;
- Teste 3 – AVGA, AVGSA, AVGSH, ADEF, AOVA e AMID.

Os resultados são iguais aos da Tabela 9 e da Tabela 10. Estes resultados parecem comprovar que as variáveis HMID, AOVA e AMID têm uma influência reduzida quando utilizadas em conjunto com as restantes variáveis.

7.3.7 Combinações com 7 variáveis

Nas combinações com 7 variáveis a melhor taxa de acerto obtida manteve-se nos 63,95%. Houve novamente 3 combinações a atingir 63,95%, mas nos 3 casos os algoritmos são diferentes, nomeadamente o Xgboost, KNN e RF. Os algoritmos KNN e RF permitem obter

maior lucro do que o Xgboost, isto acontece porque acertam em mais empates do que o Xgboost. O algoritmo RF é o que acerta em mais empates, sendo a diferença de mais de 10% comparando com os outros 2 algoritmos. Os resultados são apresentados na Tabela 12.

Tabela 12 – Melhores previsões com 13 variáveis.

Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
Xgboost	63,95%	150,28€	85,03%	10,71%	68,81%
KNN	63,95%	196,18€	85,03%	17,86%	63,30%
RF	63,95%	191,58€	80,21%	28,57%	63,30%

As 3 combinações foram feitas com as variáveis apresentadas a seguir:

- Xgboost – AVGA, AVGSA, AVGSH, ADEF, AOVA, AMID e HMID;
- KNN – AVGA, AVGSA, AVGSTH, AVGSTA, HDEF, HGOLSOFR e HWINLAST5;
- RF – AVGA, AVGSTH, AVGSTA, AOVA, HDEF, HGOLSOFR e HWINLAST5.

7.3.8 Combinações com 8 variáveis

Os testes efetuados combinando 8 variáveis, para além das 6 variáveis mais importantes utilizadas em todos os testes, permitiram aumentar a taxa de acerto para 65,26%. O melhor algoritmo foi o RF e a melhor combinação de 8 variáveis utilizou as variáveis AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HDEF, AOVA, AMID e ADEF. Este foi o modelo de previsão que obteve maior lucro e taxa de acerto, como é possível verificar na Tabela 13.

Tabela 13 – Melhor previsão com 14 variáveis.

Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
RF	65,26%	203,24€	81,28%	29,76%	65,14%

7.3.9 Combinações com 9 variáveis

Combinando os 6 atributos mais importantes com mais 9 atributos a melhor taxa de acerto obtida foi de 62,89%. Houve 5 combinações a atingir este valor, quatro delas com o algoritmo SVM e uma usando RNA. A Tabela 14 mostra os melhores resultados obtidos.

Tabela 14 – Melhores previsões com 15 variáveis.

Nº Teste	Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
1	SVM	62,89%	135,50€	89,30%	9,52%	58,72%
2	SVM	62,89%	139,66€	89,30%	7,14%	60,55%
3	RNA	62,89%	181,72€	82,35%	27,38%	56,88%
4	SVM	62,89%	131,16€	89,84%	4,76%	61,47%
5	SVM	62,89%	131,06€	89,84%	5,95%	60,55%

As 5 combinações foram feitas com as seguintes variáveis:

- Teste 1 – AVGA, AVGSA, AVGSH, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HDEF, ADEF;
- Teste 2 – AVGA, AVGSA, AVGSH, HWINLAST5, HGOLSOFR, HMID, HDEF, AOVA, AMID;
- Teste 3 – AVGA, AVGSH, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HMID, HDEF, AMID;
- Teste 4 – AVGA, AVGSA, AVGSTH, HWINLAST5, HGOLSOFR, HMID, HDEF, AMID, ADEF;
- Teste 5 – AVGA, AVGSA, AVGSTH, HWINLAST5, HGOLSOFR, HMID, AOVA, AMID, ADEF.

Os modelos de classificação que utilizaram SVM tiveram resultados idênticos. Quanto ao modelo de RNA, o lucro foi maior, bem como o número de empates corretamente previstos.

7.3.10 Combinações com 10 variáveis

Os resultados pormenorizados são apresentados na Tabela 15.

Tabela 15 – Melhores previsões com 16 variáveis.

Nº Teste	Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
1	SVM	62,36%	123,36€	88,77%	8,33%	58,72%
2	SVM	62,36%	120,46€	88,77%	7,14%	59,63%
3	SVM	62,36%	116,86€	89,30%	4,76%	60,55%
4	SVM	62,36%	123,26€	89,84%	5,95%	58,72%
5	SVM	62,36%	123,66€	89,30%	7,14%	58,72%
6	SVM	62,36%	116,76€	89,30%	4,76%	60,55%
7	SVM	62,36%	119,16€	89,84%	3,57%	60,55%
8	SVM	62,36%	116,96€	89,84%	3,57%	60,55%
9	SVM	62,36%	122,06€	89,30%	3,57%	61,47%

Nesta situação a melhor taxa de acerto foi de 62,36%. Houve 9 combinações diferentes com 10 variáveis a obter este resultado, nos 9 casos o algoritmo foi o SVM. Os resultados dos 9 testes são idênticos. Existiram ligeiras variações no lucro e na percentagem de empates acertados, mas nenhum dos modelos foi claramente superior ou inferior aos restantes.

As combinações utilizadas em cada teste foram:

- Teste 1 – AVGA, AVGSH, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HDEF, AOVA, ADEF;
- Teste 2 – AVGA, AVGSH, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, AOVA, AMID, ADEF;
- Teste 3 – AVGA, AVGSH, AVGSA, AVGSTH, HWINLAST5, HGOLSOFR, HMID, AOVA, AMID, ADEF;
- Teste 4 – AVGA, AVGSH, AVGSA, AVGSTA, HWINLAST5, HGOLSOFR, HMID, HDEF, AOVA, AMID;
- Teste 5 – AVGA, AVGSH, AVGSA, AVGSTA, HWINLAST5, HGOLSOFR, HDEF, AOVA, AMID, ADEF;
- Teste 6 – AVGA, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HMID, HDEF, AOVA, ADEF;
- Teste 7 – AVGA, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HMID, HDEF, AMID, ADEF;
- Teste 8 – AVGA, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HMID, AOVA, AMID, ADEF;
- Teste 9 – AVGSH, AVGSA, AVGSTA, HWINLAST5, HGOLSOFR, HMID, HDEF, AOVA, AMID, ADEF.

7.3.11 Combinações com 11 variáveis

No último conjunto de testes, onde se testaram modelos com as 6 variáveis mais importantes e 11 das restantes variáveis, a melhor *accuracy* obtida foi de 62,11%. Houve 4 modelos a atingir esta percentagem, todos eles utilizando o algoritmo SVM. A Tabela 16 mostra os resultados obtidos. Neste caso todos os modelos são também parecidos, as diferenças existentes são pouco significativas. As combinações de variáveis de cada modelo foram:

- Teste 1 – AVGA, AVGSH, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HMID, HDEF, AOVA, ADEF;
- Teste 2 – AVGA, AVGSH, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HDEF, AOVA, AMID, ADEF;
- Teste 3 – AVGA, AVGSH, AVGSA, AVGSTH, HWINLAST5, HGOLSOFR, HMID, HDEF, AOVA, AMID, ADEF;
- Teste 4 – AVGA, AVGSA, AVGSTH, AVGSTA, HWINLAST5, HGOLSOFR, HMID, HDEF, AOVA, AMID, ADEF;

Tabela 16 – Melhores previsões com 17 variáveis.

