FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

1x2 - Previsão de Resultados de Jogos de Futebol

Luís Miguel da Silva Duarte

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Orientador: Prof. Dr. Carlos Soares

Co-orientador: Jorge Teixeira

29 de Junho de 2015

© Luís Duarte, 2015

Resumo

Nos últimos anos tem havido um grande crescimento do *Data Mining* devido à perceção do potencial que esta tecnologia pode ter no mundo da medicina, indústria e também no desporto. Cada vez mais se armazenam maiores quantidades de dados e a extração de informação desses dados pode ser extremamente útil na resolução de problemas.

O futebol é o desporto mais popular no mundo e gera muitas paixões envolta dos seus resultados. A antecipação de um resultado é uma tarefa muito complexa devido à grande quantidade de fatores que influenciam um resultado.

A tecnologia *Data Mining* aplicada ao futebol é uma área que não é muito divulgada, nem existem muitos documentos que abordem este assunto. Isto pode ser explicado pelo facto de as equipas de futebol quererem manter em secretismo todas as investigações que fazem, e assim tirarem vantagem da utilização desta tecnologia. Apesar disso, este assunto gera muito interesse em todo o mundo. Fãs, comunicação social e apostadores querem cada vez mais ter informações privilegiadas do que acontece num jogo de futebol.

Nesta dissertação é abordada a aplicação de técnicas de *Data Mining* na previsão de resultados de jogos de futebol. A previsão dos resultados surge na forma "1x2"e foram aplicados a jogos da Liga Portuguesa de futebol. Como conclusão da dissertação foi desenvolvido um modelo que faz a previsão de resultados de jogos futuros. Em todo o processo foram usados diversos dados dos jogos de futebol que foram disponibilizados pelo Laboratório SAPO/ U.Porto. Com esses dados foram desenvolvidos outros dados que se adequam mais ao problema. Foram também utilizados várias algoritmos de Machine Learning como por exemplo Knn,SVM,random forest etc.

Palavras-chave: Data Mining, Machine Learning, Futebol, Algoritmos, Previsão.

Abstract

During the last years there has been a huge growth of Data Mining due to the perception of the potential that this technology can have in the world of medicine, in industry and also in sport. Each day there is a growing storage of data, being the extraction of information from these data extremely useful in solving problems.

Football is the most popular sport in the world and so it generates many passions around its overcome. The overcome prediction is a very complex task, due to the large amount of factors that influence a football game.

The data mining technology applied to football is an area that is neither very widespread, nor are there many documents that address this issue. This can be explained by the fact that football teams want to keep in all investigations that they do in secret, and thus taking advantage of using this technology .Nevertheless, this subject generates much interest worldwide. Fans, media and punters increasingly want to have inside information of what happens in a football game .

In this dissertation it is addressed the application of data mining techniques to predict the overcomes of football matches. The forecast of the overcomes appear in form "1x2"and has been applied to Portuguese Football League matches. As a conclusion to this dissertation, a model that forecasts overcomes in future games was developed. Throughout the process, several data from football games that were provided by the laboratory SAPO/ U.Porto, were used. With these information, other data was developed, which suit the problem. There were also used several Machine Learning algorithms such as Knn, SVM, random forest etc.

Key-words: Data Mining, Machine Learning, Football, Algorithms, Prediction.

Agradecimentos

Desde já quero agradecer ao professor Carlos Soares pelo apoio e oportunidade que me deu na aprendizagem de uma tecnologia completamente nova para mim. Gostaria de agradecer também ao co-orientador Jorge Teixeira e Laboratório SAPO/ U.Porto pela ajuda e disponibilidade em fornecer todas as ferramentas que necessitava.

Agradeço também aos meus pais e à minha irmã pelo apoio incondicional que me dão em todas as decisões que tomo, pelo respeito e por me terem dado a oportunidade de concluir o curso.

À minha namorada por estar presente em todas as etapas da minha vida, sejam elas boas ou más, por me aturar quando estou em dia "não"e por ter sempre uma palavra de conforto.

À minha família pela disponibilidade para ajudar e por fazerem de mim o que sou hoje. Apesar de longe estão muito perto.

Finalmente, um agradecimento a todos os colegas e amigos que se atravessaram no caminho e que me ajudaram a ultrapassar esta fase da minha vida. Levo amigos verdadeiros para a vida.

Luís Duarte

"Conhece-te a ti mesmo,
torna-te consciente de tua ignorância e será sábio."

Sócrates

Conteúdo

Agradecimentos **v**

1 Introdução **1**
1.1 Estrutura do Documento 2

2 Revisão da Literatura **3**
2.1 Introdução 3
2.2 Tarefas 5
2.3 Metodologia 6
2.4 Classificação 8
2.5 Avaliação em Classificação 9
2.6 Medidas de avaliação 9
2.7 Metodologias de estimação de desempenho 11
2.8 Data Mining para desporto 13

3 Perceção e Preparação dos Dados **25**
3.1 Introdução 25
3.2 Pré-processamento 26
3.3 Análise exploratória de dados 30

4 Modelação **39**
4.1 Metodologia 39
4.2 Resultados e Discussão geral 44
4.3 Discussão Geral 52

5 Conclusão **53**
5.1 Trabalho futuro 54

Referências **55**

Lista de Figuras

2.1 Ilustração das áreas que deram origem ao Data Mining 4
2.2 Ilustração das fases da metodologia CRISP-DM 6
2.3 Árvore de decisão de um problema de crédito 9
2.4 Ilustração da divisão dos dados e avaliação 9
2.5 Divisão do conjunto de dados através de *holdout* 11
2.6 Divisão dos dados segunda a amostragem aleatória 12
2.7 Divisão do conjunto de dados através da técnica *8-fold cross validation* 12
2.8 Divisão do conjunto de dados por *boosttrap* 13

3.1 Sequência das etapas de pré-processamento 26
3.2 Média de golos marcados pelo F.C.Porto 31
3.3 Média de golos marcados pelo Beira-Mar) 31
3.4 Pontos alcançados pelo Benfica nos 5 jogos anteriores 32
3.5 Pontos alcançados pela Olhanense nos 5 jogos anteriores 32
3.6 Percentagem de vitórias do F.C.Porto no confronto direto 33
3.7 Percentagem de vitórias do V.Setúbal no confronto direto 33
3.8 Diferença de pontos entre as equipa consoante o resultado final 34
3.9 Média de golos marcados pela equipa visitada e o resultado final 34
3.10 Gráfico que relaciona a percentagem de empates no confronto direto entre as duas equipas nos anos anteriores e o resultado final do jogo 35
3.11 Gráfico boxplot que ilustra a percentagem de vitórias da equipa visitante em relação ao resultado final do jogo 36
3.12 Média de golos marcados pela equipa visitante nos jogos fora em relação ao resultado final 36
3.13 Gráfico que relaciona a posição da equipa da casa no campeonato e o resultado final 37

4.1 Divisão dos dados de treino e teste 42
4.2 Evolução dos dados em *Growing Window* 43
4.3 Evolução dos dados em *Sliding Window* 43
4.4 Sequência de processos de cada iteração 44
4.5 Desempenhos dos 8 algoritmos com datasets diferentes 47
4.6 Desempenhos dos 8 algoritmos com *datasets* iguais 48
4.7 Resultado da aplicação do teste de Nemenyi 49
4.8 Taxa de acerto por jornada do algoritmo Knn 51
4.9 Taxa de acerto por jornada do algoritmo SVM com *kernel* gaussiano 51

Lista de Tabelas

2.1 Matriz de confusão de previsão de resultado 10
2.2 Identificação numérica dos documentos 14
2.3 Identificação da aplicação dos documentos 15
2.4 Tipos de tarefas implementadas nos documentos 16
2.5 Avaliação dos documentos 17

3.1 Número de atributos em cada conjunto de dados da 1º e 2º iteração 30

4.1 Comparação de desempenhos dos diferentes algoritmos em *Growing Window* e *Sliding Window* 45
4.2 Comparação dos desempenhos dos algoritmos com conjunto de dados diferentes . 45
4.3 Matriz de confusão do algoritmo Knn 46
4.4 Matriz de confusão do algoritmo JRip 46
4.5 Matriz de confusão do algoritmo random forest 46
4.6 Desempenhos dos três algoritmos selecionados na segunda iteração 49
4.7 Desempenhos dos três algoritmos selecionados na segunda iteração 50
4.8 Desempenho dos algoritmos na primeira e segunda volta do campeonato 52

Abreviaturas e Símbolos

Knn	*k-nearest neighbors*
SVM	*Support vector machine*
RF	*Random Forest*
NB	*Naive Bayes*
NN	*Neuronal Network*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
RMS	*Root Mean Square Error*
MSE	*Mean Square Error*
MAE	*Mean Absolute Error*
GW	*Growing Window*
SW	*Sliding Window*
W	*Win*
L	*Loss*
D	*Draw*
FIFA	*Fédération Internationale de Football Association*
EPL	*English Premier League*
NRL	*Australian National Rugby League*
AFL	*Australian Football League*
SR	*Super Rugby*

Capítulo 1

Introdução

Nos últimos anos houve uma grande evolução das técnicas de recolha, armazenamento e transferências de grandes volumes de dados. Os dados não têm valor nenhum se não houver mecanismos que consigam extrair conhecimento dos mesmos. Perante esta necessidade surgiu o *Data Mining* [1]. *Data mining* é uma abordagem para problemas de apoio à decisão que envolve técnicas, ferramentas matemáticas e estatísticas para extração de conhecimento em grandes quantidades de dados. Essas informações podem ser extraídas através de padrões, associações, anomalias relevantes etc. As técnicas de *Data mining* têm sido aplicadas com sucesso em áreas como indústria automóvel, como por exemplo no planeamento de produção, desenvolvimento e analise de falhas, no retalho para identificarem a preferência dos clientes, na indústria das telecomunicações para identificar fraudes, melhorar estratégias de *marketing* e identificar falhas na rede entre outros [2].

O futebol é um dos desportos mais populares do mundo. Sendo assim, a perceção do jogo e a previsão dos resultados tornou-se interessante para vários grupos de pessoas, incluindo fãs, apostadores [3], treinadores e comunicação social [4]. A moral de uma equipa (ou jogador), aptidões físicas, fadiga, lesões, objetivos são exemplos de fatores que influenciam um jogo de futebol. Estes fatores tornam a previsão do resultado final uma tarefa difícil e, por isso, interessante. Mesmo para os especialistas na área, como jornalistas, comentadores, treinadores prever um resultado de um jogo é uma tarefa complexa [5].

O objetivo principal desta dissertação é desenvolver um estudo empírico de uma abordagem de *Data Mining* no âmbito do problema que é a previsão do resultado em jogos de futebol. Este problema foi abordado como sendo um problema de classificação, em que o objetivo é prever o resultado do jogo em forma de "1x2"(vitória, empate ou derrota). Hoje em dia existe a aquisição de grandes quantidades de dados sobre cada jogo, jogador, equipa, treinos etc. No entanto, há pouca informação acerca do uso desses dados por parte das equipas, embora não haja dúvidas de que já perceberam a potencialidade da exploração desses dados para seu benefício [6]. Nesta dissertação foi aplicada a metodologia CRISP-DM para a resolução do problema. Foram realizadas duas iterações desta metodologia que permitiu uma evolução na aplicação dos métodos. Durante este percurso foram abordadas diversas variantes de algoritmos, variáveis e métodos de preparação. No fim de todo o processo foi desenvolvido um modelo final tendo em conta as características

e algoritmo que obtiverem melhores resultados. O modelo final poderá assim ser utilizado na previsão de resultados de jogos de outras ligas de futebol. Os dados utilizados na dissertação são de várias épocas da liga portuguesa e foram disponibilizados pelo Laboratório SAPO/ U.Porto. Todo o processo realizado nesta dissertação foi feito com recurso à linguagem R, que é uma linguagem de programação orientada para análise estatística e produção de gráficos [7]. O código criado em todas as fases do processo tem como base o pacote Caret. Este pacote fornece ferramentas de divisão dos dados, pré-processamento,seleção de atributos,modelação etc [8].

1.1 Estrutura do Documento

O presente documento encontra-se dividido em cinco capítulos: introdução, revisão da literatura, perceção e preparação dos dados, modelação e conclusão. Na revisão da literatura é feita uma introdução ao *Data Mining*, assim como a discusão do trabalho já realizado da sua aplicação no âmbito do desporto. O Capítulo 3 contem todo o processo de receção e análise e preparação dos dados fornecidos. No capítulo de modelação é explicado todos os algoritmos usados, como foram feitas as divisões dos dados, bem como os resultados alcançados. No último capítulo é feita uma conclusão de todo o trabalho realizado, e dar a conhecer perspetivas de trabalho futuro.

Capítulo 2

Revisão da Literatura

2.1 Introdução

Em computação, muitos dos problemas são resolvidos através de algoritmos que especificam, passo a passo, como resolver um problema. Porém, não é fácil para uma máquina realizar algumas das simples tarefas que o cérebro humano realiza no seu dia-a-dia, como por exemplo, perceber a tendência das equipas num jogo de futebol, o estado anímico dos jogadores, qual a equipa que está a jogar melhor etc. Cada vez mais treinadores, jogadores, fãs, apostadores dão maior importância à realização destas mesmas tarefas para que possam perceber melhor o que se passa num jogo de futebol e obter melhor rendimento dentro do campo [2].

As técnicas de inteligência artificial têm sido utilizadas com sucesso num grande número de problemas reais. Em particular, tendo em conta o volume crescente de dados que são recolhidos das atividades desportivas, a extração de conhecimento útil para o apoio à decisão com abordagens como estatística ou *Data Mining*, tem-se tornado cada vez mais interessante [2].

ML (*Machine learning*) é uma área de investigação das ciências de computação bem reconhecida e com muitos anos de investigação. Esta área implica o estudo e desenvolvimento de algoritmos que conseguem extrair conhecimento automaticamente a partir de experiências passadas, sem intervenção humana [9].

DM (*Data Mining*) é um conceito relativamente recente na área das ciências da computação. Com a crescente complexidade dos problemas e volume de dados, é essencial tentar aproveitar a máxima informação contida nesses dados e descobrir se há algum conhecimento contido neles [10]. De acordo com Witten and Frank (2005), a definição de *Data Mining* é: " ...Resolução de problemas através da análise de dados existentes nas bases de dados." [9]

Data Mining tornou-se assim uma ajuda fundamental para uma melhor compreensão dos problemas por parte dos seres humanos.

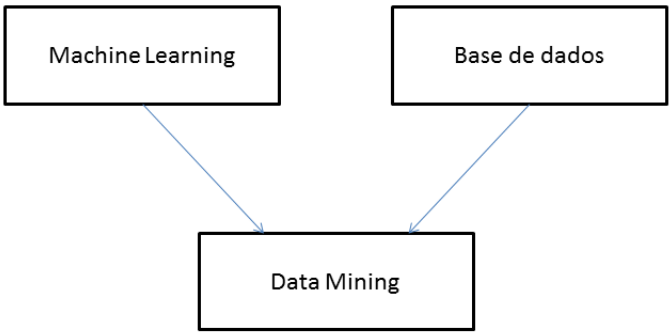

Figura 2.1: Ilustração das áreas que deram origem ao Data Mining

O processo de *Data Mining* aplica algoritmos de *ML* para a resolução de problemas com grandes quantidades de dados. Como se pode ver na Figura 2.1, o *Data Mining* junta o melhor de dois mundos, por um lado o processo geral e computacional (algoritmos de ML), por outro lado o processo específico e racional de acordo com o problema (tratamento dos dados)[1].Os algoritmos de ML mais usados em problemas de *Data Mining* são redes neuronais, algoritmos genéticos, árvores de decisão e *SVM* (*Support Vector Machine*)[11].

A capacidade de aprendizagem é considerada essencial para um comportamento inteligente. Atividades como memorizar, observar e explorar situações para aprender factos, melhorar habilidades cognitivas através da prática, organizar conhecimento novo e utilizar representações apropriadas, podem ser consideradas atividades relacionadas com aprendizagem. Em aprendizagem computacional, os computadores são programados para aprender com a experiência passada. Para tal, empregam um princípio de inferência denominado de indução, no qual se obtém conclusões genéricas a partir de um conjunto particular de registos [2]. Sendo X os registos do conjunto de dados e i o número do registo, xi representa o i-ésimo registo do conjunto de dados. Dado um conjunto de dados N=(xi,f(xi)), f representa o atributo alvo como se pode ver na equação 2.1.

$$yi = f(xi) \tag{2.1}$$

Uma técnica *Data Mining* de previsão H aprende a prever o valor de f para novos registos de X . Sendo assim, os valores do atributo alvo previsto podem ser dados por yp, como pode ser visto na equação 2.2 .

$$yp = h(xi) \tag{2.2}$$

Um dos requisitos muito importantes nos algoritmos de *Data Mining* é a capacidade de lidar com a falta de dados ou dados imperfeitos. Como a ausência de dados ou a presença de dados imperfeitos pode influenciar a eficácia do algoritmo é aconselhável o uso de técnicas de pré-processamento para a identificação e minimização da ocorrência desses problemas [2].

2.2 Tarefas

As tarefas de aprendizagem computacional podem ser divididas em duas categorias, tarefas de previsão e descrição.

As tarefas de previsão ou preditivas constroem um ou mais modelos com o objetivo de prever o valor do atributo alvo em novos dados [12]. Estas tarefas seguem o paradigma da aprendizagem supervisionada. Na aprendizagem supervisionada, os dados incluem um atributo-alvo, cujos valores podem ser estimados utilizando os atributos de entrada do registo . O objetivo de um algoritmo de *Data Mining* utilizado nestas tarefas é aprender, a partir de um conjunto de dados, um modelo ou hipótese capaz de relacionar os valores dos atributos de entrada do registo com o valor do atributo-alvo [2].

As tarefas de descrição ou descritivas descrevem o conjunto de dados de uma maneira simples e concisa apresentando as propriedades gerais dos dados [12]. Associação e *Clustering* são duas das tarefas descritivas existentes. As tarefas de descrição seguem o paradigma da aprendizagem não-supervisionada em que não existe um atributo-alvo. Ao contrário das tarefas supervisionadas, as tarefas não-supervisionadas podem ser mais difíceis de avaliar, dado que não existe um atributo-alvo com o qual se pode comparar e assim avaliar o desempenho do modelo [2].

