TÉCNICO LISBOA

# Previsão de Resultados para Desportos Coletivos

**Alexandre Figueiredo Pereira**

Dissertação para obtenção do Grau de Mestre em

**Engenharia de Telecomunicações e Informática**

Orientador: Professora Cláudia Martins Antunes

**Júri**

Presidente: Professor Ricardo Jorge Fernandes Chaves

Orientador: Professora Cláudia Martins Antunes

Vogais: Professor Miguel Ângelo Marques de Matos

Professor Rui Miguel Carrasqueiro Henriques

**Julho de 2018**

# Resumo

Cada vez são criados e recolhidos mais dados em todas as áreas, no desporto não é diferente. Acredita-se que com as análises de desempenho das equipas e dos jogadores, os apostadores podem prever melhor os resultados dos jogos. O objetivo deste trabalho é definir e explorar diferentes modelos de previsão de resultados para desportos coletivos, em particular para o basquetebol (NBA e NCAA – masculino e feminino) e o futebol (Inglaterra e Portugal). Os modelos são depois integrados num protótipo de uma aplicação móvel para que um possível cliente possa ter acesso às previsões efetuadas. Neste documento, é feita uma revisão da literatura que descreve algumas abordagens anteriores no mesmo domínio, descrevendo-se ainda as estatísticas mais usuais nas duas modalidades desportivas referidas. Os diferentes modelos explorados são treinados com Naïve Bayes, Random Forests e Perceptrão Multicamada, na presença de dois tipos de dados, os dados que resumem o jogo (equipas que se defrontam e resultado alcançado) habitualmente utilizados, os dados anteriores juntamente com alguns dados de contexto, como são a caracterização do local e data do evento e os treinadores e jogadores de cada equipa. É ainda apresentado o protótipo da aplicação móvel desenvolvido, mostrando-se que as previsões efetuadas apresentam acertos na ordem dos 70% para a NCAA e 60% para a NBA, e de cerca de 60% para o futebol português e aproximadamente de 54% para o futebol inglês.

# Abstract

Every time more data is created and collected in all areas, in sport it is no different. It is believed that with the analysis of the performance of teams and players, bettors can better predict the results of the games. The objective of this work is to define and explore different models of prediction of results for collective sports, in particular for basketball (NBA and NCAA - men and women) and football (England and Portugal). The models are then integrated into a prototype of a mobile application so that a prospective customer can access the predictions made. In this document, a review of the literature is described that describes some previous approaches in the same domain, also describing the most usual statistics in the two mentioned sports modalities. The different exploited models are trained with Naïve Bayes, Random Forests and Multi-layer Perceptron, in the presence of two types of data, the data that summarizes the game (teams that face and result achieved) usually used, previous data together with some data context and how they characterize the venue and date of the event and the coaches and players of each team. It is also presented the prototype of the mobile application developed, showing that the predictions made correct in the order of 70% for the NCAA and 60% for the NBA, and about 60% for Portuguese football and approximately 54% for the English football.

# Palavras Chave

Previsão de Resultados
Desportos coletivos
Futebol
NBA
NCAA
Modelos de aprendizagem

# Índice

# Lista de Figuras

# Lista de Tabelas

# Acrónimos

AST Assistências
EPL English Premier League
MLP Perceptrão Multicamadas
NBA National Basketball League
NCAA National Collegiate Athletic Association
PER Player Effiency Rating
PLP Primeira Liga de Portugal
SGD Gradiente Estocástico Descendente
SVM Máquina de Vetores de Suporte
TCP Transporte de Controlo de Transmissão
TO turnovers

# 1 INTRODUÇÃO

Desde há muito se fala sobre pessoas fazerem apostas, os chamados jogos de sorte ou azar. Uma das apostas mais comuns é determinar o vencedor de um evento desportivo. Este tipo de jogo deu origem ao estudo de estatísticas por parte dos apostadores para determinarem qual a sua melhor opção. Assim, com o avançar do tempo, esse estudo evoluiu até a um ponto onde a utilização das diversas variáveis envolventes nos jogos e a utilização das ferramentas de *ciência de dados* são indispensáveis.

Neste trabalho, o objetivo é definir e explorar diferentes modelos de previsão de resultados em desportos coletivos, nomeadamente basquetebol e futebol. Sendo os casos de estudo a principal liga americana de basquetebol, a National Basketball Association (NBA), e as principais competições de basquetebol universitário norte-americano, masculino e feminino, a NCAA March Madness, e ainda nas equipas que disputam as principais ligas de futebol inglesas e portuguesas, English Premier League (EPL) e Liga NOS, ou Primeira Liga (PLP), respetivamente. Os resultados dos casos de estudo são alcançados através da melhoria de modelos de previsão. Assim, para obter percentagens de acerto mais altas é necessário ter em conta algum conhecimento de domínio disponível e perceber quais os atributos relevantes que conduzem a esse objetivo.

Apesar de muitos trabalhos terem já sido feitos relativos a este tema, são poucos os que apresentam resultados interessantes para que possam ser considerados bons do ponto de vista do apostador e da sociedade em geral. Muitos dos estudos, que têm como base as modalidades a serem utilizadas neste trabalho, não ultrapassam os 65% de previsões corretas. Recentemente, o Massachusetts Institute of Technology apresentou uma ferramenta de previsão de resultados para

o Mundial de Futebol a decorrer no presente ano, 2018, na Rússia, mostrando assim a relevância que este tema tem na sociedade e na área académica. [34]

Para além dos modelos de previsão, é apresentado ainda um protótipo de aplicação móvel, construída para sistemas Android, para um potencial cliente poder ter acesso às previsões. Desta forma, é necessário um servidor para poder a realização de testes, completando assim o protótipo de uma ferramenta mais complexa.

Posto tudo isto, o documento tem mais 5 capítulos. O primeiro, capítulo 2, apresentado no seguimento, “Trabalho Relacionado”, dá a conhecer diversos modelos de previsão utilizados no estado da arte, subcapítulo 2.1, e ainda diversos pontos importantes de cada modalidade como conhecimento de domínio, subcapítulo 2.2.

No capítulo 3, dividido por 4 subcapítulos, é dado a conhecer todo o trabalho relativo à ciência de dados. No primeiro subcapítulo, subcapítulo 3.1, é feita a descrição dos dados recolhidos e utilizados nas duas modalidades em estudo, basquetebol e futebol. No subcapítulo 3.2, é apresentado o trabalho realizado e os resultados da mesma numa primeira abordagem nas diversas modalidades e competições, com a construção de diferente conjuntos de dados e ainda a utilização de diferentes algoritmos de classificação. No terceiro subcapítulo, subcapítulo 3.3, é apresentado o trabalho realizado e os resultados da mesma numa segunda abordagem nas diversas modalidades e competições, onde são adicionados novos dados para a construção dos conjuntos de dados a utilizar e consequentemente a construção de novos modelos. No último subcapítulo, subcapítulo 3.4, é feita a avaliação do protótipo da ferramenta.

No capítulo 4, “Protótipo de Ferramenta para Previsão de Resultados Desportivos”, é apresentada a arquitetura da mesma, subcapítulo 4.1, é também apresentada uma solução para fazer a previsão dos jogadores que iniciam os jogos em cada equipa, ajudando à previsão em tempo real, subcapítulo 4.2, depois é exposta a aplicação móvel, desenvolvida em Android, onde um utilizador da aplicação poderá aceder às previsões futuras e às feitas anteriormente, subcapítulo 4.3.

No capítulo 5 são feitas as conclusões e ainda são apresentadas diversas possibilidades com o intuito de melhorar o protótipo.

# 2 TRABALHO RELACIONADO

Num passado relativamente recente, se uma equipa quisesse preparar um jogo poderia por exemplo ver vídeos das outras equipas, ou até de um jogador em pormenor. Exemplo são os casos no futebol em que os guarda-redes para conhecer melhor os marcadores de penáltis dos seus adversários vêm as suas marcações em jogos anteriores. No entanto, a tecnologia evoluiu e é agora possível ter acesso a diversas informações sobre as equipas. Este desenvolvimento influenciou diversos mundos desde os desportivos aos dos adeptos, pois permite saber aspetos importantes desde as estratégias das equipas à condição dos jogadores, que nele vão intervir.

Apesar dos inúmeros desportos coletivos existentes, vamos só abordar dois, basquetebol e futebol. Na primeira modalidade, o campeonato escolhido foi a National Basketball Association (NBA), a principal liga norte-americana e a mais acompanhada a nível mundial, onde em 2017 as "*The Finals*", fase final do *Playoff* que atribui o vencedor da competição, teve uma média de 20,4 milhões telespetadores [7]. Na segunda modalidade, a previsão é efetuada para as equipas que se encontram a disputar os principais campeonatos nacionais, EPL e PLP, tendo todos os jogos das competições nacionais, excluindo a supertaça, e os jogos disputados nas competições internacionais.

Na NBA, cada equipa tem um mínimo de 82 jogos oficiais, em cada época, sendo chamado a esse conjunto Fase Regular. No total existem 30 equipas, dispostas em duas conferências, a Este e a Oeste, onde figuram 15 equipas em cada uma. No final da Fase Regular são apuradas 8 equipas de cada conferência para os *Playoff*, onde existem confrontos diretos entre dois conjuntos. Para se qualificarem para a próxima ronda é necessário ganhar 4 jogos, num máximo de 7. Depois de determinarem os vencedores de cada conferência é então feita a "*The Finals*" para apurar o vencedor da competição. Com isto cada equipa pode fazer entre um mínimo de 4

jogos e um máximo de 28 jogos adicionais à Fase Regular. No caso dos *Playoff*, os locais dos jogos são distribuídos da seguinte forma, 4 jogos em casa para a equipa com melhor registo da Fase Regular e os outros 3 fora. Enquanto que na Fase Regular o número de confrontos entre duas equipas pode ser entre 2 a 4 jogos, dependendo das regras e da organização do calendário para aquela época [10].

No futebol, em particular dos países selecionados, são utilizados os principais campeonatos, e as taças mais importantes, a nacional e a da liga. Os campeonatos nacionais são jogados em duas voltas, assim cada equipa recebe e desloca-se ao terreno dos seus outros oponentes. Em Inglaterra, existem 20 equipas o que perfaz um total de 38 jogos, já em Portugal, não é tão linear, pois dependendo da época houve mudanças ao que toca ao número de equipas no campeonato nacional, nunca havendo mais de 18 equipas, ou seja, mais do que 34 jogos. Em Inglaterra, as taças são sempre jogadas a eliminar, com jogos sorteados, quer em fases com apenas um jogo ou com dois (as equipas recebem e deslocam-se ao terreno do adversário). Em Portugal, a Taça de Portugal é jogada sempre a eliminar, apesar de muito recentemente terem alterado a “Meia-Final” para ser jogada a duas mãos, e Taça da Liga é jogada com diferentes métodos, com fases de grupos e com fases a eliminar, tal deve-se ao facto de ser uma competição recente e que ao longo dos anos levou vários ajustes até se encontrar um modelo que desportivamente e economicamente fosse mais adequado à realidade do país. As competições europeias, nomeadamente, a Liga dos Campeões e a Liga Europa, outrora denominada Taça UEFA, têm diversas fases, onde há apenas uma fase de grupos, com equipas que se qualificaram para ela através de eliminatórias a duas mãos, seguindo-se depois com fases a eliminar com dois jogos entre os conjuntos que se defrontam e a final que se disputa apenas num confronto. Quer as taças nacionais, quer as competições europeias têm em comum o facto de o jogo da final ser num estádio neutro, dessa forma não atribui qualquer favoritismo, através do fator casa, a nenhuma das equipas. No entanto, poderá acontecer que uma das equipas presentes nessa final seja a organizadora da final, como foi o caso do Sporting Clube de Portugal, onde na época de 2004/2005, defrontou o CSKA de Moscovo no Estádio José de Alvalade, Lisboa [13].

Um dos grandes desafios é saber como conciliar as estatísticas envolvidas em cada jogo e de cada modalidade sem adaptar o modelo, que determina o resultado do jogo, para cada uma das modalidades. Assim, é necessário entender que estatísticas podem ser equivalentes em cada modalidade. Por exemplo, na NBA existe uma medida apelidada de “PER” (Player Efficiency Rating), criada por John Hollinger da ESPN [8], que mede a produtividade de um jogador, quer num jogo ou num conjunto de jogos. No futebol não existe uma fórmula específica para saber num único número a produtividade de um jogador, apesar de haver diversas organizações a providenciar essas classificações. Apesar de este caso ser complexo, visto que depende de um cálculo, num caso simples como é a tabela de estatística individual de cada jogador, apresentada

em qualquer página web, na NBA chamada de Box Score. Na NBA existem 20 colunas de dados importantes, enquanto que no futebol não existe um consenso. Assim não existe uma ligação direta entre os atributos que podem ser utilizados nas diferentes modalidades. Para além disto é possível que até alguns eventos numa modalidade não sejam possíveis noutro desporto. É exemplo um evento que existe no futebol, mas que não existe no basquetebol de forma direta. No futebol, existem cartões, amarelos e vermelhos, que penalizam os jogadores, ou até treinadores, por faltas ou linguagem abusiva, no entanto, no basquetebol, não existem cartões, mas na 6ª falta cometida por um jogador, este é expulso, o equivalente a um cartão vermelho no futebol, mas não é a única forma de expulsão nesta modalidade, ainda existem as desqualificantes, técnicas e anti-desportivas, estas duas últimas são necessárias duas atribuições.

Outro desafio passa por saber quais as estatísticas que serão necessárias para obter um modelo mais correto, desde as estatísticas de equipa ou jogador até às condições atmosféricas. São exemplo o histórico de uma equipa com determinado adversário, os últimos cinco jogos realizados por qualquer das equipas, o número de golos que determinado jogador marca em cada jogo, a composição das equipas iniciais, cinco jogadores no basquetebol e onze no futebol, a temperatura, ou outros. Por outras palavras, que dados escolher para alimentar as bases de dados que serão usadas para treinar os modelos de previsão.

Tudo isto resulta no último desafio, escolher o melhor modelo para prever um resultado. No passado, Dean Oliver apelidou um estudo, sobre basquetebol, de “Four Factors of Basketball Success” [9], em que determina que estatísticas e com que pesos podemos saber o sucesso de uma equipa. Este estudo foi muito utilizado para diversas análises, todas em basquetebol, o que torna difícil a sua adaptação para uma modalidade diferente devido às estatísticas particulares que utiliza. No próximo capítulo irá ser apresentado um olhar pelos modelos utilizados noutros trabalhos semelhantes a este, onde não existe uma análise transversal a um conjunto de diferentes modalidades.

## 2.1 MODELOS DE PREVISÃO NO ESTADO DA ARTE

Este capítulo incide na pesquisa efetuada de forma a descobrir e dar a conhecer diversos trabalhos realizados sobre esta temática. Todos têm em comum o facto de terem sido feitos para um determinado desporto e não para um conjunto de modalidades. Esta secção foi dividida em quatro partes: a primeira, “*previsão de vencedor*”, onde é apresentado o conjunto de estudos que têm como objetivo identificar o resultado de um jogo, esses resultados podem surgir em dois moldes diferentes, sendo o primeiro, “*Vitória, Derrota e Empate*”, e o segundo, “*Vitória e Derrota*”; a segunda, “*previsão de resultado*”, onde é apresentado o resultado do jogo depois de ser

calculada a probabilidade de acontecerem diferentes resultados com base em golos/pontos; a terceira, "*rating*", onde a determinação do resultado do jogo tem a particularidade de os intervenientes estarem categorizados com uma determinada pontuação atribuída e a quarta, "*playoff*", onde o objetivo é determinar que equipas vão aos "*playoff"*.

### 2.1.1 PREVISÃO DE VENCEDOR

No trabalho feito por Hulmer e Fernandez [1] para a previsão de resultados no campeonato inglês, English Premier League (EPL), o melhor resultado a que chegaram foi de um erro de 48% para o conjunto de teste e de 50% para o conjunto de treino, utilizando o método de Gradiente Estocástico Descendente (SGD). Foram também testados os seguintes métodos, Naïve Bayes, Hidden Markov Model, Máquina de Suporte Vectorial usando um kernel RBF(SVM), Random Forest e com kernel linear, sendo estes apresentados do pior para o melhor. Sendo que os resultados parecem atribuir ao SGD a percentagem de erro mais baixa e ao de Naïve Bayes a mais alta. Com o objetivo de treinar o modelo, o conjunto de dados construído cobre 10 anos de jogos da EPL, desde a época 2002/2003 à época 2011/2012. Já o conjunto de teste contém todos os jogos de duas épocas da EPL, épocas 2012-2013 e 2013-2014. Constate-se que para cada jogo, o conjunto contém a equipa da casa, a equipa forasteira, o resultado, o vencedor e o número de golos de cada equipa.

No estudo desenvolvido por Tsakonas et. al. [3], para a Primeira Liga da Ucrânia, chegaram a valores de acerto de 83,8% no conjunto de treino (105 jogos) e de 64,28% no conjunto de teste (70 jogos) ao usar um modelo de Programação Genética. Enquanto que para os modelos de Redes Neuronais e Fuzzy as percentagens de acerto são de 64% (média entre acerto com conjunto de treino e de teste). Os conjuntos de dados são compostos por cinco características. A primeira é a diferença entre jogadores disponíveis entre a equipa da casa e a de fora, a segunda é a diferença resultante do número de golos marcados nos últimos cinco jogos em casa, por parte da equipa que recebe, e o número de golos marcados em igual número de jogos fora, por parte da equipa forasteira, como terceira característica, é a diferença das posições ocupadas pelas duas equipas, a quarta característica é uma medida que utiliza os pontos que cada equipa tem no campeonato e ainda o número de jogos que fizeram em casa, para o caso da equipa da casa, e fora, para o caso da equipa visitante, por último, vem a diferença de golos entre as duas equipas no total de jogos realizados nas últimas dez épocas entre elas.

Para Miljkovic et. al. [5], o foco não é a EPL, nem o futebol, mas sim a NBA. A utilização do modelo de Naïve Bayes, para classificar um conjunto de teste com 778 jogos, relativos à época de 2009-2010, conduziu a uma previsão acertada do vencedor de cada jogo de 67%. Nesse conjunto,

cada jogo é apoiado por diversas estatísticas das duas equipas que compõem o jogo, divididos em duas tabelas, a primeira com a estatística de jogo, onde figuram por exemplo o número de pontos por jogo, ressaltos por jogo, etc, de cada equipa e a segunda com os registos de cada equipa sobre as vitórias e derrotas que tem, seja ao longo da época, seja nos últimos jogos em casa, para a equipa da casa, ou para os jogos fora, para a equipa de fora. Na primeira tabela figuram dezoito estatísticas e na segunda catorze.

No estudo feito para o Mundial de Futebol de 2006, realizado na Alemanha, por Huang et. al. [11], foi possível ter um acerto de 76,9%, utilizando o modelo de Perceptrão Multicamadas (MLP) auxiliado pelo algoritmo de retropropagação. Este estudo tem uma percentagem tão alta, relativamente a outros trabalhos, devido a não contemplar os jogos que resultaram em empate. É exemplo o jogo entre as seleções nacionais de Itália e França, jogo pertencente ao conjunto de teste, tal como todos os jogos a eliminar, em que o resultado final foi de empate e que o modelo determinou que o vencedor seria a Itália, assim o expectável seria contabilizar o resultado como errado, mas foi desvalorizado pois o objetivo do trabalho não é identificar empates. O conjunto de treino elenca o conjunto de jogos, das diversas fases da competição, que antecede a fase em que o jogo se insere, com isto, não é possível determinar qualquer vencedor de um jogo da fase de grupos, para além desta regra para os diversos conjuntos de treino, na fase de grupos só são tidos em conta os jogos das equipas que obtiveram 3 vitórias ou 3 derrotas. Os jogos presentes nos conjuntos de dados são apoiados com 8 estatísticas, selecionadas entre 17 possíveis, são elas os números de: golos marcados, remates efetuados, remates à baliza, cantos, livres diretos, livres indiretos, posse de bola e faltas sofridas.

Continuando com o caso individual da NBA, Richardson et. al. [6] concluíram que para o trabalho que realizaram os modelos de Naïve Bayes e de Regressão Linear são os que melhores resultados apresentam para as épocas estudadas, de 2009 até 2013. Por outro lado, o modelo de árvores de decisão apresenta os piores resultados alcançados. Sendo que para isso, foram utilizados para conjunto de treino os dados das épocas anteriores à que se pretende prever, por exemplo para saber em relação à época de 2010, foram usadas para treino as épocas de 2008 e 2009. Os conjuntos foram guardados numa base de dados relacional onde existem 5 tabelas diferentes. A primeira, é relativa aos dados gerais de um jogo, ou seja, o nome das equipas, pontuação, identificador do jogo e ainda a época em questão. A segunda, contém o identificador de jogo, os identificadores de cada jogador e ainda mais dezoito outras características. A terceira diz respeito às estatísticas médias de cada jogador, assim, para além do identificador de cada jogador, existem, pontos por jogo, ressaltos por jogo, etc, para cada época feita, no total são cinquenta colunas. A quarta, que de certo modo está ligada à terceira, contém cinquenta e oito novos atributos por jogador, tais como a idade, o nome do atleta, a posição a que joga, o total de

ressaltos, roubos de bolas, etc. Por último, existe uma tabela, com apenas dois atributos, com o nome abreviado e completo de cada equipa, que está ligada à primeira.

No trabalho efetuado por Beckler et. al. [14], de modo a prever os resultados dos jogos da NBA, foi construído um conjunto de dados com estatísticas, quer das equipas, quer dos jogadores, ao longo das épocas de estudo, de 1991/1992 até 1996/1997. A metodologia tomou dois caminhos diferentes, independentemente dos métodos de previsão utilizados. Uma em que tinha em conta apenas os dados estatísticos da época anterior à que o jogo em que se pretende determinar o vencedor. A outra era juntar às estatísticas da época anterior, como no na metodologia anterior, os dados estatísticos da presente época até ao dia do jogo que se pretende determinar o vencedor. Foram utilizados quatro modelos para prever os resultados, Regressão Linear, Regressão Logística, Redes Neuronais e SVM. O primeiro modelo foi o que apresentou melhores resultados, podendo chegar aos 73% de previsões corretas para a primeira metodologia e 72% para a segunda metodologia. Nenhum dos outros modelos foi capaz de chegar aos 70% de acerto. Neste trabalho foi também apresentado um método para determinar quais os melhores jogadores e as suas posições com base nas suas estatísticas.

Para Puranmalka [15], o caminho a seguir para a previsão correta dos vencedores dos jogos da NBA é utilizando SVMs, com auxílio de estatísticas não só da equipa, como é exemplo a frequência de lançamentos de três pontos, mas também com estatística de jogadores, mais precisamente dos sete que mais tempo de jogo têm, com maior ênfase para a estatística de eficiência e tipos de jogadas que estes fazem, exemplo a percentagem real de lançamento (TS%). Os resultados deste trabalho situam-se entre 67% e 73% de acerto, ao longo de 10 anos, de 2003 a 2012.

### 2.1.2 PREVISÃO DE RESULTADO

Na pesquisa efetuada por Boldrin [2] para determinar a ordem da tabela classificativa no final da época, foram utilizados três diferentes modelos, todos utilizando modelos de Poisson. O primeiro modelo assume que a performance das equipas é independente do seu adversário. O segundo assume a dependência entre as equipas que participam no jogo. E o terceiro foi construído com base no segundo modelo, mas adicionando o fator casa e fora. Para suportar os modelos, foi construído um conjunto de dados cobrindo 5 épocas, desde a época 2011/2012 até à de 2014/2015. Sendo que para cada jogo foram recolhidas diversas informações, as equipas participantes, casa e fora, os golos marcados por cada uma delas e o vencedor. E ainda 3 parâmetros auxiliares para o cálculo do λ de cada equipa (Visitado e Visitante). Todos os modelos são utilizados para determinar a probabilidade de todos os resultados possíveis, com λ a

representar o número de golos que cada equipa marca, com um máximo de 6 golos por equipa, concluindo-se depois a probabilidade total de cada equipa vencer ou de ocorrer um empate. O terceiro modelo foi o mais eficaz, com uma percentagem de acerto de 53,8%, em 119 jogos testados para a época de 2016/2017, na EPL

### 2.1.3 RATINGS

Para Sathe et. al. [4], onde é efetuada a previsão de resultados da EPL, o modelo que garante melhores resultados é o modelo SVM, apresentando 59,9%, o modelo de NaïveBayes apresenta um acerto de 55% e o modelo Random Forests apresenta um acerto de 50%. Neste trabalho, o conjunto de dados é composto pelas equipas intervenientes nos diversos jogos, jogos esses recolhidos por 10 temporadas, as avaliações para cada uma das equipas, através de uma média feita à avaliação individual de cada um dos seus jogadores, e foi calculado para cada uma delas um quociente entre os jogos ganhos e o número de jogos feitos, em casa por parte do conjunto caseiro e fora por parte do conjunto forasteiro.