Nº Teste	Algoritmo	Accuracy	Lucro	% Vit. Casa	% Empates	% Vit. Fora
1	SVM	62,11%	113,86€	88,77%	5,95%	59,63%
2	SVM	62,11%	115,06€	88,77%	7,14%	58,72%
3	SVM	62,11%	114,66€	89,30%	5,95%	58,72%
4	SVM	62,11%	112,96€	89,84%	2,38%	60,55%

7.4 Análise de resultados

A abordagem tomada de testar diferentes combinações de variáveis obteve bons resultados. Em todos os casos os melhores modelos obtidos atingiram maiores taxas de acerto do que o modelo inicial, cujo valor era de 61,32%. Os algoritmos que permitiram atingir os melhores modelos nos diferentes casos foram o SVM, RF, Xgboost e RNA. O melhor modelo foi obtido testando combinações de 8 variáveis com as 6 variáveis identificadas como mais importantes, tendo portanto um total de 14 variáveis. O melhor modelo utilizou o algoritmo RF tal como foi descrito na secção 7.3.8. Este modelo obteve uma taxa de acerto de 65,26% e uma margem de lucro de 26,74%. Comparando com o modelo inicial a taxa de acerto subiu em quase 4%. A percentagem de apostas corretas em vitórias da equipa da casa diminuiu 7%, mas aumentou 26% nas apostas em empates e perto de 7% nas vitórias de equipas visitantes. A margem de lucro subiu 14%. Esta subida justifica-se pela subida do número de empates corretamente previstos. As apostas em empates costumam ter *odds* mais elevadas do que as apostas em vitórias de uma das equipas, por isso o lucro obtido foi maior. Para além de obter melhores resultados este modelo de previsão é melhor balanceado na previsão das diferentes classes do que o inicial. Este foi portanto o modelo escolhido para utilizar no sistema de apoio à decisão. A Tabela 17 apresenta em detalhe as medidas utilizadas para avaliar o modelo de previsão. Estas medidas foram a taxa de acerto, precisão, *recall*, margem de lucro e macro média da precisão e *recall*. A macro média é calculada fazendo a média da precisão e *recall* para cada classe, neste caso a vitória da equipa da casa, empate e vitória da equipa visitante. Analisando os resultados, consegue-se verificar que a macro média da precisão e *recall* estão perto do valor da taxa de acerto, por isso o modelo de previsão é equilibrado.

Tabela 17 – Descrição do melhor modelo de previsão.

Algoritmo	Random Forest
Variáveis utilizadas	*Odds* - B365H, B365D, B365A. Equipa da casa - HOVA, HDEF, HWINLAST5, HGOLSOFR. Equipa visitante - AVGH, AVGSTH, AVGSTA, AOVA, AMID, ADEF, AGOLSOFR.
Taxa de acerto	65,26%
Macro média Precisão	61,40%
Macro média *Recall*	58,73%
Margem de lucro	26,74%

Para além de calcular o desempenho do modelo na sua globalidade, foi também calculado o desempenho individual para cada classe, tal como é mostrado na Tabela 18. Foram calculadas as medidas precisão e *recall*. A precisão é a percentagem de previsões corretas dentro das previsões de cada classe. No caso dos empates, houve 50 previsões de empates, mas só 25 eram realmente empates por isso a precisão é de 50%. O *recall* corresponde à percentagem de apostas corretas para cada classe. No caso dos empates, havia no conjunto de teste 84 empates, tendo-se previsto corretamente 25 empates, por isso o *recall* é de 29,76%. Apesar de este valor não ser elevado é bastante aceitável visto ser difícil prever corretamente se um jogo de futebol vai acabar empatado.

Tabela 18 – Desempenho do melhor modelo por cada classe.

1- Vitória da equipa da casa	Precisão	68,47%
	Recall	81,29%
2- Empate	Precisão	50,00%
	Recall	29,76%
3- Vitória da equipa visitante	Precisão	65,74%
	Recall	65,14%

Para além de ter sido feita a análise global do modelo de previsão em toda a época foi também feita a sua avaliação em cada jornada. Cada jogo do conjunto de teste da época 2016/2017 tinha a identificação da jornada em que ocorreu, o que permitiu fazer esta análise. Esta análise era crucial para verificar se o modelo de previsão tinha um desempenho constante. Um modelo deste tipo não deve atingir taxas de acerto muito baixas, caso contrário pode haver grandes prejuízos numa dada jornada, sendo o objetivo do sistema proposto precisamente o contrário. A Tabela 19 mostra o desempenho do modelo de previsão em cada jornada. O modelo atingiu taxas de acerto de 60% em 10 jornadas, de 70% em 10

jornadas, de 80% em 7 jornadas, de 90% em 3 jornadas, de 50% em 5 jornadas, de 40% em apenas 1 jornada e de 30% em 2 jornadas. O modelo teve portanto um bom desempenho ao longo de toda a época, apenas tendo taxas de acerto inferiores a 50% em 3 jornadas. Nas 38 jornadas haveria lucro em 28 jornadas e prejuízo em 10 jornadas, obtendo-se lucro em 73% das jornadas.

Tabela 19 – Resultados das previsões em todas as jornadas da época 2016/17.

Primeira metade da época				**Segunda metade da época**		
Jornada	**Accuracy**	**Margem de Lucro**		**Jornada**	**Accuracy**	**Margem de Lucro**
1	70%	93%		20	50%	-1,3%
2	90%	125,9%		21	70%	69,7%
3	60%	6,3%		22	60%	16%
4	60%	21,8%		23	30%	-22%
5	70%	17,1%		24	70%	20,8%
6	50%	-4,4%		25	60%	3,8%
7	60%	38%		26	80%	53,6%
8	60%	14,4%		27	70%	31,3%
9	50%	-4,9%		28	60%	-0,2%
10	80%	48%		29	70%	48,6%
11	50%	15,8%		30	70%	56%
12	50%	-14,2%		31	80%	39,9%
13	70%	34,7%		32	60%	-7,1%
14	60%	39,1%		33	80%	47%
15	70%	48,2%		34	30%	-45,2%
16	80%	48,3%		35	40%	-38,8%
17	80%	46,9%		36	90%	57,1%
18	80%	50,8%		37	90%	56,5%
19	70%	20,7%		38	60%	-15%

Os resultados obtidos comprovam que o modelo de previsão desenvolvido tem a capacidade de ser usado num sistema de apoio à decisão de apostas de futebol. Há também que realçar que o modelo obteve melhor taxa de acerto e margem de lucro do que os modelos dos casos de estudo já existentes. A margem de lucro de 26,74% é maior do que os 26% alcançados em (Zan, 2017). O modelo acertou em 65,26% dos jogos, mais do que os 65% obtidos em

(Zuccolotto *et al.*, 2014). Salienta-se que esta diferença não é residual uma vez que o número de jogos testados nesta tese é maior do que nesses casos de estudo.

Apesar de em 2 casos de estudo, nomeadamente (Shin e Gasparyan, 2016) e (Igiri e Nwachukwu, 2014) a taxa de acerto ser superior a 70%, esses 2 casos são de dimensão reduzida e pouco detalhados, tal como explicado no capítulo 3. Não é portanto possível fazer uma comparação justa entre estes 2 casos de estudo e o presente trabalho. Assim, excluindo esses 2 casos de estudo, o modelo de previsão desenvolvido foi o que obteve melhor taxa de acerto.

8 Sistema de Apoio à Decisão

A criação de um modelo de previsão consistente com uma elevada taxa de acerto permitiu criar um sistema de apoio à decisão para indicar as previsões para os jogos da liga inglesa. Um dos objetivos desta tese de mestrado era desenvolver uma solução que pudesse ser usada na prática, para além de tentar desenvolver um bom modelo de previsão de jogos. Este é também um aspeto diferenciador relativo a trabalhos anteriores. Enquanto que nos anteriores casos de estudo apenas se desenvolveram modelos de previsão de jogos, neste trabalho foi desenvolvido um sistema de apoio à decisão que utiliza um modelo de previsão para ajudar os utilizadores a fazer apostas com base nessas previsões.

8.1 Arquitetura

O Sistema de Apoio à Decisão desenvolvido incorpora 3 componentes, tal como foi descrito no capítulo 4. A Figura 19 mostra um diagrama que ajuda a compreender melhor o funcionamento do SAD e a interligação dos componentes. O utilizador tem acesso à aplicação, que foi desenvolvida em Java, onde lhe é apresentada a previsão para os jogos da liga inglesa. A aplicação do cliente serve apenas como interface e faz a ligação ao servidor para receber as previsões dos jogos. A interface foi desenvolvida com JavaFX, uma plataforma que permite desenvolver de forma fácil interfaces em linguagem Java [65]. No servidor estão os dois outros componentes do sistema, um desses componentes é o sistema de previsão *Data Mining* que faz as previsões usando o modelo de previsão *Random Forest* descrito no capítulo 7. O outro componente é a aplicação web que recolhe a informação já guardada relativa aos jogos a prever e depois envia essa informação ao utilizador.