As tarefas preditivas ou supervisionadas podem ser divididas em dois categorias, tarefas de classificação e tarefas de regressão. As tarefas preditivas de classificação têm por objetivo identificar qual o valor do atributo-alvo a que pertence um determinado registo. Como é uma tarefa de classificação, o atributo-alvo é uma classe (atributo discreto). Nestas tarefas, os registos incluídos no conjunto de dados de treino contém as classes dos registos o que permite ao modelo aprender a classificar os novos dados [13]. A tarefa de classificação pode ser usada para identificar o resultado de um jogo de futebol em relação à equipa da casa, que pode estar incluído numa de três classes: vitória, empate, derrota. Nas tarefas preditivas de regressão o processo é semelhante à classificação. A diferença reside no facto de a identificação do atributo-alvo do registo não ser feita através de uma classe mas sim de um valor numérico [13]. Como exemplo de uma tarefa de regressão seria o número de golos de um jogo de futebol.

As tarefas descritivas ou não supervisionadas podem ser de *clustering*, associação ou sumarização. O *clustering* ou agrupamento tem como objetivo identificar o agrupamento natural de um conjunto de registos. Um agrupamento (ou *cluster*) é um conjunto de registos cuja semelhança entre registos do mesmo grupo é elevada e a semelhança entre registos de grupos diferentes é reduzida [14]. Esta tarefa difere da classificação pois não necessita que os registos sejam previamente classificados. A associação consiste em encontrar padrões frequentes de relacionamento entre os valores dos atributos de um conjunto de dados. A sumarização tem como objetivo encontrar uma descrição simples e compacta de um conjunto de dados [2].

2.3 Metodologia

Poder-se-ia pensar que, ter acesso aos dados e escolhendo um algoritmo facilmente se achava a solução para um determinado problema. Como foi dito anteriormente, o processo *Data Mining* dá uma enorme importância ao contexto do problema. Não interessa encontrar uma solução que não se enquadre no problema. Por tudo isto, é necessário definir uma metodologia para que o processo *Data Mining* passe pelas diversas fases desde a perceção do problema até à sua resolução. A metodologia escolhida foi CRISP-DM (*Cross-Industry Standard Process for Data Mining*). O atual modelo de processo CRISP-DM fornece uma visão geral do ciclo de vida de um projeto de *Data Mining*. Este processo contém as fases do projeto, as respetivas tarefas e as relações entre as tarefas. O ciclo de vida de um projeto de *Data Mining* é composto por seis fases. A sequência das fases não é rígida, dado que pode ser necessário voltar a uma fase que já foi contemplada no passado. O resultado de cada fase determina, qual a fase ou tarefa que será realizada de seguida [15]. O ciclo de vida de um projeto pode ser visto na figura 2.2.

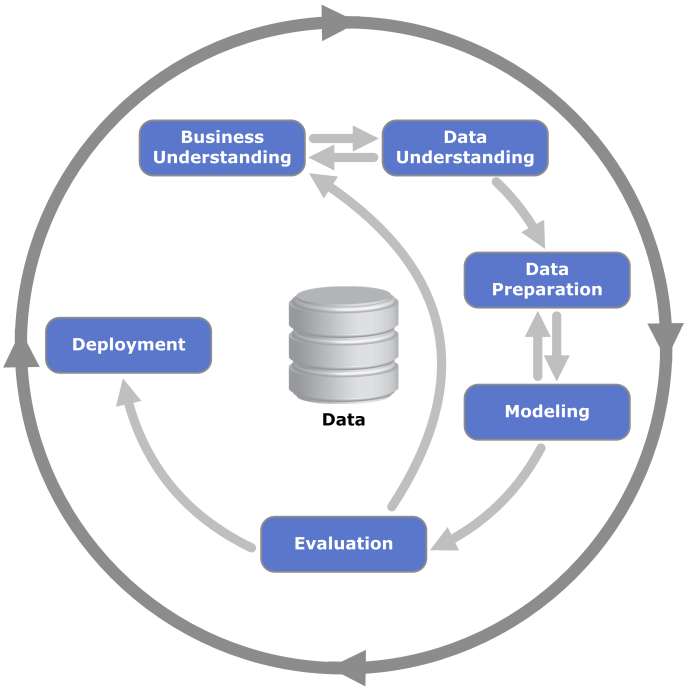

Figura 2.2: Ilustração das fases da metodologia CRISP-DM
Fonte: [16]

Uma vez encontrada uma solução, não significa que o processo de *Data Mining* termina. Esta solução pode desencadear questões novas mais focadas nos detalhes, que podem ser resolvidas devido à experiência anterior [15].

As fases do CRISP-DM são:

Entendimento do negócio

Esta fase inicial do projeto baseia-se na compreensão do objetivo do projeto e requisitos de uma perspetiva de negócio. Uma vez definido o problema numa perspetiva de negócio, este deve ser transformado num problema de *Data Mining* e deve ser desenhado um plano preliminar de implementação para alcançar os objetivos do projeto [15]. Devem, também, ser definidos os critérios de sucesso e as medidas de avaliação dos resultados [14].

Compreensão dos dados

Esta fase inicia-se com a aquisição e armazenamento dos dados. De seguida, inicia-se o processamento dos mesmos com atividades que permitem aumentar a familiaridade com os dados, identificar problemas na qualidade dos mesmos e detetar subconjuntos de dados interessantes e relevantes para o problema em questão [15]. Este processo pode ser feito com recurso a análises estatísticas, gráficos etc. A compreensão dos dados é essencial para o sucesso de todo o projeto.

Preparação dos dados

A preparação dos dados inclui a execução de um conjunto de atividades necessárias para que, tendo um conjunto de dados no seu estado original, se obtenha um conjunto de dados final que será utilizado na fase de modelação. Nesta fase são executadas tarefas de limpeza de dados, tratamento de dados, seleção de atributos e transformação de dados [15].

Modelação

Nesta fase vários algoritmos são selecionados e os seus parâmetros são calibrados para valores ótimos. Normalmente para o mesmo problema de *Data Mining* existem vários algoritmos que podem ser aplicados na procura da melhor solução. Alguns algoritmos requerem tipos de dados específicos. Por isso, por vezes é necessário recuar até a fase de preparação dos dados para que estes sejam alterados de acordo com o algoritmo a utilizar [15]. A busca pelos modelos com resultados ótimos pode levar a este retrocesso seja feito várias vezes. Por fim, são desenvolvidos processos para avaliar o desempenho e qualidade dos modelos.

Avaliação

Nesta fase, já é expectável que haja um modelo construído que tenha um desempenho razoável. Antes de avançar para a fase final de desenvolvimento, é importante avaliar as propriedades do modelo para ter a certeza que os objetivos do negócio são cumpridos. O objetivo chave desta fase é determinar se há aspetos importantes de negócio que não ainda foram considerados. No final desta fase, deve ser tomada uma decisão acerca dos resultados alcançados [15].

Desenvolvimento

A criação do modelo não significa que seja o fim do projeto. Mesmo que o objetivo do modelo seja aumentar o conhecimento pelos dados, esse conhecimento tem que ser organizado e

apresentado de uma forma que o cliente o possa usar. Dependendo dos requisitos, a fase de desenvolvimento pode ser tão simples como gerar um relatório sobre os dados, aplicar os modelos a um novo conjunto de dados etc [15]. Pode também implicar soluções mais complexas, como introduzir soluções e estratégias de desenvolvimento na área de negócio em estudo, de acordo com os resultados obtidos [14].

2.4 Classificação

Tendo um conjunto de dados com registos pré-classificados (cuja classe é conhecida), a classificação consiste na construção de um modelo que seja capaz de classificar automaticamente os novos registos (cuja classe é desconhecida) de acordo com os seus atributos. A tarefa de classificação iniciasse com a construção de um modelo através de dados históricos. O objetivo das tarefas de classificação é gerar modelos que preveem com precisão a classe alvo, para cada registo de um novo conjunto de dados. O que o algoritmo de classificação faz é tentar encontrar uma fronteira de decisão que separe os registos das diferentes classes [2].Alguns exemplos da aplicação de métodos de classificação podem ser a classificação de tendências dos mercados financeiros ou a identificação automática de objetos de interesses em imagens [17]. Um bom modelo de classificação é aquele que consegue distinguir os exemplos para todas as classes. Para se saber se o modelo implementado é um bom classificador ou não, é necessário medir o desempenho do modelo.

Árvores de decisão, redes neuronais, regras de decisão são alguns dos algoritmos de classificação usados.

2.4.1 Árvores de decisão

O algoritmo Árvore de Decisão funciona como um fluxograma em forma de árvore. Cada nó interno (não folha) indica um teste feito sobre um atributo (por exemplo, idade > 18), cada ramificação representa o resultado do teste e as folhas indicam a classe a qual o registo pertence. O nó que está no nível superior é o nó raiz [12].Querendo classificar um novo registo cujo y é desconhecido, o que a árvore de decisão faz é testar cada um dos atributos desse registo nos seus nós internos. De acordo com os atributos do registo, é traçado um caminho desde o nó raiz até a uma folha que irá atribuir uma classe ao registo. Devido aos testes efetuados em cada nó interno, as árvores de decisão são facilmente convertidas em regras de classificação. O algoritmo de árvore de decisão é muito popular porque é uma técnica simples, não requer parâmetros de configuração e funciona bem com grandes volumes de dados. Também a sua representação em forma de árvore é geralmente fácil de assimilar por parte dos seres humanos, para além de ter uma boa taxa de acerto. Os algoritmos de árvore de decisão para classificação têm sido usados na área da medicina, produção, análise financeira, astronomia, biologia molecular etc [12]. A figura 2.3 ilustra um exemplo de uma árvore de decisão.

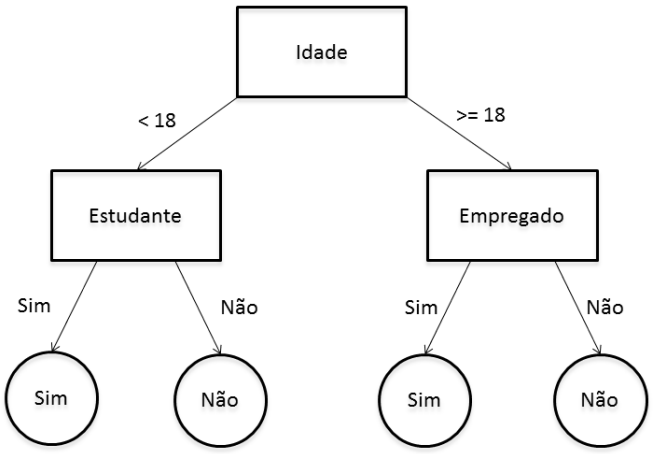

Figura 2.3: Árvore de decisão de um problema de crédito

2.5 Avaliação em Classificação

Os modelos desenvolvidos foram construídos com base em dados históricos, mas serão aplicados a dados com novos registos. Por isso, é necessário avaliar os modelos para verificar o seu desempenho quando confrontados com novos dados. A avaliação do desempenho do modelo é feita dividindo o conjunto de dados em duas partes, conjunto de dados de treino e conjunto de dados de teste. Esta divisão pode ser feita usando várias técnicas, como por exemplo *holdout*, *cross validation* etc. O conjunto de dados de treino são os dados históricos e servem para a construção do modelo. O desempenho do modelo é avaliado segundo uma medida de avaliação selecionada e usando a previsão feita com o conjunto de dados de teste [13]. Na figura 2.4 pode-se ver a divisão dos dados, a construção do modelo e a sua avaliação.

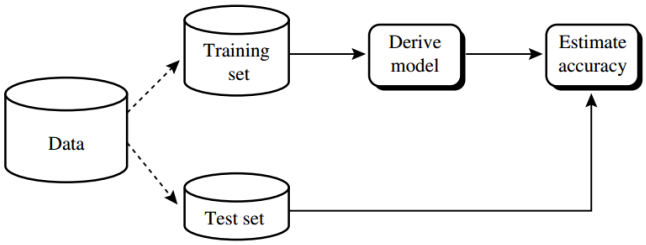

Figura 2.4: Ilustração da divisão dos dados e avaliação
Fonte: [12]

2.6 Medidas de avaliação

Existem diversas medidas de avaliação dos modelos que permitem quantificar o desempenho de cada um deles. Uma das métricas de desempenho usada na avaliação de um modelo é a taxa de

acerto ou *accuracy*, que é dada pela equação 2.3.

$$acc(h) = \frac{1}{n}\sum_{i=1}^{n} I(yi = h(xi)) \tag{2.3}$$

Tendo N exemplos de avaliação, a taxa de acerto é equivale à proporção de exemplos classificados corretamente pelo classificador H. Se a classe que foi prevista pelo modelo de classificação for igual à classe real então I(yi=h(xi))=1, caso contrário I(yi=h(xi))=0. Este tipo de medida equivale ao uso da função de custo 0-1 [2].

A taxa de acerto varia entre zero e um, em que os melhores modelos de classificação têm uma taxa de erro próxima de um. Uma métrica também muito usada é a taxa de erro. A taxa de erro é o complemento da taxa de acerto, como se pode ver na equação 2.4. Neste caso, os modelos de classificação com melhor desempenho são os que obtêm valores próximos de zero [2].

$$err(h) = 1 - acc(h) \tag{2.4}$$

Outra alternativa para avaliar o desempenho é a matriz de confusão. A matriz de confusão é uma tabela para visualização dos resultados tipicamente usada em classificação. Para um conjunto de dados, as colunas dessa matriz representam as classes verdadeiras, e as linhas, as classes previstas pelo classificador. A diagonal representa os acertos do classificador, enquanto os outros elementos correspondem aos erros cometidos nas suas predições. Através desta matriz consegue-se perceber quais as classes em que o algoritmo tem maior dificuldade de previsão [2].

Tabela 2.1: Matriz de confusão de previsão de resultado

Classe de previsão	Classe verdadeira		
	Vitória	Empate	Derrota
Vitória	8	2	1
Empate	1	7	1
Derrota	2	2	5

A tabela 2.1 representa um exemplo de uma matriz de confusão de o número de jogos de futebol que acabou em vitória, empate ou derrota da equipa da casa. A diagonal a verde representa os jogos acertados pelo modelo, e a vermelho será os jogos em que o modelo errou na sua previsão. Neste exemplo, o modelo teria acertado em 20 dos 29 jogos, o que perfaz uma taxa de acerto de aproximadamente 69%. Pode-se concluir também que o modelo teria previsto corretamente 8 das 11 vitórias da equipa da casa, apresentado uma taxa de acerto de aproximadamente 72% nas vitórias.

Apesar das métricas acima serem as mais conhecidas, existem ainda a taxa de sensibilidade, precisão, entre outras. Os modelos podem ainda ser avaliados com outros tipos de desempenho, tais como o tempo de classificação ou criação do modelo, questões de estabilidade etc.

2.7 Metodologias de estimação de desempenho

Como foi dito anteriormente, para se poder avaliar o desempenho de um modelo é necessário ter dados de treino em que o valor da variável y é conhecido, e ter dados de teste para validar o modelo. Essa divisão dos dados é necessária para que o modelo não fique dependente de um conjunto de dados específico e, ao ser submetido a outros conjuntos (com valores diferentes dos usados na construção e validação do modelo), apresente resultados insatisfatórios. Neste caso, o modelo pode simplesmente "memorizar" os dados de treino. Isto implicaria que fosse 100% preciso em prever o valor de y para esses dados, mas provavelmente muito impreciso a prever o valor de y para novos dados. Este efeito é chamado ajustamento aos dados de treino ou *overfitting*. À medida que se aumenta a precisão do modelo para um conjunto de dados específico, perde-se a precisão para outros conjuntos [13]. Os dados de teste permitem avaliar a capacidade do modelo em prever o valor de y para dados diferentes dos dados de treino. Esta divisão dos dados pode ser feita de várias formas:

- *Holdout*

  Tendo um conjunto de dados N, divide-se numa proporção P para treino e (1-P) para teste. Esta abordagem é adequada quando há um grande volume de dados. Quando o volume de dados é pequeno, ou usamos poucos dados para treino (prejudicando o modelo) ou para teste (prejudicando a qualidade da estimativa do desempenho). Outro problema com esta abordagem prende-se no facto de uma classe pode ficar muito representada num conjunto de dados e pouco representada no outro. A figura 2.5 ilustra a divisão dos dados segundo a técnica *holdout*.

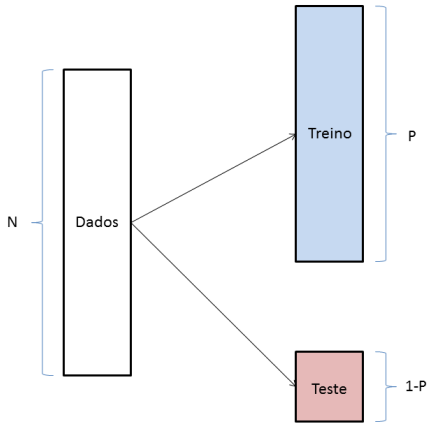

Figura 2.5: Divisão do conjunto de dados através de *holdout*

- Amostragem aleatória

  O método *holdout* faz com que os resultados sejam muito dependentes da partição escolhida para teste. A amostragem aleatória contraria esta dependência executando o método *holdout* diversas vezes com partições de teste aleatórias. Os resultados deste método são dados pela média dos diferentes testes [2].