No estudo apresentado na página web do “MIT Technology Review” [34], o modelo utilizado para fazer a previsão do vencedor do Campeonato do Mundo de Futebol de 2018 emprega a técnica de enseble de Random Forests. Utilizando classes relativas ao país, PIB per capita e à população, a classificação de cada país no ranking da FIFA [35] e ainda a média de idades dos convocados para a equipa nacional, o número de jogadores que participaram na Liga dos Campeões, competição europeia de clubes, e outras relativas à composição da seleção nacional.

### 2.1.4 PLAYOFF

Em Hoffman et. al. [12], é feita a previsão de que equipas vão ao Playoff da NBA. Para tal, foi utilizada a Análise de Componentes Principais (PCA) para saber quais as estatísticas que melhor ajudam a atingir o objetivo. De 12 estatísticas escolhidas, é concluído que há cinco que são deveras importantes: pontos marcados por jogo, pontos sofridos por jogo, percentagem de lançamentos realizados com sucesso, a diferença entre o número de perdas de bola (turnovers) defensivas, ou seja, que o adversário comete, e atacantes, ou seja que a própria equipa comete, e ainda o registo de vitórias, na fase regular, da época anterior. Com isto, foi possível classificar corretamente 26 equipas, de um total de 29, ou seja 89.64% de acerto, através do método estatístico de análise discriminante. Ao contrário dos estudos apresentados anteriormente, neste

trabalho não é calculada a percentagem de acerto com base nos jogos acertados, mas sim num resultado geral que se traduz na conquista de um objetivo de época.

## 2.2 CONHECIMENTO DE DOMINIO

Neste capítulo irá ser feita uma abordagem às características de cada desporto em estudo e que estatísticas o complementam, de forma a que os analistas façam as suas abordagens aos diferentes jogos, quer para prever vencedor, quer para analisar a performance recente de uma equipa.

Para cada uma das modalidades existem dados e estatísticas que podem ser ou não compatíveis. Na modalidade de futebol ou em qualquer modalidade praticada ao ar livre, as condições climatéricas podem ser um fator importante no desempenho de certos jogadores, o que influencia o jogo de uma equipa, podendo ou não alterar o resultado final do jogo. Por outro lado, numa modalidade que se pratica dentro de um pavilhão, em especial a NBA, não sofre com condições climatéricas, pois os pavilhões são climatizados para que as condições de jogo sejam as ideais para o desporto.

Para além do conjunto de dados e estatísticas de cada modalidade, todas as equipas são alvos de uma métrica de avaliação que tem a seguinte função:

$$avaliação = pos(t,e)$$

Sendo que *pos* é a posição que a equipa *t* ocupou no final da época anterior, *e*, à que o jogo está inserido. Quanto menor a avaliação melhor é a equipa.

### 2.2.1 BASQUETEBOL

Esta modalidade foi criada em dezembro de 1891, por um professor canadiano de educação física, James Naismith, que trabalhava na cidade de Springfield, Massachusetts, Estados Unidos da América. A criação deve-se ao facto de ser impossível fazer desporto na rua, tal como o futebol americano ou atletismo o são, devido ao frio e à existência de neve nos meses mais frios [17].

O basquetebol é acompanhado com um conjunto de regras [16] que o tornam interessante para qualquer aficionado de desporto. Dessas regras, as mais importantes são: o facto de o jogo

estar dividido em quarto partes; a número ilimitado de substituições e a possibilidade de um jogador poder ser substituído o número de vezes que o treinador pretenda; a possibilidade de o treinador poder pedir pausas técnicas, com a duração de um minuto, para corrigir a equipa ou até para descansar; haver um tempo máximo para que uma equipa realize um ataque, 24 segundos, ou seja realizar um lançamento que seja convertido em pontos ou que, mesmo que seja falhado, tenha tocado no aro que constitui o cesto (caso seja falhado e a equipa recupere a posse de bola, o tempo de ataque é reposto). Outra regra importante é o caso de a bola não estar jogável os tempos de ataque e o de jogo são parados e iniciados assim que for reposta em jogo. Outras, não tão importantes, mas que distinguem o basquetebol de outras modalidades são os critérios para faltas assinaladas e as admoestações resultantes. Um critério pode ser por falta de fair-play, que resulta em falta técnica, que neste caso a equipa perde a posse de bola, caso a detenha, e ainda são atribuídos lançamentos livres, que caso convertidos resultam na conquista de 1 ponto. Outro critério é a falta pessoal, que é quando um jogador entra em contacto excessivo com outro, na qual pode resultar a atribuição de lançamentos livres, caso o jogador atacante que sofre falta esteja num momento de lançamento ou caso a equipa a que o jogador que cometeu a falta tenha atingido o número máximo de faltas que pode fazer por cada período. Todas estas regras aumentam a competitividade das equipas, pois para além de terem mais momentos de descanso, também têm momentos para fazer ajustes táticos.

Em termos estatísticos, cada jogo tem duas *box scores*, uma correspondente a cada equipa. Uma *box score* é uma tabela, pode ser visto um exemplo na figura 1, que contém 20 estatísticas (colunas) para cada jogador que jogou nesse jogo (linhas).

**Phoenix Suns (0-1)** Share & more ▾ Glossary

| | Basic Box Score Stats | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Starters** | **MP** | **FG** | **FGA** | **FG%** | **3P** | **3PA** | **3P%** | **FT** | **FTA** | **FT%** | **ORB** | **DRB** | **TRB** | **AST** | **STL** | **BLK** | **TOV** | **PF** | **PTS** | **+/-** |
| Shawn Marion | 37:43 | 6 | 14 | .429 | 1 | 5 | .200 | 3 | 3 | 1.000 | 1 | 6 | 7 | 2 | 1 | 4 | 2 | 1 | 16 | 0 |
| Steve Nash | 34:12 | 6 | 15 | .400 | 3 | 6 | .500 | 0 | 0 | | 0 | 3 | 3 | 13 | 0 | 0 | 4 | 1 | 15 | +1 |
| Raja Bell | 32:37 | 5 | 12 | .417 | 2 | 7 | .286 | 0 | 0 | | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 1 | 12 | +1 |
| Boris Diaw | 29:34 | 2 | 2 | 1.000 | 0 | 0 | | 0 | 0 | | 2 | 3 | 5 | 4 | 0 | 0 | 4 | 6 | 4 | -16 |
| Kurt Thomas | 25:23 | 5 | 6 | .833 | 0 | 0 | | 2 | 2 | 1.000 | 1 | 7 | 8 | 1 | 0 | 0 | 1 | 4 | 12 | +3 |
| **Reserves** | **MP** | **FG** | **FGA** | **FG%** | **3P** | **3PA** | **3P%** | **FT** | **FTA** | **FT%** | **ORB** | **DRB** | **TRB** | **AST** | **STL** | **BLK** | **TOV** | **PF** | **PTS** | **+/-** |
| Leandro Barbosa | 37:19 | 9 | 14 | .643 | 6 | 8 | .750 | 6 | 8 | .750 | 0 | 1 | 1 | 4 | 3 | 0 | 4 | 4 | 30 | -6 |
| Marcus Banks | 19:06 | 4 | 6 | .667 | 0 | 2 | .000 | 0 | 0 | | 0 | 3 | 3 | 2 | 1 | 0 | 4 | 3 | 8 | -8 |
| James Jones | 12:28 | 1 | 6 | .167 | 1 | 2 | .500 | 0 | 0 | | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3 | 3 | -13 |
| Amar'e Stoudemire | 11:38 | 2 | 2 | 1.000 | 0 | 0 | | 2 | 4 | .500 | 0 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 6 | -2 |
| **Team Totals** | **240** | **40** | **77** | **.519** | **13** | **30** | **.433** | **13** | **17** | **.765** | **4** | **25** | **29** | **29** | **7** | **5** | **21** | **25** | **106** | |

**Figura 1.** *Box Score* da Equipa dos Phoenix Suns, jogo realizado a 31 de outubro de 2006 contra LA Lakers. Fonte: Basketball-Reference.com

Para além desta *Box Score*, que apenas apresenta a estatística simples, ainda existe uma outra, *Advanced Box Score*, Figura 2, que contém estatísticas mais avançadas, como é exemplo a *True Shooting Percentage* (TS%) dada pela seguinte fórmula:

$$TS\% = \frac{PTS}{2\left(FGA + (0{,}44 \times FTA)\right)} \times 100$$

| | Advanced Box Score Stats | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Starters** | **MP** ▼ | **TS%** | **eFG%** | **3PAr** | **FTr** | **ORB%** | **DRB%** | **TRB%** | **AST%** | **STL%** | **BLK%** | **TOV%** | **USG%** | **ORtg** | **DRtg** |
| Shawn Marion | 37:43 | .522 | .464 | .357 | .214 | 3.6 | 20.6 | 12.4 | 7.9 | 1.3 | 7.2 | 11.5 | 20.9 | 97 | 109 |
| Leandro Barbosa | 37:19 | .856 | .857 | .571 | .571 | 0.0 | 3.5 | 1.8 | 18.1 | 3.9 | 0.0 | 18.6 | 26.2 | 130 | 112 |
| Steve Nash | 34:12 | .500 | .500 | .400 | .000 | 0.0 | 11.4 | 5.8 | 57.8 | 0.0 | 0.0 | 21.1 | 25.3 | 105 | 119 |
| Raja Bell | 32:37 | .500 | .500 | .583 | .000 | 0.0 | 4.0 | 2.0 | 9.0 | 0.0 | 0.0 | 0.0 | 16.7 | 101 | 120 |
| Boris Diaw | 29:34 | 1.000 | 1.000 | .000 | .000 | 9.3 | 13.2 | 11.3 | 17.7 | 0.0 | 0.0 | 66.7 | 9.2 | 97 | 118 |
| Kurt Thomas | 25:23 | .872 | .833 | .000 | .333 | 5.4 | 35.8 | 21.0 | 6.2 | 0.0 | 0.0 | 12.7 | 14.1 | 151 | 113 |
| Marcus Banks | 19:06 | .667 | .667 | .333 | .000 | 0.0 | 20.4 | 10.5 | 16.8 | 2.5 | 0.0 | 40.0 | 23.8 | 77 | 111 |
| James Jones | 12:28 | .250 | .250 | .333 | .000 | 0.0 | 0.0 | 0.0 | 0.0 | 3.8 | 0.0 | 0.0 | 21.9 | 50 | 112 |
| Amar'e Stoudemire | 11:38 | .798 | 1.000 | .000 | 2.000 | 0.0 | 11.2 | 5.7 | 13.0 | 4.1 | 5.8 | 34.7 | 22.5 | 99 | 105 |
| **Team Totals** | **240** | **.627** | **.604** | **.390** | **.221** | **11.4** | **67.6** | **40.3** | **72.5** | **7.0** | **7.0** | **19.9** | **100.0** | **106.0** | **114.0** |

**Figura 2.** *Advanced Box Score* da equipa dos Phoenix Suns, jogo realizado a 31 de outubro de 2006 contra LA Lakers. Fonte: Basketball-Reference.com

As estatísticas de TS%, eFG%, 3PAr e Ftr são estatísticas relativas às preferências e performances dos lançamentos que cada jogador efetuou. As estatísticas ORB%, DRB% e TRB% representam a percentagem de ressaltos que cada jogador ganhou, sendo ORB% para ressaltos ofensivos, DRB% para ressaltos defensivos e TRB% para o total de ressaltos. A soma da ORB% de uma equipa e a DRB% da outra equipa é igual a 100%, tal como acontece com a soma das TRB% das duas equipas. AST% é usada para calcular a percentagem de lançamentos convertido com assistência do jogador. STL% é usada para determinar a percentagem de posses de bola do adversário que terminaram com roubo de bola por parte de determinado jogador. BLK% é utilizada para saber a percentagem de tentativas de lançamento de 2 pontos, por parte do adversário, que terminaram com determinado jogador a fazer um desarme de lançamento. TOV% é a percentagem de posses de bola perdidas, seja por passes errados ou por perda da bola num drible, por cada 100 posses de bola. USG% é a percentagem de tempo que cada jogador teve durante o jogo. Já ORtg mede a performance atacante de um jogador, enquanto que DRtg mede a performance defensiva de um jogador.

De seguida irão ser apresentados diversos gráficos de distribuição das classes em análise. Essas classes, recolhidas ao longo de 11 épocas, 2006/2007 a 2016/2017, são o “Mês”, o “Dia da

Semana" e "Hora" em que os jogos foram realizados e "Vitória ou Derrota" por parte da equipa da casa.

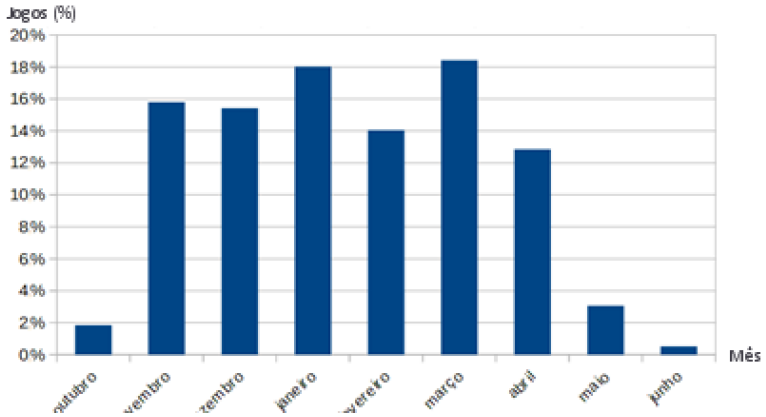


**Figura 3.** Distribuição dos jogos pelos meses

Pela Figura 3, podemos ver que os meses com mais jogos são os de janeiro e março. Tal deve-se ao facto de em fevereiro ser realizado o evento do "*All-Star Weekend*", pois ao longo de cerca de uma semana não são realizados jogos da liga, assim, cerca de 300 jogos tiveram de ser remarcados para os meses mais próximos. Em meados de abril iniciam-se os *Playoff*, o que leva a que haja uma grande quebra no número de jogos a serem realizados em abril e ainda uma quebra maior quer no mês de maio, quer no mês de junho.

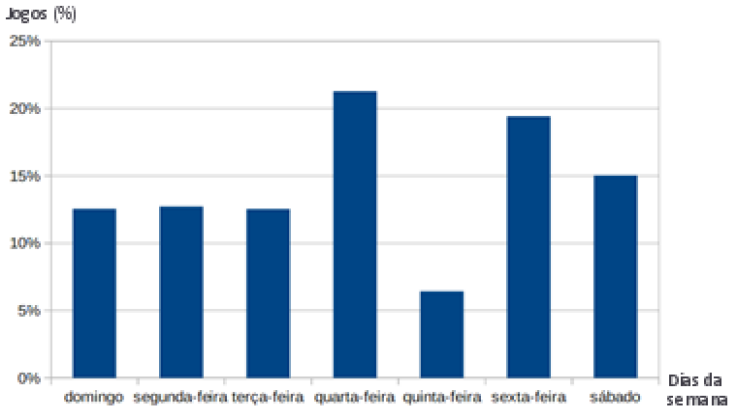


**Figura 4.** Distribuição dos jogos pelos dias da semana

Todas as equipas têm de realizar pelo menos dois jogos por semana, sendo que um deles tem de ser ao fim de semana e outro durante a semana, no entanto podem jogar quatro jogos durante o mesmo período [19]. Com suporte do Figura 4, podemos concluir que os jogos durante a semana são mais comuns às quartas-feiras e às sextas-feiras, já no fim de semana, o dia que mais jogos tem é o sábado. As quintas-feiras são os dias com menos jogos devido ao facto de as

equipas terem de descansar ou fazer grandes viagens para jogarem entre jogos às quartas e sextas-feiras ou sábados.

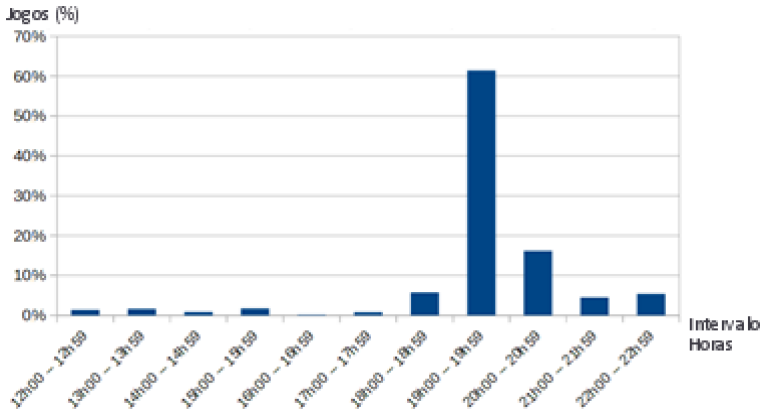


**Figura 5.** Distribuição dos jogos por horas

Devido ao número de jogos realizados durante a semana ser muito mais elevado do que ao fim de semana, tal como apresentado no Figura 4, é expectável que os jogos tenham uma hora de início depois da hora de saída do trabalho da população ativa. Assim, com apoio do Figura 5, é possível ver que mais de 60% dos jogos são iniciados no intervalo de tempo das 19h00 até às 19h59. Também é possível concluir que entre as 20h00 e as 22h59 se iniciam mais jogos do que entre 12h00 e as 18h59

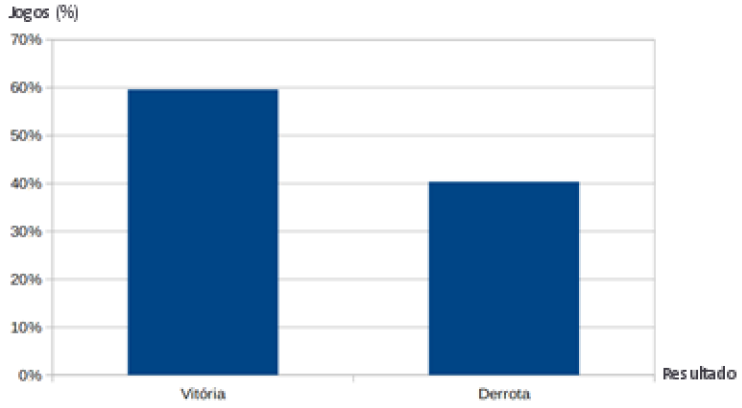


**Figura 6.** Distribuição de vitórias e derrotas com fator casa

Com base no Figura 6, é possível concluir, como esperado, que as equipas que jogam em casa ganham cerca de 60% dos seus jogos. Para este caso não são contabilizados os jogos que no fim do período regulamentar se encontram empatados, pois na NBA todos os jogos têm de ter um vencedor, mesmo que para isso seja necessário prolongamento.

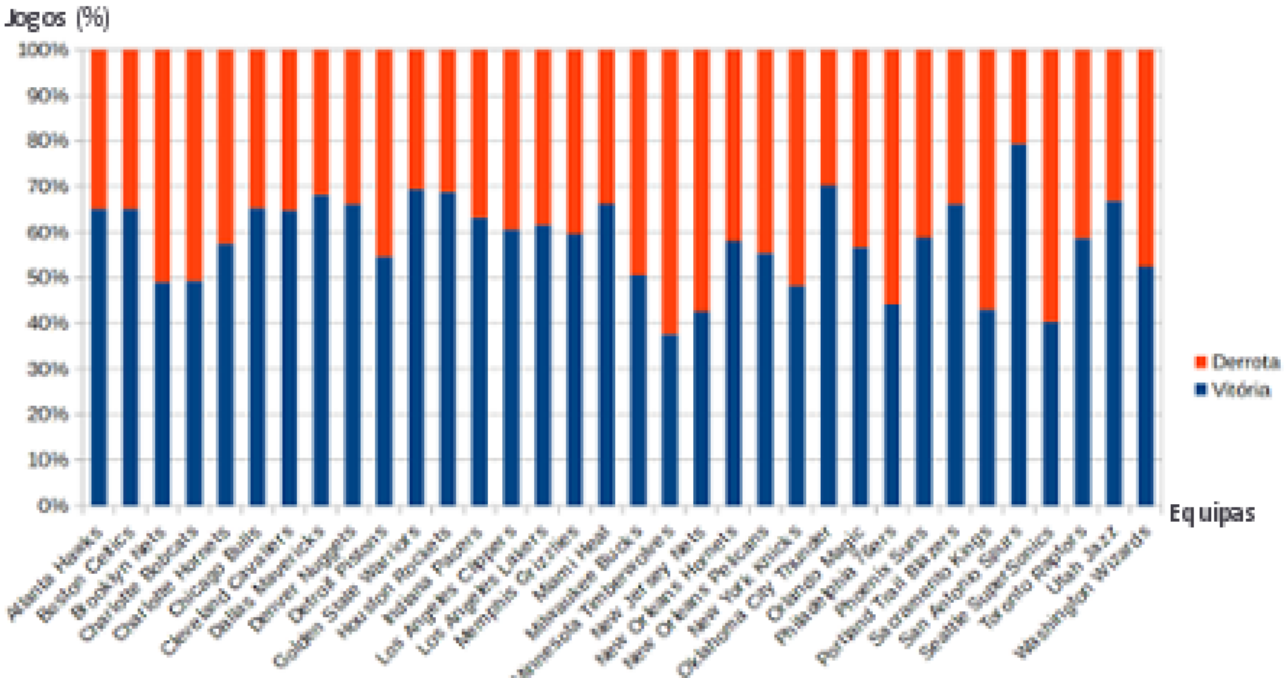


**Figura 7.** Distribuição de vitórias e derrotas com fator casa, por equipa

Pelo Figura 7, em compensação ao Figura 6, é reforçada a ideia de que a equipa que joga no seu recinto desportivo tem uma percentagem de vitória sempre superior a 50%. De notar que, a equipa dos San Antonio Spurs é a única a ultrapassar os 70% de vitórias, quando joga em casa, o que se pode traduzir num elevado número de previsão de falsos positivos. O aparecimento de falsos positivos prende-se com o facto de equipas, como a dos San Antonio Spurs, terem uma percentagem muito alto de um certo acontecimento, o que conduz a que a previsão siga a tendência regista no passado, mesmo que erradas. Existem 17 equipas que apresentam uma percentagem de vitórias abaixo dos 60%.

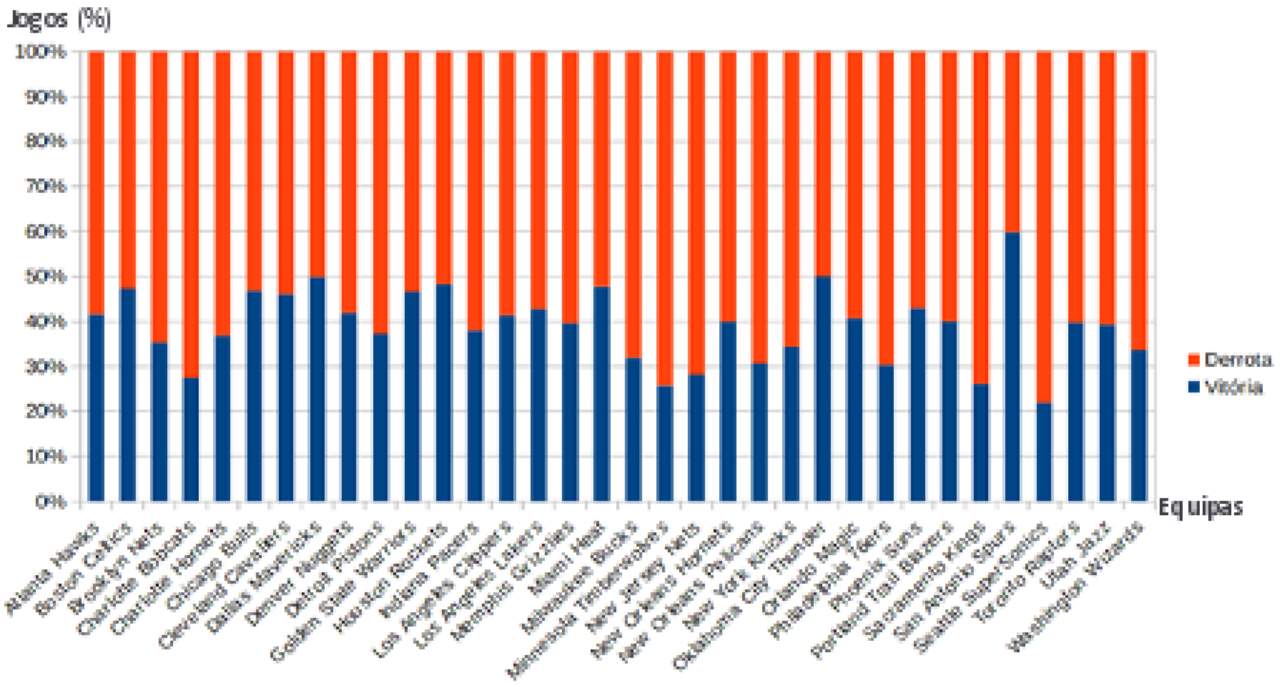


**Figura 8.** Distribuição de vitórias e derrotas sem fator casa, por equipa

As equipas têm um desempenho nos jogos fora pior (Figura 8) do que nos jogos em casa. Tal como nos jogos em casa, a equipa com mais vitórias são os San Antonio Spurs. Para as equipas com percentagens de derrotas superiores a 70% é esperado que o número de falsos positivos seja superior a equipas onde o conjunto de treino seja mais variado.