Em cada semana de jogos o sistema de previsão é utilizado para prever os jogos dessa semana. O sistema obtém os dados dos jogos a prever, faz as previsões e guarda-as. Idealmente as previsões seriam gravadas numa base de dados, no entanto, não foi criada uma base de dados no âmbito desta tese de mestrado e as previsões são gravadas em ficheiros "csv". Quando um utilizador abre o SAD e faz o pedido das previsões a aplicação web apenas tem de ler a informação com as previsões dos jogos e enviá-la à aplicação do cliente. Deste modo o utilizador obtém as informações de forma rápida. Caso o sistema de previsão fosse executado sempre que um cliente pede as previsões dos jogos o envio seria demorado, podendo tornar-se muito demorado se houvesse vários utilizadores a usar o SAD ao mesmo tempo.

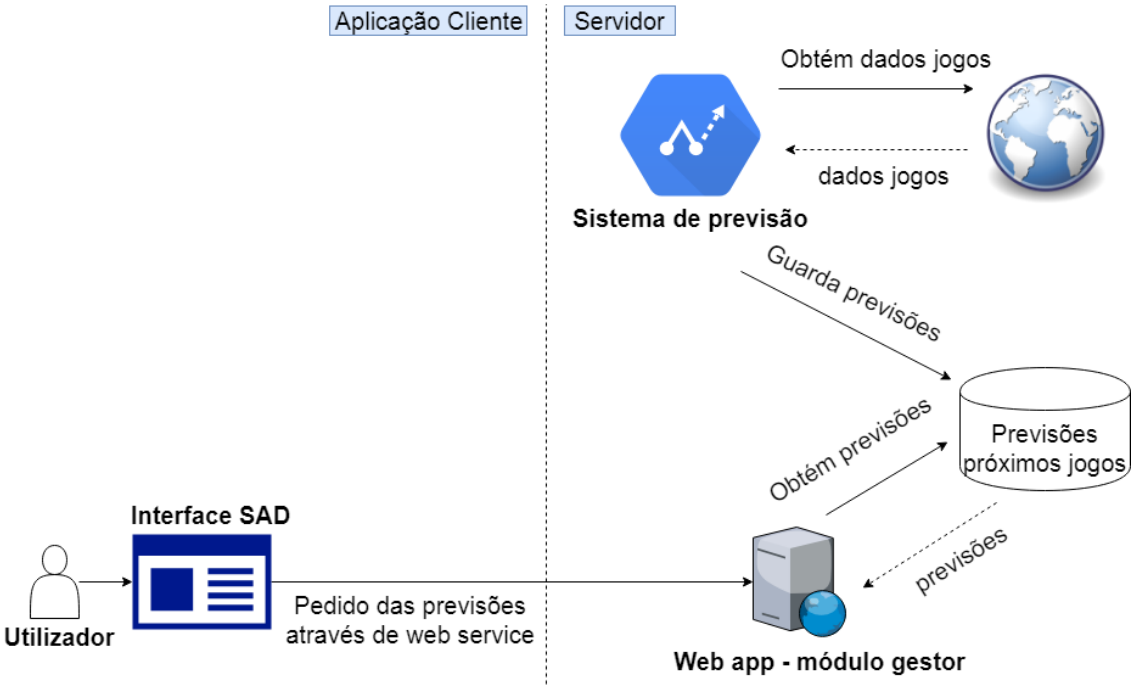

Figura 19 – Diagrama de funcionamento do SAD.

A Figura 20 contém o diagrama de classes UML da aplicação do cliente. A classe "Jogos" contém a informação detalhada de cada jogo, as equipas, previsão de resultado e análise de risco. A classe "LerPrevisoes" é usada para ler e processar a informação recebida do servidor. A classe "FXML Document Controller" é a classe que utiliza a "LerPrevisoes" para receber a informação dos jogos e coloca essa informação na interface da aplicação.

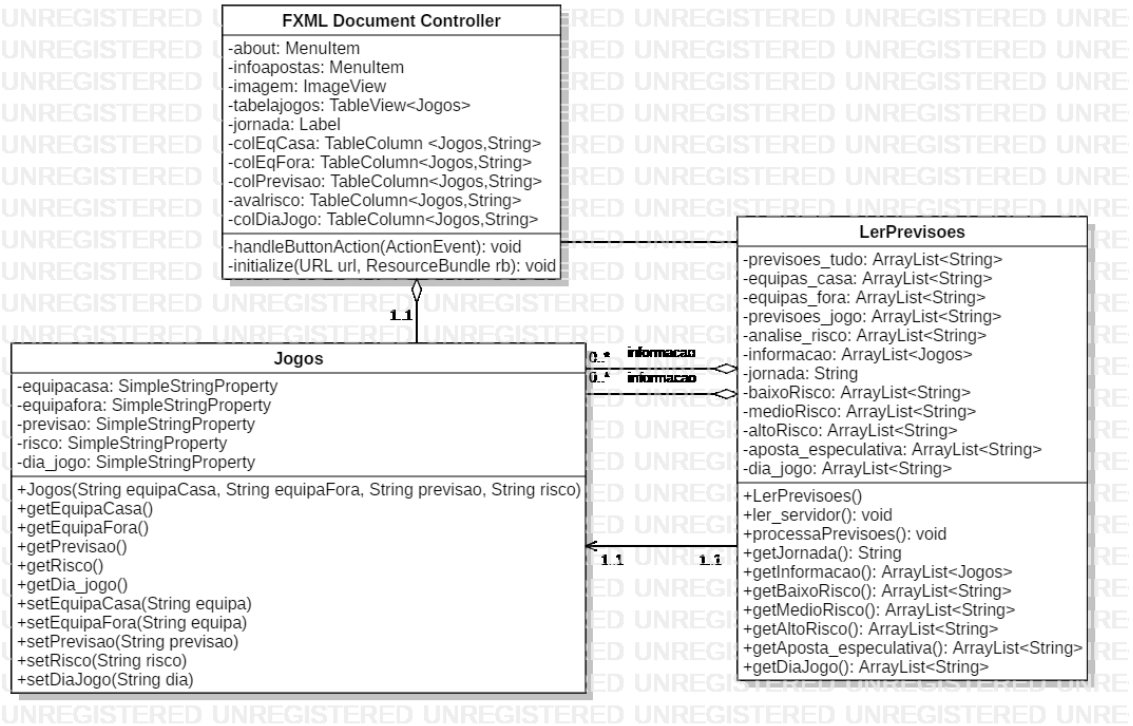

Figura 20 – Diagrama de classes da aplicação cliente.

8.2 Comunicação entre componentes

Para receber as previsões dos jogos a aplicação cliente faz um pedido GET [66] ao servidor através de um RESTful *Web Service* [67]. O servidor foi criado numa máquina local, o pedido é feito através do url:

- http://localhost:8080/WebApplication1/webresources/generic/Usuario/get

O pedido é feito usando o nome de utilizador e palavra-passe do cliente. O nome de utilizador e palavra-passe são encriptados para garantir segurança e confidencialidade dos dados. Desta forma consegue-se garantir que só pessoas autorizadas têm acesso às previsões dos jogos. Para além disso o código fonte que faz a previsão dos jogos fica no servidor e portanto o utilizador não tem acesso a esse código. Como há uma separação entre a interface e o servidor é também possível, caso seja necessário, alterar a interface e utilizar outras tecnologias como uma aplicação web.

8.3 Apoio à Decisão

O sistema desenvolvido informa os utilizadores do risco de fazer uma dada aposta, para além de lhes indicar qual a previsão do resultado para cada jogo. Isto permite fornecer aos utilizadores um apoio à decisão que lhes possibilita avaliar de forma mais detalhada se devem ou não fazer uma determinada aposta.
A avaliação de risco é feita com base na probabilidade do resultado previsto pelo modelo de previsão *Data Mining* ocorrer. Este processo ocorre de acordo com o que é explicado no diagrama da Figura 21. Para um dado jogo, como o Arsenal vs Liverpool, o modelo de previsão *Data Mining* faz a previsão do resultado, podendo a previsão ser que o Arsenal vai ganhar o jogo. A partir daí vai ser verificada qual a probabilidade de o Arsenal ganhar esse jogo. Caso a probabilidade de o Arsenal ganhar seja superior a 70% a aposta é de baixo risco, visto que a probabilidade de acertar a aposta é mais do dobro da de falhar. Caso a probabilidade esteja entre 52% e 70% o risco é médio, visto a probabilidade de acertar ser superior à de falhar. Se a probabilidade de o Arsenal ganhar estiver entre 34% e 52% considera-se que a aposta é de alto risco, visto as hipóteses de acertar a aposta serem idênticas às de falhar. Se a probabilidade do Arsenal ganhar for inferior a 34% a aposta é considerada como especulativa, neste caso não se aconselha apostar neste jogo. A probabilidade de acertar no resultado de um jogo de futebol ao acaso é de 33,3%, por isso se a probabilidade do resultado previsto for menor que 34% o risco de falhar é demasiado elevado e por isso não se deve apostar nesses jogos. A probabilidade serve apenas para indicar o risco da aposta e a indicação da aposta a fazer é sempre dada de acordo com o que é previsto pelo modelo *Data Mining*.