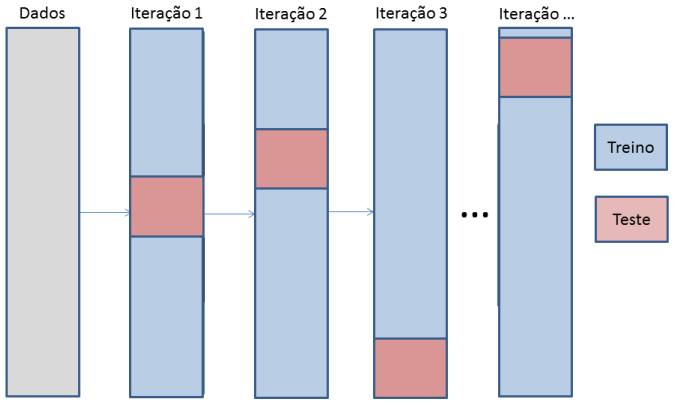

Figura 2.6: Divisão dos dados segunda a amostragem aleatória

- *N-fold cross-validation*

  O conjunto de exemplos é dividido em N subconjuntos de tamanho aproximadamente igual. Uma das partições é usada para teste, enquanto as restantes são utilizadas no treino do método. Este processo é realizado N vezes, utilizando em cada ciclo uma partição diferente para teste. O desempenho final é dado pela média dos desempenhos observados sobre cada subconjunto de teste [2]. A figura 2.7 ilustra o método *8-fold cross validation*.

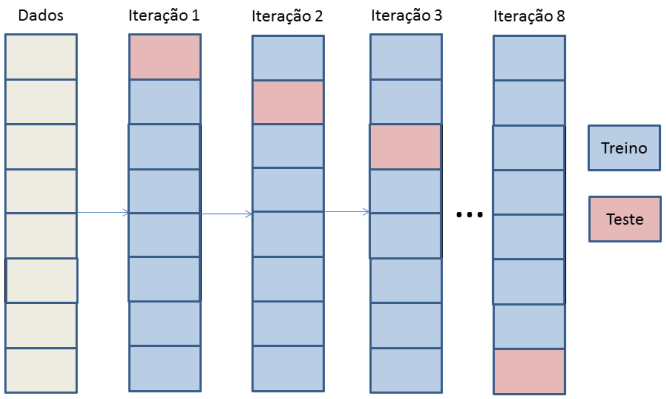

Figura 2.7: Divisão do conjunto de dados através da técnica *8-fold cross validation*

- *Leave-one-out*

  É um caso particular do método *cross-validation*. Em cada ciclo um exemplo é separado para teste, enquanto os restantes são usados no treino. O desempenho é dado pela soma dos desempenhos verificados para cada exemplo de teste individual [2]. Não é possível estratificação e é computacionalmente muito pesado.

- *Bootstrap*

  Neste método, são gerados N subconjuntos de treino a partir do conjunto de exemplos original. Os exemplos são amostrados aleatoriamente desse conjunto, com reposição. O resultado é dado pela média do desempenho em cada subconjunto de teste. Esta abordagem é adequada quando o volume de dados é pequeno [2].

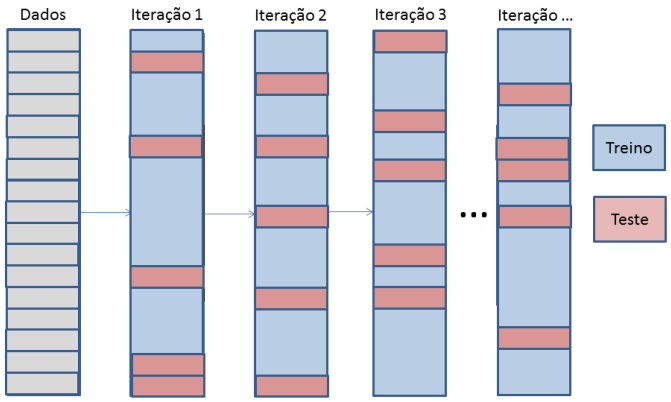

Figura 2.8: Divisão do conjunto de dados por *boosttrap*

Nos métodos que envolvem médias de desempenho, deve-se reportar também os valores do desvio padrão que poderá ser um indicativo de sensibilidade dos objetos usados no treino. Um desvio padrão elevado indica uma alta variância nos resultados ou seja, uma instabilidade do modelo perante mudança nos objetos [2].

2.8 Data Mining para desporto

Existem já uma quantidade significativa de trabalhos relativos à aplicação de técnicas de *data mining* em desporto. Foram analisados alguns relativos ao futebol mas não só, para termos uma visão alargada do que já foi feito e de quais as potencialidades. Os trabalhos podem ser caracterizados de acordo com as seguintes dimensões: trabalho estudado, objetivo do estudo, objetivo de *Data Mining*, desporto, tarefa de *Data Mining*, tipo de atributos, torneio, medida de avaliação, metodologia de estimação de desempenho, validação e algoritmos que foram estudados nos trabalhos.

2.8.1 Dimensões

A maior parte dos trabalhos estudados foram caracterizados segundo dez dimensões. Os trabalhos cujo ponto de interesse se focava em poucas dimensões não foram incluídos nas tabelas.

Devido à grande dimensão das tabelas 2.3,2.4 e 2.5, foi atribuído um número a cada documento para o identificar como é visível na figura 2.2.

Tabela 2.2: Identificação numérica dos documentos

Título	Documento
Compound Framework for Sports Prediction:The Case of Study	1
Football Predictions based on a fuzzy model genetic and neural tuning	2
Soft Computing-Based Result Prediction of Football Games	3
Predicting football results using Bayesian nets and other machine learning	4
Predicting football scores using machine learn techniques	5
Applying Data Mining techniques to Football Data from European Championships	6
An improved prediction system for football a match result	7
Football result prediction with Bayesian Network in Spanish League- Barcelona Team	8
Bayesian hierarchical model for the prediction of football results	9
A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup	10
A simlulation model for football championships	11
Soccer Match Result Prediction using Neural Networks	12
Artificial Intelligence in Sports Prediction	13
A Comparative Study on Neural Network based Soccer Result Prediction	14
A Neural Network Method for Prediction of 2006 World Cup Football Game	15
Predicting Soccer Match Results in the English Premier League	16
Predicting the outcome of NBA playoffs using the Naïve Bayes Algorithms	17

Tabela 2.3: Identificação da aplicação dos documentos

Doc.	Objetivo do Estudo	Desporto	Torneio
1	.Simulação de um torneio	Futebol	Mundial 2002
2	.Previsão de resultado final	Futebol	Liga Finlandesa
3	.Comparação de técnicas data mining .Simulação de um torneio	Futebol	Liga Ucraniana
4	.Comparação de técnicas data mining .Estudo de fatores	Futebol	Liga Inglesa Equipa:Tottenham
5	.Comparação de técnicas data mining .Estudo de fatores	Futebol	Liga dos Campeões
6	.Trabalho exploratório	Futebol	Liga Portuguesa, Inglesa, Espanhola, Italiana, Francesa, Alemã
7	.Comparação de técnicas data mining	Futebol	Liga Inglesa
8	.Previsão do resultado de um jogo de futebol	Futebol	Liga Espanhola Equipa: Barcelona
9	.Estudo de fatores .Previsão do resultado final	Futebol	Liga Italiana
10	.Simulação de um torneio	Futebol	Mundial 2006
11	.Simulação de um torneio	Futebol	Euro 2000
12	.Comparação de técnicas data mining	Futebol	Liga Inglesa
13	.Avaliar desempenhos em cada um dos desportos	Futebol e Rugby	NRL AFL Liga Inglesa
14	.Comparação de técnicas data mining	Futebol	Liga Italiana
15	.Previsão de resultado final	Futebol	Mundial 2006
16	.Comparação de técnicas data mining	Futebol	Liga Inglesa
17	.Simulação de um torneio	Basquetebol	NBA

Tabela 2.4: Tipos de tarefas implementadas nos documentos

Doc.	Objectivos de Data Mining	Tarefa de Data Mining	Tipos de Atributos
1	Previsão do resultado de um jogo e vencedor de torneio	Classificação	.Características do jogo atual .Características físicas e mentais .Organização da equipa
2	Previsão de resultado	Classificação	.Histórico de jogos .Fator Casa
3	Previsão de resultado e vencedor de campeonato	Classificação	.Característica do jogo atual .Histórico de golos .Ranking .Desempenho recente
4	Previsão de resultado	Classificação	.Característica do jogo atual .Fator Casa .Ranking .Características físicas
5	Previsão de resultado	Classificação	.Desempenho recente .Rankings .Característica do jogo atual .Histórico de golos
6	Previsão de resultado	Classificação	.Característica do jogo atual .Histórico de golos .Histórico de jogos
7	previsão do resultado	Classificação	.Característica do jogo atual .Histórico de golos .Organização da equipa
8	previsão do resultado	Classificação	.Características do jogo atual .Histórico de jogos e golos .Desempenho recente .Características físicas e mentais
9	previsão do resultado	Classificação	.Organização da equipa .Histórico de golos .Fator casa
10	Previsão de resultado final e vencedor de torneio	Regressão	.Opinião de especialistas .Ranking .Fator Casa
11	Previsão de resultado final e vencedor de torneio	Regressão	.Organização da equipa .Ranking
12	Previsão de resultado final	Classificação	.Histórico de jogos .Histórico de golos .Desempenho Recente
13	Previsão de resultado final	Classificação	.Histórico de jogos .Desempenho recente .Ranking
14	Previsão de resultado	Classificação	.Histórico de jogos .Desempenho recente .Ranking
15	Previsão de resultado	Classificação	.Histórico de golos .Histórico de jogos .Características dos jogos anteriores
16	Previsão de resultado	Classificação	.Característica do jogo atual .Desempenho recente .Ranking
17	Previsão de resultado	Classificação	.Histórico de jogos .Fator casa

Tabela 2.5: Avaliação dos documentos

Doc.	Medida de Avaliação	Metodologia de Estimação do Desempenho	Validação	Algoritmos
1	.MAE .RMS	-	.Comparação de desempenho com outros modelos .Comparação com resultados reais	.Redes Bayesianas .Regras de decisão
2	Taxa de acerto	Holdout/ Growing Window	.Comparação com resultados reais	.Modelo fuzzy
3	.Taxa de acerto .MSE	Holdout/ Growing Window	Comparação com a classificação real	.Modelo Fuzzy .Redes neuronais .Algoritmo genético
4	Taxa de acerto	Holdout/ Growing Window	.Comparação de desempenho com outros modelos	.Redes Bayesianas .Árvore de decisão .Naive Bayes .Knn
5	Taxa de acerto	10 fold cross-validation	.Comparação de desempenho com outros modelos	.Naive Bayes .Redes Bayesianas .LogistBoost .Knn .Redes neuronais
6	Taxa de acerto	Holdout	-	.Árvore de decisão
7	Taxa de acerto	-	.Comparação de desempenho com outros modelos	.Redes Neuronais .Regressão logística
8	Taxa de acerto	-	.Comparação com resultados reais	.Rede Bayesiana
9	Taxa de acerto	-	.Comparação de desempenho com outros modelos	.Rede Bayesiana
10	Taxa de acerto	-	.Comparação de desempenho com outros modelos	.Abordagem Bayesiana
11	-	-	.Comparação com resultados reais	-
12	Taxa de acerto	10-fold cross-validation	.Comparação de desempenho com outros modelos	.Redes neuronais .Arvores de decisão .KNN .Naive Bayes
13	Taxa de acerto	-	Comparação com previsões de especialistas	.Redes Neuronais
14	.Taxa de acerto .MSE	Holdout/ Growing Window	.Comparação de desempenho com outros modelos	.Redes Neuronais
15	Taxa de acerto	-	-	.Redes Neuronais
16	Taxa de erro	Holdout/ Growing Window	Comparação com previsões de especialistas	.Naive Bayes .SVM Gaussiano .SVM Linear .Random forest
17	Taxa de acerto	Holdout/ Growing Window	.Comparação com resultados reais	.Naive Bayes

A tabela 2.3 caracteriza os documentos pela aplicação do mesmo.

A tabela 2.4 mostra o tipo de tarefas que são executadas em cada trabalho.

A tabela 2.5 ilustra como é feita a avaliação dos modelos, bem como os algoritmos que foram usados.

2.8.1.1 Objetivo do estudo

Existem fundamentalmente cinco objetivos dos trabalhos estudados, simulação de um torneio ou jogo de futebol, previsão de resultados, comparação de técnicas de *Data Mining*, deteção de fatores relevantes num jogo e uso da mesma técnica em diferentes desportos. A simulação de um torneio tem como objetivo simular todos os jogos do torneio e determinar o vencedor do mesmo. Normalmente os torneios têm duas fases distintas: uma fase de grupos em que os dois primeiros de cada grupo passam para uma fase a eliminar. Posto isto, a maior parte dos documentos estudados divide o seu estudo em duas fases, em que na primeira fase são utilizados dados referentes aos torneios anteriores, e na fase posterior já é incluído o rendimento de cada equipa na fase anterior (ex. [5],[18]).

Alguns trabalhos têm como propósito a previsão do resultado final de um jogo de futebol. A previsão do resultado é feita tanto através do número de golos marcados por cada uma das equipas, como também de uma forma categórica em relação à equipa da casa (vitória, empate ou derrota) (ex.[19], [20], [21]).

A comparação de técnicas de *Data Mining* aplicadas ao futebol, também é um dos objetivos propostos por diversos trabalhos, entre eles [22] e [23]. Nestes trabalhos são usadas diversas técnicas de *Data Mining* na previsão de resultados e posteriormente são comparados os desempenhos de cada uma das técnicas.

Existem diversos fatores num jogo de futebol que influenciam o resultado final do mesmo. [24] e [25] centram a sua atenção em busca desses fatores. Embora tenham feito um grande esforço nessa procura, o que a maior parte dos documentos diz é que não é totalmente claro os fatores que influenciam o jogo, pois o futebol depende de fatores muito raros (quando comparado com outros desporto como o basquetebol ou ténis) e a falha de um simples jogador pode ditar o resultado de um jogo.

Alguns trabalhos focam-se ainda na descoberta de conhecimento acerca do futebol através de análises exploratórias. O trabalho [26] optou pelo uso das redes neuronais no futebol e no *rugby*, avaliando assim os desempenhos em cada um dos desportos.

2.8.1.2 Objetivos de data mining

Os documentos estudados centram-se maioritariamente em dois objetivos de *Data Mining*, previsão do resultado de um jogo de futebol ou previsão do vencedor de um campeonato ou torneio. Um torneio normalmente tem diversas fases. A fase de grupos e a fase a eliminar dos torneios obrigam a diferentes análises de um jogo de futebol. Por exemplo, uma equipa normalmente joga mais “descontraída” na fase de grupos porque sabe que vai fazer três jogos, ao contrário da fase a

eliminar que se perder é eliminada do torneio. Este "estado de espirito", o cansaço, a moral são tudo fatores analisados nas diferentes fases dos torneios.

2.8.1.3 Desporto

A maior parte dos trabalhos selecionados para o estudo do problema *Data Mining* centraram-se num único desporto, o futebol. A escolha quase exclusiva dos trabalhos relacionados com futebol deve-se ao facto de cada desporto ter as suas características e particularidades, e nem todos os estudos se podem adaptar a este ou aquele desporto. Comparando dois desportos coletivos muito populares, como o futebol e o basquetebol, encontra-se diferenças significativas na estrutura dos dois jogos, desde o número de jogadores em campo (22 no futebol contra 10 no basquetebol), duração de um jogo, divisão do tempo jogado, número de substituições, tipo de resultado etc. Todas estas características tornam muito difícil adaptar estudos de um desporto para o outro. No entanto, [26] faz a análise do desempenho da mesma técnica *Data Mining* entre dois desportos, como é o caso do futebol e o *rugby*. De realçar que existem em grande quantidade documentos que implementam técnicas *Data Mining* nos desportos Norte-Americano, como por exemplo o basquetebol estudado por [27].

2.8.1.4 Tarefas de data mining

Os problemas são tipicamente abordados como uma de duas tarefas:

- Classificação

  Quando o atributo alvo identifica as categorias às quais os registos pertencem, é chamado de classe e assume valores discretos. A previsão de vitória, empate ou derrota de uma equipa em que o atributo alvo só tem três classes possíveis é um exemplo de um problema de classificação (ex.[28], [29]).

- Regressão

  O atributo alvo é descritivo por valores numéricos contínuos, como é o caso de [30] em que é previsto o vencedor de um torneio através de probabilidades ou [31] que prevê o resultado de um jogo através dos golos marcados por cada uma das equipas.

2.8.1.5 Tipos de Atributos (Features)

Os fatores que influenciam um jogo de futebol são a questão central com que a maior parte dos investigadores dos trabalhos estudados se deparam. Para além de não serem totalmente consensuais, há muitos fatores que são difíceis de quantificar numericamente. Exemplos dessa dificuldade é o caso da fadiga, a moral da equipa, a reação a um golo marcado ou sofrido etc. Uma simples má decisão de um jogador pode ditar o resultado de um jogo [24]. Os fatores usados podem ser organizados de acordo com os seguintes grupos:

- Histórico de jogos (ex. [31], [26])

  Descreve o desempenho da equipa em todos os jogos realizados até ao momento, seja em casa ou fora.

- Histórico de golos (ex. [28], [21])

  É uma característica que refere o número de golos marcados e sofridos nos jogos anteriores, bem como as respetivas médias.

- Desempenho recente (ex. [24], [29])

  Descreve a "forma" atual através do desempenho da mesma nos últimos jogos. Por norma, são usados os últimos 4,5 ou 6 jogos para analisar a forma da equipa.

- Características físicas da equipa (ex. [22], [3])

  Insere-se aqui o trabalho de equipa, resistência, esforço, agressividade, a força do ataque e defesa.

- Características mentais da equipa (ex. [5], [3])

  É uma característica difícil de caracterizar. Inclui a reação às adversidades, concentração, se é uma equipa ofensiva ou defensiva.

- Organização da equipa (ex. [25], [18])

  Inclui as formações táticas, modelo de jogo, tendências, pontos fortes e fracos das equipas em cada jogo.

- Características dos jogos anteriores (ex. [21])

  Este fator engloba todas as estatísticas dos jogos anteriores, como por exemplo as faltas, cantos, posse de bola, etc.

- Característica do jogo atual (ex. [24], [20])

  Contem a localização do jogo, a reputação das equipas, jogadores do onze inicial, jogadores lesionados ou suspensos, condições climatéricas etc.

- Fator Casa (ex. [30], [27])

  O fator casa é um termo usado para descrever a vantagem que a equipa da casa tem pelo facto de jogar no seu estádio. O fator casa foi estabelecido como um importante fator que influencia o resultado final. A sua existência afeta jogadores, treinadores, árbitros, fãs e até a

comunicação social. As causas para esta vantagem podem ter a ver com o público no estádio, viagens, domínio territorial, fatores psicológicos etc [32]. Nas competições desportivas, a equipa da casa ganha mais de 50% dos jogos em casa. No futebol a percentagem de vitória da equipa da casa é de aproximadamente 64,5% [33].

