CURSO DE CIÊNCIA DA COMPUTAÇÃO

Henrique Luis Schmidt

USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NO AUXÍLIO EM PREVISÃO DE RESULTADOS DE PARTIDAS DE FUTEBOL

Santa Cruz do Sul

2017

Henrique Luis Schmidt

USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NO AUXÍLIO EM PREVISÃO DE RESULTADOS DE PARTIDAS DE FUTEBOL

Trabalho de Conclusão apresentado ao Curso de Ciência da Computação da Universidade de Santa Cruz do Sul, como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Profa. Dra. Rejane Frozza

Santa Cruz do Sul
2017

RESUMO

O futebol é um dos esportes mais populares do mundo, o mercado de apostas neste esporte é uma indústria multibilionária e auxiliar na previsão do resultado de uma partida de futebol não é algo trivial, pois há diversos fatores que o influenciam. Por isso, a previsão de resultados de futebol é um problema de pesquisa importante e o uso de técnicas de aprendizado de máquina é indicado. Segundo (MITCHELL, 1997), algoritmos de aprendizado de máquina são especialmente úteis em domínios onde há uma grande base de dados e regularidades ocultas podem ser descobertas. Neste contexto, o objetivo do trabalho é desenvolver um sistema que possa auxiliar na previsão do resultado de uma partida de futebol, tendo como saída o resultado com maior chance de ocorrência para uma partida. A metodologia definida para este trabalho envolveu uma pesquisa bibliográfica para encontrar trabalhos relacionados, a análise destes para definir as técnicas, os atributos utilizados e a validação do sistema desenvolvido, que foi feita comparando resultados ocorridos com as previsões. Para desenvolver este sistema, foram utilizados dados estatísticos de partidas de equipes do campeonato inglês e dados de habilidade de clubes e jogadores do jogo de computador Fifa. As técnicas Random Forest, Support Vector Machine e Redes Neurais foram desenvolvidas e o melhor resultado atingido foi 58,77% de acertos na previsão dos resultados das partidas.

Palavras-chave: Aprendizado de máquina, futebol, redes neurais artificiais, random forest, support vector machines.

ABSTRACT

Football is one of the most popular sports in the world, the betting market in this sport is a multibillionaire industry and helping the prediction of a football match is not something easy, because there are many factors that influence it. For all these reasons, the prediction of football matches is an important research problem and the machine learning techniques usage is indicated. According to (MITCHELL, 1997), machine learning aghoritms are specially useful when there is a big database and hidden regularities can be found. In this context, this work goal is to develop a system that can help predicting the result of a football match, having as an output the match result with greater chances to occur. The methodoly defined to this work involved a bibliographic research to found related works, the analysis of these works to define the techniques, the attributes and the system validation, that was made comparing the real results with the predicted ones. To develop this system, statistical data from matches in the Premier League and hability data from players in the computer game Fifa were used. The techniques Random Forest, Support Vector Machine and Artificial Neural Networks were developed and the best result was 58.77% of accuracy in the match predictions.

Keywords: Machine learning, football, artificial neural networks, random forest, support vector machines.

LISTA DE ILUSTRAÇÕES

Figura 1 Etapas de estudo para a fundamentação teórica 11
Figura 2 Procedimentos metodológicos 34
Figura 3 Fluxograma do trabalho desenvolvido 37
Figura 4 Modelo relacional do banco de dados a ser utilizado 39
Figura 5 Modelagem da tabela partida 40
Figura 6 Modelagem da tabela Equipe 41
Figura 7 Modelagem da tabela Dados_equipe_fifa 41
Figura 8 Modelagem da tabela Liga 42
Figura 9 Modelagem da tabela Temporada 42
Figura 10 Modelagem da tabela Classificação 43
Figura 11 Arquivos com dados estatísticos de partidas de futebol 45
Figura 12 Site mostrando dados de habilidade da equipe Manchester City 46
Figura 13 Arquivo mostrando odds de diferentes empresas para uma partida 47
Figura 14 Fluxo de execução da técnica Rede Neural Artificial 51
Figura 15 Fluxo de execução da técnica Random Forest 52
Figura 16 Exemplo de classificação utilizando SVM 53
Figura 17 Fluxo de execução da técnica Support Vector Machine 54
Figura 18 Fluxo de execução da técnica k-fold cross validation 56
Figura 19 Interface inicial do sistema desenvolvido 58
Figura 20 Interface para cálculo de porcentagem de acertos 59
Figura 21 Configuração avançada da técnica Random Forest 60
Figura 22 Configuração avançada da técnica Random Forest 61
Figura 23 Configuração avançada da técnica Redes Neurais Artificiais 61
Figura 24 Seleção de partidas por equipe 62
Figura 25 Seleção de partidas por período de datas 63
Figura 26 Exibição de partidas fitradas 64
Figura 27 Configuração da predição de partidas de futebol 65
Figura 28 Exibição da previsão de resultados 66
Figura 29 Comando para busca dos dados de execução 67
Figura 30 Tabela para armazenar dados da execução 5 67