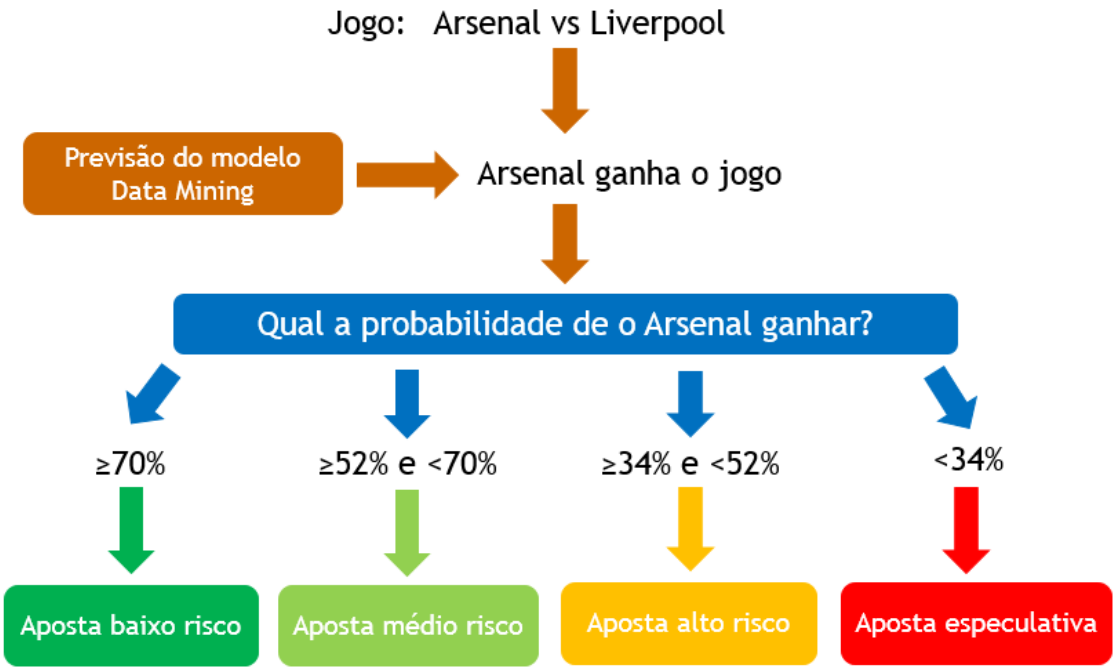

Figura 21 – Diagrama do processo de apoio à decisão

Ao determinar a probabilidade do resultado previsto ocorrer está-se não só a dar mais informação ao utilizador, mas também a tornar o SAD mais consistente visto existir uma dupla análise para determinar qual será o resultado final dos jogos. O sistema permite evitar prejuízos quando o resultado final do jogo é muito difícil de prever. Por exemplo, se o modelo *Data Mining* fizer uma previsão incorreta e a aposta indicada for uma aposta especulativa não há prejuízo, visto que não é suposto apostar em apostas especulativas, já que a probabilidade de acertar é muito reduzida. Assim, ao fazer a avaliação de risco das apostas, está-se a colmatar eventuais falhas do modelo de previsão.

A interface da aplicação é mostrada na Figura 22. Na interface é indicada a jornada para a qual é feita a previsão e as previsões feitas, salientando aquelas em que o risco é baixo ou médio. É também indicada a avaliação de risco para cada jogo e o dia em que se realiza cada jogo. Um apostador deve apostar num jogo até ao início desse jogo. A Figura 22 mostra as previsões para a 4ª jornada da liga inglesa da época 2018/2019, disputada nos dias 1 e 2 de Setembro de 2018. Quando um utilizador abre a aplicação são-lhe apresentadas as previsões para os jogos da jornada seguinte ao dia em que está a ser usada a aplicação. Por exemplo, a 3ª jornada da liga inglesa terminou no dia 27 de Agosto de 2018. Se um utilizador usar o SAD entre os dias 28 de Agosto e 2 de Setembro são-lhe apresentadas as previsões mostradas na Figura 22, referentes aos jogos da 4ª jornada. A partir do dia 3 de Setembro, o dia seguinte a terminar a 4ª jornada, já serão apresentadas as previsões para a 5ª jornada, jogada entre os dias 15 e 17 de Setembro de 2018. Uma vez que só existem *odds* disponíveis para os jogos da jornada seguinte, e o sistema de previsão utiliza as *odds* para fazer as previsões, só pode ser feita a previsão para a jornada seguinte. Através do menu "Ajuda" o utilizador tem acesso à explicação sobre como funciona a análise de risco das apostas, tal como é mostrado na Figura 23.