- Ranking (ex. [23], [34])

  Analisa a classificação atual no campeonato no caso das equipas, e no FIFA *world ranking* no caso das seleções.

2.8.1.6 Torneio

As previsões dos jogos abordados nos trabalhos inserem-se numa de três categorias: Jogos de um torneio com fase de grupo e a fase posterior é a eliminar, como é os casos dos Campeonato do Mundo, Campeonato da Europa e a Liga dos Campeões [30] e [5].

Jogos de um torneio tipo Campeonato, em que as equipas jogam todas contra todas e no fim a equipa que acumular mais pontos ganha. A liga que gera mais interesse nas técnicas de *Data Mining* é a liga inglesa, talvez porque seja a liga mais vista em todo o mundo [28]. Foram estudadas ainda a liga portuguesa, espanhola, italiana, francesa, alemã, finlandesa e ucraniana na Europa e a liga brasileira (Ex. [20] e [35]).

Jogos de uma equipa específica, como é o caso de [24] que analisa a equipa do Tottenham nas épocas 1995/1996 e 1996/1997 e o [3] que analisa e prevê os resultados do FC Barcelona na época 2008/2009 da liga espanhola.

2.8.1.7 Medida de avaliação

A medida de avaliação usada em quase todos os trabalhos que lidam com problemas de classificação é a taxa de acerto (*accuracy*). Em problemas de regressão são usadas outras medidas de avaliação, nomeadamente o RMS (*root mean square error*) e o MSE (*mean square error*). Quando existem problemas de regressão em que os modelos têm como atributo alvo um valor numérico, é difícil dizer avaliar se o valor previsto está correto. Nestes casos, em vez de se verificar se o valor está correto, foca-se na "distância" entre o valor previsto e o valor real. Existem funções que medem o erro entre o valor real e o valor previsto, tais como o MAE (*mean absolute error*), MSE,RMS etc [12]. [5] usa o RMS como forma de avaliar o seu algoritmo, enquanto [23] e [19] usam o MSE.

2.8.1.8 Metodologia de estimação

Poucos documentos explicam o tipo de metodologia de estimação aplicada aos seus trabalhos. Dois documentos fazem referência ao *N-fold cross-validation*. Um deles usa mesmo o *10-fold cross-validation*. O conjunto é dividido em 10 partes e em cada ciclo um dos conjuntos serve de teste e os restantes servem de treino. [34] usa o método *holdout* e *Growing Window* como

metodologia de estimação. Neste trabalho o conjunto de dados de treino é constituído por jogos de 10 épocas (2002/2003 até 2011/2012), e o conjunto de dados de teste contém jogos de 2 épocas (2012/2013 e 2013/2014).

2.8.1.9 Validação

Como forma de validar os seus modelos, os diferentes documentos usam comparação com resultados reais, comparação de desempenho com outros modelos e comparação com as previsões de especialistas (humanos).

Alguns trabalhos validam os seus modelos comparando as suas previsões com os resultados reais dos jogos. [5] e [19] usam os resultados reais como base para a validação dos seus próprios modelos. [3] comparou os seus resultados com os resultados reais dos jogos do Barcelona na época 2008/2009. Outros trabalhos optam por validar os seus modelos comparando os desempenhos com outros modelos já implementados anteriormente (ex. [25], [29]). [28] compara o desempenho do seu algoritmo com os algoritmos implementados por [36]. Também foi verificado que alguns trabalhos comparam o desempenho dos seus modelos com os resultados dos jogos previstos por especialistas em futebol. Por exemplo, [26] compara o seu desempenho com as previsões de seres humanos. As previsões feitas por especialistas também podem ser vistas como um algoritmo de previsão.

2.8.1.10 Técnicas *data mining*

Diferentes técnicas de previsão *Data Mining* foram abordadas nos documentos estudados. Pode-se organizar algumas das técnicas de previsão da seguinte forma:

Modelo baseado em distancia

- Knn (Ex. [22] e [24])

Modelos probabilísticos

- Naive Bayes (Ex. [29] e [27])
- Redes Bayesianas (Ex. [3] e [25])

Modelos baseados em procura

- Árvores de decisão (Ex. [20] e [29])
- Regras de decisão (Ex. [5])

Modelos baseados em otimização

- Redes Neuronais (Ex. [23] e [21])

- SVM (Ex. [34])

Além do estudo destas técnicas de *Data Mining* para previsão de resultados de jogos de futebol, um assunto que está muito em voga hoje em dia, e que vai crescendo de dia para dia é as apostas desportivas. Também existem alguns trabalhos que abordam a utilização das apostas desportivas como base de previsões de resultados, nomeadamente [37] e [38].

2.8.2 Resultados

A maioria dos trabalhos estudados optaram por utilizar tarefas de classificação para a resolução dos seus problemas de previsão. Também a grande maioria dos trabalhos têm o futebol como desporto alvo, porque cada desporto é único e tem as suas características próprias. O tipo de competição estudada recaiu maioritariamente sobre grandes torneios mundiais, como é o caso do Campeonato do Mundo, Campeonato da Europa e a Liga dos Campeões e também sobre as principais ligas europeias como a portuguesa, espanhola, inglesa, italiana, alemã e francesa. A medida de avaliação mais utilizada foi a taxa de acerto, embora existam alguns trabalhos que usam os erros MAE,MSE e RMS. Na escolha da metodologia de estimação de desempenho houve uma distinção entre duas soluções escolhidas pelos trabalhos. Uma das soluções escolhidas foi o uso da técnica *N-fold cross validation* que não tem em conta o fator tempo. A outra solução recaiu sobre a valorização do tempo. Os trabalhos que valorizaram o fator tempo usaram conjunto de dados de treino que temporalmente são anteriores ao conjunto de dados de teste. O método de validação mais frequente foi a comparação do desempenho dos seus modelos com modelos desenvolvidos anteriormente.

Capítulo 3

Perceção e Preparação dos Dados

3.1 Introdução

Os dados disponíveis para a realização do estudo empírico foram disponibilizados pelo Laboratório SAPO/ U.Porto através de um ficheiro csv com aproximadamente setenta e cinco mil registos. Cada registo deste conjunto de dados informa a ocorrência de um evento no jogo de futebol. Os eventos podem ser do tipo golo, autogolo, substituição, cartão amarelo, duplo cartão amarelo e cartão vermelho. Neste conjunto de dados estão contemplados todos os jogos referentes a cinco épocas de futebol do campeonato português e cinco épocas do campeonato inglês. As épocas referidas são 2009/2010, 2010/2011, 2011/2012, 2012/2013 e 2013/2014. Cada evento registado contém a informação do jogo, evento, a equipa e do jogador(es) responsáveis pela ocorrência do evento.

Os dados disponibilizados sobre o jogo são:

- CompetitionID – atributo numérico que tem como objetivo identificar a competição
- FullName - atributo nominal, que contem o nome da competição.
- Match_MatchID - atributo numérico que identifica o jogo.
- Match_FixtureDate – atributo que contem a data do jogo.
- Stadium_Stadium - atributo numérico que identifica o estádio.
- Stadium_Name - atributo nominal, que contem o nome do estádio.
- HomeTeam_TeamID - atributo numérico que identifica a equipa visitada.
- HomeTeam_Name - atributo nominal que identifica a equipa visitada.
- AwayTeam_TeamID – atributo numérico que identifica a equipa visitante.
- AwayTeam_Name – atributo nominal que identifica a equipa visitante pelo nome.

Os dados disponibilizados sobre o evento são:

- Occurrence_OccurrenceID – atributo numérico que identifica o evento ocorrido.
- Occurrence_OccurrenceTypeID – atributo numérico que classifica o tipo de evento.
- Occurrence_Minute - atributo numérico que dá a informação do minuto em que ocorreu o evento .
- Occurrence_SourceTeamID – atributo numérico que identifica a equipa que deu origem ao evento.
- Occurrence_SourceTeamName - atributo nominal que identifica a equipa que deu origem ao evento pelo nome.
- Occurrence_TargetTeamID - atributo numérico que identifica a equipa que foi alvo do evento.
- Occurrence_TargetTeamName - atributo nominal que identifica a equipa que foi alvo do evento pelo nome.

3.2 Pré-processamento

A etapa de pré-processamento dos dados é muito importante pois tem como funções a receção, preparação e transformação dos dados. A figura 3.1 mostra o ciclo das diferentes etapas de pré-processamento.

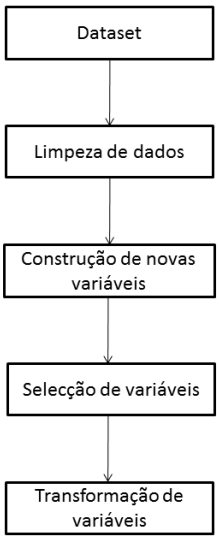

Figura 3.1: Sequência das etapas de pré-processamento

3.2.1 Preparação dos dados

O conjunto de dados disponibilizado foi analisado com o intuito de fazer uma limpeza aos dados. A limpeza aos dados passa por encontrar e eliminar dados inconsistentes, redundantes, duplicados etc. A análise foi feita com base numa única época para reduzir a quantidade de dados e facilitar a sua compreensão.

Duas das características do conjunto de dados são a equipa que deu origem ao evento e a equipa que foi alvo do evento. Estas duas características são usadas unicamente quando o evento é uma substituição. Neste caso as duas características têm o mesmo valor porque a equipa que faz a substituição dos seus jogadores é a mesma. Trata-se de duas características redundantes em que uma delas não acrescenta conhecimento nenhum ao conjunto de dados. Sendo assim, uma das características foi eliminada.

Para a deteção de eventos duplicados, foram comparados o número de ocorrências de determinado evento no conjunto de dados com a realidade. Por exemplo, foram comparados o número total de golos de uma determinada época com o número real de golos dessa época, e verificou-se que o número de golos registados pelo conjunto de dados era superior à realidade, logo existem dados duplicados. Existem dois tipos de dados duplicados. Os dados duplicados que são fáceis de detetar por terem a mesma identificação e facilmente são eliminados. Por outro lado, existem os dados duplicados que tendo identificações diferentes são o mesmo evento, o que os torna de difícil deteção e eliminação. Sendo assim, não se procedeu à eliminação de todos os dados duplicados, tendo porem a noção que os dados não estão totalmente de acordo com a realidade.

Foram detetados ainda alguns campos sem valores (*NULL*). Este facto não teve um impacto muito grande porque esses campos não foram detetados como sendo campos com grande significado no contexto do problema.

3.2.2 Engenharia de Variáveis

Esta etapa consiste em criar novas variáveis a partir de variáveis existentes. Esta operação permite que as novas variáveis demonstrem relacionamentos entre as variáveis existentes.

3.2.2.1 1ª Iteração

O conjunto de dados disponibilizado não é adequado ao problema em questão, uma vez que, o objetivo deste estudo é a previsão de resultados de jogos de futebol e não a previsão da ocorrência de eventos.

Numa primeira fase optou-se por usar unicamente os eventos que influenciam diretamente o resultado de um jogo, ou seja, os golos, deixando de parte os restantes eventos como substituições ou cartões. Posto isto, com base nos eventos "golo" foi criada uma tabela nova que contempla todos os resultados dos jogos de futebol nas épocas mencionadas em cima. A tabela de resultados contêm como campos a competição em questão, a identificação do jogo, a identificação da equipa visitada e visitante, o número de golos marcados por cada uma das equipas, a data e a jornada. A tabela de resultados serviu de base para a criação de novas variáveis na primeira iteração.

Embora dispondo de todos os jogos de dez campeonatos diferentes, numa primeira análise foi decidido trabalhar unicamente com um campeonato para que fosse mais fácil a compreensão e análise dos dados. O campeonato português 2012/2013 foi o escolhido. Esta escolha deveu-se ao facto de se dispor de uma época posterior (2013/2014) para avaliação do modelo escolhido, bem como pelo facto de se dispor de dados de três épocas anteriores.

O conjunto de dados construído para a primeira iteração tem como registos (linhas) os jogos que se pretende analisar (campeonato português 2012/2013) e como campos (colunas) algumas características relativas a cada jogo. Essas características podem ser divididas em quatro tipos de classes. As características relativas ao jogo em questão, à equipa visitada, à equipa visitante e ao histórico de confronto direto entre as equipas. As características do jogo definem as equipas que se defrontam, a jornada em questão, a data da partida e a diferença de pontos entre as equipas à data do jogo. As características que descrevem cada uma das equipas definem a "força" do ataque e da defesa, apetência para ganhar jogos e a sua "forma" à data do jogo.

A "força" do ataque da equipa visitada é caracterizado através da média de golos marcados nos jogos que disputou em casa, bem como a média de golos marcados em todos os jogos anteriores (casa e fora). A "força" do ataque da equipa visitante é caracterizado através da média de golos marcados nos jogos que disputou fora de casa, bem como a média de golos marcados em todos os jogos anteriores (casa e fora). A "força" da defesa de cada uma das equipas é caracterizada de forma semelhante à "força" do ataque substituindo os golos marcados por golos sofridos. A apetência para ganhar jogos é caracterizada através da percentagem de vitórias de cada uma das equipas, tanto em casa como fora de casa.

A "forma" de uma equipa pode variar em poucos jogos, por isso, é caracterizada pela performance da equipa no último jogo, nos últimos três jogos e nos últimos cinco jogos. Esta distinção do espaço temporal pode permitir analisar, entre outras coisas, se a equipa está num ciclo positivo de resultados, se inverteu um ciclo negativo de resultados etc.

O confronto direto entre as equipas nos anos anteriores também pode ser um fator importante que influencia um jogo de futebol, assim como o fator casa. Neste sentido, o histórico de confrontos entre as equipas são caracterizados pela percentagem de vitórias da equipa visitada e pela percentagem de empates que se verificaram nas três épocas anteriores, tanto no total de jogos, como somente em jogos no estádio em que se realiza o jogo.

3.2.2.2 2ª Iteração

Após os resultados da primeira iteração, é necessário voltar à preparação de novos dados para que se possa obter melhores resultados. O objetivo da segunda iteração é pensar novamente no problema e criar novas variáveis. Essas variáveis surgiram de novos dados com mais informação e têm como objetivo complementar as variáveis da primeira iteração.

Os novos dados disponibilizados pelo Laboratório SAPO/ U.Porto contêm aproximadamente quatro mil e quinhentos registos, em que cada registo é um jogo. Os novos dados contêm registos de novas épocas do campeonato português e sete épocas do campeonato inglês. As épocas do campeonato português são entre 2006/2007 até 2014/2015, e as épocas do campeonato inglês são

entre 2008/2009 até 2014/2015. Os novos dados forneceram novas informações sobre os jogos, entre elas, a posse de bola de cada uma das equipas, cantos, faltas, remates à baliza e remates para fora.

Com os novos dados estão disponíveis não só mais informação sobre cada jogo como também mais jogos de épocas que não estavam disponíveis no conjunto de dados anterior, caso das épocas 2006/2007 até 2008/20019 do campeonato português. No entanto, esta segunda iteração foi pensada como complemento da primeira e optou-se por continuar o estudo com as mesmas épocas da iteração anterior.

O conjunto de dados construído para a segunda iteração tem como base o conjunto de dados usado na primeira iteração com o acrescento de novas características que entretanto foram criadas. As novas características são a posição em que terminou o ano anterior a equipa visitada e visitante, posição atual da equipa na presente época, performance do ataque e defesa, comportamento relativo das equipas nos jogos anteriores, posse de bola média e remates à baliza.

A posição atual da equipa no presente campeonato pode estar contida num de três grupos. O primeiro grupo contém as equipas da parte de cima da tabela que se situam entre o primeiro e a quinta posição. O segundo grupo contém as equipas que se situam entre a sexta e a décima primeira posição, e por fim, o terceiro grupo contém as equipas que se situam entre a décima segunda e a décima sexta posição. Este escalonamento permite agrupar as equipas que lutam pelos mesmos objetivos.

A performance do ataque e da defesa é um complemento às características da primeira iteração que continha a "força" do ataque e da defesa. Este complemento é dado pela quantidade de jogos sem sofrer e marcar golos de cada uma das equipas. A quantidade de jogos sem sofrer golos reforça a qualidade da defesa, e a quantidade de jogos sem marcar reforça o défice do ataque.

O comportamento relativo das equipas nos jogos anteriores inclui nas mesmas características a apetência das equipas para ganhar, empatar e perder jogos. A primeira característica revela quanto mais forte é a equipa visitada, fazendo a relação de jogos ganhos pela equipa visitada com os jogos perdidos pela equipa visitante. A segunda característica faz a relação de empates entre as duas equipas. Estas características são interessantes porque relacionam diretamente uma equipa com a outra, dando uma noção de superioridade ou inferioridade.

Com todas as novas características incluídas, o conjunto de dados da segunda iteração tem mais 12 campos que o da primeira iteração.

3.2.3 Seleção de variáveis

Depois de construir o conjunto de dados com todas as características necessárias, é essencial fazer um pré-processamento dos dados para perceber se há características correlacionadas, com um único valor etc. O pré-processamento pode ser feito através de métodos de filtro ou *wrapper*. A opção escolhida recaiu sobre os métodos de filtro uma vez que os métodos de *wrapper* podem provocar *overfitting* devido à quantidade reduzida de registos. O método de filtro utilizado foi a identificação de características correlacionadas. Com este método é possível eliminar as características que evoluem de forma semelhante e que não acrescentam informação nenhuma ao

conjunto de dados. Posto isto foi criado mais três conjuntos de dados. O primeiro contém todas as características com correlação inferior a “0,9”, o segundo com correlação inferior a “0,8” e o terceiro com correlação inferior a “0,7”. Cada iteração passou a dispor de quatro conjuntos de dados que integraram a fase de modelação. A tabela 3.1 mostra o número de características de cada conjunto de dados usado na modelação.