### 2.2.2 FUTEBOL

Diz-se que a modalidade do Futebol foi criada em 1863 na Inglaterra, apesar de haver relatos anteriores sobre a realização de jogos semelhantes, em diversos locais do mundo [18].

Neste desporto existe um grande acompanhamento por parte da população mundial. Tal deve-se a eventos particulares, desde as rivalidades entre os diversos clubes, como o Real Madrid CF com o FC Barcelona, até à existência de grandes personalidades que são também imagens de marcas, exemplo de Cristiano Ronaldo, que faz campanhas publicitárias quer para produtos de beleza quer de roupa. Para o sucesso, contribui também o ambiente vivido no estádio, em particular nos grandes estádios como Old Trafford, em Manchester, Inglaterra, ou o Estádio do Maracanã, no Rio de Janeiro, Brasil. E mais recentemente, toda a loucura no mercado de transferências, em que clubes pagam quantias elevadas de dinheiro por jogadores, tem aumentado o interesse no Futebol.

Para além de todas estas características, existem regras que ajudam a que o futebol seja fortemente acompanhado. O facto de o contacto físico ser permitido até um ponto em que não seja considerado uma agressão ou que de alguma forma possa resultar nalgum tipo de jogada que não tenha como intenção jogar a bola. O número de jogadores em campo e as dimensões deste tornam o posicionamento dos jogadores mais disperso, tornando a leitura do jogo mais fácil, sem que os jogadores estejam demasiado afastados entre si. Apesar de um jogo ter cerca de 90 minutos de duração, ao contrário do basquetebol, este não tem paragens no tempo devido a lesões ou outro tipo de pausas, a não ser um intervalo, que divide o jogo em duas partes, cada uma com 45 minutos. Cada equipa tem a possibilidade de realizar três substituições, sendo que um jogador que saiu não pode voltar a entrar.

Em termos estatísticos, cada jogo tem diversas estatísticas das equipas, as principais são o número de golos, número de remates efetuados, número de remates à baliza, percentagem de passes concretizados, percentagem de duelos aéreos ganhos, fintas bem-sucedidas, cortes bem feitos, percentagem de posse de bola, o número de faltas cometidas e as admoestações recebidas (cartões amarelos e vermelhos). Para além destas estatísticas por equipa, também é possível saber que jogadores participaram no jogo e as estatísticas respetivas. Essas estatísticas são por exemplo o número de golos marcados, número de remates feitos, número de remates enquadrados com a baliza, a percentagem de passes concretizados, o número de passes chave concretizados, as admoestações recebidas, se esteve nalguma substituição, para entrar ou sair do jogo.

De seguida irão ser apresentados diversos gráficos de distribuição das classes em análise para Inglaterra e Portugal. Essas classes, recolhidas ao longo de 10 épocas, 2007/2008 a 2016/2017, são o “Mês”, o “Dia da Semana” e “Hora” em que os jogos foram realizados e “Vitória ou Derrota” por parte da equipa da casa.

#### *2.2.2.1 Inglaterra*

De seguida, procede-se à demonstração e análise da distribuição das diferentes classes, no contexto inglês.

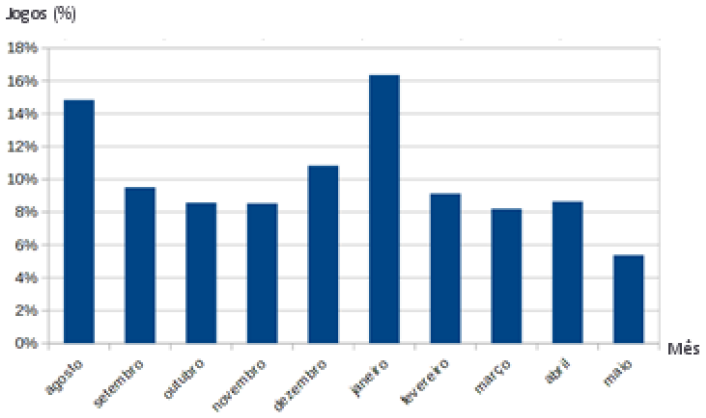


**Figura 9.** Distribuição dos jogos pelos meses

Em Inglaterra, os meses com mais jogos são os de janeiro e agosto. O mês com mais jogos é o de janeiro, com cerca de 1000 jogos jogados ao longos de 10 anos. Este mês tem a calendarização normal da EPL e ainda fases eliminatórias das taças nacionais. O mês de agosto tem mais jogos que o normal visto que para além do arranque da EPL, ainda há jogos da Taça da Liga, o que aumenta bastante o número de jogos relativamente aos outros meses. Já dezembro tem alguns jogos a mais, pois para além da planificação normal, tem ainda o chamado “Boxing Day”, no dia 26 de dezembro, que adiciona uma ronda da liga ao calendário, ou seja, mais 10 jogos nesse mês, em cada ano.

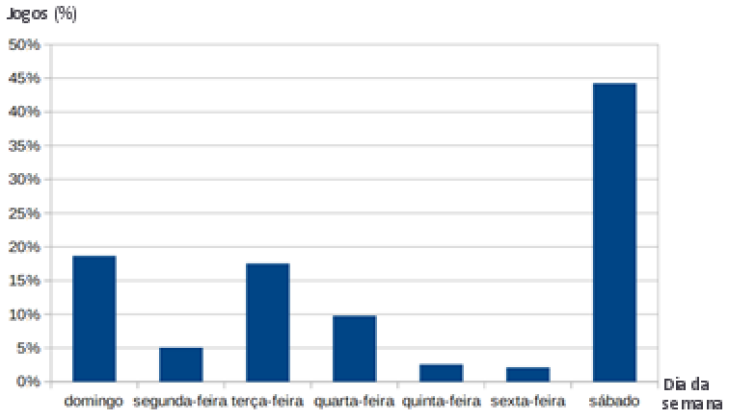


**Figura 10.** Distribuição dos jogos pelos dias da semana

Como esperado, os jogos aos sábados são os mais representativos, sendo cerca de 45% da amostra. Tal deve-se ao facto da EPL ter tradicionalmente um maior número de jogos realizados ao sábado, pois é a competição com mais jogos ao longo dos 10 anos. Nos domingos, os jogos também são maioritariamente da EPL, pois os 10 jogos semanais não são todos feitos ao sábado. Já as terças-feiras têm um grande número de jogos devido ao conjunto de jogos da Liga dos Campeões, taças nacionais e ainda alguns jogos da EPL. Nas quartas-feiras, grande parte

dos jogos são relativos à Liga dos Campeões, enquanto que nas quintas-feiras os jogos são maioritariamente da Liga Europa.

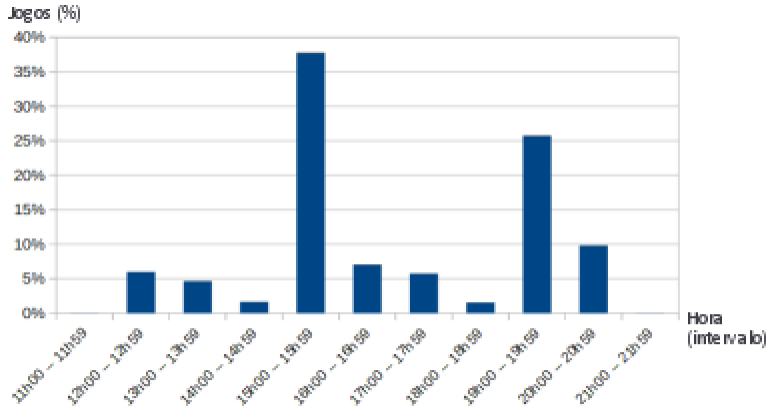


**Figura 11.** Distribuição dos jogos por horas

Como esperado, e até por motivos sociais, a maioria dos jogos inicia-se à tarde, em particular entre as 15h00 e as 15h59. Pelo Figura 11, pode ser observado que, para além do horário anterior, entre as 19h00 e as 19h59 existem uma quantidade significativa, comparativamente com os restantes intervalos temporais, pois pelas 19h45 acontece a grande maioria dos jogos da Liga dos Campeões.

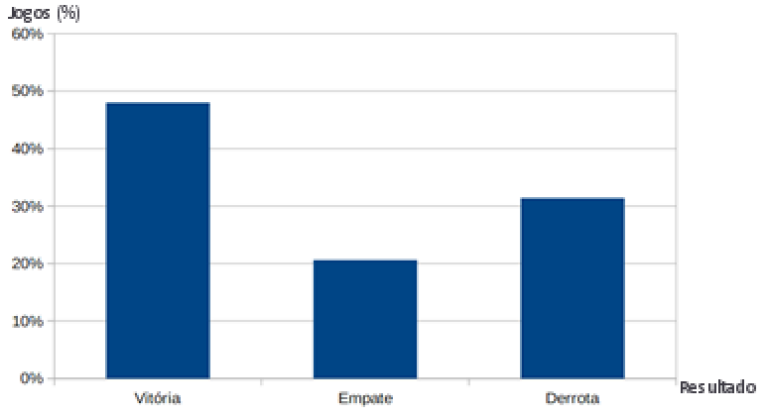


**Figura 12.** Distribuição de vitórias, empates e derrotas com fator casa

Naturalmente, as equipas que jogam em casa têm um número de vitórias significativo, perto dos 50% da amostra. Equipas como Manchester United FC, entre outras, terão uma percentagem superior aos 50%.

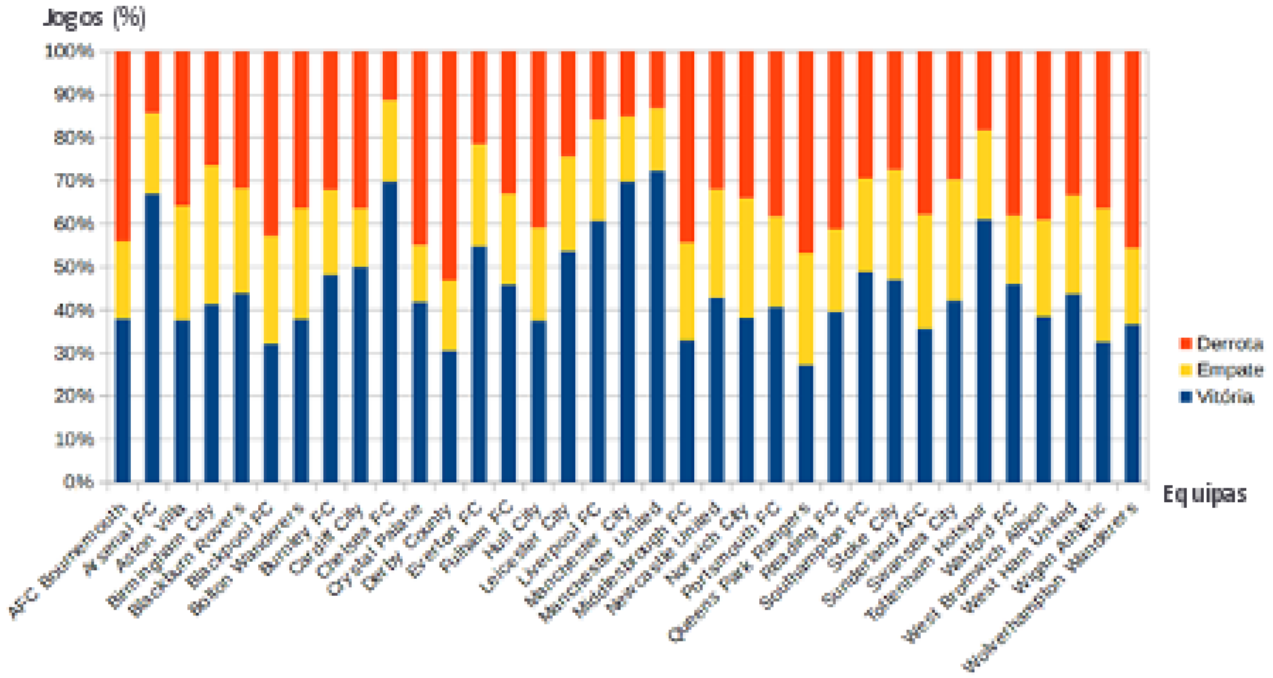


**Figura 13.** Distribuição de vitórias, empates e derrotas com fator casa, para as equipas participantes da EPL em todas as competições que jogaram

Pelo Figura 13, e ao contrário do que seria expectável, apenas 8 das 35 equipas, que disputaram a EPL entre 2007/2008 e 2016/2017, têm uma percentagem de vitórias superior a 50%. Algo que vai de encontro ao resultado apresentado pelo Figura 12. Através da demonstração do desempenho por equipa, é possível ter uma noção realista de quais são as equipas mais fortes a jogar em casa, fator que pode muitas vezes ser importante para assegurar a continuidade na competição, exemplo do Derby County que tem uma percentagem de derrotas acima de 50%, resultante de um ano na EPL que terminou com a descida de escalão. Para equipas como o Chelsea FC, Manchester City e Manchester United, que apresentam cerca de 70% de vitórias nos seus jogos em casa, o resultado esperado dos futuros jogos será expectável de “Vitória”, o que pode traduzir-se num maior número de falsos positivos.

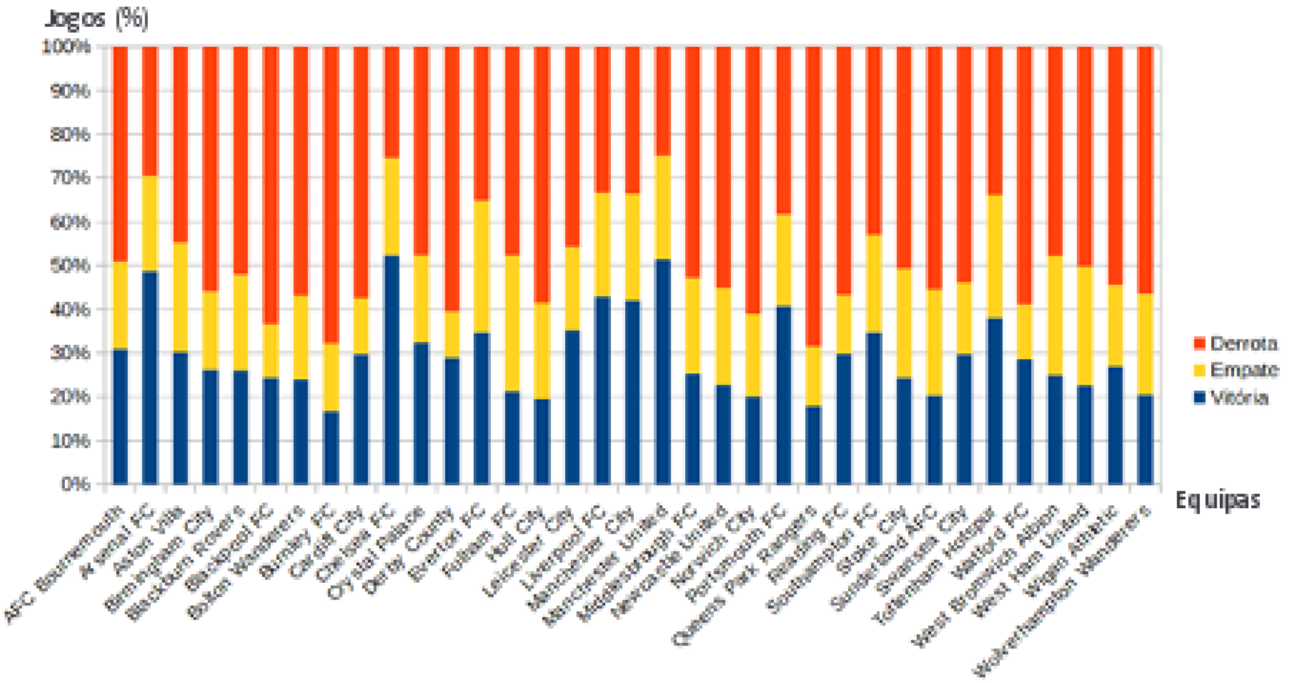


**Figura 14.** Distribuição de vitórias, empates e derrotas sem fator casa, para as equipas participantes da EPL em todas as competições que jogaram

Pelo Figura 14 é possível observar o desempenho das diversas equipas que participaram na EPL ao longo das épocas em estudo. As equipas que têm uma percentagem de derrotas acima de 50% estiveram no segundo escalão das ligas nacionais. Já as equipas com percentagem superior a 50% sagraram-se campeãs da EPL pelo menos duas vezes. Estas conclusões podem ajudar os modelos para preverem os resultados.

#### *2.2.2.2 Portugal*

De seguida, procede-se à demonstração e análise da distribuição das diferentes classes, no contexto português.

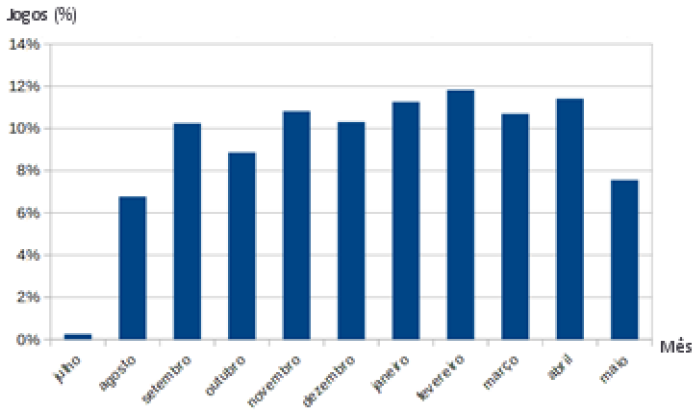


**Figura 15.** Distribuição dos jogos pelos meses

Com apoio do Figura 15, é possível observar que os meses com maior número de jogos têm são janeiro, fevereiro e março. Tal deve-se ao facto de as pausas efetuadas para as seleções nacionais jogarem serem no final de agosto e início de setembro, no mês de outubro e no mês de março. Já em dezembro, as equipas portuguesas fazem uma pause na semana do Natal, o que aumenta o número de jogos nos restantes meses. Julho tem um número muito baixo de jogos oficiais, pois apenas estão contemplados os jogos da Taça da Liga. Em agosto, o número de jogos, apesar de mais elevado, é baixo pois não são contemplados os jogos antes da terceira eliminatória da competição, pois até aí as equipas da PLP não jogam. Apesar de tudo, é apresentada uma distribuição de jogos equitativa, pois 7 meses têm entre 10% e 12% do total de jogos, individualmente.

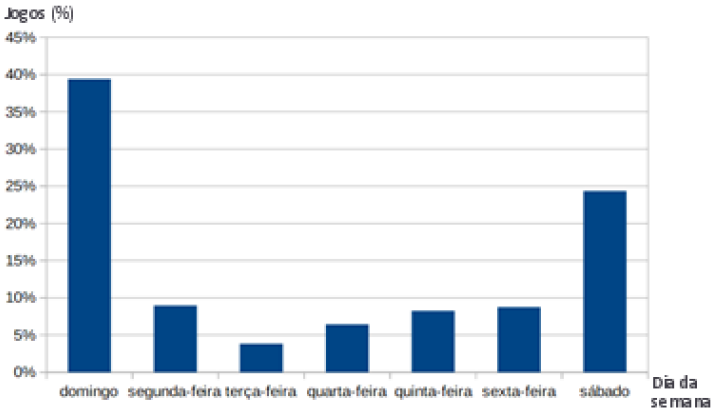


**Figura 16.** Distribuição dos jogos pelos dias da semana

Todas as jornadas da PLP são jogadas entre sexta-feira e segunda-feira, com um maior número de jogos ao domingo e ao sábado, o que é confirmado com o Figura 16. Já os jogos às terças-feiras são causa da disputa da competição europeia, a Liga dos Campeões, e nas quartas-feiras é superior, pois apesar de ter o mesmo número de jogos da Liga dos Campeões que as

terças-feiras, ainda tem taças nacionais, Taça de Portugal e Taça da Liga. O facto de as quintas-feiras terem mais jogos que as terças-feiras e as quartas-feiras são devido ao facto de os jogos da Liga Europa serem feitos nesse dia, pois Portugal tem mais vagas nesta competição europeia do que na Liga dos Campeões.

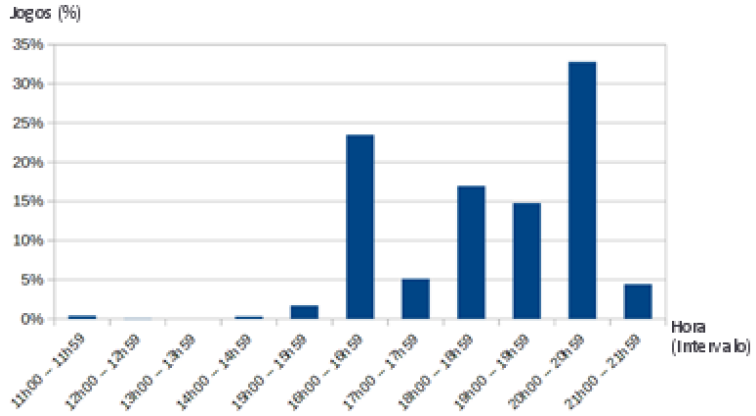


**Figura 17.** Distribuição dos jogos por horas

Os jogos das equipas portuguesas são, maioritariamente, iniciados em quatro horários diferentes. O primeiro, entre as 20h00 e as 20h59, pois o horário nobre das cadeias televisivas portuguesas inicia-se às 20h00, o que, em conjunto com o facto de o futebol ser o desporto rei em Portugal, aumenta o número de jogos a essa hora, e ainda a disputa de jogos nesse horário na Liga Europa. Em segundo estão os jogos que iniciam entre as 16h00 e as 16h59, pois é um horário criado para que as equipas com uma massa associativa menor, exemplo Moreirense FC, possa ter mais pessoas nos seus estádios. De seguida, os jogos começados entre as 18h00 e as 18h59 têm como base a política de atração de público aos estádios e ainda os jogos disputados na Liga Europa. Por último, a importância do horário entre as 19h00 e as 19h59 deve-se ao facto da disputa de jogos da Liga dos Campeões nesse intervalo de tempo.

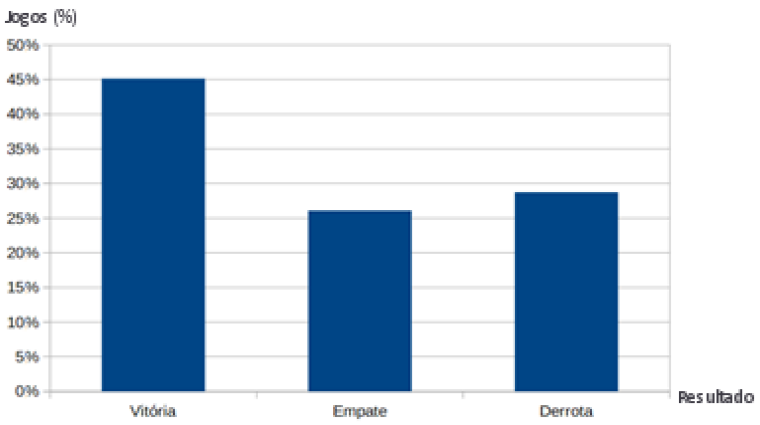


**Figura 18.** Distribuição de vitórias, empates e derrotas com fator casa

Com apoio do Figura 18, podemos concluir que apesar das equipas que jogam no seu estádio terem um número de vitórias superior ao número de empates ou ao número de derrotas, não chegam a ter 50% de vitórias nos jogos que disputam em casa. Claro que para equipas com avaliações, normalmente, superiores, exemplo do Sport Lisboa e Benfica, tal não acontece.