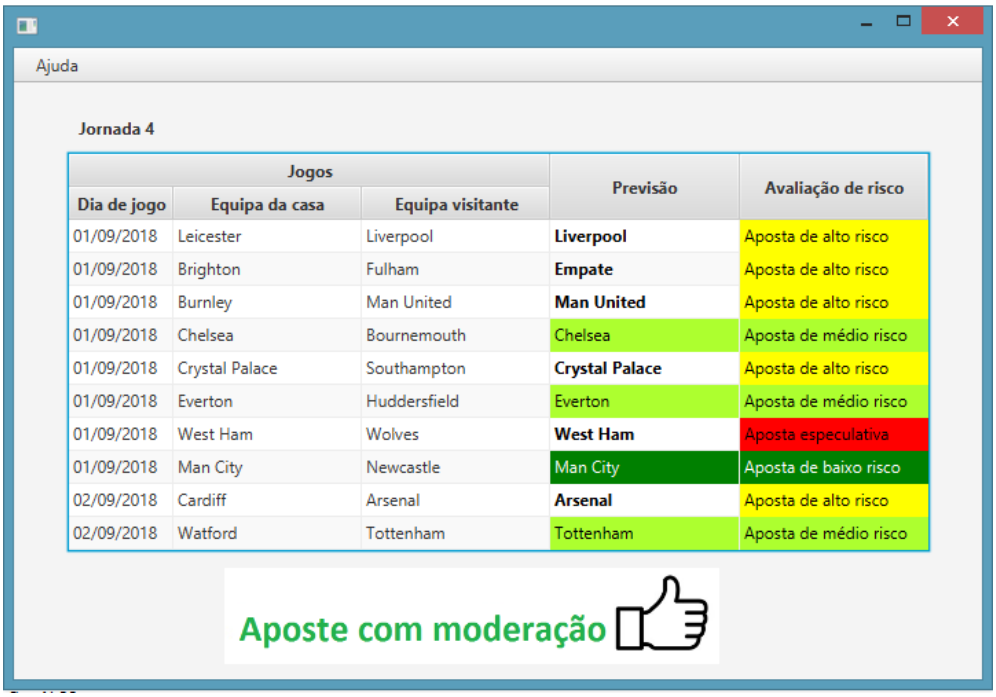

Figura 22 – Interface do SAD – previsão de resultados

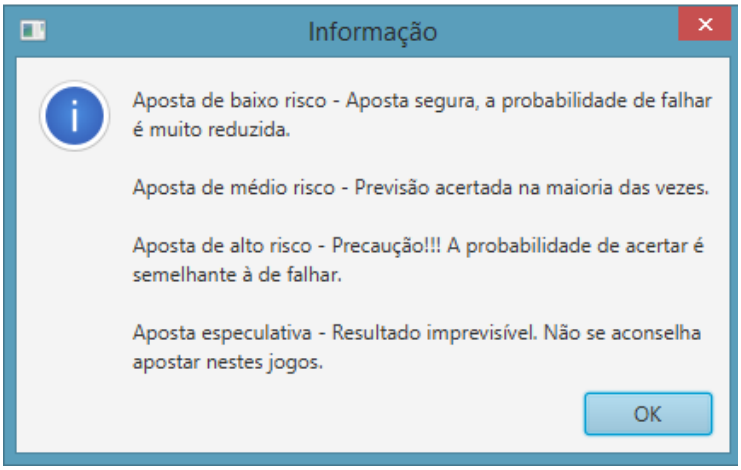

Figura 23 – Interface do SAD – informação sobre análise de risco

8.4 Testes

Foram efetuados testes ao SAD para garantir que este funcionava de uma forma correta. Na aplicação cliente, ao inicializar a aplicação, é verificado se é possível conectar ao servidor. Caso não seja possível é apresentada uma mensagem de erro para informar o utilizador. O código utilizado para fazer este teste foi implementado no método "testConnection()", mostrado na Figura 24. Também é verificada a informação obtida do servidor. Caso uma das previsões recebidas do servidor não aponte uma informação correta a previsão para esse jogo não é mostrada. Por exemplo, para o jogo Arsenal vs Liverpool, se a previsão for "Chelsea" é porque houve um erro e por isso não se apresenta essa previsão.

```
111     public boolean testConnection() {
112         HttpExemplo http = new HttpExemplo();
113 
114         String url = "http://localhost:8080/WebApplication1/webresources";
115 
116         boolean server_on = true;
117 
118         try {
119             int teste = http.testServer(url);
120         } catch (Exception ex) {
121             Alert alert = new Alert(AlertType.ERROR);
122             alert.setTitle("Erro");
123             alert.setHeaderText(null);
124             alert.setContentText("Não foi possível contactar o servidor. Tente mais tarde.");
125 
126             alert.showAndWait();
127             server_on = false;
128         }
129         return server_on;
130     }
```

Figura 24 – Código usado para testar a conexão ao servidor.

9 Conclusão

Esta tese tratou de um problema numa área de negócio que tem atualmente uma crescente procura, as apostas desportivas. O objetivo passou por fazer a correta previsão dos resultados de jogos de futebol, utilizando para isso *Data Mining*. Este trabalho englobou por isso duas vertentes relevantes quer na parte comercial, quer na parte tecnológica.

Uma das fases cruciais deste trabalho foi a recolha de dados. No trabalho foram utilizados dados de duas fontes diferentes, uma para a obtenção de dados estatísticos relativos aos jogos e a outra com dados relativos às equipas. Antes da construção dos modelos de previsão foi feita uma fase de análise e processamento dos dados, esta etapa é bastante importante num projeto de *Data Mining* e permitiu fazer asserções sobre as variáveis a utilizar para prever os resultados. O estudo efetuado comparou vários algoritmos de modo a criar o melhor modelo de previsão possível e o algoritmo que comprovou ser o melhor foi o *Random Forest*. O modelo de previsão foi testado em todos os jogos da época 2016/2017 da Liga Inglesa, o total de jogos de teste foi de 380. O modelo foi treinado com dados de 4 épocas, da época 2012/2013 à 2015/2016, sendo portanto os dados de treino e teste distintos. O número de jogos testados foi bastante maior do que o número de jogos testado em trabalhos anteriores. Para além disso, como o modelo de previsão criado foi testado numa época inteira foi possível fazer uma avaliação detalhada do comportamento do modelo de previsão ao longo das várias jornadas da época.

A percentagem de jogos corretamente acertados pelo modelo foi de 65,26%, sendo superior à percentagem atingida noutros trabalhos. A margem de lucro obtida também foi superior à dos casos de estudo referenciados. Neste trabalho foi ainda determinado o lucro que se obteria fazendo apostas com base nas previsões feitas, algo que não é feito na maioria dos casos de estudo nesta área. A avaliação feita foi bastante pormenorizada, tendo sido detalhada a taxa de acerto de jogos e margem de lucro que se obteria em cada semana de apostas.

Além de se desenvolver um modelo de previsão foi ainda criado um SAD para ajudar os apostadores a acertar mais vezes nos resultados dos jogos da Liga Inglesa. Isto permitiu que o projeto desenvolvido tivesse uma componente mais prática, fazendo com que o modelo inferido a partir dos dados possa ser facilmente usado por utilizadores não peritos. Para que a ajuda dada aos utilizadores fosse mais completa o SAD faz a avaliação de risco das apostas a fazer com base na probabilidade de ocorrência dos resultados previstos pelo modelo *Data Mining*. Deste modo o utilizador sabe se a aposta tem maior ou menor risco tendo assim um maior apoio na obtenção de lucros nas apostas desportivas.

O projeto desenvolvido atingiu os objetivos propostos e foi possível criar uma solução capaz de responder ao problema da falta de ajuda e apoio à decisão quando se fazem apostas desportivas. Foi demonstrada uma estratégia para prever resultados de jogos de futebol e

comprovou-se que com a utilização de estatísticas de jogos anteriores em conjunto com técnicas de *Data Mining* é possível fazer previsões corretas de futuros jogos.

9.1 Trabalho futuro

Uma das maiores dificuldades em trabalhos na área de *Data Mining* é a obtenção de dados. O trabalho desenvolvido poderia ser melhorado caso houvesse mais dados disponíveis. Em outros casos de estudo foram utilizados dados como o confronto direto entre as equipas para prever os resultados dos jogos. Estas estatísticas poderiam ajudar a aumentar a taxa de acerto dos modelos de previsão. Outras informações seriam também relevantes como saber se há jogadores lesionados nas equipas. Quando um bom jogador se lesiona o rendimento da sua equipa em geral diminui. No futuro, caso se consiga obter estes dados, será possível melhorar o modelo de previsão e consequentemente o SAD.

O SAD desenvolvido, apesar de estar funcional, ainda não está pronto para ser usado como uma aplicação comercial. Para que isso fosse possível seria necessário fazer mais testes à aplicação, quer do lado do cliente quer do lado do servidor, para garantir que não ocorreriam quaisquer erros a executar a aplicação. Era também importante criar uma base de dados para guardar os dados dos jogos previstos da melhor forma possível de modo a serem facilmente usados em previsões de épocas futuras. A interface da aplicação, apesar de identificar claramente as apostas que devem ser feitas e cumprir os objetivos do SAD, poderia ser mais apelativa, algo indispensável numa aplicação comercial.

Referências

(Awoyemi *et al.*, 2017) Awoyemi, J. O., Adetunmbi, A. O. and Oluwadare, S. A., 2017 'Credit card fraud detection using machine learning techniques: A comparative analysis', *2017 International Conference on Computing Networking and Informatics (ICCNI)*. Lagos, Nigéria. doi: 10.1109/ICCNI.2017.8123782. Disponível em: https://ieeexplore.ieee.org/document/8123782/.

(Bittencourt, 2003) Bittencourt, H. R., 2003. Regressão logística politômica: revisão teórica e aplicações. *Acta Scientiae,* 5**,** 77-86. Disponível em: http://www.periodicos.ulbra.br/index.php/acta/article/view/146.

(Boldrin, 2017) Boldrin, B., 2017 *Predicting the result of english premier league soccer games with the use of Poisson models*. STETSON UNIVERSITY. Disponível em: http://www2.stetson.edu/~efriedma/research/boldrin.pdf.

(Bunker e Thabtah, 2017) Bunker, R. P. and Thabtah, F., 2017. A machine learning framework for sport result prediction. *Applied Computing and Informatics*. Disponível em: http://www.sciencedirect.com/science/article/pii/S2210832717301485.

(Cañizares *et al.*, 2017) Cañizares, P. C., Merayo, M. G., Núñez, M. and Suárez-Paniagua, V., 2017 *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*. 2017/09//. pp. 572-576.

(Cao, 2012) Cao, C., 2012. Sports Data Mining Technology Used in Basketball Outcome Prediction.

(Chapman *et al.*, 2000) Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, 2000 'Step-by-step data mining guide'. 2000. Disponível em: https://www.the-modeling-agency.com/crisp-dm.pdf.

(Chen e Guestrin, 2016) Chen, T. and Guestrin, C., 2016 'XGBoost: A Scalable Tree Boosting System', *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA. doi: 10.1145/2939672.2939785. Disponível em: http://doi.acm.org/10.1145/2939672.2939785.