Tabela 3.1: Número de atributos em cada conjunto de dados da 1º e 2º iteração

Conjunto de dados	1ª Iteração	2ª Iteração
Completo	51	63
Correlação inferior a "0,9"	38	48
Correlação inferior a "0,8"	24	35
Correlação inferior a "0,7"	20	31

3.2.4 Transformação de variáveis

Alguns algoritmos não lidam bem com variáveis nominais. Sendo assim, essas variáveis têm que ser transformadas em variáveis numéricas. A “forma” atual da equipa é caracterizada, entre outras, por uma variável nominal que dita a performance da equipa nos últimos cinco jogos. Por exemplo, se a equipa ganhou o ultimo jogo, empatou o penúltimo e perdeu os restantes, a variável “ forma” seria “L L L D W”. Foram criadas duas variáveis numéricas que substituíssem esta variável nominal. As variáveis criadas foram o número de vitórias e o número de pontos alcançados nesses cinco jogos. Seguindo o exemplo acima, o número de vitórias seria 1 e o número de pontos seria 4. A transformação também foi feita para as restantes variáveis que ditam a “forma” da equipa.

3.3 Análise exploratória de dados

A EDA (*exploratory data analysis*) é uma abordagem para análise de dados através maioritariamente de gráficos que permite a perceção dos dados, deteção de *outliers*, extração variáveis importantes etc. Esta abordagem foi realizada sobre o conjunto de dados construído na primeira e segunda iteração. Nos gráficos quando aparece “W” significa vitória da equipa da casa, “D” significa empate e “L” significa derrota da equipa da casa. As características que envolvem os golos marcados e sofridos foram as primeiras a serem analisadas. Neste ponto optou-se por analisar os dados dos dois extremos da tabela classificativa. Sendo assim, comparou-se os golos marcados pelo FCPorto (1º classificado) e Beira-Mar (16º classificado).

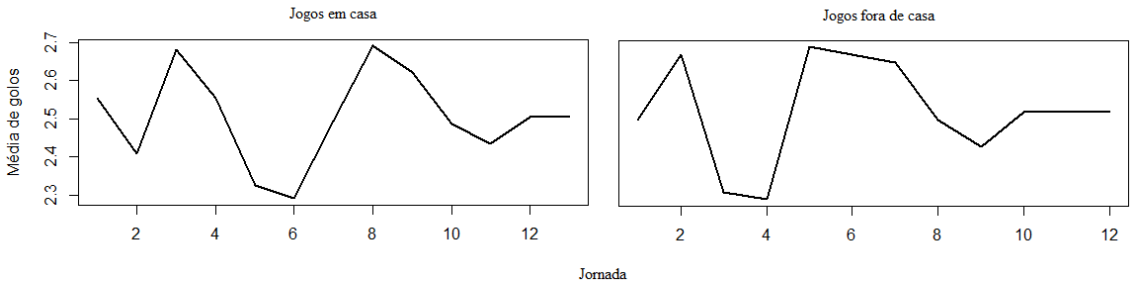

Figura 3.2: Média de golos marcados pelo F.C.Porto

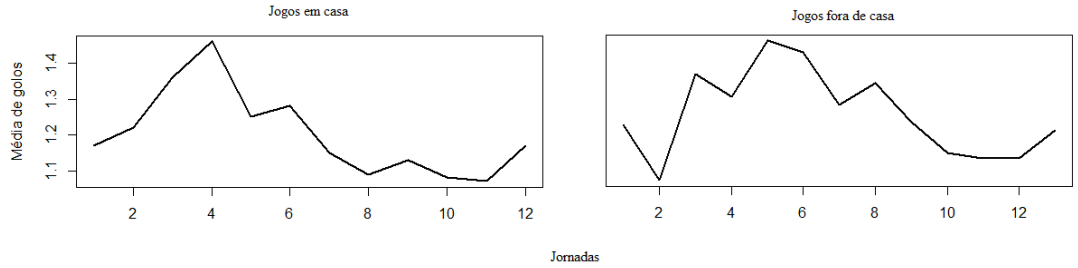

Figura 3.3: Média de golos marcados pelo Beira-Mar)

Os gráficos da figura 3.2 e 3.3 representam os golos marcados pelo FCPorto e o Beira-Mar respetivamente. O gráfico da direita diz respeito aos jogos realizados na condição de visitado e os gráficos da esquerda dizem respeitos aos jogos disputados como visitante. Como se pode ver na figura 1, o FCPorto teve uma quebra de golos marcados a meio da época tanto nos jogos em casa como fora. O facto de a quebra se dar em jogos realizados no mês de Dezembro, podem explicar que neste mês das festividades os jogadores baixam os seus níveis de concentração. Analisando os jogos do Beira-Mar, pode-se verificar que o número de golos marcados baixou na segunda metade da época. Esta quebra pode ser explicada pela saída de jogadores importantes no marcado de transferências de Janeiro ou pela maior competitividade do campeonato na segunda volta. Comparando a média de golos marcados das duas equipas facilmente se vê a grande diferença que existe, chegando por vezes a ser mais de um golo por jogo. A segunda análise foi feita considerando a performance da equipa nos últimos jogos 5 jogos que realizou. Esta análise é feita com base no número de pontos que a equipa conseguiu obter nos últimos 5 jogos. Neste caso, optou-se por fazer a análise com o Benfica (2º classificado) e o Olhanense (14º classificado). A figura 3.4 representa os pontos alcançados pelo Benfica e a figura 3.5 os pontos alcançados pelo Olhanense.

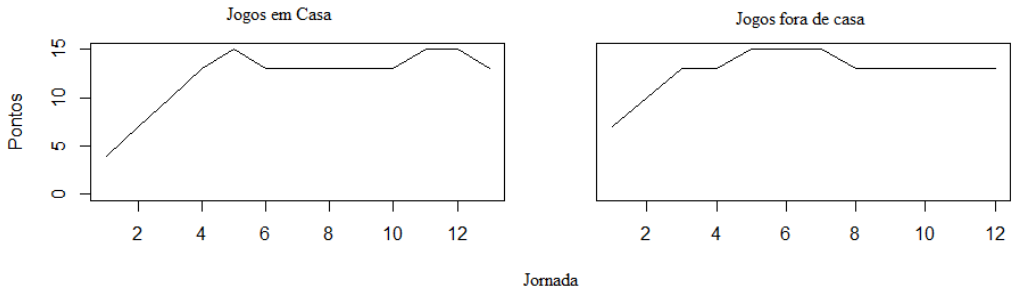

Figura 3.4: Pontos alcançados pelo Benfica nos 5 jogos anteriores
[**?**]

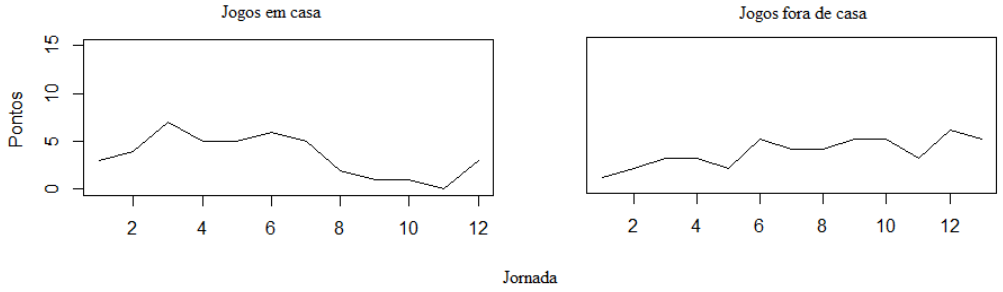

Figura 3.5: Pontos alcançados pela Olhanense nos 5 jogos anteriores

Como seria de esperar há uma grande diferença de pontos alcançados nos últimos 5 jogos pelas duas equipas.

A figura 3.4 mostra que o percurso do Benfica foi muito regular ao longo da época tanto em casa como fora. No entanto, numa observação mais pormenorizada é possível observar que os picos de forma nos jogos em casa e fora aconteceram em períodos diferentes. No percurso em casa o Benfica atingiu o pico de forma no inicio e no fim da época. Em contrapartida, no percurso fora de casa o Benfica atingiu o seu pico a meio da época. Também é possível observar que a taxa de vitórias foi sempre muito alta ao longo da época. Esta observação pode corroborar uma velha máxima do futebol que diz que uma equipa que tem vindo a ganhar tem mais probabilidade de ganhar o próximo jogo.

A figura 3.5 revela que o percurso da Olhanense nos jogos em casa e fora foi totalmente oposto. Consegue-se verificar que a Olhanense realizou um início de campeonato bastante promissor em casa mas foi decaindo à medida que a época se aproximava do fim. Nos jogos fora de casa verificou-se a situação contrária, a Olhanense obteve melhor desempenho nos últimos jogos do campeonato.

A terceira análise aborda duas propriedades importantes num jogo de futebol, o confronto direto e o fator "casa". Esta análise difere das anteriores porque não há uma alteração dos dados

durante a época uma vez que se refere a dados de épocas anteriores. Esta análise aborda a percentagem de vitórias no seu estádio que uma equipa tem quando defronta o seu adversário atual. Foram usadas uma equipa que é forte em casa (FCPorto) e uma equipa que é menos forte (V.Setubal).

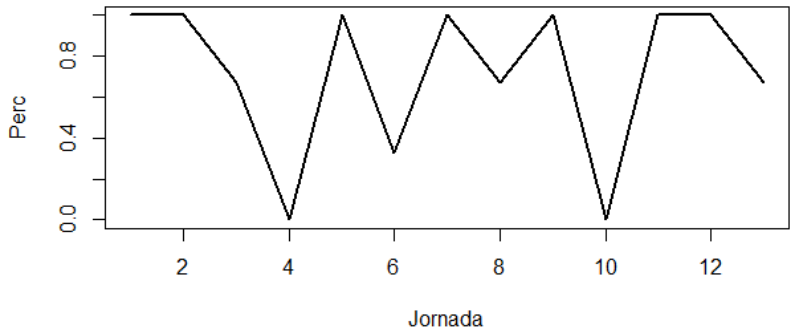

Figura 3.6: Percentagem de vitórias do F.C.Porto no confronto direto

A figura 3.6 mostra a percentagem de vitórias no confronto direto do FCPorto em casa. Analisando o gráfico verificasse que o FCPorto é realmente forte em casa, mas vê-se que na jornada 4 e 10 a percentagem de vitorias é zero o que causou desconfiança dado o seu historial em casa. Após uma análise mais profunda conclui-se que aqueles jogos são realizados contra equipa recém-promovidas à primeira liga, logo não existe histórico de confrontos. Esta análise permitiu perceber que se uma das características do conjunto de dados é o confronto direto, então tem que se retirar os jogos que incluem equipas recém-promovidas.

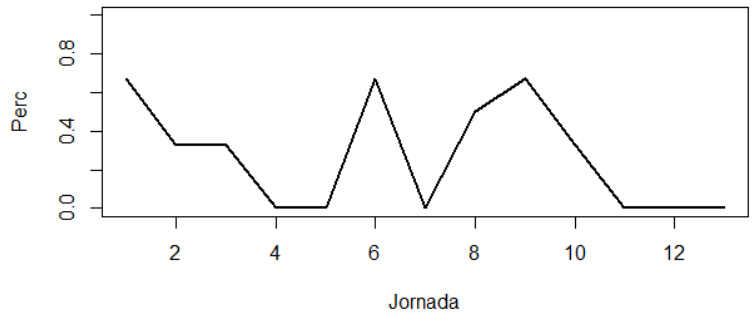

Figura 3.7: Percentagem de vitórias do V.Setúbal no confronto direto

A figura 3.7 mostra a percentagem de vitórias no confronto direto do V.Setúbal em casa. O gráfico da figura mostra que a percentagem de vitórias do V.Setúbal é realmente baixa, sendo maioritariamente abaixo de 50%. As análises descritas anteriormente foram realizadas considerando-se as características e equipas selecionadas. No entanto, é necessário fazer uma análise em que se

relaciona certas características com o resultado final do jogo. A figura 3.8 relaciona a diferença de pontos entre a equipa visitada e visitante com o resultado final.

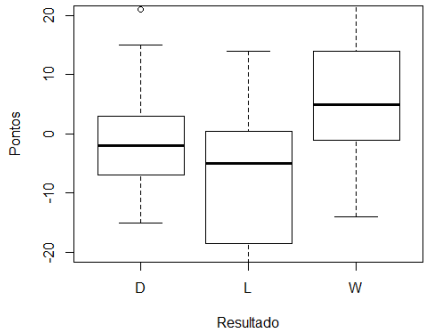

Figura 3.8: Diferença de pontos entre as equipa consoante o resultado final

Na figura 3.8 pode-se verificar que os jogos que acabaram em vitórias da equipa da casa têm, normalmente uma diferença pontual positiva, o que significa que a equipa da casa tem mais pontos que a equipa visitante. Pode-se verificar ainda que a maioria dos jogos que acabam empatados tem uma diferença pontual à volta de zero. Por outro lado, os jogos que acabam em derrota da equipa da casa, na sua maioria tem uma diferença pontual negativa. Neste caso a equipa visitante tem mais pontos e está melhor classificada que a equipa da casa.

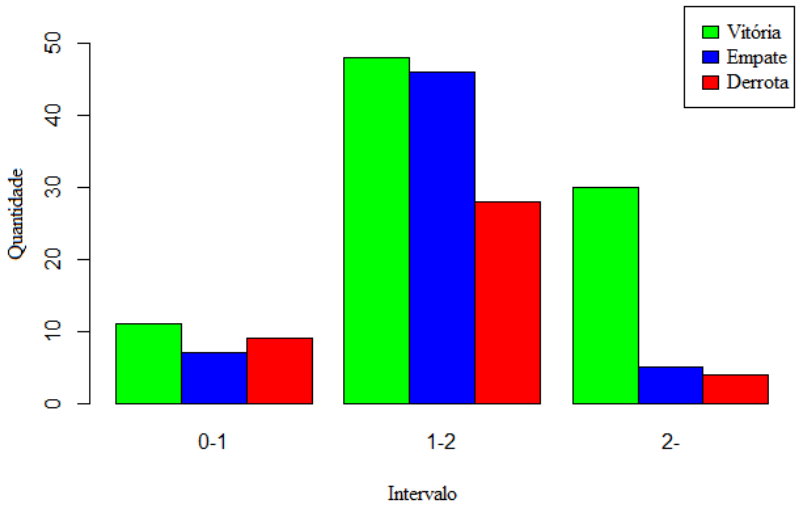

Figura 3.9: Média de golos marcados pela equipa visitada e o resultado final

Observando o gráfico da figura 3.9, retira-se que quando a equipa da casa tem uma média de golos marcados superior a dois o resultado que acontece mais frequentemente é a vitória da equipa da casa. Neste caso a vitória da equipa da casa surge em aproximadamente trinta jogos, enquanto o empate e a derrota surgem abaixo da dezena de jogos. Quando a média de golos marcados

pela equipa da casa é inferior a dois golos existe um equilíbrio entre vitória, empate e derrota no resultado final.

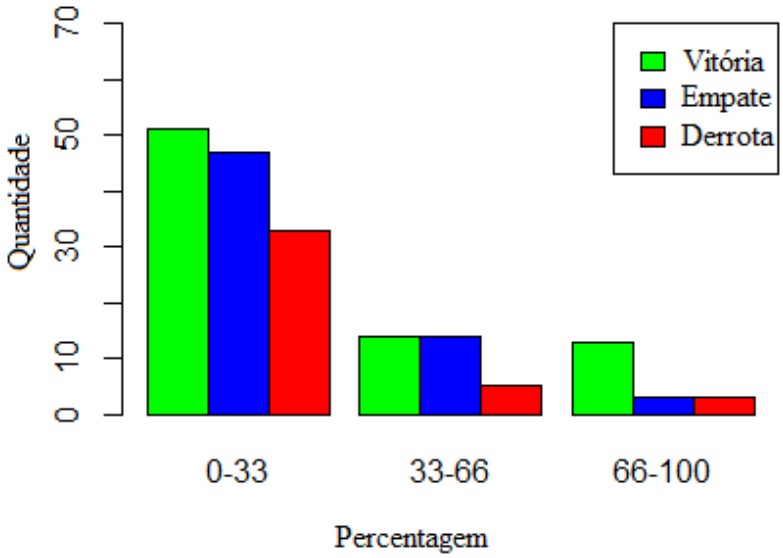

Figura 3.10: Gráfico que relaciona a percentagem de empates no confronto direto entre as duas equipas nos anos anteriores e o resultado final do jogo

Poder-se-ia pensar que quanto maior a percentagem de empates no confronto direto maior seria número de jogos que terminam empatados. No entanto, isso não acontece. Como se pode ver na figura 3.10, se a percentagem de empates no confronto direto estiver entre 33% e 66% a quantidade de jogos que acabam empatados é a mesma dos jogos que acabam em vitórias da equipa da casa. O surpreendente é que quando a percentagem de empates no confronto direto está entre 66% e 100%, o número de jogos que terminam com um empate é inferior ao número de jogos que terminam com a vitória da equipa da casa. Esta diferença pode dever-se ao número reduzido de amostras que se tem do confronto direto. Quando a percentagem de empates no confronto direto está entre 33% e 66%, esses jogos na sua maioria têm 5 ou 6 jogos de confronto direto. Ao invés, os jogos cuja percentagem de empates está entre 66% e 100% têm na sua maioria 1 ou 3 jogos de confronto direto. Esta amostra reduzida pode explicar a elevada percentagem de empates que não se reflete no resultado final.

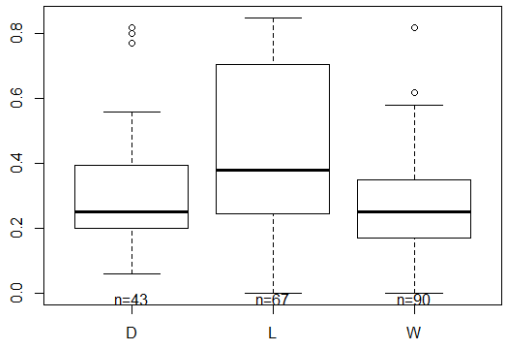

Figura 3.11: Gráfico boxplot que ilustra a percentagem de vitórias da equipa visitante em relação ao resultado final do jogo

Na figura 3.11 pode-se ver que quando a percentagem de vitórias da equipa visitante é alta o resultado é normalmente "L" a vitória dessa mesma equipa. No entanto, a diferença entre os *boxplots* de "W" e "D" não são muito acentuados. Este fenómeno pode ser explicado pelo desequilíbrio do futebol português. Os dois primeiros classificados dominaram o campeonato, e nos jogos fora de casa ganharam a maioria o que fez com que o *boxplot* "L" ficasse desequilibrado. Nos restantes jogos, verificasse que na maioria dos jogos que terminam com empate ou vitória da equipa visitada, a equipa visitante tem uma percentagem de vitória baixa. Também é possível identificar *outliers* nos *boxplots* "W" e "D".