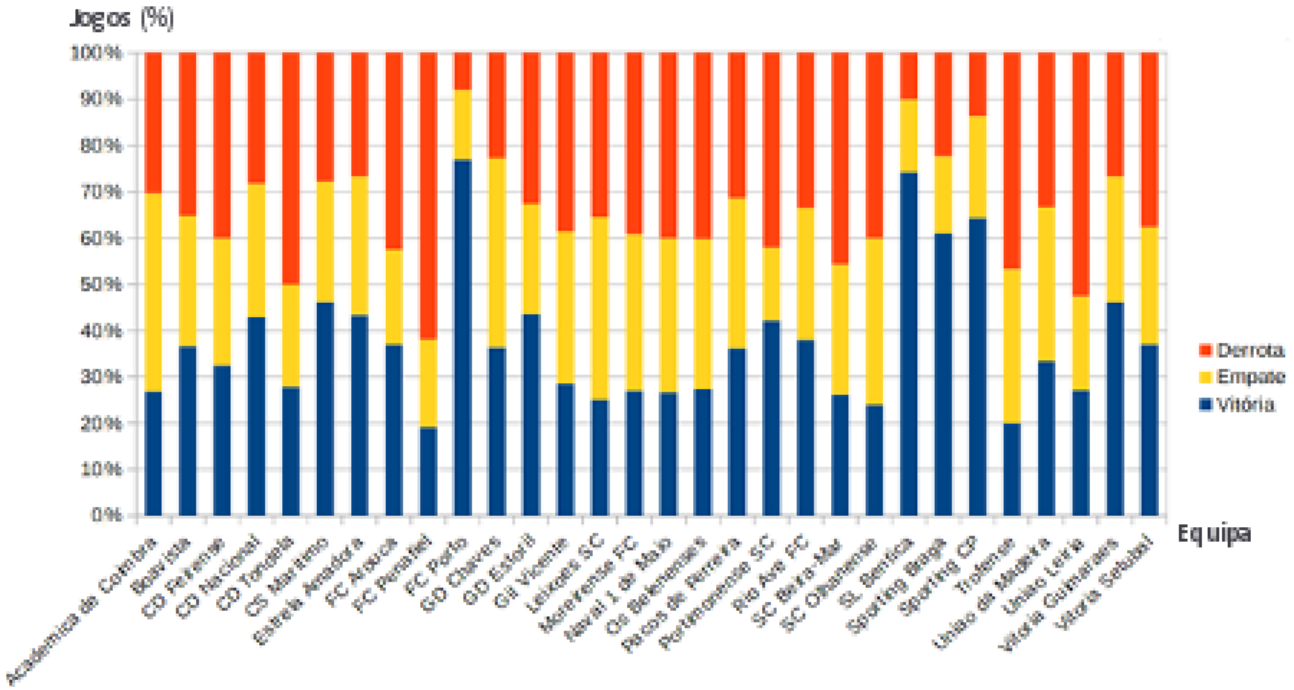


**Figura 19.** Distribuição de vitórias, empates e derrotas com fator casa, para as equipas participantes da Primeira Liga (Portugal), em todas as competições que jogaram

Pelo Figura 19, podemos concluir que em 30 equipas que disputaram a Primeira Liga há 4 equipas que se destacam em Portugal, são o FC Porto, SL Benfica, Sporting Braga e Sporting CP. As primeiras duas são as únicas que ultrapassam os 70% de vitórias a jogar no seu estádio e são também as únicas que se sagraram campeãs entre 2007/2008 e 2016/2017. Já as equipas que têm menos de 30% de vitórias em casa, no total 12 equipas, estiveram pelo menos um ano em ligas inferiores.

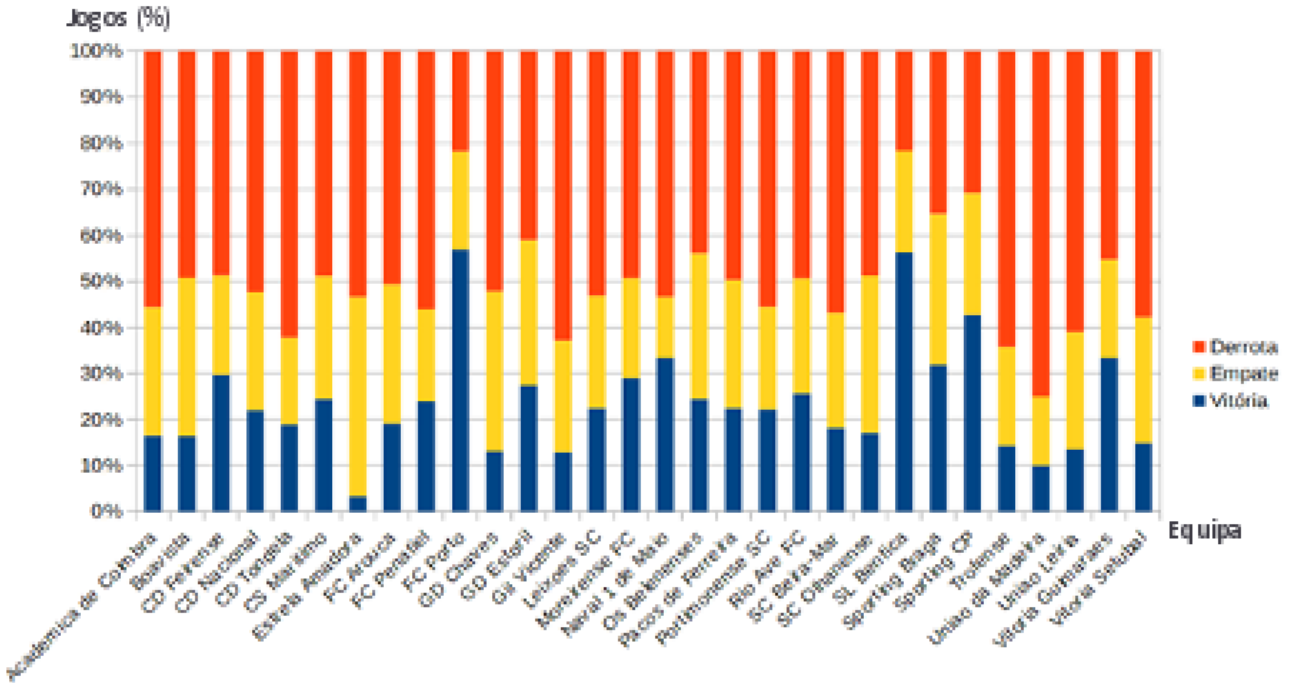


**Figura 20.** Distribuição de vitórias, empates e derrotas sem fator casa, para as equipas participantes da Primeira Liga (Portugal), em todas as competições que jogaram

Para equipas como o FC Porto e SL Benfica, analisando o Figura 20 em conjunto com o Figura 19, é reforçada a ideia de que estes dois clubes podem ser considerados de topo em relação a todos os outros em Portugal. Existem apenas 6 equipas com mais de 30% de vitórias nos jogos fora. A equipa da União da Madeira é a que tem uma percentagem de derrotas maior, no entanto é o Estrela da Amadora que tem o pior registo de vitórias e que tem o maior registo de empates.

## 2.3 RESUMO

Inicialmente, são apresentados diversos estudos feitos por outros autores de forma a prever o resultado de encontros desportivos, ao longo de um determinado tempo. Esses trabalhos incidem nas modalidades estudadas neste trabalho, basquetebol e futebol. Os diversos estudos utilizam uma grande diversidade de modelos construídos para prever os resultados dos jogos a prever. Os modelos apresentam diferentes conjuntos de dados e também de técnicas de classificação, como por exemplo Naïve Bayes, modelos com distribuição Poisson, Redes Neuronais, entre outros.

Apesar das diferenças entre as modalidades e ligas em estudo, os jogos disputados pelas equipas que a constituem são feitos em qualquer dia. No entanto, no caso do futebol a maioria dos jogos são disputados no sábado e no domingo. No basquetebol e no futebol, as equipas que jogam em casa geralmente não perdem os encontros disputados nas suas instalações, apesar de que no futebol a percentagem de empates pode ser muito próxima à percentagem de jogos ganhos.

# 3 MODELOS DE PREVISÃO

## 3.1 DESCRIÇÃO DOS DADOS

Neste capítulo serão descritos os conjuntos de dados coligidos e ainda apresentados os resultados preliminares do desempenho de modelos treinados usando as técnicas Naïve Bayes e Random Forests, recorrendo a validação cruzada (k-folds cross-validation, com k=10).

Foram criados 5 conjuntos de dados para fazer a previsão dos jogos. O primeiro conjunto de dados, “DADOS DE EQUIPA”, apenas contempla o nome das equipas e o resultado. O segundo, “DADOS DE EQUIPA E TEMPORAIS”, tem adicionalmente os dados temporais do jogo, ou seja, o mês, dia da semana e hora de início. O terceiro, “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS”, resultado de juntar a “DADOS DE EQUIPA E TEMPORAIS” a altitude e as condições atmosféricas, quando necessárias, ou seja, em jogos realizados a céu aberto. Ao quarto, “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS, E AVALIAÇÃO DE EQUIPA”, foi adicionada a avaliação de ambas as equipas, calculadas pela fórmula no capítulo 2.3, a disputar o respetivo jogo. Por último, “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA”, é a combinação de “DADOS DE EQUIPA” e dos dados adicionais de “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS, E AVALIAÇÃO DE EQUIPA”, ou seja, é o conjunto do nome das equipas e da respetiva avaliação, tal como mostrado na Tabela 1.

| | Equipas e Resultados | Dados Temporais | Condições Atmosféricas | Avaliação das Equipas |
|---|---|---|---|---|
| “DADOS DE EQUIPA” | X | | | |
| “DADOS DE EQUIPA E TEMPORAIS” | X | X | | |
| “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS” | X | X | X | |
| “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS, E AVALIAÇÃO DE EQUIPA” | X | X | X | X |
| “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” | X | | | X |

**Tabela 1.** Conjuntos de dados e respetivos atributos

### 3.1.1 DADOS DE TREINO

Cada conjunto de dados criado tem as suas particularidades e foi recolhido de fontes diferentes. Essas particularidades são existentes devido ao facto, de tal como já foi dito anteriormente, de cada modalidade ter um conjunto de estatísticas diferentes, tal como será dado a conhecer no próximo capítulo. Para além das fontes diferentes, é necessário o tratamento dos mesmos, garantindo assim que os dados não estão errados ou em branco.

#### *3.1.1.1 BASQUETEBOL*

O conjunto de dados correspondente à principal liga norte-americana para a modalidade de basquetebol é composto com um total de 14062 jogos. Todos esses jogos inserem-se no total de confrontos existentes desde a época de 2006/2007 à época de 2016/2017, ou seja, onze anos de competição. No conjunto de dados estão contidos os jogos da Fase Regular e dos *Playoffs*.

Apesar de serem fases distintas, com uma mentalidade, por parte dos jogadores e das equipas técnicas, diferentes, no geral, por vezes acontecem alguns acontecimentos não esperados na Fase Regular. Isso pode dever-se, por exemplo, ao facto de um certo jogo ser decisivo para uma equipa alcançar um lugar nos *Playoffs*.

Já no caso das competições universitárias norte-americanas em basquetebol, NCAA Division I March Madness, para masculinos e para femininos, os conjuntos de dados com as informações relativas às equipas, aos jogadores e aos jogos que foram realizados, quer na Fase Regular que em torneios secundários, foram disponibilizados pela NCAA.

Assim, o trabalho a realizar prendia-se com perceber que atributos seriam de interesse ter no conjunto de dados para treino e para teste. Alguns desses atributos foram utilizados diretamente no conjunto de dados para construir o modelo pretendido, como por exemplo o identificador da equipa. Por outro lado, outros tiveram de ser calculados, como por exemplo o rácio de assistências e turnovers, $racio_{ast/TO} = \frac{ast}{TO}$.

Para quaisquer estatísticas dos diversos jogadores, sejam elas simples, como o número de pontos, ou calculadas como o Player Efficiency Rating (PER) ou até o número de minutos jogados por cada jogador, foi necessário iterar diversos ficheiros com todos os eventos dos jogos disputados na fase regular das épocas 2010 a 2018. Cada um desses ficheiros, um por cada época, tem cerca de 2,5 milhões de eventos.

Para saber o número de minutos que cada jogador fez numa determinada época é necessário determinar o momento de jogo em que entrou e saiu, sendo que se iniciar ou terminar o jogo em campo não existe um evento dedicado a esse momento. De forma a determinar o PER foi necessário utilizar a fórmula criada por John Hollinger [8].

#### *3.1.1.3 FUTEBOL*

O conjunto de dados correspondente aos dois países em estudo, Inglaterra e Portugal, tem os jogos em que as equipas das principais ligas nacionais participaram, EPL e PLP, respetivamente, entre as épocas de 2007/2008 até à época de 20016/2017, nas diversas competições oficiais, seja nas competições nacionais ou nas competições europeias.

A diferença de jogos em cada conjunto é grande. São 6083 jogos no contexto inglês e 3162 jogos no contexto português. Esta diferença prende-se muito ao facto da principal liga inglesa ser constituída por 20 equipas, enquanto que a principal liga portuguesa teve durante as diferentes épocas um máximo de 18 equipas, tendo havido diversos anos com apenas 16 equipas.

## 3.2 PRIMEIRA ABORDAGEM

### 3.2.1 NAÏVE BAYES

O classificador de Naïve Bayes é um classificador probabilístico. Este classificador trabalha com probabilidades condicionadas, ou seja, com o cálculo de um evento ocorrer dado que houve um determinado acontecimento anterior. Com o Naïve Bayes é assumido que todos os atributos que descrevem os dados são condicionalmente independentes entre si [20].

Na próxima tabela é possível ver as percentagens de acerto da previsão do vencedor, utilizando para teste o conjunto de treino, depois de treinar o modelo de Naïve Bayes com o conjunto de treino. Para o futebol, há duas colunas para cada país, uma com “1X2” e outra com “12”. A primeira tem em conta todos os jogos do conjunto, quer para treino quer para previsão, já a segunda não tem em conta os jogos que tiveram como resultado final um empate. Como no basquetebol tem de haver sempre um vencedor em todos os jogos, seja apurado em tempo regulamentar ou em prolongamento, não se coloca esta questão.

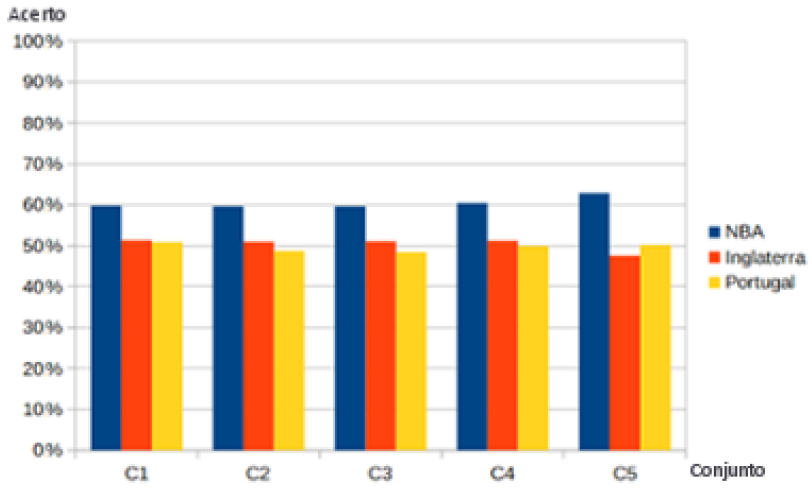


**Figura 21.** Resultado das previsões (NaïveBayes)

Pela Figura 21, é possível concluir que com as informações presentes em “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” se consegue alcançar os melhores resultados para a NBA, enquanto que “DADOS DE EQUIPA” obtém os melhores resultados para a modalidade de futebol. É também possível observar que a diferença de resultados entre “DADOS DE EQUIPA E TEMPORAIS” e “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS” é desprezável, e que estes dois conjuntos são piores que “DADOS DE EQUIPA”. Para “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA”, nos casos da NBA e das equipas inglesas a previsão torna-se mais correta, comparativamente a “DADOS DE EQUIPA E TEMPORAIS”, “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS” e “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS, E AVALIAÇÃO DE EQUIPA”, no entanto para as equipas portuguesas tal não acontece, tornando-se no pior resultado obtido no conjunto de testes. Com esta tentativa, os resultados ficam aquém do que os autores dos trabalhos apresentados no capítulo 2.

### 3.2.2 RANDOM FORESTS

O Random Forests é uma técnica de classificação por combinação de modelos (ensemble method), baseado em árvores de decisão. As árvores usadas são modelos fracos, que combinados por mecanismos de votação atingem níveis de desempenho consideráveis. A diversidades destas árvores é garantida pela seleção de parte dos atributos disponíveis em cada momento da criação das árvores.

Para o treino com este modelo, o máximo de profundidade das árvores é de duas camadas.

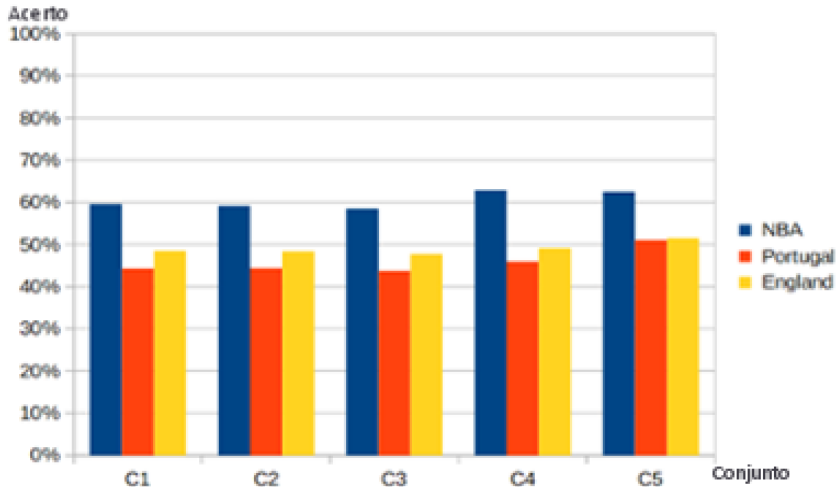


**Figura 22.** Resultado das previsões (Random Forest)

Pela Figura 22, é possível observar que com os atributos presentes em “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS, E AVALIAÇÃO DE EQUIPA” se consegue alcançar os melhores resultados, ainda que de forma pouco significativa quando comparado com os outros casos, independentemente do desporto ou do país a observar. É também possível observar que a diferença de resultados entre “DADOS DE EQUIPA”, “DADOS DE EQUIPA E TEMPORAIS” e “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS” é desprezável, apesar dos resultados em “DADOS DE EQUIPA” serem melhores que em “DADOS DE EQUIPA E TEMPORAIS” e em “DADOS DE EQUIPA E TEMPORAIS” serem melhores que em “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS”. Para o caso de “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA”, os contextos relacionados com futebol tiveram uma subida bastante expressiva no acerto das previsões, especialmente em Portugal. Quer para Portugal, quer para Inglaterra, este foi o único caso em que a previsão passou os 50% de acerto. No mesmo caso, a NBA apresenta um resultado superior a 60%, no entanto, é ligeiramente inferior ao obtido em “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS, E AVALIAÇÃO DE EQUIPA”, podendo ser desprezada

esta diferença. Com esta tentativa, os resultados ficam aquém do que os autores dos trabalhos apresentados no capítulo 2.

### 3.2.3 REDES NEURONAIS - MLP

O perceptrão multi-camada (MLP) é uma rede neuronal que contém uma ou mais camadas de unidades de processamento ou nós escondidas, que não são parte da entrada nem da saída. É de notar o alto grau de conectividade que estas redes apresentam, através das inúmeras ligações entre os nós de cada camada com a seguinte, não existindo ciclos nos grafos correspondentes. A figura 3 mostra um grafo da arquitetura de um perceptrão multicamadas com *n* entradas, duas camadas escondidas e uma camada de saída. [28]

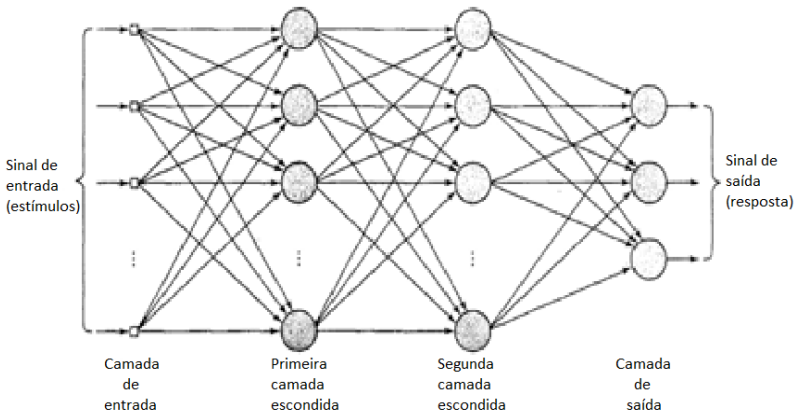


**Figura 23.** MLP com 2 camadas escondidas, 1 camada de saída

Tal como Sheela et. al. referiram no seu trabalho [25] não é necessário um número elevado de nós para construir a rede neuronal pretendida. Outro problema para além de saber aquele número é saber quantas camadas ocultas são necessárias para construir o melhor modelo. Heaton [26] diz que 2 ou menos camadas ocultas são necessárias para conjunto de dados simples e que o número de nós utilizados deverá estar compreendido entre o número de entradas e saída da rede, por exemplo.

De forma a testar qual o melhor caminho a seguir foram testadas diversas configurações para criar um modelo com MLP, em cross-validation com 10 partições. Tendo como número de camadas ocultas 1 ou 2, um número de nós em cada uma delas entre 5 e 9, 12 entradas e 1 saída, e ainda utilizando a otimização “Adam”, proposta por Kingma et. al. [27].

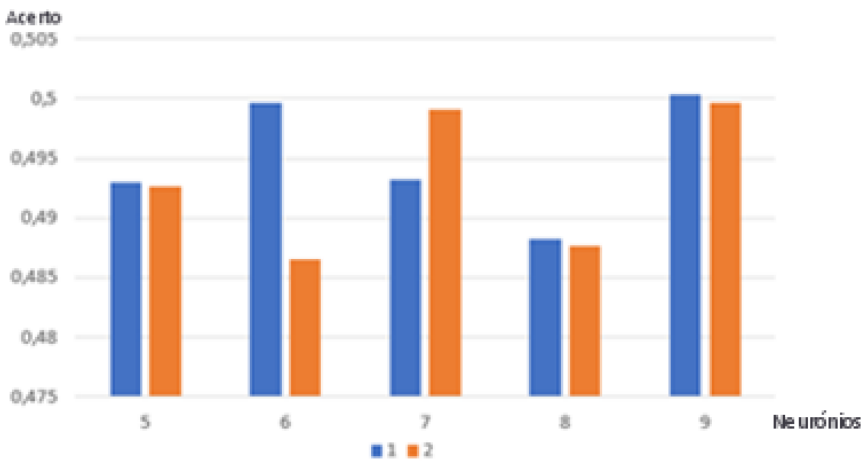


**Figura 24.** Resultado das previsões MLP

É possível concluir, com auxílio da Figura 24, que para este caso, onde apenas os jogos da Liga Portuguesa são considerados, e com um número de entradas reduzido, que não é necessário utilizar duas camadas ocultas para construir a rede neuronal pretendida. É também possível concluir que 9 é o número de nós que permite alcançar melhores resultados. Heaton diz também que outro caminho a seguir para determinar o número de nós a utilizar é escolher2/3 do total de entradas, adicionado ao total de saídas. Assim, 9 seria o número escolhido, e também aquele que nos oferece melhor precisão, único a ultrapassar os 50% de acerto.