(Dasgupta, 2018) Dasgupta, N., 2018 *Practical Big Data Analytics*. Packt Publishing Ltd.

(Duarte, 2015) Duarte, L., 2015 *1X2 – Previsão de Resultados de Jogos de Futebol*. FEUP. Disponível em: https://sigarra.up.pt/feup/pt/pub_geral.pub_view?pi_pub_base_id=35444&pi_pub_r1_id

(Dusza *et al.*, 2016) Dusza, K., Korda, D., Kozłowski, K., Szwej, B., Kozielski, M., Michalak, M., Sikora, M. and Wróbel, L., 2016 'Application of RapidMiner and R environments to dangerous seismic events prediction', *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. Gdansk, Polónia. Disponível em: https://ieeexplore.ieee.org/document/7733247/.

(Fayyad *et al.*, 1996) Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. From data mining to knowledge discovery in databases. *Advances in Knowledge Discovery and Data Mining,* 17**,** 1-36.

(Frank *et al.*, 2016) Frank, E., Hall, M. A. and Witten, I. H., 2016 'The WEKA Workbench'. 2016. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

(Gomes *et al.*, 2015) Gomes, J., Portela, P. and Santos, M. F., 2015. Decision Support System for predicting Football Game result. 348-353.

(Hackathorn e Keen, 1981) Hackathorn, R. D. and Keen, P. G. W., 1981. Organizational Strategies for Personal Computing in Decision Support Systems. *MIS Q.,* 5**,** 21-27. Disponível em: http://dx.doi.org/10.2307/249288.

(Haghighat *et al.*, 2013) Haghighat, M., Rastegari, H. and Nourafza, N., 2013. A Review of Data Mining Techniques for Result Prediction in Sports. 2.

(Holsapple, 2008) Holsapple, C., 2008 'DSS Architecture and Types', in *Handbook on Decision Support Systems 1 - Basic Themes | Frada Burstein | Springer*.

(Hwang *et al.*, 2018) Hwang, C., Chen, M. S., Shih, C. M., Chen, H. Y. and Liu, W. K., 2018 'Apply Scikit-Learn in Python to Analyze Driver Behavior Based on OBD Data', *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. Cracóvia, Polónia. doi: 10.1109/WAINA.2018.00159. Disponível em: https://ieeexplore.ieee.org/document/8418144/.

(Igiri e Nwachukwu, 2014) Igiri, C. P. and Nwachukwu, E. O., 2014. An Improved Prediction System for Football a Match Result. *IOSR Journal of Engineerin,* 04**,** 12-20. Disponível em: https://www.researchgate.net/publication/273164409_An_Improved_Prediction_System_for_Football_a_Match_Result.

(Karkuzhali e Manimegalai, 2017) Karkuzhali, S. and Manimegalai, D., 2017. Computational intelligence-based decision support system for glaucoma detection. *Biomedical Research (0970-938X),* 28**,** 4737. Disponível em: http://search.ebscohost.com/login.aspx?direct=true&site=eds-live&db=a9h&AN=124198397.

(Koen *et al.*, 2014) Koen, P. A., Bertels, H. M. J. and Kleinschmidt, E. J., 2014. Managing the Front End of Innovation—Part II: Results from a Three-Year Study. *Research-Technology Management,* 57**,** 25-35. Disponível em: http://www.tandfonline.com/doi/abs/10.5437/08956308X5703199.

(Kumar, 2013) Kumar, G., 2013 *Machine Learning for Soccer Analytics*. Master thesis. KU Leuven. Disponível em: https://www.researchgate.net/publication/257048220_Machine_Learning_for_Soccer_Analytics.

(Kursa e Rudnicki, 2010) Kursa, M. B. and Rudnicki, W. R., 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software*. Disponível em: https://www.jstatsoft.org/article/view/v036i11.

(Larose, 2014) Larose, D. T., 2014 *Discovering Knowledge in Data: An Introduction to Data Mining*. 2 edn.

(Maimon e Rokach, 2010) Maimon, O. and Rokach, L., 2010 *Data Mining and Knowledge Discovery Handbook*. 2nd edn. Springer Publishing Company, Incorporated.

(Mohri *et al.*, 2012) Mohri, M., Rostamizadeh, A. and Talwalkar, A., 2012 *Foundations of Machine Learning*. The MIT Press.

(Nicola *et al.*, 2012) Nicola, S., Ferreira, E. P. and Ferreira, J. J. P., 2012. A novel framework for modeling value for the customer, an essay on negotiation. *International Journal of Information Technology & Decision Making,* 11**,** 661-703. Disponível em: http://www.worldscientific.com/doi/abs/10.1142/S0219622012500162.

(Osterwalder *et al.*, 2010) Osterwalder, A., Pigneur, Y. and Clark, T., 2010 *Business model generation: a handbook for visionaries, game changers, and challengers*.

(Pettersson e Nyquist, 2017) Pettersson, D. and Nyquist, R., 2017 *Football Match Prediction using Deep Learning*. Master thesis. Chalmers University of Technology. Disponível em: http://publications.lib.chalmers.se/records/fulltext/250411/250411.pdf.

(Prasetio e Harlili, 2016) Prasetio, D. and Harlili, D., 2016 *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. 2016/08. pp. 1-5.

(Refaeilzadeh *et al.*, 2009) Refaeilzadeh, P., Tang, L. and Liu, H., 2009 'Cross-Validation', in Liu, L. and ÖZsu, M. T. (eds.) *Encyclopedia of Database Systems*. Springer US, pp. 532-538.

(Rodrigues, 2016) Rodrigues, H., 2016 *Ferramenta para Text Mining em Textos completos*. Faculdade de Engenharia. Disponível em: http://hdl.handle.net/10216/85394.

(Ruiz *et al.*, 2017) Ruiz, H., Power, P., Wei, X. and Lucey, P., 2017, 2017. New York, NY, USA: ACM, pp. 1991-2000. Disponível em: http://doi.acm.org/10.1145/3097983.3098121 (Acedido: 2017/12/27).

(Sauter, 2011) Sauter, V. L., 2011 *Decision Support Systems for Business Intelligence*. John Wiley & Sons, Inc.

(Schumaker *et al.*, 2010) Schumaker, R. P., Solieman, O. K. and Chen, H., 2010. Sports Data Mining. *Information Systems Journal,* 26**,** 15-22. Disponível em: http://www.springerlink.com/index/10.1007/978-1-4419-6730-5.

(Shalabi *et al.*, 2006) Shalabi, L. A., Shaaban, Z. and Kasasbeh, B., 2006. Data Mining: A Preprocessing Engine. *Journal of Computer Science,* 2**,** 735-739.

(Shin e Gasparyan, 2016) Shin, J. and Gasparyan, R., 2016 'A novel way to Soccer Match Prediction'. 2016. Disponível em: http://cs229.stanford.edu/proj2014/Jongho%20Shin,%20Robert%20Gasparyan,%20A%20novel%20way%20to%20Soccer%20Match%20Prediction.pdf.

(Tan *et al.*, 2005) Tan, P.-N., Steinbach, M. and Kumar, V., 2005 *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

(Tan *et al.*, 2013) Tan, P.-N., Steinbach, M. and Kumar, V., 2013 *Introduction to Data Mining: Pearson New International Edition*. 1 edn.

(Turban, 2011) Turban, E., 2011. Decision Support and Business Intelligence Systems, 9/E. *Prentice Hall***,** 696.

(Vercellis, 2011) Vercellis, C., 2011 *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons.

(Wienclaw, 2013) Wienclaw, R. A., 2013. Decision Support Systems. *Research Starters: Business (Online Edition)*. Disponível em: http://connection.ebscohost.com/c/essays/27577746/decision-support-systems.

(Wirth e Hipp, 2000) Wirth, R. and Hipp, J., 2000 'CRISP-DM: Towards a standard process model for data mining', *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. 2000. pp. 29-39. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.198.5133.

(Woodall, 2003) Woodall, T., 2003. Conceptualising 'value for the customer': an attributional, structural and dispositional analysis. *Academy of Marketing Science Review,* 2003. Disponível em: http://www.amsreview.org/articles/woodall12-2003.pdf.