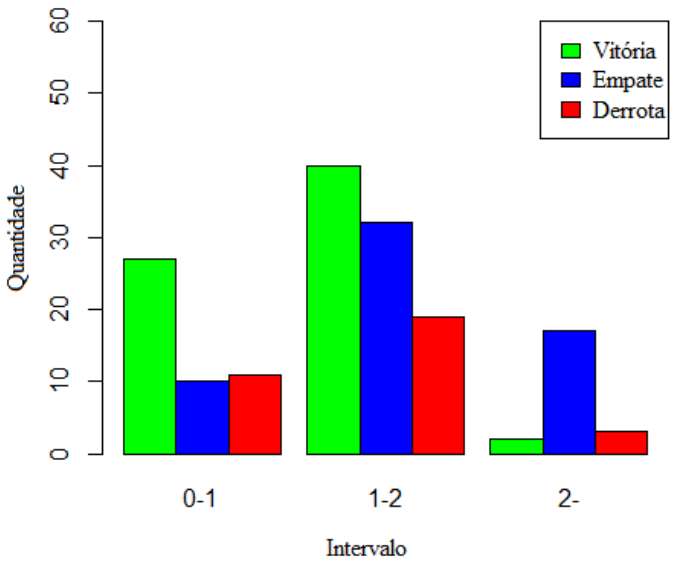

Figura 3.12: Média de golos marcados pela equipa visitante nos jogos fora em relação ao resultado final

O que é surpreendente é que quanto melhor média de golos marcados a equipa visitante tem, maior é o número de empates no final da partida como se pode ver na figura3.12. Era esperado

que o resultado predominante neste caso fosse a vitória dessa mesma equipa. Este facto pode ser explicado pelo fator casa que equilibra os jogos.

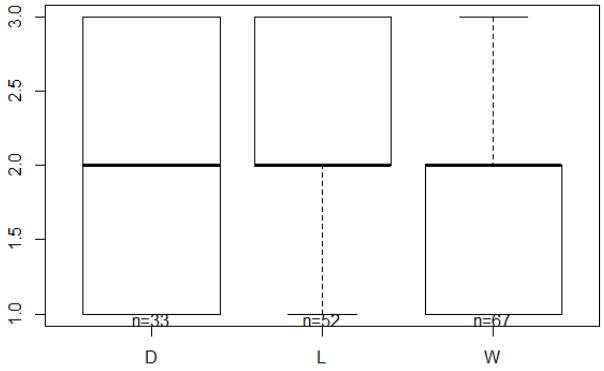

Figura 3.13: Gráfico que relaciona a posição da equipa da casa no campeonato e o resultado final

Foi explicado anteriormente que de acordo com a sua posição na tabela classificativa, a equipa integra um de três grupos. Como se pode ver na figura 3.13 há um grande número de jogos que termina com a vitória da equipa da casa e essa mesma equipa pertence ao grupo um (primeiros 5 classificados) ou dois (entre sexto e décimo classificado). Constatasse ainda que nas derrotas sofridas pela equipa da casa, na maioria dos jogos essa mesma equipa pertencia ao grupo dois ou três (últimos 5 classificados).

Capítulo 4

Modelação

O capítulo de modelação está dividido em três secções: metodologia, resultados e discussão geral. Neste capitulo o termo inglês "*dataset*" terá o mesmo significado que o termo "conjunto de dados" anteriormente usado.

4.1 Metodologia

4.1.1 Dataset

Na etapa de modelação foram usados quatro *datasets* de comparação que diferem apenas nas variáveis que os compõem. O *dataset* completo contém todas as variáveis criadas na etapa de preparação dos dados. Os restantes três *datasets* são o resultado da seleção de variáveis através da procura de correlação entre elas. Os três *datasets* são compostos pelas variáveis com correlação inferior a "0,9","0,8" e "0,7".

Como foi explicado anteriormente, os *datasets* integrados na etapa de modelação incluem os jogos referentes ao campeonato português 2012/2013. Nesse campeonato foram disputados 240 jogos distribuídos por 30 jornadas.

Cada jogo dispõe de informação relativa ao próprio jogo, à equipa visitada, à equipa visitante e ao confronto direto. A grande parte das variáveis é relativa às duas equipas, e têm como origem os jogos que essas equipas realizaram nas jornadas anteriores. Foram excluídos os jogos das 5 primeiras jornadas porque contém informação sustentada em poucos jogos. Sendo assim, os *datasets* passaram a ser compostos por 200 jogos.

Na etapa de análise de dados foi concluído que os atributos do confronto direto estavam incompletos. As equipas que estão pela primeira vez na primeira liga não tem qualquer histórico de jogos anteriores com as restantes equipas, logo, não existem dados de confronto direto e esses campos estão vazios. Para contornar esta adversidade decidiu-se retirar 48 jogos que incluíam equipas recém-promovidas e assim o *dataset* passou a ter 152 jogos.

Sendo assim, na fase de modelação estão dispostos quatro *datasets* com 152 jogos cada um.

4.1.2 Algoritmos

O modelo escolhido para a previsão de resultados de jogos de futebol tem que ter por base um algoritmo de *Data mining*. C5.0, random forest, Knn, Jrip, SVM com *kernel* linear, SVM com *kernel* gaussiano, naive bayes e redes neuronais foram os algoritmos escolhidos para a seleção do melhor modelo.

4.1.2.1 C5.0

C5.0 é um classificador construído com base no algoritmo C4.5. O algoritmo C5.0 tem melhor desempenho que o C4.5 em diversas características, entre elas, a velocidade de processamento, uso eficiente da memória etc. Este algoritmo tem 2 versões, uma versão baseada em árvores de decisão, e tem outra versão baseada em regras. Uma árvore de decisão usa a estratégia de dividir para conquistar para resolver um problema de decisão. Um problema complexo é dividido em problemas mais simples, aos quais é aplicada recursivamente a mesma estratégia [2].

4.1.2.2 JRip

JRip ou RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) foi proposto por William W. Cohen. É um classificador baseado em regras. As regras de decisão são baseadas na forma se A então B em que A é um conjunto de condições. As regras de decisão e as árvores de decisão são bastante idênticas na forma de representar generalizações dos objetos. Consequentemente, ambas definem superfícies de decisão semelhantes [2].

4.1.2.3 Random forest

O random forest é um algoritmo de combinação de outros algoritmos (*ensemble*), que combina uma grande quantidade de árvores de decisão independentes. As árvores de decisão são normalmente geradas através do método *bagging*. Este método permite que as árvores de decisão sejam criadas a partir de diferentes subconjuntos de dados de treino aleatoriamente gerados com reposição. A classificação dos dados é feita por votação, cada árvore de decisão vota numa classe e a classe com mais votos será a resposta dada pelo algoritmo [39].

4.1.2.4 Knn

O Knn é um algoritmo de reconhecimento de padrões, que tem variações definidas pelo número de vizinhos considerados. Cada objeto representa um ponto no espaço de entrada de acordo com os seus atributos. Para a classificação de um objeto cuja classe é desconhecida, é calculada a distância entre o novo objeto e os objetos classificados anteriormente. A distância entre os objetos é calculada através de uma métrica. Normalmente, a distância euclidiana é a métrica usada por este algoritmo. Os k objetos de treino mais próximos do novo objeto votam numa classe e a classe mais votada é atribuída a esse objeto [2].

4.1.2.5 SVM

As SVM são baseadas na teoria de aprendizagem estatística desenvolvida por Vapnik (1995). As principais características das SVM são a boa capacidade de generalização, robustez com grandes dimensões de dados, convexidade da função objetivo e base teórica bem estabelecida dentro da matemática e estatística [40].

- *kernel* linear

  Separação de objetos pertencentes às classes através de fronteiras lineares. São eficazes na classificação de conjuntos de dados linearmente separáveis ou que possuam uma distribuição aproximadamente linear [2].

- *kernel* gaussiano(Não-linear)

  Há alguns casos em que não é possível dividir satisfatoriamente os dados por um hiperplano. As SVMs não lineares lidam com este problema mapeando o conjunto de treino do seu espaço original, para um novo espaço de maior dimensão. Nesta nova dimensão o conjunto de treino já pode ser separado por uma SVM Linear [2].

4.1.2.6 Naive Bayes

O classificador naive bayes pertence à família dos classificadores probabilísticos e é baseado no teorema de Bayes. O teorema de Bayes permite calcular a probabilidade de um evento à posteriori utilizando as probabilidades à priori. Este classificador assume que os valores dos atributos de um objeto são independentes entre si. Este facto ajuda a combater os problemas derivados da maldição de dimensionalidade que são recorrentes em muitos algoritmos, ou seja, este classificador lida bem com grandes quantidades de dados.

4.1.2.7 Redes Neuronais

O desenvolvimento das redes neuronais artificiais tem como inspiração a estrutura e funcionamento do sistema nervoso. As redes neuronais são sistemas computacionais distribuídos compostos de unidades de processamento simples, densamente interconectadas. Estas unidades, conhecidas como neurónios artificiais, computam funções matemáticas. As unidades estão dispostas em varias camadas e interligadas por conexões que possuem pesos associados que ponderam a entrada recebida por cada neurónio na rede. Os pesos têm os seus valores ajustados num processo de aprendizagem e codificam o conhecimento adquirido pela rede [2].

4.1.3 Avaliação

4.1.3.1 Divisão dos dados

Dispondo de quatro *datasets* com 152 jogos do campeonato português 2012/2013, é necessário iniciar uma nova etapa de modelação que será o treino e avaliação dos algoritmos. Foi determinado

que o algoritmo iria analisar os jogos jornada a jornada, formando assim grupos de 8 jogos. Uma vez que o objetivo é prever os resultados dos jogos de uma jornada de cada vez, os dados foram divididos em dados de treino e dados de teste. A divisão em dados de treino e teste pode ser feita de diversas maneiras, como por exemplo a divisão de dados baseada no resultado do jogo. Esta divisão assegurava que as três classes (vitória, empate e derrota) estivessem balanceadas nos dados de treino. No entanto, o fator tempo é muito importante na resolução deste problema de previsão. Com a divisão dos dados baseada no resultado do jogo corre-se o risco de tentar prever um acontecimento em que os dados de treino são dados posteriores aos dados de teste. Por esta razão, o tipo de divisão escolhida para este problema de previsão foi baseada em serie temporais em que os dados de teste são sempre posteriores aos dados de treino.

Os dados de treino incluem todos os jogos desde a jornada 6 até à jornada anterior que se pretende prever o resultado. Os dados de teste incluem todos os jogos da jornada de previsão.

Figura 4.1: Divisão dos dados de treino e teste

A figura 4.1 ilustra um exemplo da previsão de resultados dos jogos da jornada 10. Neste caso, os dados de treino seriam os jogos das jornadas 6,7,8 e 9, enquanto os dados de teste seriam os jogos da jornada 10.

No *package Caret* existe a função *createTimeSlices* que faz a divisão de dados de treino e teste baseada em series temporais. Contudo, esta função só permite que a cada iteração de treino/teste se avance um único objeto (1 jogo) e, tal como foi definido anteriormente, o objetivo deste problema é que se avance uma jornada (8 jogos/objetos). Inicialmente pensou-se em alterar a função *createTimeSlices* para que pudesse avançar 8 objeto em vez de um. No entanto, devido à eliminação de determinados jogos, a maioria das jornadas passam a ter 6 jogos, mas quando as equipas recém-promovidas jogam uma contra a outra, a jornada passa a ter 7 jogos de análise. Visto isto, e perante a possibilidade de cada jornada não ter um número fixo de jogos (tanto pode ter 6 ou 7 jogos) foi impossível alterar a função *createTimeSlices* do *Caret*. Assim, optou-se por criar uma função de raiz que definisse quais os objetos de treino e teste consoante a jornada de previsão.

4.1.3.2 Geração de modelos

O *dataset* contém dados de 24 jornadas. Contudo, só se pode fazer previsão de 23 jornadas porque a primeira jornada do *dataset* (jornada 6) não contém jornadas anteriores que possam servir de treino. Sendo assim, para que se possam prever os resultados de todos os jogos do *dataset*

é necessário realizar 23 iterações. Cada iteração prevê os resultados de uma jornada diferente. Optou-se por analisar duas estratégias de evolução dos dados, *Growing Window* e *Sliding Window*.

Growing Window define a jornada 6 como inicial para todas as iterações. Nesta estratégia, o *dataset* de treino vai sempre crescendo até a previsão da última jornada.

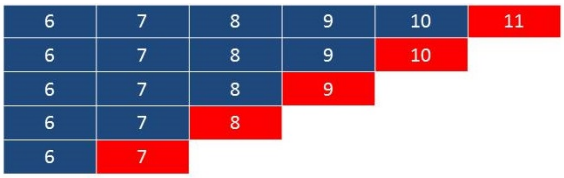

Figura 4.2: Evolução dos dados em *Growing Window*

A figura 4.2 ilustra as cinco primeiras iterações de implementação do modelo. Em azul estão marcadas as jornadas de treino, e a vermelho estão marcadas as jornadas de teste. No *Sliding Window* a jornada inicial varia com as transições de cada iteração. Nesta estratégia, o *dataset* de treino contem sempre a mesma quantidade de jogos.

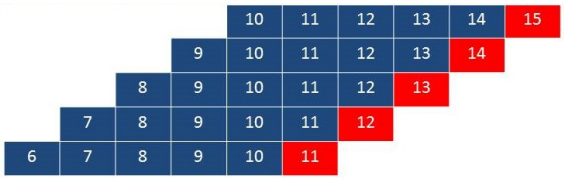

Figura 4.3: Evolução dos dados em *Sliding Window*

A figura 4.3 ilustra cinco iterações usando a estratégia *Sliding Window*. O *dataset* de treino é sempre composto pelas 5 jornadas anteriores à jornada de teste. Com os *datasets* de treino e teste definidos para as 23 iterações, é necessário aplicar os algoritmos e ajustar os seus parâmetros. Consoante a escolha de um certo algoritmo, poderá haver diversos parâmetros que se podem mudar e que alteram o comportamento do algoritmo. Na fase de treino são escolhidos os valores dos parâmetros que obtêm melhor desempenho, criando assim um modelo. O desempenho é medido através dos dados de avaliação que foram definidos.

4.1.3.3 Medidas de Avaliação

A avaliação de um modelo pode ser analisada consoante vários aspetos, como por exemplo a taxa de acerto, o tempo de computação ou aprendizagem, compreensibilidade do conhecimento, requisitos de armazenamento do modelo, entre outros. A *accuracy* ou taxa de acerto foi a métrica de desempenho usada na avaliação do modelo. Neste caso, a avaliação do modelo é analisada pelo desempenho do modelo na classificação de novos exemplos, ou seja, exemplos do *dataset* de teste.

A taxa de acerto é calculada através de uma matriz de confusão. A matriz de confusão faz o emparelhamento entre os resultados previstos pelo modelo e os resultados reais dos jogos em

questão. Assim, para cada iteração é criada uma matriz de confusão que dá a informação da taxa de acerto para aquela jornada. Após 23 iterações, é feita a média de todas as taxas de acerto calculadas anteriormente e é assumido esse valor como a taxa de acerto desse modelo.

Embora não fosse o aspeto principal a ter em conta na escolha do modelo, também foi calculado o tempo de computação de cada modelo.

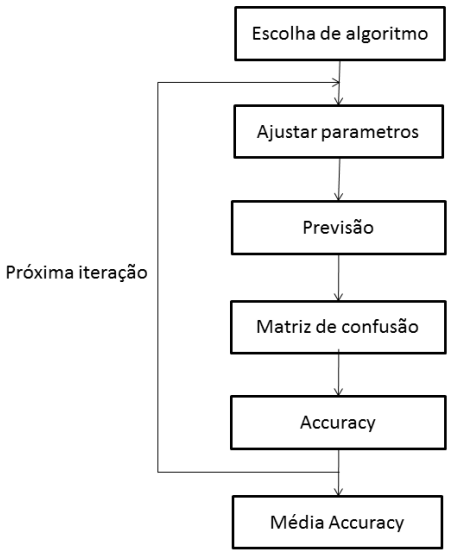

Figura 4.4: Sequência de processos de cada iteração

A figura 4.4 ilustra todo o processo realizado desde a escolha do algoritmo até à medição do seu desempenho. O ciclo é realizado por cada uma das 23 iterações.

4.2 Resultados e Discussão geral

A métrica que define o desempenho de cada modelo implementado é a taxa de acerto. O resultado do desempenho do modelo é calculado através da média dos desempenhos das 23 jornadas. Os *datasets* usados nas análises têm todos os mesmos 152 jogos. Desses 152 jogos, é feita a previsão de 146 jogos porque não é feita a previsão da primeira jornada do *dataset*. Nos 146 jogos há 65 vitórias da equipa da casa, 50 empates e 31 derrotas. A baseline dos *datasets* é aproximadamente 44%.

4.2.1 1ª Iteração: Conjunto de Dados

Na primeira iteração foi construído quatro *datasets* que já foram descritos anteriormente. Esses *datasets* foram testados com 8 algoritmos para se perceber qual a melhor relação entre *datasets* e algoritmos.

Numa primeira análise foi feita a comparação dos desempenhos quando se aplica a estratégia de evolução dos dados *Growing Window* e *Sliding Window*. Foram comparadas poucas amostras das duas estratégias, o que não permitiu concluir que uma estratégia se adequa mais que a outra. No entanto, esta comparação permitiu fazer uma escolha da estratégia a utilizar no desenvolvimento do estudo. A comparação foi feita aplicando o *dataset* completo aos 8 algoritmos selecionados.