Testando agora para todos os conjuntos de dados anteriormente propostos, de “DADOS DE EQUIPA” a “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA”, utilizando a mesma ideia que anteriormente exposto temos o seguinte gráfico:

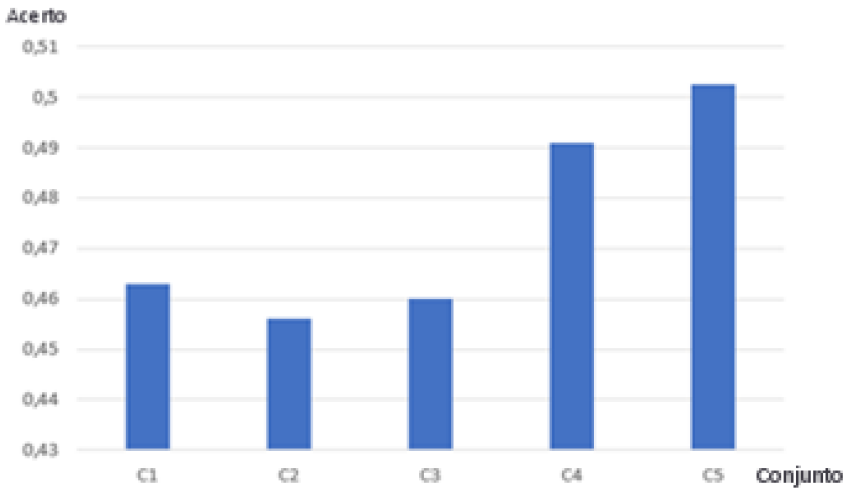


**Figura 25.** Resultado das previsões (MLP – Com diversas configurações)

Tal como visto para as Random Forests, também com MLP, com o conjunto de dados “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” a apresentar os melhores resultados, permitindo ao modelo construído ultrapassar os 50% de acerto. Este modelo permite estar equiparado em termos de performance ao construído com Random Forests para o mesmo conjunto de dados, sendo ambos os melhores para a respetiva técnica.

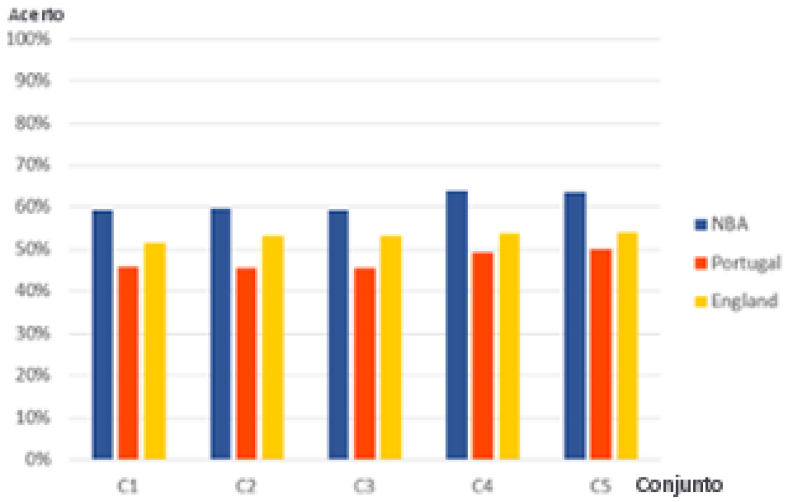


**Figura 26.** Resultado das previsões (MLP)

Pela Figura 26, é possível observar que com os atributos presentes em “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS, E AVALIAÇÃO DE EQUIPA” se alcançam os melhores resultados para a NBA. No entanto, com “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” os valores de previsão não diferem muito, diferença inferior a 0.3%. Já para o caso do futebol em Portugal e na Inglaterra, atingem-se os melhores resultados com ao utilizar os atributos referidos para “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA”. No caso dos diferentes modelos construídos para o contexto inglês não se observa uma melhoria significativo, sendo a diferença entre o melhor e o pior de apenas cerca de 2.5%, sendo com “DADOS DE EQUIPA” o pior e com “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” o melhor. No contexto português o melhor modelo é obtido ao utilizar “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA”, com cerca de 50%, o único aproximar-se desse valor. De forma geral, é possível concluir que utilizando “DADOS DE EQUIPA”, “DADOS DE EQUIPA E TEMPORAIS” ou “DADOS DE EQUIPA, TEMPORAIS E AMBIENTAIS” são obtidos os piores acertos das previsões realizadas. Com esta tentativa, os resultados ficam aquém do que os autores dos trabalhos apresentados no capítulo 2.

### 3.2.4 BASQUETEBOL UNIVERSITÁRIO – EUA – NCAA MARCH MADNESS

Em fevereiro de 2018 surgiu a hipótese de entrar em dois concursos patrocinados pela Google Cloud, em conjunto com a National Collegiate Athletic Association (NCAA), Google Cloud & NCAA ML Competition 2018-Men's e Google Cloud & NCAA ML Competition 2018-Women's, a realizar na plataforma Kaggle.

Os concursos têm contextos semelhantes pois a modalidade, o número de equipas presentes, e as fases e a forma como se disputam são as mesmas. Com as características a serem basquetebol universitário, 64 equipas a disputar cada competição e 6 rondas a eliminar, respetivamente. A maior diferença é relativa ao género, pois uma competição é para atletas masculinos e a outra para femininos. As equipas que constituem os torneios, NCAA Division I March Madness Men e NCAA Division I March Madness Women, podem variar, visto que todos os anos as equipas necessitam de se qualificar para esta fase, sendo esta fase o equivalente aos Playoff disputados na NBA.

Os dois concursos têm como objetivo não só determinar qual a equipa que ganha determinado jogo, mas também qual a probabilidade que esse acontecimento tem.

As datas importantes para o torneio masculino foram [22]:

- 10 de março – Submissão de previsões de jogos realizados nas edições jogadas anteriormente do torneio
- 11 de março – 68 equipas dadas a conhecer (Selection Sunday)
- 15 de março, até às 15h – submissão final com a previsão para os jogos desta edição

As datas importantes para o torneio feminino foram [21]:

- 11 de março – Submissão de previsões de jogos realizados nas edições jogadas anteriormente do torneio
- 12 de março – 64 equipas dadas a conhecer
- 16 de março, até às 15h – submissão final com a previsão para os jogos desta edição

Devido a estas datas, a previsão dos jogos não pode ser feita ronda a ronda, mas sim antes do torneio iniciar. Este facto faz com que o modelo não possa ser alterado, ou seja, não possibilitando o conjunto de treino com mais acontecimentos e alterar o conjunto de teste, caso necessário, sendo exemplo o facto de um determinado jogador estar lesionado.

### *3.2.4.1 Construção de Modelos*

Numa primeira fase os modelos construídos diferiam apenas no conjunto de dados fornecido e qual a profundidade das árvores de decisão que compunham o algoritmo de Random Forests.

Numa segunda fase foi testada a possibilidade de dividir as equipas pela performance das mesmas nas edições do torneio de 2010 até 2017. Sendo essa divisão feita em três partes. A primeiro com as 5 equipas com mais vitórias nesses anos, TOP 5. A segunda sendo as equipas medianas. Aqui, fazem parte todas as equipas que obtiveram mais de 1 vitória, no caso da NCAA masculina, e 4 vitórias, na NCAA feminina, excluindo, naturalmente, as equipas que já se encontram selecionadas na primeira divisão, Middle.

Por último, a terceira divisão tem as restantes equipas, ou seja, todas as equipas que tiveram 3 ou menos vitórias durante o período selecionado, Bottom. Foi construído um modelo de previsão em que era calculada a probabilidade de determinada equipa vencer outra através do cálculo dessa previsão com uma ou duas Árvores de Decisão. Caso as equipas façam parte da mesma divisão de equipas não é necessário recalcular a probabilidade, caso seja diferente são calculadas duas probabilidades diferentes, utilizando a média das duas como probabilidade final. As três Árvores de Decisão criadas são treinadas consoante o contexto dessas equipas e com profundidades e atributos diferentes entre elas e adequados a cada caso, Diagrama 1.

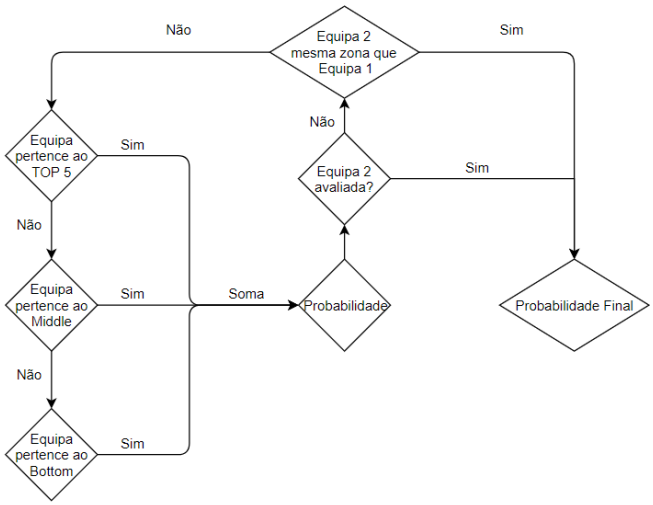


**Figura 27.** Algoritmo para cálculo de probabilidade de vencedor

Numa terceira fase, esta apenas utilizada no torneio de basquetebol universitário feminino, foi usado o algoritmo AdaBoost, com RandomForests, sem que fosse tida em conta a diferença de performance entre as equipas.

Em qualquer um dos modelos testados, a probabilidade é calculada sempre de forma a saber qual a probabilidade da equipa com um identificador mais baixo ganhar à equipa com identificador mais alto.

#### *3.2.4.2 Ficheiro de Submissão*

Para cada um dos concursos podíamos escolher 2 ficheiros, formato .csv, que continham todas as possibilidades de jogos que pudessem acontecer, ou seja $68 \times 67 \div 2 = 2278$ jogos para a competição masculina [24], visto que antes do torneio têm um último qualificador com 8 equipas, qualificando-se 4 para o torneio, ou $64 \times 63 \div 2 = 2016$ jogos na competição feminina [23].

Cada linha do ficheiro representa um jogo, sendo representado pelo identificador do jogo, id, e pela probabilidade da equipa com menor identificador ganhar à equipa com identificador maior, pred. O identificador de jogo resulta da concatenação da época, identificador de equipa mais baixo, seguido do identificador de equipa mais alto.

Exemplo:

```
id,pred
2018_1104_1112,0.244857862099
2018_1104_1113,0.567567567568
2018_1104_1116,0.466843893073
…
```

A escolha dos ficheiros a submeter para avaliação prendeu-se com a identificação do modelo apresentava melhores resultados com o uso de Random Forests e com o ensemble de árvores de decisão criadas de acordo com o ranking das equipas. Sendo que no caso da competição feminina para o primeiro caso foi utilizado o algoritmo AdaBoost, que se verificou obter melhores resultados na validação do modelo do que o uso do algoritmo tradicional de Random Forests.

#### *3.2.4.3 Avaliação*

Em ambas as competições, as submissões foram avaliadas por Log Loss:

$$LogLoss = -\frac{1}{n}\sum_{1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

Com $n$ a ser o número de jogos avaliados, $y_i$ a ser 1, caso a equipa com identificador mais baixo ganhe, ou 0, caso perca, $\hat{y}_i$ é a probabilidade de vitória calculada para a equipa com identificador mais baixo, e a ser utilizado um logaritmo natural (base e).

Esta avaliação é melhor quanto menor for o seu valor. No entanto, esta avaliação não traduz o número de jogos corretamente previstos, contrariamente ao que se pretende neste trabalho. Assim, irão ser fornecidas quer a avaliação $LogLoss$, quer a percentagem de vencedores corretamente previstos para ambos os concursos.

#### *3.2.4.3.1 Google Cloud & NCAA ML Competition 2018-Men's*

O gráfico 24 mostra o valor de *LogLoss* e a percentagem de acerto, precisão, de cada um dos modelos, onde o eixo vertical à esquerda representa o valor referente ao *LogLoss*, quanto menor melhor e o eixo vertical à direita é a percentagem de acerto de cada um dos modelos.

Como é possível concluir pela Figura 28, o melhor modelo, Modelo 1, usa a técnica de Random Forests, com máximo de profundidade das árvores de 3 e com os seguintes atributos, identificadores das equipas, seeds das equipas e a sua diferença e o rácio de assistências e turnovers de cada uma das equipas, e obteve-se um erro de 0.602905, o que proporcionou a seguinte classificação, 361 de 934 (Top 39%). A percentagem de acerto no vencedor dos jogos foi de $69.84\%$. Assim, este resultado ultrapassa a maioria dos estudos feitos pelos autores presentes no capítulo 2.

Já o modelo submetido com a divisão das equipas pela sua performance, Modelo 2, obteve um erro de 0.613724 e uma percentagem de acerto no vencedor dos jogos de $66.67\%$.

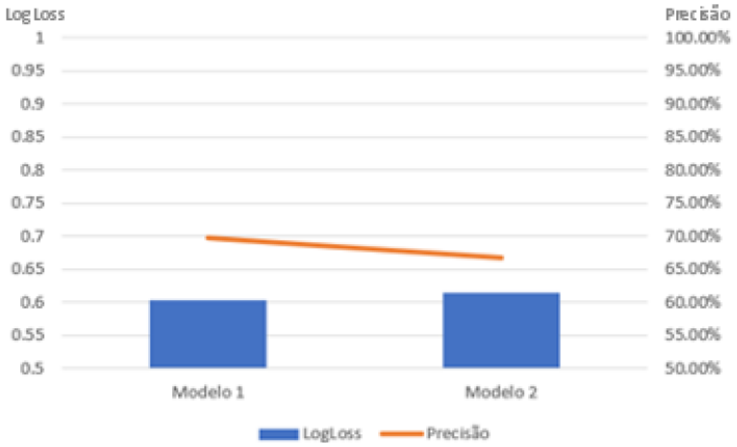

**Figura 28.** LogLoss e Precisão dos Modelos para NCAA 2018 – Men's

Para as cinco equipas com mais vitórias entre as edições de 2010 e 2017 a árvores de decisão criada tem dois nós de profundidade, utilizando a diferença do lugar de qualificação das equipas, a diferença de pontos marcados e sofridos por parte da equipa com maior identificador e ainda o número de ressaltos ofensivos da equipa com identificador mais baixo.

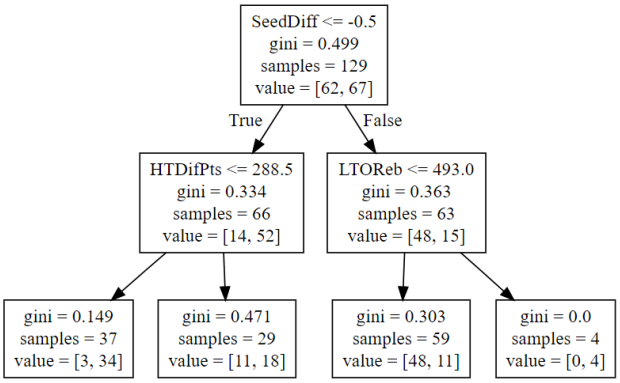


**Figura 29.** NCAA - Árvore de Decisão para equipas TOP 5

Tal como na árvore anterior, a árvore de decisão criada depois de treino para as equipas medianas, com mais de 1 vitória nas últimas 8 edições da NCAA, as características de diferença do lugar de qualificação das equipas e o número de ressaltos ofensivos da equipa com identificador mais baixo estão presentes. As diferenças desta árvore (Imagem 2) comparativamente com a anterior (Imagem 1) são relativas à profundidade, 3 nós de profundidade, o treinador da equipa com identificador mais pequeno e o uso do rácio de assistência e turnovers da equipa com identificador mais elevado.

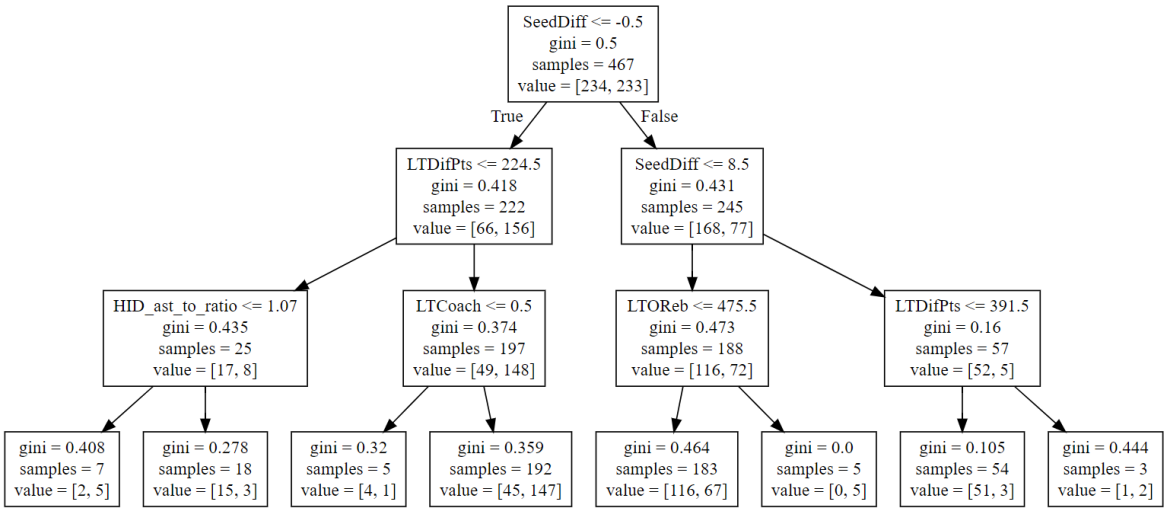


**Figura 30.** NCAA - Árvore de Decisão para equipas com performance mediana

Para a última árvore deste ensemble de árvores de decisão, e treinada com os resultados das piores equipas em termos de vitórias, todas as que alcançaram menos de duas vitórias, desde 2010 na NCAA, a profundidade volta a ser de 2, e o uso da diferença de pontual da equipa com identificador menor também é utilizada. O atributo de diferença pontual para a equipa com identificador maior é também utilizado, tal como o número de vitórias da equipa com menor identificador.

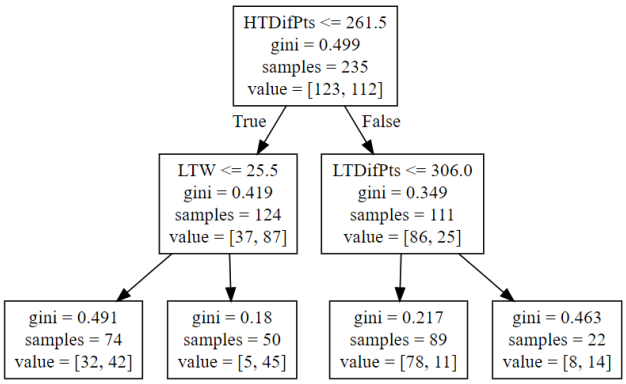


**Figura 31.** NCAA - Árvore de Decisão para equipas menos vitoriosas

Nesta edição, houve diversos resultados surpreendentes. São exemplo o facto de ter sido a primeira vez na história da competição que uma equipa com lugar de qualificação (seed) número 16 ter ganho à equipa com o lugar de qualificação número 1.

As avaliações traduzem de forma igual qual o melhor modelo, sendo o primeiro descrito o que obtém melhor avaliação pela fórmula $LogLoss$ e pela percentagem de acerto no vencedor dos jogos.

A performance atingida na fase de validação foi de 0.66 no modelo que utiliza Random Forests e de 0.79 no modelo que difere as equipas pelas suas performances. Estes números são superiores aos resultados apresentados para a edição do torneio desta última época.

### *3.2.4.3.2 Google Cloud & NCAA ML Competition 2018-Women's*

O gráfico 25 mostra o valor de *LogLoss* e a percentagem de acerto, precisão, de cada um dos modelos, onde o eixo vertical à esquerda representa o valor referente ao *LogLoss*, quanto menor melhor e o eixo vertical à direita é a percentagem de acerto de cada um dos modelos.

Como é possível concluir pela Figura 32, o melhor modelo usa a técnica AdaBoost, com Classificador de Árvores de Decisão, com máximo de profundidade de 2, com um erro de 0.586361, o que proporcionou a seguinte classificação, 381 de 505 (Top 76%). A percentagem de acerto no vencedor dos jogos foi de 66.67%.

Já o modelo submetido com a divisão de equipas pela sua performance apresenta um erro de 0.613288 e uma percentagem de acerto no vencedor dos jogos de 69.84%. Para as equipas consideradas as 5 melhores, a árvore de decisão apenas usa o atributo da diferença da posição de qualificação entre a equipa com identificador mais baixo e a equipa com identificador mais alto.

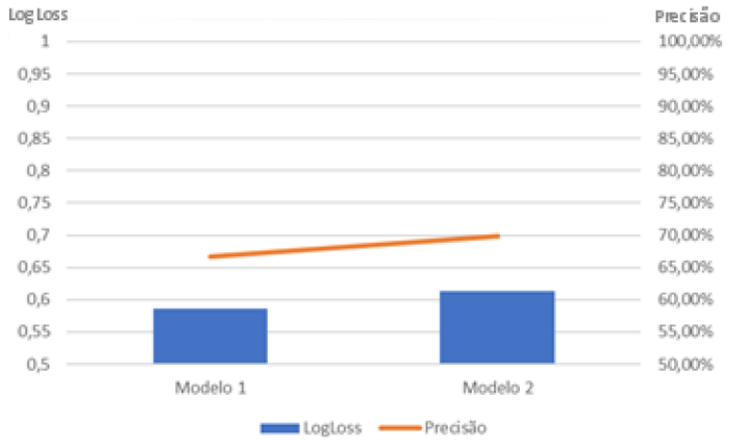


**Figura 32.** LogLoss e Precisão dos Modelos para NCAA 2018 – Women's

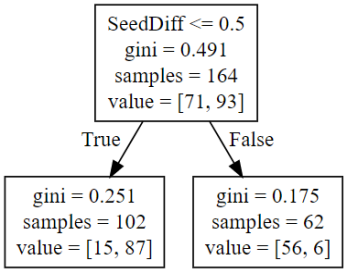


**Figura 33.** WNCAA - Árvore de Decisão para equipas TOP 5

Para as equipas medianas, a árvore de decisão utiliza também apenas a diferença de posições tal como descrito no caso anterior, no entanto tem uma profundidade de 2 unidades.

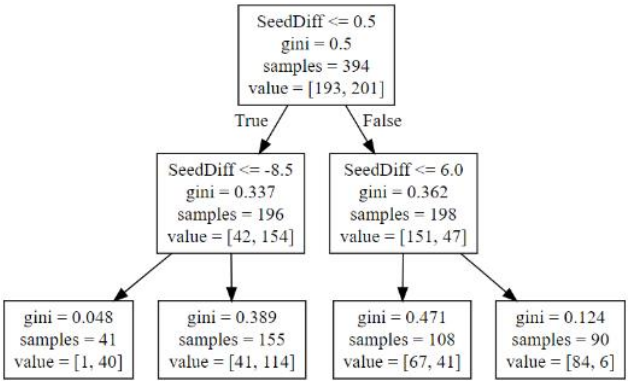


**Figura 34.** WNCAA - Árvore de Decisão para equipas medianas

Para as equipas consideradas mais fracas, devido ao número de vitórias na WNCAA nos últimos 8 anos, a árvore de decisão tem a mesma profundidade da árvore criada para as equipas medianas. Nesta árvore de decisão, as características utilizadas são a diferença dos lugares de qualificação, tal como nas duas outras árvores criadas para os outros grupos de equipas participantes no torneio, o lugar de qualificação da equipa com identificador mais baixo e ainda o rácio de assistências e turnovers da equipa com identificador maior. Assim, o resultado ultrapassa a maioria dos estudos feitos pelos autores presentes no capítulo 2.