(Zan, 2017) Zan, T. v. d., 2017 *Predicting the outcome of soccer matches in order to make money with betting*. Universidade Roterdão. Disponível em: https://thesis.eur.nl/pub/37404

(Zeithaml, 1988) Zeithaml, V. A., 1988. Consumer Perceptions of Price, Quality, and Value: A Means-End Model and Synthesis of Evidence. *Journal of Marketing,* 52**,** 2-22. Disponível em: http://www.jstor.org/stable/1251446.

(Zuccolotto *et al.*, 2014) Zuccolotto, P., Carpita, M., Sandri, M. and Simonetto, A., 2014 'Football Mining with R', in *Data Mining Applications with R*.

Referências Web

[1] Lusa, «Apostas - Placard já conquistou 920 mil apostadores e distribuiu 200 milhões em prémios», *Diário de Notícias*, 2016. [Online]. Disponível em: https://www.dn.pt/sociedade/interior/placard-ja-conquistou-920-mil-apostadores-e-distribuiu-200-milhoes-em-premios-5378654.html. [Acedido: 31-Dez-2017]

[2] SoccerVista, «SoccerVista - football results, predictions and betting picks», *SoccerVista*, 2018. [Online]. Disponível em: http://www.soccervista.com. [Acedido: 10-Fev-2018]

[3] Vitibet, «Apostas desportivas,resultado ao vivo,apostas», *Vitibet*, 2018. [Online]. Disponível em: http://www.vitibet.com/index.php?lang=pt. [Acedido: 10-Fev-2018]

[4] ZuluBet, «ZuluBet - Previsões de Futebol», *Zulubet*, 2018. [Online]. Disponível em: http://pt.zulubet.com/. [Acedido: 10-Fev-2018]

[5] Forebet, «Mathematical football predictions, Tips, Statistics, Previews.», *Forebet*, 2018. [Online]. Disponível em: https://www.forebet.com//. [Acedido: 10-Fev-2018]

[6] BettingTips, «Dicas de Apostas», *GooglePlay*, 2018. [Online]. Disponível em: https://play.google.com/store/apps/details?id=com.betsoft.bettingtipsapp. [Acedido: 11-Fev-2018]

[7] BetPredictor, «Bet Predictor na App Store», *Itunes*, 2018. [Online]. Disponível em: https://itunes.apple.com/pt/app/bet-predictor/id700908899?mt=8. [Acedido: 11-Fev-2018]

[8] Pinnacle, «Explicação dos tipos básicos de apostas», *Pinnacle*, 2017. [Online]. Disponível em: https://www.pinnacle.com/pt/betting-articles/Betting-Strategy/basic-bet-types-explained/AS7JD8W7JUM6DBKN. [Acedido: 23-Dez-2017]

[9] Faustino, «O que são Odds? Análise de diferentes tipos de Odds», *Asmelhoresapostasonline*, 2012. [Online]. Disponível em: http://www.asmelhoresapostasonline.com/o-que-sao-odds-analise-de-diferentes-tipos-de-odds/. [Acedido: 23-Dez-2017]

[10] Ramos, «Teoria das probabilidades», 2013. [Online]. Disponível em: http://pwp.net.ipl.pt/deetc.isel/pramos/MA/p_e/Resumo_-_Teoria_das_probabilidades.pdf

[11] bet.pt, «Apostas Desportivas | bet.pt», *Bet*, 2018. [Online]. Disponível em: https://www.bet.pt/apostas-desportivas/. [Acedido: 31-Jul-2018]

[12] Betclic, «Apostas Online e Casino | Betclic», *Betclic*, 2018. [Online]. Disponível em: https://www.betclic.pt/. [Acedido: 31-Jul-2018]

[13] Jogos Santa Casa, «Jogos Santa Casa - Placard», 2018. [Online]. Disponível em: https://www.jogossantacasa.pt/web/Placard. [Acedido: 13-Set-2018]

[14] P. Crisóstomo, «Santa Casa da Misericórdia. Apostas desportivas: um negócio de muitos milhões mas de valor incerto no país», *Público*, 2017. [Online]. Disponível em: https://www.publico.pt/2017/02/08/economia/noticia/apostas-desportivas-um-negocio-de-muitos-milhoes-mas-de-valor-incerto-no-pais-1761230. [Acedido: 31-Dez-2017]

[15] L. Villalobos, «Apostas online. Apostas de futebol valem 43% do jogo "online"», *Público*, 2017. [Online]. Disponível em: https://www.publico.pt/2017/11/05/desporto/noticia/apostas-de-futebol-valem-43-do-jogo-online-1791431. [Acedido: 31-Dez-2017]

[16] Martins, «Porque é que 400 mil apostam no Placard?», *ESEV*, 2016. [Online]. Disponível em: http://www.esev.ipv.pt/dacomunicacao/index.php/2016/07/05/porque-e-que-400-mil-apostam-no-placard/. [Acedido: 31-Dez-2017]

[17] EstorilSolCasinos, «EstorilSolCasinos.pt», *EstorilSolCasinos.pt*, 2018. [Online]. Disponível em: https://www.estorilsolcasinos.pt. [Acedido: 31-Jul-2018]

[18] CasinoPortugal, «CASINO PORTUGAL - Apostas Desportivas e Casino online», *Casino Portugal*, 2018. [Online]. Disponível em: https://www.casinoportugal.pt. [Acedido: 31-Jul-2018]

[19] NossaAposta, «Apostas Desportivas | Nossa Aposta |», *Nossa Aposta*, 2018. [Online]. Disponível em: https://www.nossaaposta.pt/. [Acedido: 31-Jul-2018]

[20] Placard.pt, «Placard.pt - aposta na emoção do desporto. Aqui, jogas em casa.», *Placard.pt*, 2018. [Online]. Disponível em: https://apostas.placard.pt. [Acedido: 31-Jul-2018]

[21] L. Villalobos, «Licenças para jogo online perto de duplicar», *Público*, 12-Jun-2018. [Online]. Disponível em: https://www.publico.pt/2018/06/12/economia/noticia/licencas-para-jogo-online-perto-de-duplicar-1834065. [Acedido: 31-Jul-2018]

[22] Observador, «Cada jornada movimenta 340 milhões de euros em apostas», *Observador*, 29-Jan-2018. [Online]. Disponível em: https://observador.pt/2018/01/29/cada-jornada-movimenta-340-milhoes-de-euros-em-apostas/. [Acedido: 31-Jul-2018]

[23] L. James, «bet365 success as profits hit £514 million», *Stokesentinel*, 07-Nov-2017. [Online]. Disponível em: http://www.stokesentinel.co.uk/news/bet365-success-profits-hit-514-735052. [Acedido: 31-Dez-2017]

[24] Research and Markets ltd, «Worldwide Gambling Market - Forecasts, 2016-2022», *Researchandmarkets*, 2016. [Online]. Disponível em: https://www.researchandmarkets.com/reports/3769265/worldwide-gambling-market-by-types-digital. [Acedido: 27-Dez-2017]

[25] A. Suspiro, «FMI. Crescimento de Portugal», *Observador*, 2017. [Online]. Disponível em: http://observador.pt/2017/10/10/fmi-crescimento-de-portugal-acelera-para-25-este-ano-mas-trava-em-2018/. [Acedido: 27-Dez-2017]

[26] SokkerPRO, «SokkerPRO ® – Software de prognóstico de futebol», *Sokkerpro*, 2018. [Online]. Disponível em: http://sokkerpro.com/pt/. [Acedido: 31-Dez-2017]

[27] Strategyzer, «Strategyzer | Resources», *Strategyzer*, 2018. [Online]. Disponível em: https://strategyzer.com/platform/resources. [Acedido: 01-Jan-2018]

[28] StatisticsSolutions, «Conduct and Interpret a Multinomial Logistic Regression», *Statistics Solutions*, 2018. [Online]. Disponível em: http://www.statisticssolutions.com/mlr/. [Acedido: 01-Ago-2018]

[29] Joglekar, Sachin, «Logistic Regression», *Sachin Joglekar's blog*. 16-Ago-2015 [Online]. Disponível em: https://codesachin.wordpress.com/2015/08/16/logistic-regression-for-dummies/. [Acedido: 04-Set-2018]

[30] Universidade de São Paulo, «Regressão Logística». 2016 [Online]. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf

[31] Statisticshowto, «Multinomial Logistic Regression: Definition and Examples», *Statistics How To*, 2017. [Online]. Disponível em: http://www.statisticshowto.com/multinomial-logistic-regression/. [Acedido: 13-Set-2018]