Tabela 4.1: Comparação de desempenhos dos diferentes algoritmos em *Growing Window* e *Sliding Window*

	C5.0	RF	Knn	Jrip	SVM Linear	SVM gauss.	NB	NN
G.W.	0.45689	0.52378	0.45134	0.50277	0.47031	0.51976	0.45859	0.44960
S.W.	0.45378	0.48755	0.44824	0.46654	0.49722	0.51894	0.48447	0.43304

Na tabela 4.1 podem ser observadas as comparações feitas aos desempenhos de cada algoritmo aplicando a estratégia *Growing Window* e *Sliding Window*. Fazendo uma análise por algoritmo, pode-se observar que a estratégia *Growing Window* tem melhor desempenho em 4 algoritmos (RF,Knn,JRip,NN), o *Sliding Window* tem melhor desempenho em 2 algoritmos (SVM Linear,NB) e os algoritmos C5.0 e SVM gaussiano apresentam resultados idênticos. Feita esta análise, foi decidido usar a estratégia *Growing Window* no estudo completo. Analisando ainda a tabela 1 pode-se observar que com o *dataset* completo, o algoritmo *random forest* tem um desempenho interessante face aos restantes algoritmos. O algoritmo *random forest* aplicando *Growing Window* tem aproximadamente mais 8 pontos percentuais face à baseline.

A segunda análise é feita tendo em conta os quatro *datasets* construídos na primeira iteração, tanto o *dataset* completo que foi usado na análise anterior, como os *datasets* em que foram aplicados filtros de correlação.

Tabela 4.2: Comparação dos desempenhos dos algoritmos com conjunto de dados diferentes

Dataset	C5.0	RF	Knn	Jrip	SVM Lin.	SVM gauss.	NB	NN
Completo	0.4568	0.5237	0.4513	0.5027	0.4703	0.5197	0.4585	0.4496
Correlação <"0,9"	0.3654	0.5238	0.5082	0.4927	0.4596	0.5155	0.4741	0.4648
Correlação <"0,8"	0.4389	0.5155	0.5507	0.5579	0.4689	0.5217	0.5300	0.5072
Correlação <"0,7"	0.4151	0.5300	0.5869	0.5072	0.4472	0.5234	0.4658	0.4927

Na tabela 4.2 podem ser observadas os desempenhos dos 8 algoritmos selecionados com os quatro *datasets* diferentes. O melhor desempenho obtido por um modelo foi de 58% de taxa de acerto. Este desempenho foi obtido aplicando o algoritmo Knn ao *dataset* que contem os atributos com correlação inferior a “0,7” pontos. Este modelo conseguiu um desempenho superior à baseline de 14 pontos percentuais.

Analisando ainda a tabela 4.2 é notório que alguns modelos reagem melhor quando utilizado um filtro de correlação nos atributos. Os modelos RF,Knn, JRip, SVM gaussiano ,NB,NN obtiveram os seus melhores desempenhos com correlação feita, somente os modelos C5.0 e SVM Linear obtiveram melhores desempenhos com o *dataset* completo.

Três modelos (RF,Knn,SVM gaussiano) obtiveram o seu melhor desempenho com correlação inferior a "0,7", outros 3 (JRip,NB,NN) com correlação inferior a "0,8" e 2 (C5.0,SVM Linear) com o *dataset* completo. Conclui-se assim que o *dataset* original tem demasiados atributos redundantes para a maior parte dos algoritmos selecionados.

Dispondo dos desempenhos de cada modelo é importante perceber quais as razões que levam os algoritmos a errar. Desta feita, fez-se a seleção dos três melhores desempenhos de algoritmos diferentes e analisou-se as respetivas matrizes de confusão.

Tabela 4.3: Matriz de confusão do algoritmo Knn

	Empate	Derrota	Vitória
Empate	7	3	7
Derrota	17	32	13
Vitória	6	12	43

Na tabela 4.3 pode ser observado a matriz de confusão que se refere à aplicação do algoritmo Knn ao *dataset* com correlação inferior a "0,7". O desempenho deste modelo tem uma taxa de acerto de aproximadamente 58,69%. As colunas indicam os jogos previstos pelo modelo e as linhas indicam os resultados reais desses mesmos jogos. A tabela 4.3 mostra que o modelo tem uma taxa de acerto de aproximadamente 68% na previsão de vitórias e empates e de 23% na previsão de empates. Há uma clara diferença de desempenho do modelo na previsão de vitória e derrotas em relação aos empates.

Tabela 4.4: Matriz de confusão do algoritmo JRip

	Empate	Derrota	Vitória
Empate	5	1	2
Derrota	16	28	16
Vitória	9	18	45

Tabela 4.5: Matriz de confusão do algoritmo random forest

	Empate	Derrota	Vitória
Empate	2	4	7
Derrota	15	29	13
Vitória	13	14	43

Nas tabelas 4.4 e 4.5 pode ser observado as matrizes de confusão que se referem à aplicação do algoritmo JRip ao *dataset* com correlação inferior a "0,8" e à aplicação do algoritmo *random forest* ao *dataset* com correlação inferior a "0,7", respetivamente.

As tabelas 4.4 e 4.5 mostram que a tendência da taxa de acerto verificada na tabela 4.3 se mantém. A taxa de acerto nas vitórias mostradas pelas tabelas 4.4 e 4.5 são de 71% e 68% respetivamente, enquanto a taxa de acerto das derrotas é de 59% e 61%. A previsão dos empates tem uma taxa de acerto de 16% e 6% nas tabelas 4.4 e 4.5,respetivamente, o que constitui um

desempenho muito abaixo dos restantes resultados. Pode-se assumir que os diferentes algoritmos têm uma clara dificuldade na previsão de empates tendo acertado em todos eles menos de 25% dos empates totais. Esta dificuldade pode ser explicada pelo facto de os empates constituírem a classe minoritária no *dataset*, com cerca de 21%. Outra explicação pode ser o fator casa. Como se pode observar nas tabelas 4.3,4.4 e 4.5, quando o resultado real do jogo foi empate os algoritmos preveem maioritariamente a derrota da equipa da casa. Esta tendência pode dever-se ao facto de o fator casa não ter tanta influência no modelo como devia, ou seja, o fator casa muitas vezes equilibra um jogo em que a equipa visitante é superior à equipa da casa e o resultado final termina num empate.

4.2.1.1 Significância Estatística dos Resultados

Uma vez concluída a analise dos resultados é importante perceber até que ponto estes resultados são confiáveis ou ocorreram por acaso. Determinar se um modelo é melhor que o outro pelo simples exame de inferioridade/superioridade de médias não é aconselhável. Muitas vezes as diferenças verificadas não são significativas, e pode-se considerar os desempenhos obtidos equivalentes. É necessário realizar um teste de hipóteses para a comparação dos desempenhos dos modelos que estão a ser analisados [2]. Foi aplicado o teste de Friedman. Este teste é baseado na comparação de *rankings* de desempenho [2]. O teste de Friedman avaliou a significância dos 8 modelos tendo em conta a sua taxa de acerto em cada jornada. Foram feitas duas análises em que os modelos selecionados para cada algoritmo foram escolhidos de maneira diferente.

Numa primeira fase foram escolhidos os modelos com melhor desempenho em cada algoritmo. Com esta escolha, há a possibilidade de existir modelos que foram construídos com *datasets* diferentes, o que pode ser um erro de avaliação.

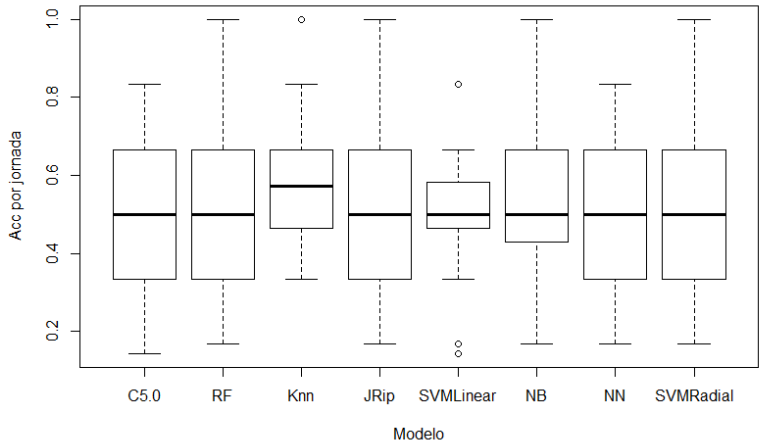

Figura 4.5: Desempenhos dos 8 algoritmos com datasets diferentes

A figura 4.5 ilustra a taxa de acerto de cada um dos algoritmos selecionados. No campo da estatística, são formuladas hipóteses acerca de uma dada amostra, estas hipóteses são submetidas a determinados testes. A hipótese a ser testada designa-se por hipótese nula, a hipótese alternativa é a conclusão a que se chega quando a hipótese nula é rejeitada [41].

No teste de Friedman, a hipótese nula será a aleatoriedade da amostra e a hipótese alternativa será a não aleatoriedade da amostra.

A probabilidade de significância (ρ) obtida pela aplicação do teste de Friedman foi de 0,7492. Assumindo um nível de significância de 10%, o teste de Friedman não rejeitou a hipótese nula porque $\rho > \alpha$ =0,1. Não é detetado qualquer diferença entre as taxas de acerto dos diferentes modelos, sendo assim nenhuma conclusão se pode tirar acerca das performances dos algoritmos.

Na segunda fase foram escolhidos os modelos dos diferentes algoritmos com o mesmo *dataset*, o *dataset* com correlação inferior a "0,7" foi o escolhido.

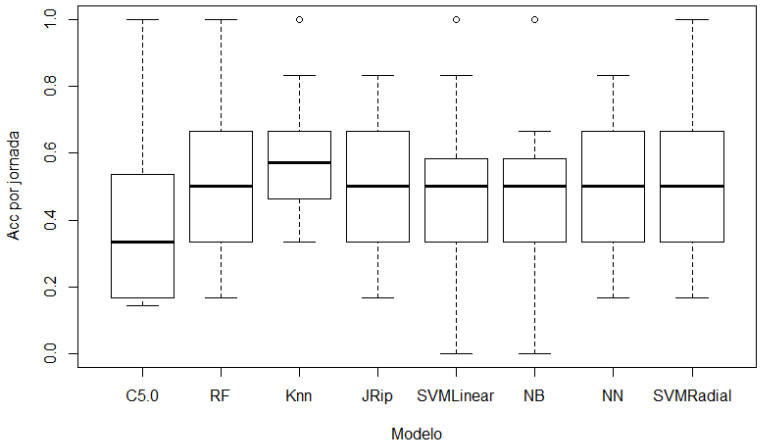

Figura 4.6: Desempenhos dos 8 algoritmos com *datasets* iguais

O desempenho dos diferentes modelos selecionados pode ser visto na figura 4.6. Nesta segunda fase os resultados obtidos pela aplicação do teste de Friedman foram diferentes. A probabilidade de significância obtida foi de 0,01701. Mantendo o nível de significância em 10%, o resultado demonstra que teste de Friedman rejeitou a hipótese nula ($\rho<\alpha$=0.1), podendo-se assumir que existem diferenças significativas entre os modelos. Para que se detete quais os grupos de modelos que diferem entre si, é necessário fazer um teste *post-hoc* que irá realizar múltiplas comparações entre os modelos. O teste escolhido foi o teste Nemenyi.

	C5.0	RF	Knn	JRip	SVMLinear	NB	NN
RF	0.41037812	NA	NA	NA	NA	NA	NA
Knn	0.08108606	0.9953144	NA	NA	NA	NA	NA
JRip	0.85091031	0.9972211	0.8363929	NA	NA	NA	NA
SVMLinear	0.99636741	0.8646799	0.3909728	0.9972211	NA	NA	NA
NB	0.99636741	0.8646799	0.3909728	0.9972211	1.0000000	NA	NA
NN	0.80520421	0.9988651	0.8776857	1.0000000	0.9940306	0.9940306	NA
SVMRadial	0.20925302	0.9999625	0.9999354	0.9660388	0.6570722	0.6570722	0.9794705

Figura 4.7: Resultado da aplicação do teste de Nemenyi
Fonte:

Na figura 4.7 pode-se ver os resultados após aplicação do teste de Nemenyi. Os únicos dois modelos que pertencem a grupos diferentes são os modelos baseados nos algoritmos C5.0 e Knn, sendo assim pode-se afirmar que a diferença de performance entre estes dois modelos é confiável e verdadeira. Pode-se afirmar ainda, que os restantes modelos pertencem ao mesmo grupo.

4.2.2 2ª Iteração: Novo Conjunto de Dados

Após a discussão e análise dos resultados da primeira iteração, um dos objetivos da segunda iteração era melhorar o desempenho dos modelos com a adição de novas características aos *datasets*. Na segunda iteração decidiu-se reduzir o número de algoritmos a serem testados com base nos resultados da primeira iteração. Os algoritmos selecionados para a fase de modelação da segunda iteração foram o knn , *random forest* e o SVM com *kernel* gaussiano.

O algoritmo knn foi escolhido por ter o melhor desempenho de todos os algoritmos e também por ser um algoritmo de *Data Mining* clássico. A escolha do *random forest* surgiu pelo seu bom desempenho na primeira iteração e também por ser um algoritmo de combinação de outros algoritmos. O algoritmo SVM com *kernel* gaussiano foi escolhido também pelo seu bom desempenho, pela robustez com grandes dimensões de dados e boa capacidade de generalização.

Tabela 4.6: Desempenhos dos três algoritmos selecionados na segunda iteração

Dataset	RF	Knn	SVM gaussiano
Completo	0.4865424	0.5020704	0.5082816
Correlação inferior a "0,9"	0.4875776	0.4720497	0.5300207
Correlação inferior a "0,8"	0.5082816	0.4699793	0.5424431
Correlação inferior a "0,7"	0.5082816	0.4689441	0.5207039

Na tabela 4.6 está ilustrado o desempenho de cada um dos três algoritmos aplicados aos quatros *datasets* construídos na segunda iteração. Como se pode observar, o desempenho dos algoritmos RF e Knn baixaram em relação aos desempenhos da primeira iteração. Com a descida do desempenho dos algoritmos, pode-se concluir que as características adicionadas na segunda iteração não se demonstraram relevantes perante o problema de previsão de resultados. Em contrapartida, o desempenho do algoritmo SVM com *kernel* gaussiano melhorou ligeiramente a sua performance.

No geral, o desempenho dos algoritmos não melhorou como era o objetivo da segunda iteração, o que significa que as características adicionadas não constituíram uma mais-valia em relação às

características anteriores. De salientar ainda, que os algoritmos continuaram a ter problemas na previsão dos empates como resultado final.

4.2.3 3ª Iteração: Aumento do Conjunto de Dados

Da primeira iteração para a segunda iteração não houve evolução da taxa de acerto, por isso, foi decidido fazer uma terceira e última iteração. Perante a permanência do problema de previsão dos empates, foram pensadas duas soluções para que o problema fosse ultrapassado. Uma solução seria o balanceamento das classes e a outra solução seria aumentar o *dataset* de treino.

O empate está presente em 21% de todos os resultados, o que, apesar de ser a classe minoritária não justifica de todo um balanceamento dos dados. Por esta razão, optou-se pelo aumento de jogos no *dataset* de treino. Este aumento do número de jogos foi conseguido pela inclusão no *dataset* de treino dos jogos das três épocas anteriores. Assim, no *dataset* de treino passaram a constar jogos das épocas 2009/20010, 2010/2011, 2011/2012 e 2012/13. O *dataset* de treino aumentou de 152 para 607 jogos.

Tabela 4.7: Desempenhos dos três algoritmos selecionados na segunda iteração

Dataset	RF	Knn	SVM gaussiano
Completo	0.5217391	0.5942029	0.5641822
Correlação inferior a "0,9"	0.5424431	0.4937888	0.5931677
Correlação inferior a "0,8"	0.563147	0.5351967	0.5714286
Correlação inferior a "0,7"	0.5776398	0.5766046	0.5724638

A tabela 4.7 apresenta as taxas de acerto de cada algoritmo utilizando os *datasets* construídos na terceira iteração. De uma forma geral, a inclusão de mais dados de treino permitiu que os algoritmos melhorassem os seus desempenhos. A taxa de acerto da maioria dos algoritmos testados está acima dos 50%. Apesar da melhoria, o melhor desempenho tem uma taxa de acerto de aproximadamente 59%, o que constitui uma subida de 1% em relação ao melhor desempenho efetuado nas iterações anteriores.

Após a avaliação do desempenho dos algoritmos, foi feita uma análise ao desempenho na primeira e segunda volta do campeonato dos dois melhores algoritmos (Knn e SVM gaussiano).