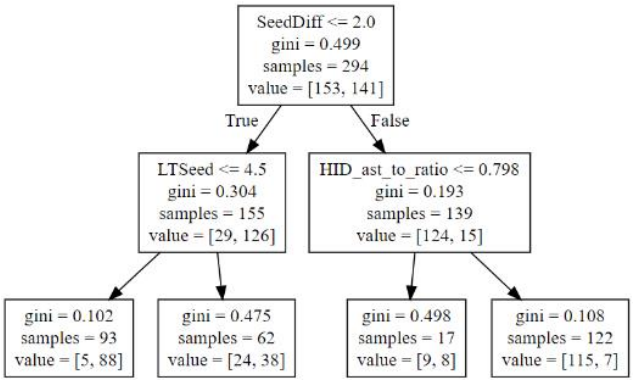


**Figura 35.** WNCAA - Árvore de Decisão para equipas menos vitoriosas

Nesta edição houve algumas surpresas nalguns confrontos, o exemplo das equipas de Buffalo a vencer duas rondas como equipa mais fraca teoricamente. E ainda as equipas de Virginia, Creighton, Florida Gulf Coast, Central Michigan e Minnesota a ganharem a primeira ronda, com nenhuma delas a ser a favorita à vitória.

As avaliações não indicam de forma igual qual o melhor modelo, sendo que o primeiro descrito o que obtém melhor avaliação pela fórmula $LogLoss$, no entanto o segundo modelo apresenta um maior número de vencedores quando comparado com o primeiro modelo. Posto isto, o melhor modelo, dado o objetivo deste trabalho seria o segundo modelo apresentado, contrariando o que seria melhor para o concurso.

A performance atingida na fase de validação foi de 0.71 no modelo que utiliza AdaBoost e de 0.82 no modelo que difere as equipas pelas suas performances. Estes números são superiores aos resultados apresentados para a edição do torneio desta última época.

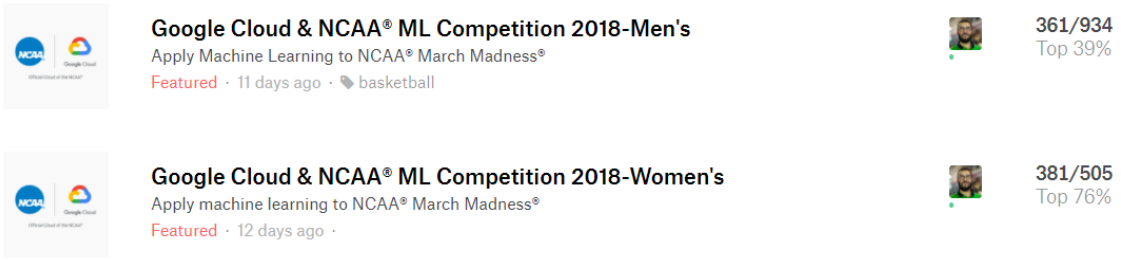


**Figura 36.** Captura de ecrã da página pessoal – visitada a 13 de abril de 2018

## 3.3 SEGUNDA ABORDAGEM

Neste subcapítulo, são apresentadas novas abordagens e novos modelos para as diferentes modalidades, cobrindo em basquetebol a principal liga norte-americana, a NBA, e em futebol a EPL e a LPL.

### 3.3.1 BASQUETEBOL SÉNIOR – USA – NBA

Neste ponto, o objetivo passa por prever os jogos da fase regular da principal liga norte-americana, NBA, relativos na presente época, 2017/2018.

#### *3.3.1.1 Trabalho Realizado*

Para poder fazer a previsão para a fase regular que terminou no passado mês de abril, foi necessário recolher quais os jogos que ocorreram e as informações necessárias para trabalhar sobre os mesmos. Informações essas que dizem respeito à classificação de cada equipa na temporada anterior, também na fase regular, e quais os jogadores que nesses encontros participaram, sendo utilizados apenas seis deles. A escolha dos jogadores assenta em dois casos. O primeiro, que contempla cinco jogadores, é referente aos jogadores que iniciam a partida, sendo por norma os jogadores que mais tempo de jogo têm. O segundo caso, onde é apenas selecionado um jogador, é escolhido o jogador que mais tempo jogou como suplente, sendo designado na NBA por "*Sixth Man*".

Para a construção do conjunto de dados de teste e para ir de encontro à utilização de jogadores a jogar em cada encontro, através de uma previsão com base no número de jogos que cada atleta tem "no cinco inicial", sendo escolhidos os cinco que mais jogos têm, e sendo escolhido para além destes o jogador com mais minutos para "*Sixth Man*", para cada equipa que irá disputar o jogo em questão.

O treino do modelo foi feito exclusivamente com jogos da fase regular de cada uma das épocas, pois assim não é desfigurado o contexto da competição, pois a fase regular é uma sequência de jogos que permeia a consistência ao longo do ano, enquanto que os *playoffs* são uma fase a eliminar, com rondas disputadas à melhor de 7 jogos.

#### *3.3.1.2 Avaliação*

Empregando a técnica de Random Forests e as classes apresentadas em “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” a percentagem de acerto foi de 59.11%. Enquanto que utilizando a mesma técnica e adicionando às classes utilizadas os cinco jogadores iniciais de cada equipa e o suplente principal, a percentagem de acerto foi de 60%. Este acerto não é melhor do que o que foi alcançado na primeira abordagem a esta competição e fica aquém dos resultados de outros autores.

Utilizando a técnica de MLP e utilizando as classes apresentadas em “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” a percentagem de acerto foi de 56.50%. Usando a mesma técnica para construir o modelo de previsão e adicionando às classes que constituem “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” os seis jogadores selecionados, segundo os critérios apresentados na secção anterior, a percentagem de acerto foi de 57.48%.

Com os 60% de precisão no modelo construído, a percentagem de acerto manteve-se no que já tinha sido apresentado no capítulo 2.

Na Figura 37, em baixo apresentada, é apresentado o número de vitórias, e no gráfico 27 é apresentado o número de derrotas de cada equipa, quer pela previsão, utilizando os resultados da melhor previsão feita, *Random Forests* com o uso dos atributos presentes em “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” e com a adição de classes respetivas aos jogadores, à esquerda, quer pelos resultados reais, à direita. É possível ver que mesmo com a adição dos jogadores, equipas que contrataram melhores jogadores não têm um número de jogos corretamente previstos como esperados, é exemplo o caso da equipa dos Minnesota Timberwolves que era previsto que atingissem apenas 16 vitórias, nos 82 jogos da fase regular, e conseguiram levar vencidos 47 jogos. Outro fator que o algoritmo na aprendizagem não contemplou foi o crescimento em termos de desempenho de jogadores que são jovens, sendo que a equipa onde é mais notória esse problema é a dos Philadelphia 76ers com uma previsão de 13 vitórias face às 52 alcançadas. Em qualquer um dos casos, tanto na equipa de Minnesota ou de Philadelphia, estas foram equipas que desde a época 2012/2013, ou seja, nos últimos 5 anos não tinham conseguido a qualificação para os *Playoffs*, e que este ano conseguiram.

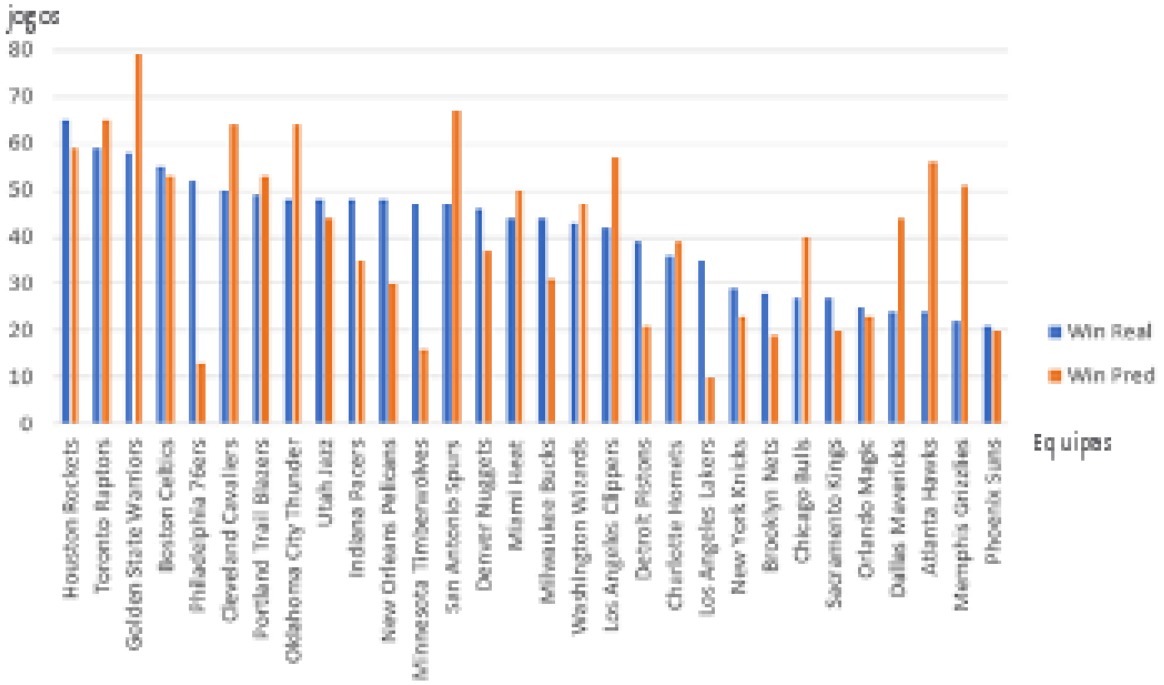


**Figura 37.** Vitórias reais e previstas – NBA

A equipa sediada em San Antonio tinha uma previsão para ter quase 70 vitórias e tal não aconteceu dado que o seu melhor jogador esteve lesionado praticamente a época inteira, o modelo ao ter em conta os jogadores que jogavam em cada um dos jogos e estando vários nomes que têm altas percentagens de vitórias nos anos que serviram para o treino do modelo, mas estando em fase descendente na carreira, resulta numa má previsão, sendo a diferença entre a previsão e o real de 20 vitórias.

A equipa de Golden State Warriors, que viria a ser novamente campeã este ano, também não correspondeu às expectativas, a previsão seria de 79 vitórias, algo que nenhuma equipa na NBA já alcançou, o que naturalmente seria uma expectativa errada à partida. No entanto devido ao cansaço acumulado das últimas épocas, pois têm estado nas "*Finals*" nos 3 anos anteriores, o que representa no mínimo 98 jogos jogados, tendo sempre ultrapassado os 100. Tendo em conta esse cansaço, o treinador procedeu a uma utilização dos jogadores suplentes de forma mais acentuada, retirando minutos aos principais jogadores

Equipas como Minnesota Timberwolves, New Orleans Pelicans, Indiana Pacers, obtiveram menos derrotas do que o previsto pois reforçaram-se com jogadores com desempenho superior à média dos jogadores da liga. Por outro lado, os Philadelphia 76ers, os Los Angeles Lakers, os Milwaukee Bucks tiveram um crescimento de performance por parte dos seus jogadores jovens, criando assim condições para diminuírem o seu número de derrotas.

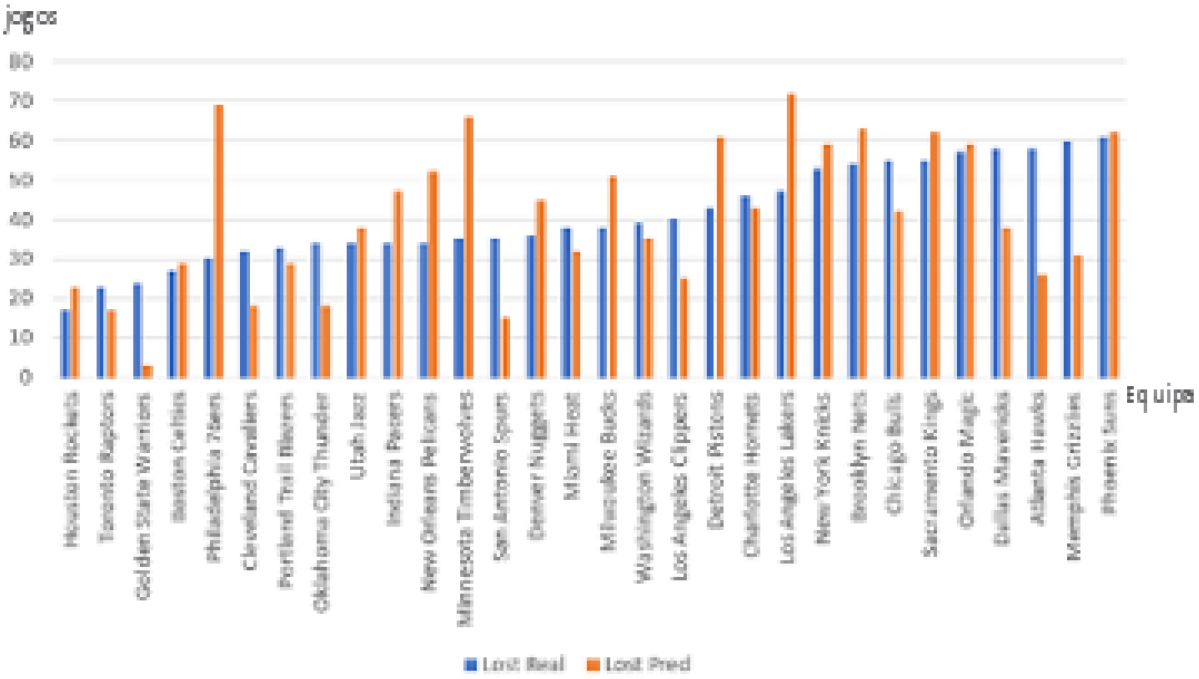


**Figura 38.** Derrotas reais e previstas - NBA

Pela previsão, as equipas que se qualificariam para os *Playoffs* na Conferência Este seriam Toronto Raptors, Cleveland Cavaliers, Atlanta Hawks, Boston Celtics, Miami Heat, Washington Wizards, Chicago Bulls e Charlotte Hornets. Na realidade as equipas presentes na última fase da época foram, Toronto Raptors, Boston Celtics, Philadelphia 76ers, Cleveland Cavaliers, Indiana Pacers, Miami Heat, Milwaukee Bucks e Washington Wizards. Ou seja, 5 idas aos Playoffs previstas de forma correta em 8 possíveis.

Pela previsão, as equipas que se qualificariam para os *Playoffs* na Conferência Oeste seriam Golden State Warriors, San Antonio Spurs, Oklahoma City Thunder, Houston Rockets, Los Angeles Clippers, Portland Trail Blazers, Memphis Grizzlies e Utah Jazz. Na realidade as equipas presentes na última fase da época foram, Houston Rockets, Golden State Warriors, Portland Trail Blazers, Oklahoma City Thunder, Utah Jazz, New Orleans Pelicans, San Antonio Spurs e Minneosta Timberwolves. Ou seja, 6 idas aos Playoffs previstas de forma correta em 8 possíveis.

### 3.3.2 FUTEBOL

Neste subcapítulo irá ser abordado o trabalho realizado e a respetiva avaliação da ferramenta na competição de futebol. Sendo que, o caso de estudo desta modalidade divide-se em duas partes, a EPL e a PLP

Com a época 2017/2018 já terminada, e pelo facto de em Inglaterra e em Portugal a modalidade com mais fãs ser o futebol, e por consequência as respetivas competições mais importantes serem as principais, EPL e PLP, então utilizou-se a ferramenta nestes contextos.

#### *3.3.2.1 Trabalho Realizado*

Os conjuntos de treino criados para as principais ligas nacionais inglesas e portuguesas reúnem dados das últimas 10 épocas, de 2007/2008 a 2016/2017. O conjunto de dados de teste, para a época desportiva de 2017/2018, em Portugal tem 306 encontros, enquanto que em Inglaterra tem 380.

Com o objetivo de poder melhorar a construção dos modelos para as duas competições foi implementado também no modelo a previsão dos jogadores que jogam no jogo a prever. Essa previsão passar por saber qual a tática que o treinador que assume as escolhas tem como preferencial e escolher os jogadores que têm mais minutos de forma a ocupar cada posição da formação escolhida pelo responsável da equipa técnica. No entanto, é necessário corrigir o caso em que é o primeiro jogo da época, onde todos os jogadores têm 0 minutos. Dessa forma é necessário fazer o levantamento de que jogadores foram mais vezes titulares na última época, quais é que já não se encontram no plantel e fazer a respetiva substituição.

#### *3.3.2.2 Avaliação*

Na principal liga portuguesa, a percentagem de acerto ao utilizar as *Random Forests* e com as classes de “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA”, sem recurso à introdução dos jogadores que constituem as equipas iniciais de cada equipa foi de 54.24%. Depois de introduzir os 22 jogadores que iniciam um jogo e utilizando novamente as *Random Forests*, com um máximo de profundidade das árvores ajustada, a percentagem de acerto é de 60.13%. Também foi testado o MLP para fazer a previsão dos jogos e as percentagens de acerto foram de 51.31% e de 53.92%, com as classes de “DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA” sem a adição dos onzes iniciais para cada equipa, e com a inclusão dos 22 jogadores nos atributos do conjunto de dados, respetivamente. Ao ultrapassar os 60% de acerto é claramente melhor do que o esperado, cerca de 51%, tal como mostrado no capítulo 2. Assim são quase 10% de diferença, sendo essa diferença positiva.

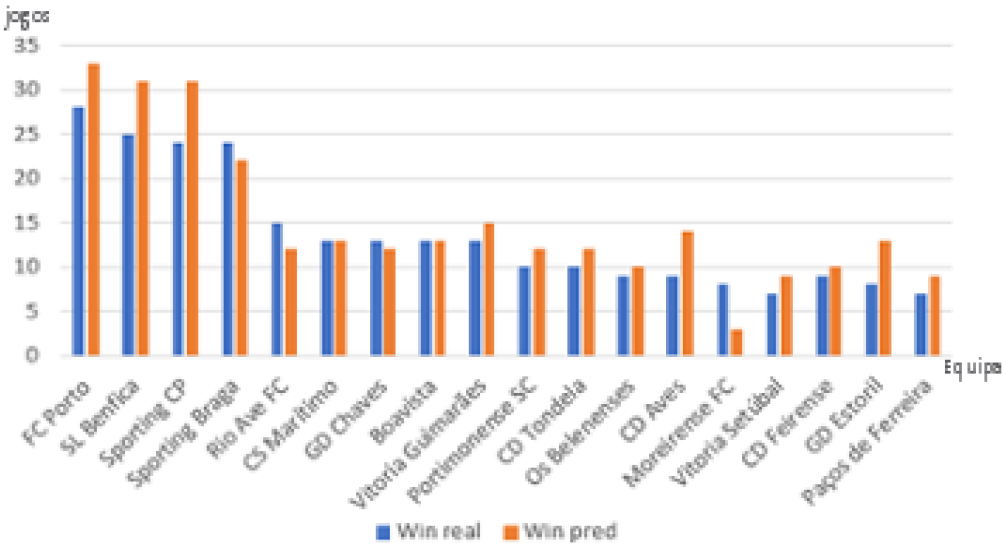


**Figura 39.** Vitórias reais e previstas – PLP

Pela Figura 39 é possível ver que para os três primeiros clubes apresentados o número de vitórias previstas é bastante superior ao que na realidade se pôde observar. Já para o caso das equipas do CS Marítimo e Boavista, o número de vitórias previstas é exatamente o mesmo. Nos casos do Moreirense FC e do Rio Ave FC, a previsão é de vitórias é claramente inferior ao que ambas as equipas alcançaram.

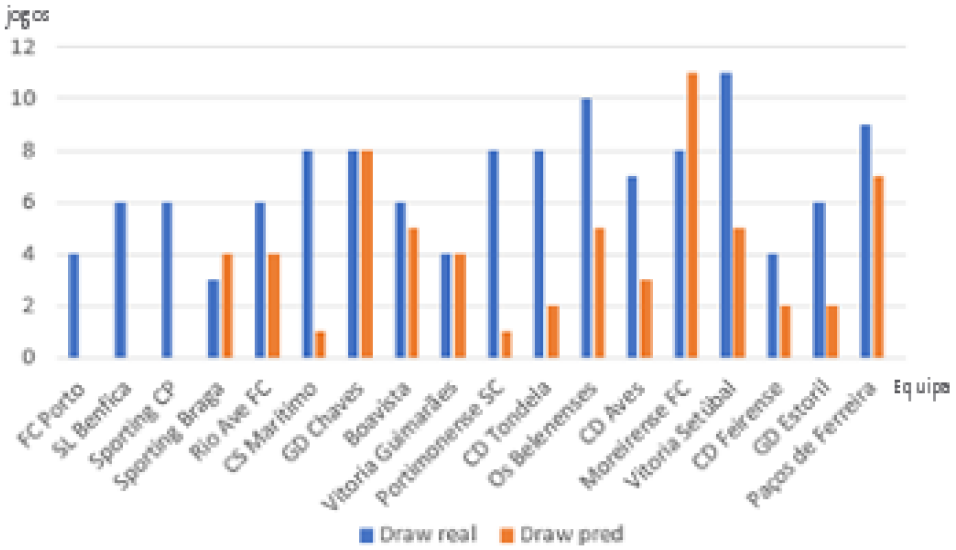


**Figura 40.** Empates reais e previstos – PLP

Com ajuda da Figura 40, apesar de terem sido previstos diversos empates, muitas das equipas obtiveram mais empates do que os previstos. No caso das 3 principais equipas portuguesas, FC Porto, SL Benfica e Sporting CP, não foram sequer previstos quaisquer empates,

muito por causa do que foi possível observar no Gráfico 17, um baixo número de empates que estas equipas tiveram nos últimos 10 anos. No entanto, o número de empates da equipa GD Chaves e da equipa do Vitória de Guimarães foi corretamente previsto. Nos casos das equipas Sporting Braga e Moreirense FC foi previsto um número de empates superior ao que aconteceu ao longo da época 2017/2018.

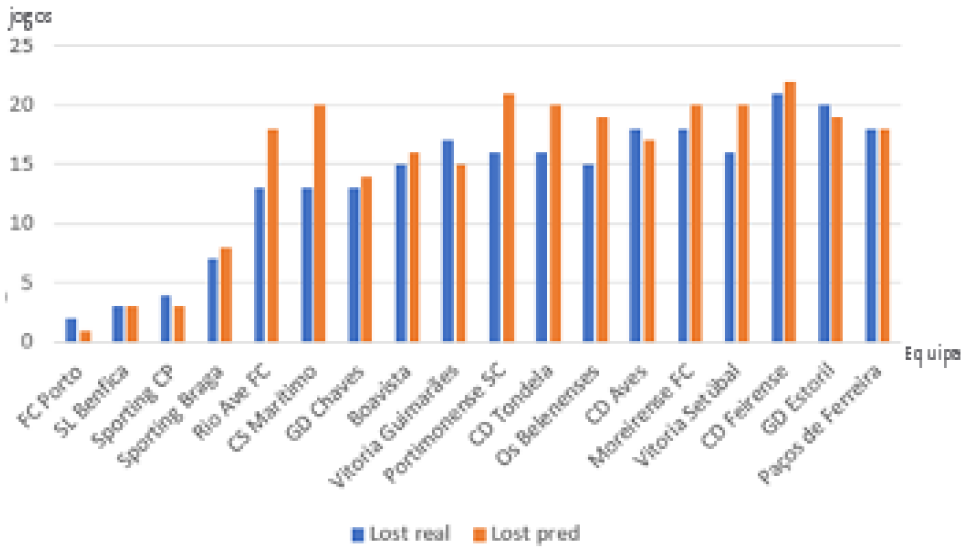


**Figura 41.** Derrotas reais e previstas – PLP

Com a ajuda da Figura 41 pode-se ver que o número de derrotas previsto é tendencialmente superior ao que foi o número de derrotas reais, muito devido ao fraco acerto do número de empates, visto no Gráfico 29. No entanto, para a equipa do SL Benfica e do Paços de Ferreiro não existe diferença entre a realidade e a previsão.