[32] M. Hahsler, «Classification - Basic Concepts, Decision Trees, and Model Evaluation», 2016. [Online]. Disponível em: http://michael.hahsler.net/SMU/EMIS7332/slides/chap4_basic_classification.pdf. [Acedido: 07-Jan-2018]

[33] Xgboost developers, «Introduction to Boosted Trees — xgboost 0.80 documentation», 2016. [Online]. Disponível em: https://xgboost.readthedocs.io/en/latest/tutorials/model.html. [Acedido: 04-Set-2018]

[34] T. Chen e C. Guestrin, *xgboost: Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, for Python, R, Java, Scala, C++ and more. Runs on single machine, Hadoop, Spark, Flink and DataFlow*. Distributed (Deep) Machine Learning Community, 2018 [Online]. Disponível em: https://github.com/dmlc/xgboost. [Acedido: 01-Ago-2018]

[35] J. Brownlee, «A Gentle Introduction to XGBoost for Applied Machine Learning», Ago-2016. [Online]. Disponível em: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/. [Acedido: 01-Ago-2018]

[36] R. Shaw, «XGBoost: A Concise Technical Overview». 2017 [Online]. Disponível em: https://www.kdnuggets.com/2017/10/xgboost-concise-technical-overview.html, https://www.kdnuggets.com/2017/10/xgboost-concise-technical-overview.html. [Acedido: 04-Set-2018]

[37] A. Samudrala, «Unveiling Mathematics Behind XGBoost». 2018 [Online]. Disponível em: https://www.kdnuggets.com/2018/08/unveiling-mathematics-behind-xgboost.html, https://www.kdnuggets.com/2018/08/unveiling-mathematics-behind-xgboost.html. [Acedido: 04-Set-2018]

[38] I. Reinstein, «XGBoost, a Top Machine Learning Method on Kaggle, Explained», Out-2017. [Online]. Disponível em: https://www.kdnuggets.com/2017/10/xgboost-top-machine-learning-method-kaggle-explained.html. [Acedido: 01-Ago-2018]

[39] N. Donges, «The Random Forest Algorithm», *Towards Data Science*, 22-Fev-2018. [Online]. Disponível em: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd. [Acedido: 01-Ago-2018]

[40] S. Ray, «6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python)», *Analytics Vidhya*. 11-Set-2017 [Online]. Disponível em: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. [Acedido: 01-Ago-2018]

[41] Statsoft, «Support Vector Machines (SVM)», *Statsoft*, 2017. [Online]. Disponível em: http://www.statsoft.com/Textbook/Support-Vector-Machines. [Acedido: 07-Jan-2018]

[42] Iddo, «Support Vector Machines explained well», *Bytesizebio*, 2014. [Online]. Disponível em: http://bytesizebio.net/2014/02/05/support-vector-machines-explained-well/. [Acedido: 07-Jan-2018]

[43] S. Sayad, «KNN Classification», 2010. [Online]. Disponível em: http://www.saedsayad.com/k_nearest_neighbors.htm. [Acedido: 01-Ago-2018]

[44] C. Cao, «Sports Data Mining Technology Used in Basketball Outcome Prediction», 2012.

[45] L. Timson, «Germany's World Cup a win for tech», *SMH*, 15-Jul-2014. [Online]. Disponível em: http://www.smh.com.au/it-pro/business-it/germanys-world-cup-a-win-for-tech-20140715-ztbx7.html. [Acedido: 02-Out-2017]

[46] P. Rosário e R. Santos, «Simulámos a Liga dos Campeões 10 mil vezes», *Observador*, 27-Set-2016. [Online]. Disponível em: http://observador.pt/especiais/simulamos-a-liga-dos-campeoes-10-mil-vezes/. [Acedido: 01-Out-2017]

[47] Richardson, «Beautiful Soup Documentation», *Crummy*, 2015. [Online]. Disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Acedido: 27-Dez-2017]

[48] Exsys Inc, «Exsys Corvid Expert System Development Tool», *Exsys*, 2011. [Online]. Disponível em: http://www.exsys.com/exsyscorvid.html. [Acedido: 27-Dez-2017]

[49] K. Gupta, «The Best Programming Languages for Data Mining», *FreelancingGig*, 07-Ago-2017. [Online]. Disponível em: https://www.freelancinggig.com/blog/2017/08/07/best-programming-languages-data-mining/. [Acedido: 07-Jan-2018]

[50] R-project, «R: What is R?», *R-project*, 2018. [Online]. Disponível em: https://www.r-project.org/about.html. [Acedido: 07-Jan-2018]

[51] Python Software Foundation, «Welcome to Python.org», *Python.org*, 2018. [Online]. Disponível em: https://www.python.org/. [Acedido: 07-Jan-2018]

[52] K. Willems, «Choosing R or Python for data analysis? An infographic», *DataCamp*, 12-Mai-2015. [Online]. Disponível em: http://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis. [Acedido: 07-Jan-2018]

[53] P. Prakash, «Compiler vs Interpreter - Difference between compiler and interpreter», *Codeforwin*, 17-Mai-2017. [Online]. Disponível em: http://codeforwin.org/2017/05/compiler-vs-interpreter.html. [Acedido: 12-Jan-2018]

[54] F. Rodrigues, «Descoberta de Conhecimento - Avaliação de Modelos», 2016. [Online]. Disponível em: https://moodle.isep.ipp.pt/mod/folder/view.php?id=49785

[55] PremierLeague, «Premier League Football News, Fixtures, Scores & Results», *Premier League*, 2018. [Online]. Disponível em: https://www.premierleague.com/. [Acedido: 18-Jun-2018]

[56] Football-data, «Football Betting | Football Results | Free Bets | Betting Odds», *Football-data*, 2018. [Online]. Disponível em: http://www.football-data.co.uk/. [Acedido: 18-Jun-2018]

[57] bet365, «bet365 - Sports Betting, Premier League, Champions League and World Cup 2018 Football Odds, plus Grand Slam Tennis Prices, Casino, Poker, Games, Vegas, Bingo», *bet365*, 2018. [Online]. Disponível em: https://www.bet365.com/en/. [Acedido: 18-Jun-2018]

[58] Sofifa, «sofifa.com», *Sofifa*, 2018. [Online]. Disponível em: https://sofifa.com/. [Acedido: 20-Jun-2018]

[59] EASports, «FIFA 19 - Football Video Game - EA SPORTS Official Site», *EA SPORTS*, 2018. [Online]. Disponível em: https://www.easports.com/fifa. [Acedido: 20-Jun-2018]

[60] M. K. C. from J. Wing *et al.*, *caret: Classification and Regression Training*. 2017 [Online]. Disponível em: https://CRAN.R-project.org/package=caret

[61] SPSS-Tutorials, «What are Z-Scores? Quick Tutorial with Examples», *SPSS-Tutorials*, 2016. [Online]. Disponível em: https://www.spss-tutorials.com/z-scores-what-and-why/. [Acedido: 15-Jul-2018]

[62] F. Rodrigues, «Descoberta de Conhecimento - Preparação dos Dados». 2016 [Online]. Disponível em: https://moodle.isep.ipp.pt/mod/folder/view.php?id=49785

[63] M. B. Kursa, «Boruta», *Boruta*, 2010. [Online]. Disponível em: https://mbq.github.io/Boruta/. [Acedido: 10-Jul-2018]

[64] M. Kuhn, *The caret Package*. 2018 [Online]. Disponível em: https://topepo.github.io/caret/recursive-feature-elimination.html. [Acedido: 30-Jul-2018]

[65] M. Pawlan, «What Is JavaFX? | JavaFX 2 Tutorials and Documentation», *What Is JavaFX?*, 2013. [Online]. Disponível em: https://docs.oracle.com/javafx/2/overview/jfxpub-overview.htm. [Acedido: 15-Ago-2018]

[66] Tutorialspoint, «HTTP Requests», *www.tutorialspoint.com*, 2018. [Online]. Disponível em: https://www.tutorialspoint.com/http/http_requests.htm. [Acedido: 15-Ago-2018]

[67] Tutorialspoint, «RESTful Web Services Tutorial», *www.tutorialspoint.com*, 2018. [Online]. Disponível em: https://www.tutorialspoint.com/restful/. [Acedido: 15-Ago-2018]