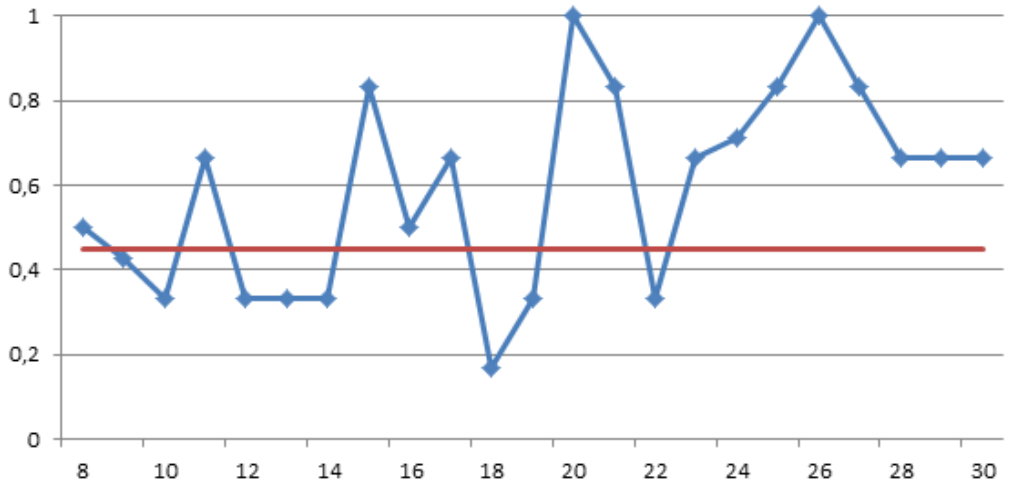

Figura 4.8: Taxa de acerto por jornada do algoritmo Knn

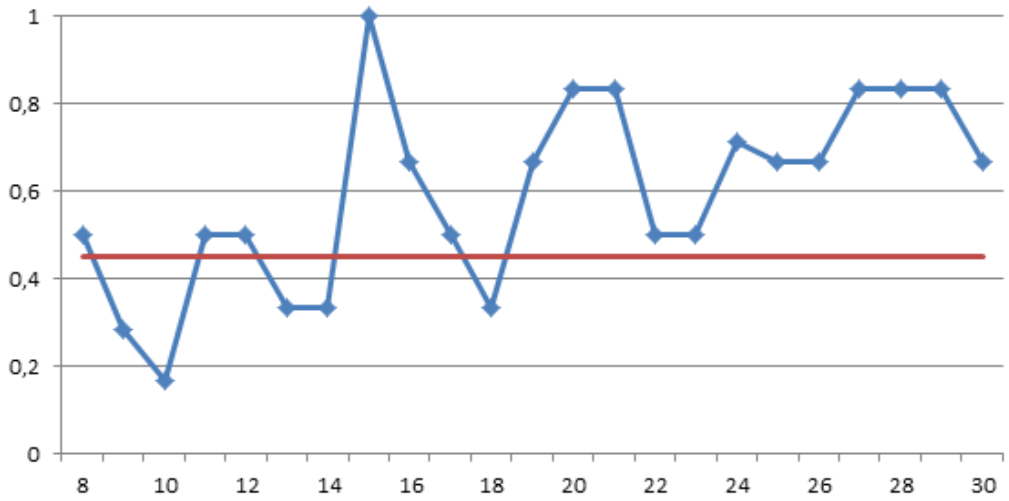

Figura 4.9: Taxa de acerto por jornada do algoritmo SVM com *kernel* gaussiano

Na figura 4.8 e 4.9 é ilustrada a taxa de acerto jornada a jornada dos dois melhores desempenhos. A vermelho está marcada a baseline de 45% dos *datasets* de treino.

Observando os gráficos da figura 4.8 e 4.9, verifica-se que o desempenho dos algoritmos vai melhorando à medida que o campeonato se aproxima do fim. Fazendo a divisão da primeira e segunda volta na jornada 15, é notório que os desempenhos na segunda volta são superiores aos desempenhos na primeira volta, apesar de conterem mais jogos.

Na tabela 4.8 pode-se ver a média das taxas de acerto na primeira e segunda volta por parte dos dois algoritmos. Contabilizando só os jogos da segunda volta, as taxas de acerto são sensivelmente 20 pontos percentuais superiores à baseline, o que constitui uma evolução face à taxa de acerto no total de jogos. Depois de concluída a terceira iteração é necessário fazer a escolha do modelo

Tabela 4.8: Desempenho dos algoritmos na primeira e segunda volta do campeonato

	Knn (59,4%)	SVM gaussiano (59,3%)
Primeira volta (8 jogos)	0.4523809	0.470238
Segunda Volta (15 jogos)	0.6698412	0.6587301

final do estudo. O modelo final escolhido foi baseado no algoritmo Knn porque foi o que obteve melhor desempenho. O *dataset* a utilizar será o *dataset* completo com todas as características.

4.2.4 Avaliação final: Conjunto de Dados Desconhecido

Uma vez escolhido o modelo final e o tipo de pré-processamento efetuado aos dados, é necessário avaliar o desempenho e a capacidade de generalização do modelo final. A avaliação é feita com base em dados novos que não foram utilizados anteriormente. A escolha recaiu sobre os dados da época 2013/2014 como já tinha sido mencionado na fase de preparação dos dados. O desempenho do modelo final com os novos dados obteve uma taxa de acerto de aproximadamente 45%.

4.3 Discussão Geral

O conjunto de dados disponível não era adequado ao problema uma vez que fazia referencia a eventos isolados dos jogos em vez das estatísticas dos mesmos. Durante o processo de preparação dos dados, essas estatísticas forma criadas e mostraram-se adequadas ao problema.

Ao longo de todo o processo, diversos modelos apresentaram desempenhos satisfatórios. Os melhores modelos alcançaram mesmo uma taxa de acerto 10% a 15% maior do que a baseline do conjunto de dados. A primeira iteração apresentou modelos com desempenhos interessantes. No entanto, os modelos da segunda iteração não apresentaram qualquer evolução no seu desempenho em relação à iteração anterior. Um dos motivos para esta estagnação pode residir na escolha incorreta dos algoritmos a utilizar.Na terceira iteração, quando foi alargado o conjunto de dados de treino os modelos apresentaram uma melhoria nos seus desempenhos.

A taxa de acerto na previsão dos empates foi sempre inferior a 25% e tornou-se num problema transversal às três iterações. Este facto foi detetado na primeira iteração e ouve um esforço para o tentar minimizar nas iterações seguintes, sem resultado.

O modelo final apresentou uma taxa de acerto de aproximadamente 45% quando confrontado com "novos"dados. Houve uma quebra no desempenho do modelo, uma vez que a sua taxa de acerto anterior era de aproximadamente 59%. Esta quebra pode indiciar que o modelo está demasiado ajustado aos dados de treino, ou seja, tem uma capacidade de generalização reduzida.

Capítulo 5

Conclusão

Esta dissertação teve como objetivo o estudo empírico de uma abordagem *Data Mining* na previsão de resultados de jogos de futebol. CRISP-DM foi a metodologia de *Data Mining* seguida. No geral, foram feitas duas iterações de todas as fases desta metodologia, embora por vezes fosse necessário voltar à etapa anterior para correção de pequenos erros. O entendimento do problema de previsão dos resultados foi feito com recurso à bibliografia estudada e descrita na revisão da literatura.

Na primeira iteração foram recebidos os dados dos jogos da liga portuguesa fornecidos pelo Laboratório SAPO/ U.Porto. Uma vez recebidos os dados, foi necessário compreender o seu conteúdo e perceber a sua qualidade. Devido à análise feita com recurso a gráficos, foram detetados dados duplicados, inconsistentes, sem informação etc. Seguiu-se a limpeza dos destes dados e construção de novas variáveis que fossem úteis de acordo com o problema. Depois da seleção de atributos com recurso a técnicas de correlação, foram construídos 4 conjuntos de dados que foram aplicados na fase de modelação. Os quatro conjunto de dados tinham níveis de correlação diferentes. Na fase de modelação foram criados modelos tendo por base 8 algoritmos e os quatro conjuntos de dados criados anteriormente. Foram avaliados o desempenhos do modelos através da taxa de acerto de cada um deles. Também foram estudadas algumas possíveis falhas nos modelos para que fossem compensadas nas iterações seguintes.

Na segunda iteração, o Laboratório SAPO/ U.Porto forneceu novos dados com mais informação sobre os jogos. Foram construídas novas variáveis e novos conjunto de dados para a fase da modelação. Com base nos resultados da iteração anterior foram escolhidos 2 algoritmos para a criação de novos modelos. Os resultados dos modelos da primeira e segunda iteração são muito idênticos, o que significa que não houve evolução de desempenho. Perante esta estagnação, decidiu-se alargar o conjunto de dados de treino. Sendo assim, o conjunto de dados de treino passou a conter jogos de 4 épocas, quando anteriormente continha jogos de uma época. Com esta alteração o desempenho dos modelos melhorou. O modelo com melhor desempenho alcançou uma taxa de acerto de 58%. Uma vez escolhido o modelo final, foi testada a capacidade de generalização do modelo com a previsão de jogos que ainda não tinham sido testados. O modelo obteve uma taxa de acerto de 45% com os novos dados. Verificou-se uma descida no desempenho do modelo, o

que pode significar que o modelo está demasiado ajustado aos dados de treino. Com o desenvolvimento de um modelo final pode afirmar-se que os objetivos desta dissertação foram alcançados e cumpridos.

5.1 Trabalho futuro

Após a realização da dissertação e análise dos resultados é possível a evolução do mesmo com vista à melhoria de desempenhos dos modelos.

Esta dissertação abordou o problema de previsão do resultado de um jogo como sendo uma de três hipóteses, vitória da equipa da casa, empate ou derrota. Das três hipóteses, a vitória da equipa da casa é o acontecimento que acontece com mais frequência. Uma abordagem futura seria encarar o problema de classificação como tendo um atributo alvo de duas classes, vitoria ou não vitória da equipa da casa. A não vitória seria a previsão de empate ou derrota. Um modelo que classifique duas classes tem tendência a ter uma taxa de acerto maior do que um que classifique três classes.

Uma outra abordagem futura poderá ser a inclusão de novos atributos com características particulares no conjunto de dados utilizado. Essas características podem ser de vários tipos. Características físicas como por exemplo fadiga, jogadores lesionados, resistência, esforço etc. Características psicológicas como por exemplo mudança de treinador, jogadores insatisfeitos, concentração, reação às adversidades etc. Características do foro técnico-tático como táticas utilizadas, modelo de jogo, pontos fortes e fracos, jogadores influentes etc. Muitas destas características são subjetivas o que torna difícil a sua caracterização, no entanto serão mais-valias na construção de novos modelos para a previsão de resultados.

Os modelos desenvolvidos tiveram melhor desempenho nos jogos da segunda volta do campeonato do que nos jogos da primeira volta. Uma abordagem futura seria implementar um modelo que fizesse a previsão do resultado somente nos jogos da segunda volta, tendo como dados de treino os jogos da primeira volta.

Referências

[1] Kj Cios, W Pedrycz, Rw Swiniarski, e La Kurgan. *Data mining: a knowledge discovery approach*. 2010. URL: http://dl.acm.org/citation.cfm?id=1941721.

[2] Ana Carolina Lorena, Katti Faceli, Márcia Oliveira, André Ponce De Leon Carvalho, e João Gama. *Extração de Conhecimento de Dados - Data Mining*. 1ª edição, 2012.

[3] Farzin Owramipur, Parinaz Eskandarian, e Faezeh Sadat Mozneb. Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. *International Journal of Computer Theory and Engineering*, 5(5):812–815, 2013. URL: http://www.ijcte.org/index.php?m=content&c=index&a=show&catid=51&id=925, doi:10.7763/IJCTE.2013.V5.802.

[4] Abraham Mehrez e Michael Y. Hu. Predictors of the outcome of a soccer game - a normative analysis illustrated for the Israeli Soccer League. *ZOR Zeitschrift f??r Operations Research Mathematical Methods of Operations Research*, 42:361–372, 1995. doi:10.1007/BF01432510.

[5] Byungho Min, Jinhyuck Kim, Chongyoun Choe, Hyeonsang Eom, Robert Ian, e Bob Mckay. A CompoundFramework for Sports Prediction: The Case Study ofFootball. *Knowledge-Based Systems or Expert Systems with Applications*, (February), 2007.

[6] J. McCullagh. Data Mining in Sport: A Neural Network Approach. *International Journal of Sports Science . . .* , 3(03):131–138, 2010. URL: http://www.worldacademicunion.com/journal/SSCI/SSCIvol04no03paper01.pdf.

[7] InfoEscola Navegando e Aprendendo. No Title. URL: http://www.infoescola.com/informatica/linguagem-de-programacao-r/.

[8] Author Max Kuhn, Jed Wing, Steve Weston, e Andre Williams. The caret Package, 2007. URL: http://topepo.github.io/caret/index.html.

[9] J Fürnkranz, D Gamberger, e N Lavrač. Foundations of rule learning. *Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki*, (2003), 2012. URL: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:No+Title#0$\delimiter"026E30F$nhttp://books.google.com/books?hl=en&lr=&id=JsTXnjWexCEC&oi=fnd&pg=PR5&dq=Foundations+of+Rule+Learning&ots=Yte7yaCOUZ&sig=WcY6bUOhz2feQ7uu70Y-2sHYBQo, doi:10.1007/978-3-540-75197-7.

[10] Sergio Navega. Princípios Essenciais do Data Mining. *Anais de Infoimagem, Cenadem*, 2002. URL: http://www.intelliwise.com/snavega/.

[11] Robert P Schumaker, Osama K Solieman, e Hsinchun Chen. Sports Data Mining. *Information Systems Journal*, 26:15–22, 2010. URL: http://www.springerlink.com/index/10.1007/978-1-4419-6730-5, doi:10.1007/978-1-4419-6730-5.

[12] Jiawei Han, Micheline Kamber, e Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 3rd edição, 2011. URL: http://www.cs.uiuc.edu/~hanj/bk3/.

[13] Co Camilo e Jc Silva. Mineração de Dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, página 29, 2009. URL: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf.

[14] Ana Cristina da Silva Anacleto. Aplicação de Técnicas de Data Mining em Extracção de Elementos de Documentos Comerciais Tese de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão. *Faculdade de Economia Universidade do Porto*, 2009. URL: http://repositorio-aberto.up.pt/bitstream/10216/22553/2/MADANAANACLETO.pdf.

[15] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, e Rüdiger Wirth. CRISP-DM 1.0: Step-by-step data mining guide. Relatório técnico, 2000.

[16] Wikipédia. Wikipédia-Cross Industry Standard Process for Data Mining. URL: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining.

[17] U Fayyad, G Piatetsky-Shapiro, e P Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–53, 1996. URL: http://www.scopus.com/scopus/inward/record.url?eid=2-s2.0-0002283033&partnerID=40&rel=R8.2.0$\delimiter"026E30F$nhttp://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230, doi:10.1609/aimag.v17i3.1230.

[18] Ruud H Koning, Michael Koolhaas, Gusta Renes, e Geert Ridder. A simulation model for football championships. *European Journal of Operational Research*, 148(October):268–276, 2003. doi:10.1016/S0377-2217(02)00683-5.

[19] a Tsakonas e G Dounias. Soft computing-based result prediction of football games. *The First International ...*, 3(May):15–21, 2002. URL: http://www.researchgate.net/publication/2560104_Soft_Computing-Based_Result_Prediction_of_Football_Games/file/79e41509b9947b0861.pdf.

[20] S Nunes e Marco Sousa. Applying data mining techniques to football data from European championships. *...da 1ª Conferência de Metodologias de ...*, (December 2005), 2006. URL: http://repositorio-aberto.up.pt/handle/10216/282.

[21] Kou-yuan Huang, Senior Member, e Wen-lung Chang. A Neural Network Method for Prediction of 2006 World Cup. *Test*, páginas 18–23, 2010.

[22] a. Joseph, N. E. Fenton, e M. Neil. Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(April 2005):544–553, 2006. doi:10.1016/j.knosys.2006.04.011.

[23] B.G. Aslan e M.M. Inceoglu. A Comparative Study on Neural Network Based Soccer Result Prediction. *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, páginas 545–550, 2007. doi:10.1109/ISDA.2007.12.

[24] Josip Hucaljuk e Alen Rakipovic. Predicting football scores using machine learning techniques. *2011 Proceedings of the 34th International Convention MIPRO*, 48:1623–1627, 2011.

[25] Gianluca Baio e M Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, páginas 1–13, 2010. URL: http://www.tandfonline.com/doi/abs/10.1080/02664760802684177, doi:10.1080/02664760802684177.

[26] Alan McCabe e Jarrod Trevathan. Artificial intelligence in sports prediction. *Proceedings - International Conference on Information Technology: New Generations, ITNG 2008*, páginas 1194–1197, 2008. doi:10.1109/ITNG.2008.203.

[27] Na Wei. Predicting the outcome of NBA playoffs using Algorithms. *University of South Florida*, páginas 2008–2009, 2011.

[28] Chinwe Peace e Enoch Okechukwu. An Improved Prediction System for Football a Match Result. 04(12):12–20, 2014.

[29] Rk Balla. Soccer match result prediction using neural networks. 2007. URL: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Soccer+Match+Result+Prediction+using+Neural+Networks#2.

[30] a K Suzuki, L E B Salasar, J G Leite, e F Louzada-Neto. A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup. *Journal of the Operational Research Society*, 61:1530–1539, 2010. doi:10.1057/jors.2009.127.

[31] a P Rotshtein, M Posner, a B Rakityanskaya, Machon Lev, e Vinnitsa National. Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41(4):619–630, 2005. doi:10.1007/s10559-005-0098-4.

[32] Richard Pollard. Home Advantage in Football: A Current Review of an Unsolved Puzzle. *The Open Sports Sciences Journal*, 1:12–14, 2008. doi:10.2174/1875399X00801010012.

[33] Anna Waters e Geoff Lovell. An Examination of the Homefield Advantage in a Professional English Soccer Team from a Psychological Standpoint. *Football Studies*, 5:46–59, 2002.

[34] Ben Ulmer e Matthew Fernandez. Predicting Soccer Match Results in the English Premier League. 2013.

[35] Alessandro Martins e Alves Uff. SIMULAÇÕES DE RESULTADO PARA O CAMPEONATO BRASILEIRO DE 2008 COM BASE EM MODELOS LOGITO. 2009.

[36] Douwe Buursma. Predicting sports events from past results. 2011.

[37] Kj Kain e Td Logan. Are sports betting markets prediction markets? Evidence from a new test. *Journal of Sports Economics*, (January), 2014. URL: http://jse.sagepub.com/content/15/1/45.short.

[38] Declan Sheridan. Modelling football match results and testing the efficiency of the betting market. 2012. URL: http://eprints.nuim.ie/4469/.

[39] Victor Luiz Guimarães Ademir Rafael Marques Guedes. SISTEMA DE IDENTIFICAÇÃO DE ÍRIS UTILIZANDO LOCAL BINARY PATTERN E RANDOM FOREST Ademir Rafael Marques Guedes , Victor Luiz Guimarães Universidade Federal de Ouro Preto (UFOP) Departamento de Computação.

[40] Ana Carolina Lorena e André C. P. L. F. De Carvalho. Introdução à Máquinas de Vetores de Suporte. *Relatórios técnicos do ICMC*, página 66, 2003.

[41] Filipe Gago Da Câmara e Dr. Osvaldo Silva. Estatística Não Paramétrica, Testes de Hipóteses e Medidas de Associação. *Estatística não paramétrica*, página 121, 2001. URL: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Estat?stica+N?o+Param?trica#1.