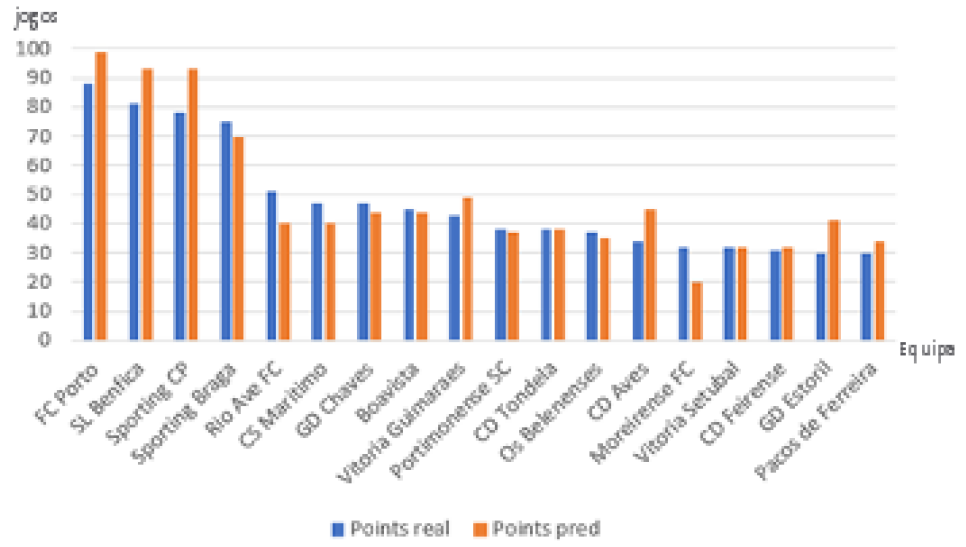


**Figura 42.** Pontos reais e previstos – PLP

Pela Figura 42 pode ser vista a classificação prevista à esquerda e a classificação real à direita. Com este recurso é possível ver que os quatro primeiros classificados são previstos corretamente apesar da diferença de pontuações previstas e reais. É também possível observar que das 18 equipas, 11 têm uma diferença pontual menor que 10 entre a previsão e a realidade. Sendo que para as equipas do CD Tondela e do Vitória de Setúbal essa diferença é mesmo nula. No caso da equipa do Rio Ave FC que teve uma época melhor do que o a previsão apontava, também obteve a segunda melhor época, a nível pontual, da sua história. O conjunto da madeira, CS Marítimo, obteve a terceira melhor marca pontual na EPL dos últimos 10 anos, sendo que nesse mesmo período a média pontual situa-se à volta dos 42 pontos. A equipa do Moreirense obteve uma das suas melhores épocas da sua história, até pelo facto de nos últimos 10 anos ter sido apenas a sua quinta participação na PLP. No lado negativo, para além dos 3 principais clubes estão o CD Aves e o GD Estoril.

Enquanto que, depois de ter sido criado o modelo, usando MLP e com as classes de "DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA", para fazer a previsão no contexto da principal liga inglesa, a percentagem de acerto é de 43.16%, menor do que os 52.10% com *Random Forests* e com as classes de "DADOS DE EQUIPA E AVALIAÇÃO DE EQUIPA". Depois de terem sido adicionados os jogadores titulares que participaram em cada um dos jogos a percentagem de acerto é de 46.84%, utilizando MLP, igualmente menor do que com para a técnica de *Random Forests* que atingiu os 53.95%. Depois de terem sido adicionados os 22 jogadores iniciais de cada jogo, foram alteradas as parametrizações de cada uma das técnicas. Este resultado de 54% é ligeiramente superior ao apresentado na primeira abordagem, onde para a EPL situava-se próximo dos 52%.

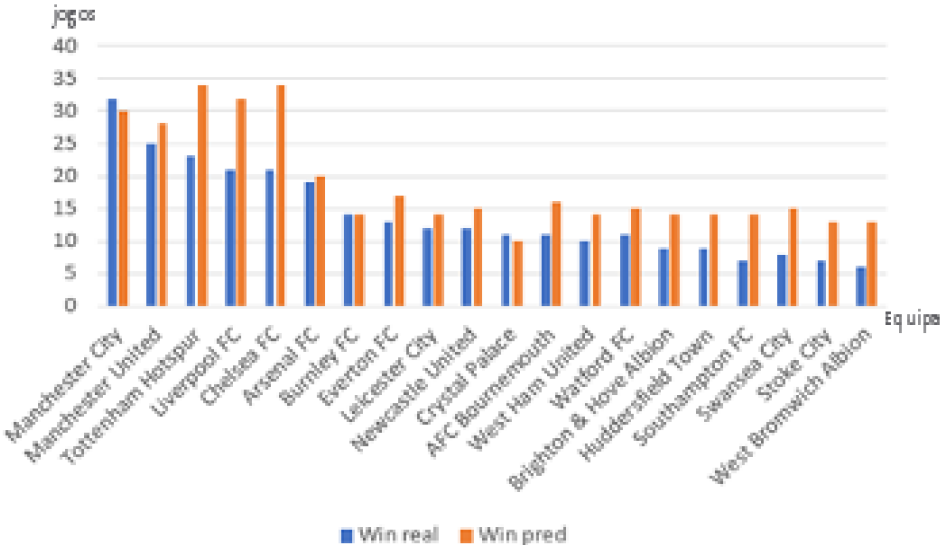


**Figura 43.** Vitórias reais e previstas – EPL

Pela Figura 43 conclui-se que equipas como Tottenham Hotspur, Liverpool FC e Chelsea FC tiveram 10 jogos previstos como vitória a mais do que o que aconteceu na realidade, ou seja cerca de 26,32% de erro. Depois há o caso da equipa do Burnley FC que obteve o mesmo número de vitórias na realidade que a previsão deu.

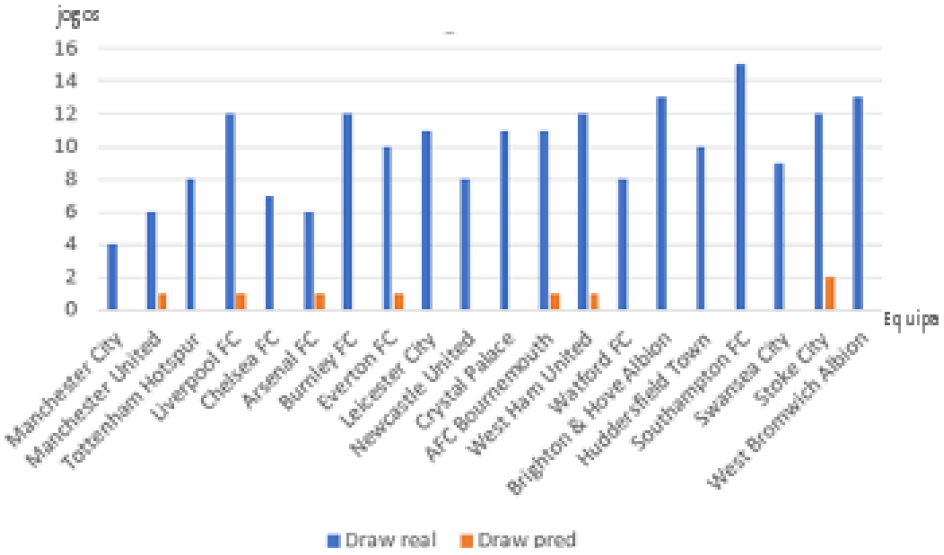


**Figura 44.** Empates reais e previstos – EPL

Um dos principais problemas na previsão da EPL é referente aos empates nos jogos disputados ao longo dos anos. Tal como é possível analisar com auxílio da Figura 44, o número de empates previstos, a laranja, é quase inexistente. É possível ver que para 9 das 20 equipas que figuraram esta época na EPL que pelo menos 10 empates não foram previstos de forma correta, o que representa um erro de pelo menos 26,32% em cada equipa.

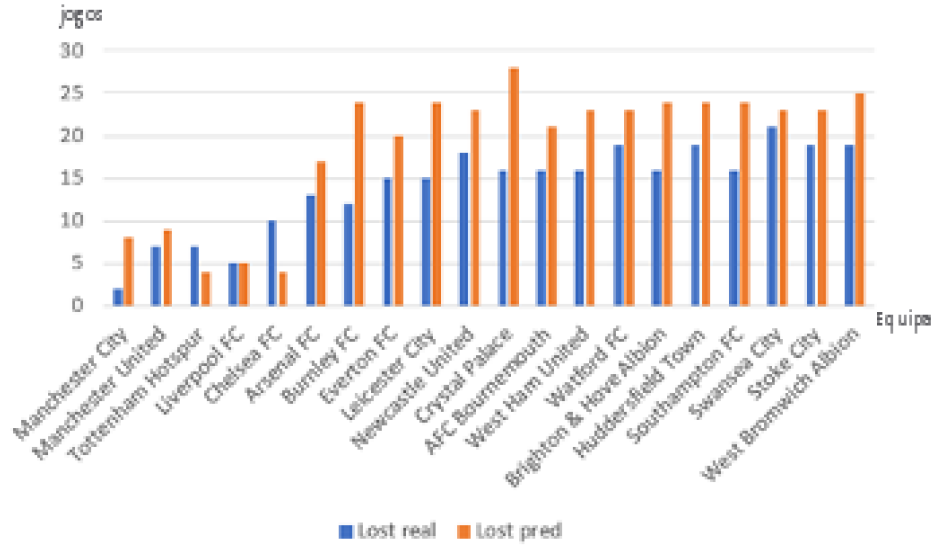


**Figura 45.** Derrotas reais e previstas – EPL

Como é possível concluir pela Figura 45 houve diversas equipas que tiveram mais derrotas do que o esperado. Mesmo com o erro associado à fraca previsão de empates, a previsão de derrotas para a equipa do Tottenham Hotspur e do Chelsea FC é inferior ao previsto, o que indica que as previsões relativas aos empates reais foram contabilizadas como vitórias, tal como pode ser visto com auxílio da Figura 43. As restantes equipas tiveram previsões de derrotas superiores às contabilizadas, maioritariamente devido à fraca precisão da previsão de empates.

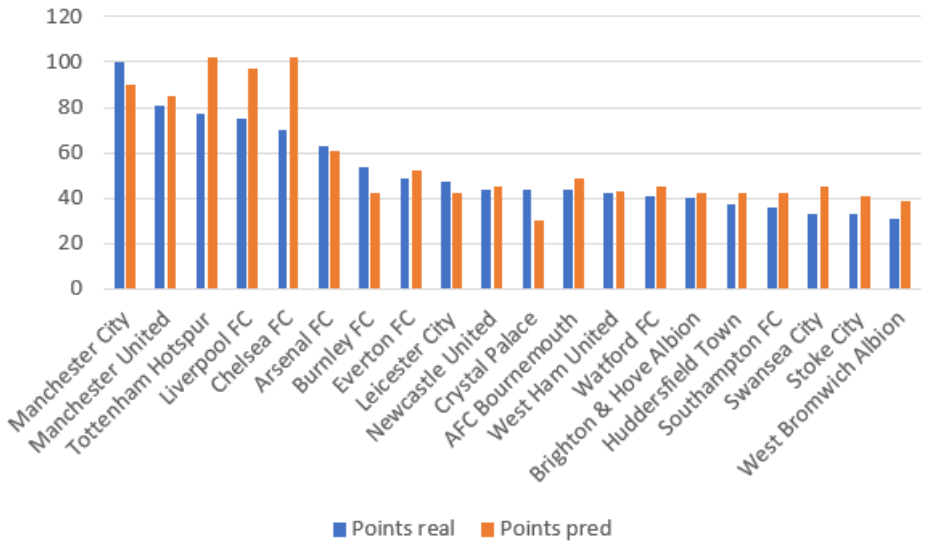


**Figura 46.** Pontos reais e previstas – EPL

Como é possível concluir através da Figura 46, as equipas presentes no Top 6 da EPL são as mesmas, quer na previsão, quer na realidade, apesar de terem uma ordenação diferente. É também possível concluir pelo mesmo recurso que das três equipas que descem de divisão (últimos 3 classificados) duas delas foram corretamente previstas.

Em 14 das 20 equipas desta edição da EPL a diferença pontual entre a realidade e a previsão não ultrapassa os 10 pontos. Das 6 que não integram esse conjunto, Tottenham Hotspur, Liverpool FC e Chelsea FC tiveram os jogos que eram empates previstos como vitórias o que fez aumentar o número de pontos previstos. Os outros 3 casos são relativos a equipas que superaram as expectativas do que era previsto para a época. O caso do campeão da edição de 2017/2018, o Manchester City, que estabeleceu um novo recorde de pontos na liga. O Burnley FC atingiu um recorde individual ao estabelecer o 7º lugar atingido esta época como melhor posição na sua história na EPL. Por último, o Crystal Palace que teve a terceira melhor época, a nível pontual, dos últimos 14 anos.

As percentagens de um modo geral têm um erro causado devido à fraca capacidade de previsão dos casos em que ocorrem empates nos jogos disputados.

Em Portugal, na PLP, houve 61 empates, dos quais 12 foram corretamente previstos, ou seja 19.67% dos empates foram previstos, ou seja, o número de empates que não foram corretamente previstos representa 33.05% da totalidade do erro. Em Inglaterra, na EPL, houve 99 empates, dos quais apenas 3, foram corretamente previstos pelo modelo criado, ou seja, apenas 3.03% corretamente previstos, o que representa 54.86% do erro total.

Portanto, pode-se considerar que pelo menos cerca de $^1/_3$ do erro, em qualquer um dos contextos é causado pela incapacidade dos modelos não serem capazes de prever os casos reais de empates.

Como é possível observar, os resultados desta segunda abordagem são melhores do que os alcançado no subcapítulo anterior. No caso da PLP, a melhoria é de cerca de 10%, enquanto que no caso da EPL, a melhoria não é tão notória, sendo à volta de 2%. Com estas melhorias, as previsões atingem os resultados apresentados pelos diversos autores no capítulo 2. Podendo-se inclusive considerar que no caso da PLP, o resultado atingido ultrapassa a maioria dos trabalhos feitos para a modalidade de futebol.

## 3.4 AVALIAÇÃO DA FERRAMENTA

A avaliação da ferramenta é feita ao trabalho efetuado no domínio da *ciência de dados*, onde a construção dos modelos, dos conjuntos de características e, por consequência, da percentagem de acerto que os mesmos providenciaram.

Assim, a avaliação do trabalho a realizado na área da *ciência de dados*, tem como base os resultados alcançados. O ideal seria ultrapassar os 70% de acerto das previsões realizadas, pois garante que qualquer cliente, que seja apostador, tenha lucros garantido. Deste modo, a ferramenta iria superar a maioria dos trabalhos realizados por outros autores. Para além disto, não devem ser descuradas todas as funcionalidades implementadas, quer as estritamente necessárias (extração, transformação, carregamento, etc…), quer as auxiliares (previsão dos jogadores que irão iniciar as partidas).

As funcionalidades necessárias para o trabalho da ciência de dados estão implementadas de forma correta, o que torna possível a previsão dos diversos jogos que compõem as diferentes competições das modalidades abordadas no trabalho. Como é possível concluir utilizando as avaliações de cada caso de estudo, NCAA March Madness, NBA, PLP ou EPL, o ideal, em termos de percentagem de acerto dos resultados dos jogos, independentemente da modalidade, é utilizar a técnica de aprendizagem *Random Forests*.

Com a Figura 47, é possível ver que as percentagens de acerto mais altas para cada uma das competições apresentadas neste capítulo. Nenhum dos casos está a baixo dos 50% acerto,

sendo que apenas 1 caso, o da EPL (Inglaterra), se situa a baixo dos 60% de acerto. De notar que para o caso referente ao basquetebol universitário norte-americano, NCAA – M para os masculinos, NCAA – F para os femininos, aproxima-se dos 70% de acerto.

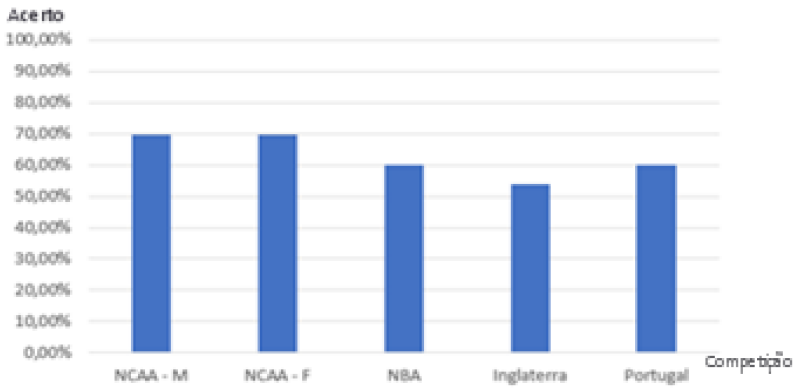


**Figura 47.** Resultado das previsões (melhores)

O tempo necessário para treinar um modelo pode ser visto na Figura 48, sendo considerado o tempo apenas dos modelos que oferecem as melhores percentagens de acerto nas previsões efetuadas. Sendo o eixo vertical o tempo em segundos. É possível concluir que os modelos testados com *Random Forests* são mais rápidos, sempre inferiores a 1 segundo, sendo contabilizado o tempo desde o treino do modelo e a previsão efetuada. Para além de serem mais rápidos, oferecem também as melhores percentagens de acerto.

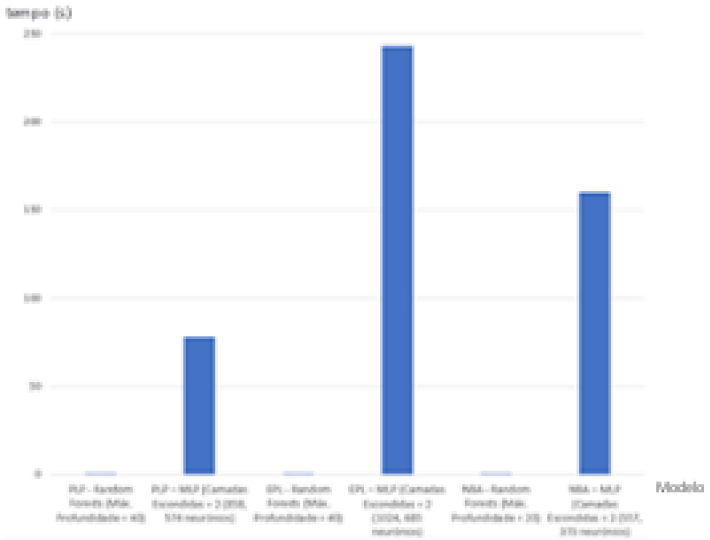


**Figura 48.** Tempos despendidos no treino e previsão dos modelos

Concluindo, quer em termos de acerto quer em termos de tempo que o modelo necessita para ser treinado e fazer a previsão, a técnica a utilizar serão as *Random Forest*, comparativamente ao MLP.

## 3.5 RESUMO

Neste capítulo são descritos os dados a utilizar na construção dos modelos de previsão. São testados diversos classificadores de treino, Naïve Bayes, Random Forests, Árvores de Decisão, MLP e AdaBoost, com diversas configurações.

Com os diferentes modelos construídos, através da variação de dados utilizados, técnicas e respetivas configurações, são diferenciadas as abordagens, separando em duas. Sendo que a segunda difere da primeira ao adicionar a previsão dos jogadores que irão participar nos jogos a prever.

Na primeira abordagem, a previsão dos jogos efetuados na NCAA March Madness, masculinos e femininos aproxima-se a 70% de acerto, superando a maioria dos resultados alcançados por outros autores, relativamente à previsão de jogos na modalidade de basquetebol. A segunda abordagem apresenta os melhores resultados em comparação aos resultados apresentados na primeira abordagem. Os resultados alcançados na segunda abordagem são de forma geral iguais aos que os autores referenciados anteriormente atingem, sendo que no caso da PLP, os resultados atingidos são melhores do que a maioria dos estudos feitos na modalidade de futebol.

# 4 PROTÓTIPO DE FERRAMENTA PARA PREVISÃO DE RESULTADOS DESPORTIVOS

O protótipo da ferramenta de previsão tem duas componentes fundamentais, cliente e servidor, que irão ser descritas no ponto seguinte. Estas duas componentes estão ligadas via Internet, através do protocolo TCP.

O protótipo é baseado num modelo de negócio do tipo *Freemium*. Este modelo leva a que certos conteúdos sejam disponibilizados gratuitamente, enquanto que outros necessitam de um pagamento, neste caso uma subscrição mensal. A parte gratuita está disponível a qualquer utilizador da aplicação terá uma previsão semanal de um jogo. A parte que necessita de subscrição está dividida em duas partes mais pequenas, um pacote em que são adicionadas todas as previsões relativas a apenas uma modalidade, enquanto que o outro pacote, com um montante a pagar pela subscrição superior, com as previsões de duas modalidades.

## 4.1 ARQUITETURA

O protótipo está dividido em duas componentes essenciais: a do servidor, “Servidor”, e a do cliente, “Cliente” (Figura 49).

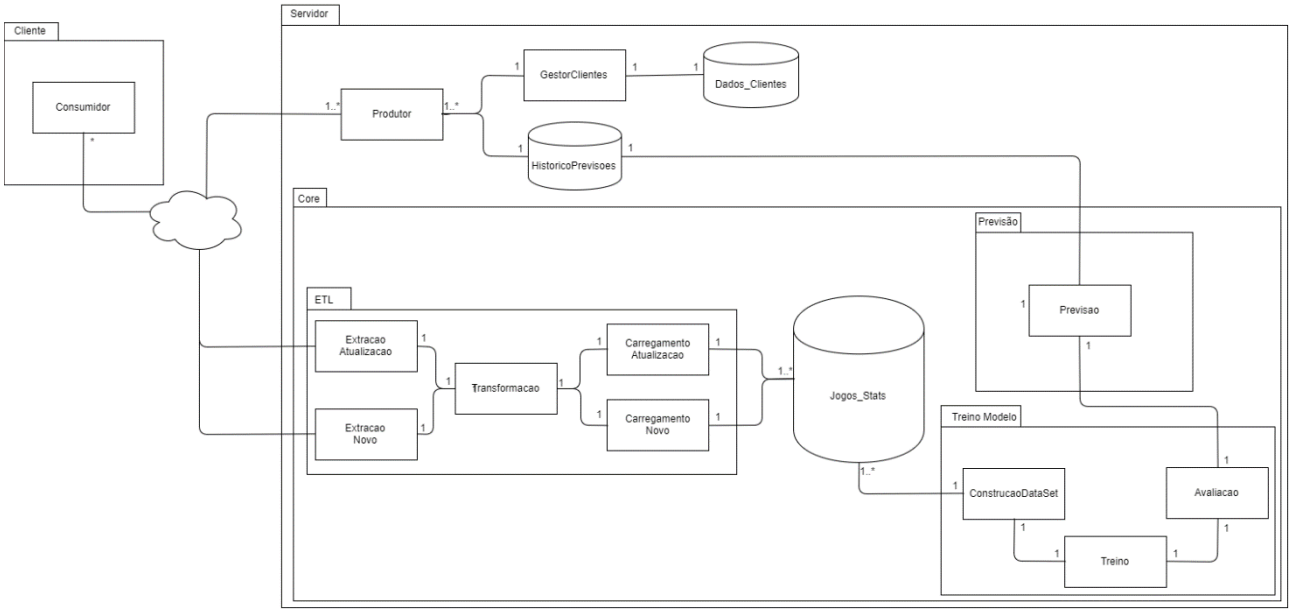


**Figura 49.** Arquitetura da solução

No "*Servidor*" existe uma partição de módulos, onde é executada grande parte do trabalho exigido na ferramenta. O módulo "*Core*" é responsável pela componente do trabalho relativa à *ciência de dados*.

A extração, transformação e carregamento, diferenciando um caso novo de um já existente, seja uma liga ou uma modalidade, são feitos no módulo de "*ETL*", em "*ExtracaoAtualização*" e "*ExtracaoNovo*", "*Transformacao*" e "*CarregamentoAtualizacao*" c "*CarregamentoNovo*", respetivamente. Na componente de extração é necessária a criação de um algoritmo de captura dos dados presentes nos sítios com os dados de cada jogo, salvo se não permitido, sendo necessário proceder a um trabalho manual. A transformação tem de ser feita depois da extração de dados para que os atributos fiquem uniformes, não sendo necessário em todas as colunas. Por último, o carregamento está divido em duas partes. A primeira, caso seja o primeiro carregamento, onde será necessário criar uma nova base de dados para a nova modalidade ou competição. A segunda, caso seja apenas uma atualização, tem de ser feita de forma cumulativa e não de substituição dos dados anteriores.

Para poder fazer o estudo relativo ao conhecimento de domínio foi necessário recolher os diversos dados e estatísticas, para os diferentes contextos. Para tal, é necessário utilizar algoritmo de extração de dados a diversas fontes, devido ao facto de as fontes serem diferentes para cada circunstância. Ou seja, para tal pode ser necessário estudar a estrutura HTML de cada site. Não fosse esta tarefa demorada, a extração de dados requer muito tempo, pois diversas páginas limitam a velocidade de acesso às suas páginas por parte de um IP que faça um determinado

número de pedidos num determinado período. Caso não seja permitida a extração de dados para fins académicos, é necessário proceder à extração de dados manualmente, o que aumenta ainda mais o tempo necessário para constituir o domínio.

Para além dos dados para treino, são também recolhidos dados sobre o domínio, relativamente às equipas e aos jogadores, completando-se assim os dados de cada equipa enquanto organização. Quanto aos jogadores que integram cada uma das equipas, os dados incluem atributos pessoais, idade, nacionalidade e altura, e as estatísticas de cada época e o clube onde jogou durante esse tempo.

O trabalho realizado no módulo "*ETL*" é armazenado numa base de dados, "*Jogos_Stats*", idealmente numa *data warehouse*, em esquema de estrela, e utilizadas no módulo de "*Treino_Modelo*".

O módulo de "*Treino_Modelo*" é responsável por determinar qual o melhor modelo a utilizar na previsão de um conjunto de jogos, utilizando um *conjunto de dados*, construído em "*ContrucaoDataSet*", utilizando os dados recolhidos e armazenados na base de dados "*Jogos_Stats*". Em seguida é necessário proceder ao treino, em "*Treino*", utilizando um método com o conjunto de dados obtido, criando assim o modelo de previsão, que é depois avaliado, em "*Avaliacao*", de modo a poder ser feita uma comparação com outros modelos e definir qual a ser utilizado para prever os resultados dos eventos futuros.

O módulo de previsão, "*Previsao*", é responsável por prever o resultado de um conjunto de jogos, utilizando o modelo adequado, previamente avaliado. Seguindo-se o armazenamento dos mesmos numa base de dados, onde estarão todos os resultados previstos e o resultado dos mesmos, em "*HistoricoPrevisoes*".

Ainda existe um processo de autenticação, onde irá ser necessário guardar, numa base de dados, os registos dos diversos utilizadores (Clientes), "*Dados_Clientes*", e acedidos através do "*GestorClientes*", gerindo o acesso aos conteúdos a que cada cliente tem acesso.

Existe também replicação nos servidores, "*Produtor*", de forma a evitar que em caso de falha de um produtor os consumidores deixem de ter acesso às previsões. O "Produtor" tem uma política de replicação passiva. Com isto, os "*Consumidores*" interatuam com um "*Produtor*" principal. Sendo que, os restantes "*Produtores*" estão de reserva, quando detetam que o principal falhou, um deles torna-se o primário. E ainda um padrão de desenho de *Publish-Subscriber*, de modo a que qualquer alteração no "*Servidor*" seja comunicada pelo "*Produtor*" ao "*Consumidor*" de modo a ter sempre as previsões atualizadas sempre que necessário.

No "*Cliente*", desenvolveu-se uma aplicação, "*Consumidor*", de preferência uma *aplicação móvel*, de modo a possibilitar o acesso em qualquer lugar, desde que haja uma conexão à Internet. Caso o utilizador esteja online, este terá sempre a última atualização disponibilizada pelo "*Produtor*". Caso esteja offline, a aplicação irá sempre mostrar os últimos dados recolhidos online.

## 4.2 PREVISÃO EM TEMPO REAL

Uma das riquezas deste protótipo é a capacidade de prever o resultado de um jogo futuro, a partir dos dados históricos da mesma modalidade, arquivados ao longo do tempo no "*Jogos_Stats*".

Para melhorar a capacidade de previsão do resultado de uma partida, foi implementado um sistema de previsão do onze titular das duas equipas, até que sejam oficializados os jogadores que jogam de início no encontro. Esta previsão teve em conta o número de minutos que cada jogador normalmente joga, a competição em que o jogo se insere e a condição física (lesões), entre outros atributos que sejam identificados como relevantes.

De modo a implementar este modulo recorreu-se a um pacote de aprendizagem automática. Dado o context do trabalho, a escolha do pacote recaiu entre o Scikit-Learn e o Apache Spark.

O *Scikit-Learn* é uma biblioteca de *Python* para trabalhar em aprendizagem automática (*machine learning*), contendo vários algoritmos de classificação, regressão e *clustering*, tais como random forests e k-means. Esta biblioteca está projetada para interagir com outras bibliotecas *Python*, nomeadamente *NumPy* e *SciPy*. Contém ferramentas simples e eficientes para análise de dados e descoberta de informação. Pode ser utilizado em diversos contextos. Esta biblioteca é de código fonte aberto e pode ser utilizada comercialmente, utilizando a licença BSD. [30]

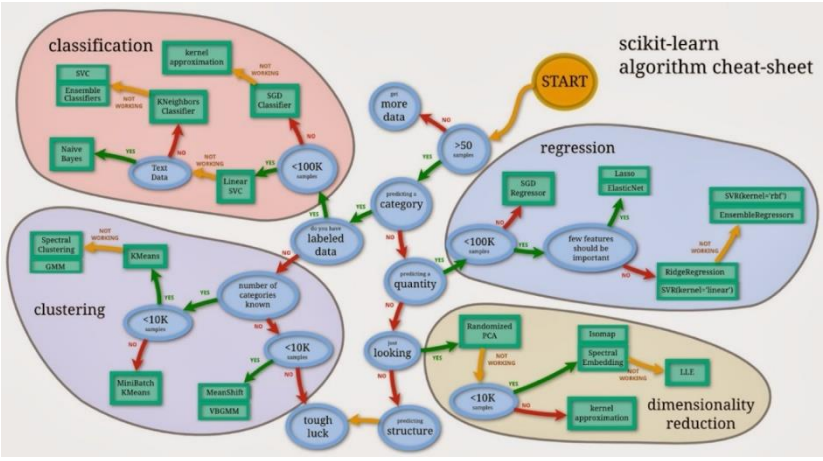


**Figura 50.** Representação do algoritmo Scikit-Learn, segundo Andreas Muller

O *Apache Spark* é uma framework de código fonte aberto para computação distribuída, e é útil para processamento rápido e generalizado. O processamento poderá ser feito em *batch* (semelhante ao *MapReduce*) e em diferentes cargas de trabalho, como *streaming*, consultas interactivas (*queries*), e aprendizagem automática.

O Apache Spark inclui várias bibliotecas para ajudar a criar aplicações de aprendizagem automática (MLlib), processamento de streams (Spark Streaming) e processamento de gráficos (GraphX) [29].

Tal como diversos artigos referem inclusive a apresentação de Ruusmann [31], o uso do Apache Spark permite trabalhar de forma mais eficiente com conjuntos de dados na ordem dos PetaBytes (PB), enquanto que o Scikit-Learn é aconselhado até à ordem dos GigaBytes (GB).

O *Scikit-Learn* tem contribuidores experientes quer em *Machine Learning* quer em desenvolvimento de *software*. Outra grande vantagem desta biblioteca é que cobre amplamente as tarefas necessárias para poder realizar um trabalho sólido em *Machine Learning*. Problemas de velocidade e escalabilidade que possam surgir podem ser na sua maioria resolvidos com um servidor com muita memória [32].

Concluindo, se o objetivo é ter uma solução distribuída ou por ter tamanhos de conjunto de dados superiores à ordem dos GB o melhor é escolher o Apache Spark, até porque contém a maioria dos recursos necessários em Machine Learning. Se o caminho a seguir passar por uma solução num sistema simples e com conjuntos de dados de tamanho até à ordem dos GB, ou ter um foco mais declarado em Ciência de Dados, o melhor é escolher trabalhar com Scikit-Learn.

Deste modo, e tendo em conta que neste trabalho os conjuntos de dados não chegam a tamanhos de 1GB, não há necessidade de instalar uma *framework* complexa para desenvolver este trabalho, e, portanto, o módulo foi desenvolvido com base no Scikit-Learn.

## 4.3 APLICAÇÃO MÓVEL

A aplicação móvel é onde o cliente/utilizador poderá ter acesso às previsões que a ferramenta de previsão gera e depois lhe é possível aceder de acordo com o plano de subscrição que este assinou.

Esta solução está implementada para Android, para uma versão mínima da API 15, Android 4.0.3 (Ice Cream Sandwich) [33]. Este recurso está divido em 2 áreas distintas, a primeira responsável pela autenticação do cliente, Log In, subcapítulo 3.3.1, e a segunda pela área de trabalho do cliente, Área de Trabalho, onde existem 3 fragmentos, “Hot Tips”, “Past Tips” e “Profile”, subcapítulo 4.3.2.

Devido ao facto de a aplicação ter necessidade de comunicar com o servidor foi necessário colocar no *AndroidManifest.xml* as permissões necessárias a aceder à internet.

### 4.3.1 LOG IN

A atividade “*LoginActivity*” é a principal da aplicação e onde o cliente tem de proceder à sua autenticação, utilizando um nome (*Enter Name*) e uma palavra passe (*Password*), Figura 51.

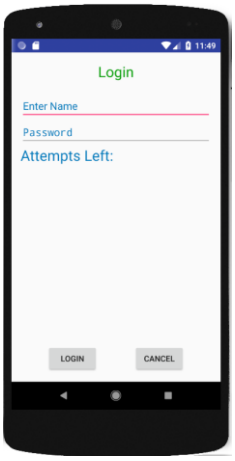


**Figura 51.** *LoginActivity*

Depois de preencher os dados necessários e ao ativar o botão “LOGIN” é feita a comunicação com o servidor de forma a validar os dados, com a palavra passe a ser cifrada para que no lado de servidor não haja possibilidade de saber o valor real da mesma. Caso o nome de utilizador esteja correto, mas a palavra passe não, o utilizador dispõe de mais 2 tentativas adicionais. No caso de o nome de utilizador ser desconhecido pelo servidor é automaticamente registado, junto da sua palavra passe, e atribuído nível de subscrição mais baixo. Quer neste último caso, quer no caso em que o utilizador é reconhecido e a palavra passe estar correta, a aplicação mostra a atividade seguinte, como é mostrada no subcapítulo seguinte.

### 4.3.2 ÁREA DE CLIENTE

A “Área de Cliente” é uma atividade que tem 3 fragmentos de forma a ter todas as informações necessárias que o cliente deve dispor, Figura 52.

O primeiro, “Hot Tips”, mostra todas as previsões de jogos ainda por decorrer nos próximos dias de acordo com a subscrição que o utilizador possui. Por exemplo, se o cliente tiver uma subscrição que contempla apenas os jogos da NBA, então não deverão ser apresentados quaisquer jogos relativos a outras ligas ou modalidades.

O segundo fragmento, “Past Tips”, tem a percentagem de previsões corretas ao longo da época e ainda uma lista das previsões dos últimos jogos concretizados, para cada uma das ligas que são alvo de estudo.

Já o terceiro fragmento, “Profile”, mostra todas as informações pessoais sobre o utilizador, o seu nome de utilizador e o tipo de subscrição que possui. Neste último fragmento é também possível alterar a subscrição efetuada. Caso ocorra uma alteração do nível de subscrição, para além do campo “Subscription Level” ser atualizado, é atualizado também o primeiro fragmento de forma ao cliente ter acesso imediato às previsões relativas à nova subscrição.

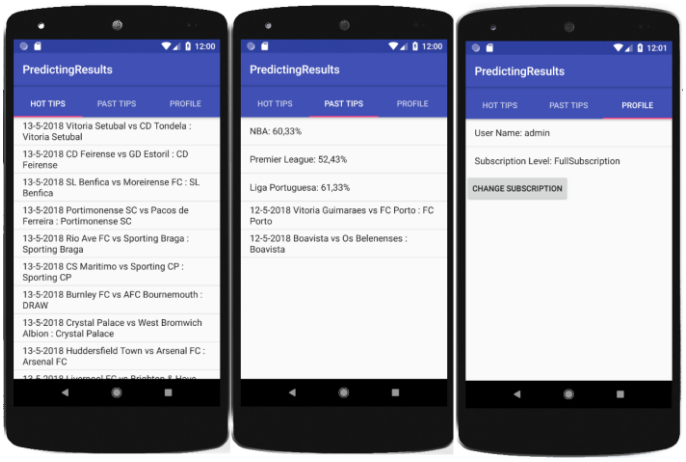


**Figura 52.** Área de trabalho (3 Fragmentos)

## 4.4 RESUMO

Neste capítulo é descrita a arquitetura que contém a ferramenta de previsão de jogos nas diferentes modalidades desportivas e do protótipo de uma aplicação móvel, para sistemas Android, para que um potencial cliente tenha acesso às previsões calculadas. É descrito também como o sistema de previsão em tempo real funciona, prevendo os jogadores que participam nos jogos que estão na lista de encontros a prever. Por último é apresentado em detalhe o protótipo da aplicação móvel, dando a conhecer dois ecrãs fundamentais, o de LogIn, onde o cliente se autentica, e o de Área de Cliente, onde o cliente tem acesso às previsões de jogos futuros, o histórico de previsões e a área pessoal.

# 5 CONCLUSÃO

## 5.1 CONCLUSÕES

Depois de terem sido estudados trabalhos realizados anteriormente, por parte de outros autores, e de terem sido estudadas as modalidades de basquetebol e de futebol, demos a conhecer as estatísticas e o domínio a serem utilizados.

Com todo o conhecimento recolhido e com a construção de conjuntos de dados, um para a NBA, outro para a EPL e outro para a PLP, foram treinados modelos com as técnicas de Naïve Bayes e Random Forests. Os melhores resultados obtidos, na generalidade, foram obtidos com a segunda técnica. Os melhores resultados para o Naïve Bayes são ligeiramente superiores a 60% para a NBA e ligeiramente superiores a 50% para o futebol. Para o treino com Random Forests, o melhor resultado para a NBA é ligeiramente superior a 60%, ainda assim, superior ao obtido no classificador anterior, e de ligeiramente superiores a 50% para a EPL e para a Primeira Liga, as três foram obtidas utilizando as mesmas características nos conjuntos de dados criados para cada um dos domínios. Já para os modelos construídos com MLP as percentagens de acerto são ligeiramente superiores a 60% para NBA, 50% para a PLP e ligeiramente inferiores a 55% para a EPL.

Depois disto e desenvolvendo os modelos, foi possível melhorar os modelos e as percentagens de acerto dos mesmos. Sendo que os melhores modelos para qualquer uma das modalidades e/ou competições foram feitos com Random Forests. Para a NBA atingiu-se os 60% de acerto, para a Primeira Liga de Portugal superou-se ligeiramente os 60% e para a English Premier League quase foi alcançado os 54% de acerto. Para além destes três contextos distintos

para as competições de basquetebol universitário, a NCAA, quer na variante de masculino, quer na variante de femininos, a percentagem de acerto rondou os 70%.

Para além disto, foi assegurada a possibilidade de um potencial apostador ter acesso às previsões através de um protótipo de uma aplicação móvel, desenvolvida para sistemas Android, estabelecendo uma comunicação com o servidor de forma distribuída e tolerante a falhas de servidor relacionadas com disponibilidade.

## 5.2 TRABALHO FUTURO

Como trabalho futuro seria necessário promover diversas mudanças de forma a melhorar a performance da aplicação e de forma a melhorar os modelos de previsão.

De forma a melhorar em termos de desempenho, seria necessário utilizar uma base de dados, com acesso por qualquer uma das réplicas de servidor, e ajustar a forma de como os pedidos de dados são efetuados. Para além da base de dados e das respetivas tabelas, seria também necessário deixar de utilizar a biblioteca de Scikit-Learn e utilizar a solução criada pela Apache, o Spark, com auxílio da biblioteca MLlib para trabalhar o problema relativo à previsão de resultados dos diversos jogos das diferentes modalidades. Estas mudanças significativas devem ser implementadas quando o número de modalidades e/ou competições fosse suficientemente grande, o que pode ser traduzido em maior sobrecarga para o sistema devido ao incremento significativo de dados. Esta solução não foi implementada pois como visto as Random Forest são suficientemente rápidas no treino e previsão dos modelos construídos com as mesmas.

Por outro lado, para promover a melhoria dos modelos de previsão, seria necessário prever de forma mais assertiva que jogadores jogam em cada jogo. Para lidar com este problema surgem duas estratégias diferentes. Na primeira, logo no início da época, seria necessário prever que jogadores jogam em cada jogo, numa previsão para a época toda. Desse modo, seria necessário criar um modelo de previsão que tivesse em conta ao fim de quantos jogos cada jogador fica indisponível para participar numa próxima partida, seja por castigo ou lesão. Ou seja, nesta estratégia, o foco não seria unicamente prever qual seria a equipa titular, mas sim individualizar cada jogador de forma a compor a equipa que iniciaria o encontro seguinte. A segunda estratégia, já num contexto de previsão de resultados jogo a jogo, seria necessário atualizar a tabela correspondente aos jogadores que compõem as equipas que hipoteticamente se iriam defrontar, num atributo que determina se o jogador está disponível ou não para jogar nessa partida. Caso um jogador esteja indisponível seria necessário então prever qual o jogador que o iria substituir. Para fazer esta atualização de cada jogador para cada jogo, poderia ser também implementado

um algoritmo de interpretação textual que dando um conjunto de informações noticiosas fizesse a atualização dos jogadores disponíveis ou convocados para cada jogo.

Outra possibilidade com o intuito de melhorar seria adicionar novos atributos, no qual dou mais importância a dois, contexto e moral. O primeiro seria relativo à importância que o jogo tem para cada equipa, como por exemplo acontece no caso de uma equipa estar a lutar por um objetivo, seja a qualificação para um *playoff*, aplicável por exemplo à NBA, ou a uma competição continental, aplicável no futebol. O segundo poderia ser feito de duas formas diferentes, a moral da equipa e a moral de cada jogador. No caso da moral da equipa poderia ser determinada de diferentes formas, ou tendo em conta eventos que possam afetar a equipa, assumindo a equipa como uma unidade, ou pelo cálculo resultante da moral dos jogadores que a constituem. Este último caso leva à segunda forma de calcular a moral, individualizando cada atleta. Para este problema seria necessário ter em atenção casos que poderiam levar a mudanças anímicas num jogador, exemplo da renovação de um contrato ou problemas com a equipa técnica, por exemplo.

# 5 Conclusão

# REFERÊNCIAS

[1] B. Ulmer e M. Fernandez, 'Predicting Soccer Match Results in the English Premier League'. http://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf, consultado a 23/09/2017.

[2] B. Boldrin, 'Predicting the Result of English Premier League Soccer Games with the Use of Poisson Models'. http://www2.stetson.edu/~efriedma/research/boldrin.pdf , consultado a 23/09/2017.

[3] A. Tsakonas, G. Dounias, S. Shtovba e V. Vivdyuk. 'Soft computingbased result prediction of football games', em The Ist International Conference on Inductive Modelling (ICIM'2002), pp. 15–23, Lviv, Ukraine, 20-25 Maio 2002.

[4] S. Sathe, D. Kasat, N. Kulkarni e R. Satao, 'Predictive Analysis of Premier League Using Machine Learning', em International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 3, pp. 4121-4124, Março 2017.

[5] D. Miljkovic, L. Gajic, A. Kovacevic e Z. Konjovic, 'The use of data mining for basketball matches outcomes prediction' em Proceedings of the 8th International Symposium on Intelligent Systems and Informatics. pp. 309–312. IEEE, 2010

[6] L. Richardson, D. Wang, C. Zhang, X. Yu, 'NBA Predictions', Dezembro 2014. http://www.stat.cmu.edu/~lrichard/links/nba_predictions.pdf, consultado a 25/09/2017.

[7] M. Brown, 'How TV Actually Lost The NBA Postseason, Even With Ratings Up For Finals', Junho 2017. https://www.forbes.com/sites/maurybrown/2017/06/14/how-tv-actually-lost-the-nba-postseason-even-with-ratings-up-for-finals/, consultado a 03/10/2017.

[8] J. Hollinger, "What is PER?", Agosto 2011.

http://www.espn.com/nba/columns/story?columnist=hollinger_john&id=2850240, consultado a 03/10/2017.
[9] “Four Factors”.
https://www.basketball-reference.com/about/factors.html, consultado a 03/10/2017.
[10] “How the NBA Schedule is Made”, in “Analytics 101”.
https://www.nbastuffer.com/analytics101/how-the-nba-schedule-is-made/, consultado a 04/10/2017.
[11] K. Huang e W. Chang, “A Neural Network Method for Prediction on 2006 World Cup Football Game”, em The 2010 International Joint Conference on Neural Networks, IEEE, 2010.
[12] L. Hoffman e M. Joseph, “A Multivariate Statistical Analysis of the NBA”.
http://www.units.miamioh.edu/sumsri/sumj/2003/NBAstats.pdf, consultado a 14/10/2017.
[13] “CSKA faz história em Lisboa”, junho 2005.
http://pt.uefa.com/uefaeuropaleague/season=2004/index.html, consultado a 14/10/2017.
[14] M. Beckler, H. Wang e M. Papamichael, “NBA Oracle”.
https://www.mbeckler.org/coursework/2008-2009/10701_report.pdf, consultado a 15/10/2017.
[15] K. Puranmalka, “Modelling the NBA to Make Better Predictions”, outubro 2013.
https://ai2-s2-pdfs.s3.amazonaws.com/375a/3918aaaaf3cdf04480a0406419f0a77640ce.pdf, consultado a 18/10/2017.
[16] “Official Rules of the National Basketball Association”.
http://www.nba.com/analysis/rules_index.html, consultado a 23/10/2017.
[17] “Basketball History: Origin of the Sport”.
http://www.thebasketballworld.com/history.htm, consultado a 23/10/2017.
[18] “History of Football – The Origins”.
http://www.fifa.com/about-fifa/who-we-are/the-game/index.html, consultado a 24/10/2017.
[19] K. Arnovitz, “Why is there 82-game schedule?”, março 2017.
http://www.espn.com/blog/truehoop/post/_/id/32294/why-is-there-an-82-game-schedule, consultado a 25/10/2017.
[20] D. J. Hand, “The Top Ten Algorithms in Data Mining”, pp. 163 – 178, Taylor & Francis Group, 2009.
[21] “Timeline”, https://www.kaggle.com/c/womens-machine-learning-competition-2018#timeline, consultado a 12/04/2018.
[22] “Timeline”, https://www.kaggle.com/c/mens-machine-learning-competition-2018#timeline, consultado a 12/04/2018.
[23] “Evaluation”, https://www.kaggle.com/c/womens-machine-learning-competition-2018#evaluation, consultado a 12/04/2018.

[24] “Evaluation”, https://www.kaggle.com/c/mens-machine-learning-competition-2018#evaluation, consultado a 12/04/2018.
[25] K. G. Sheela e S. N. Deepa, “Review on Methods to Fix Number of Hidden Neurons in Neural Networks”, Hindawi Publishing Corporation, volume 2013, artigo 425740, http://dx.doi.org/10.1155/2013/425740, consultado a 17/04/2018.
[26] J. Heaton, “The number of Hidden Layers”, junho de 2017, http://www.heatonresearch.com/2017/06/01/hidden-layers.html, consultado a 17/04/2018.
[27] D. P. Kingma e J. L. Ba, “Adam: A Methor for Stochastic Optimization”, https://arxiv.org/pdf/1412.6980.pdf, consultado a 17/04/2018.
[28] S. Haykin, “Neural Networks – a comprehensive foundation”, 2ª edição. Prentice Hall, 1999
[29] “O que é o Apache Spark?”, Amazon Web Services. https://aws.amazon.com/pt/emr/details/spark/, consultado a 09/05/2018.
[30] “scikit-learn: Machine Learning in Python”. http://scikit-learn.org/stable/, consultado a 9 de maior de 2018.
[31] V. Ruusmann, “R, Scikit-Learn and Apache Spark ML – What difference does it make?”, https://pt.slideshare.net/VilluRuusmann/r-scikitlearn-and-apache-spark-ml-what-difference-does-it-make, consultado a 10/05/2018.
[32] B. Lorica, “Six reasons why I recommend scikit-learn”. https://www.oreilly.com/ideas/six-reasons-why-i-recommend-scikit-learn, consultado a 10/05/2018.
[33] “SDK Platform Release Notes”. https://developer.android.com/studio/releases/platforms, consultado a 10/05/2018.
[34] “Machine Learning predicts World Cup Winner”, 12 de junho de 2018. https://www.technologyreview.com/s/611397/machine-learning-predicts-world-cup-winner/, consultado a 20/06/2018.
[35] “FIFA/Coca Cola World Ranking – Men’s Ranking”, https://www.fifa.com/fifa-world-ranking/ranking-table/men/index.html, consultado a 21/06/2018.