LISTA DE TABELAS

Tabela 1 Comparativo entre os trabalhos relacionados analisados 29
Tabela 2 Melhores resultados obtidos nos trabalhos relacionados 30
Tabela 3 Bibliometria realizada 33
Tabela 4 Resultados das execuções 68

SUMÁRIO

1 INTRODUÇÃO 8
2 FUNDAMENTAÇÃO TEÓRICA 10
2.1 Aprendizado de Máquina 12
2.2 Random Forest 13
2.3 Principal Component Analysis 13
2.4 Rede Bayesiana 13
2.5 Naive Bayes 14
2.6 Logit Boost 14
2.7 k-Nearest Neighbors 14
2.8 Redes Neurais Artificiais 14
2.9 Máquina de Vetores de Suporte 15
2.10 K-fold cross-validation 15
2.11 Apostas esportivas 15
3 TRABALHOS RELACIONADOS 17
3.1 pi-football: A Bayesian network model for forecasting Association Football match 17
3.2 Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries 19
3.3 Discovering the Drivers of Football Match Outcomes with Data Mining 22
3.4 Predicting football scores using machine learning techniques 24
3.5 Prediction of Football Match Results in Turkish Super League Games 25
3.6 Treinamento Supervisionado para Previsão de Partidas de Futebol: Uma Abordagem usando Dados de Videogames 27
3.7 Tabela Comparativa 28
3.8 Considerações 30
4 METODOLOGIA 32
4.1 Caracterização da pesquisa 32
4.2 Bibliometria quantitativa 32
4.3 Procedimentos Metodológicos 33
5 SISTEMA DE APRENDIZADO DE MÁQUINA NO AUXíLIO EM PREVISÃO DE RESULTADOS DE PARTIDAS DE FUTEBOL 36
5.1 Técnicas Escolhidas 38
5.2 Base de Dados 38
5.3 Modelagem do Banco de Dados 39
5.3.1 Tabela Partida 40
5.3.2 Tabela Equipe 41
5.3.3 Tabela Dados_equipe_fifa 41
5.3.4 Tabela Liga 42

5.3.5 Tabela Temporada 42
5.3.6 Tabela Classificação 42
5.4 Importação dos dados 44
5.4.1 Dados de resultados das partidas 44
5.4.2 Dados de habilidade no Fifa 45
5.5 Pré-Processamento dos Dados 47
5.5.1 Cálculo da média de odds de cada uma das equipes 47
5.5.2 Cálculo da classificação e médias de atributos das equipes em cada uma das rodadas 47
5.5.3 Cálculo de habilidades por setor para cada uma das equipes 49
5.6 Técnicas desenvolvidas 49
5.6.1 Rede Neural Artificial 49
5.6.2 Random Forest 52
5.6.3 Support Vector Machine 52
5.6.4 K-fold cross validation 54
5.6.5 Scikit-learn 56
5.6.6 Pandas 57
5.7 Interface de utilização do sistema 57
5.7.1 Interface Inicial 57
5.7.2 Cálculo de porcentagem 58
5.7.3 Configuração avançada de técnicas 59
5.7.4 Previsão de resultados de partidas de futebol 62
5.7.5 Melhoria de desempenho 66
5.8 Resultados 68
6 CONCLUSÃO 70
REFERÊNCIAS 72

1 INTRODUÇÃO

Aprendizado de máquina é o estudo de algoritmos capazes de aprender a melhorar seu desempenho em uma determinada tarefa baseado em experiência prévia. Esta área da computação está fortemente relacionada com reconhecimento de padrões e estatística (MJOLSNESS; DECOSTE, 2001).

Estes algoritmos, em conjunto com mineração de dados, têm sido utilizados em diversas áreas, afetando positivamente a vida das pessoas. Um exemplo foi desenvolvido em Sydney, na Austrália, para prever quais tubos de água possuíam maior chance de falha, diminuindo falhas de serviço para os cidadãos. Este trabalho utilizou a técnica chamada Hierarchical Beta Process (LI et al., 2014). Também foi desenvolvido um algoritmo para prever catástrofes naturais, como tornados e furacões (MCGOVERN et al., 2014).

Um grande desafio para o desenvolvimento de soluções utilizando aprendizado de máquina é a colaboração com pessoas que são experientes no domínio da aplicação que está sendo desenvolvida. É necessário o auxílio de especialistas na obtenção e entendimento dos dados (RUDIN; WAGSTAFF, 2014).

Entre as diversas áreas do conhecimento que podem prover diferentes aplicações está o esporte. Neste contexto, o futebol é o esporte mais popular do mundo e o auxílio à previsão de resultados é importante para fãs de esporte, técnicos de time, jornalistas e apostadores. O mercado de apostas está crescendo muito e é uma indústria multibilionária, por isso, há pesquisas acadêmicas e organizações que focam em soluções para prever resultados, com a possibilidade de obter lucro neste mercado (CONSTANTINOU; FENTON; NEIL, 2013).

Assim, o trabalho une as áreas de esporte e aprendizado de máquina, a fim de obter avaliações satisfatórias (que são de interesse dos usuários) no auxílio em previsão de resultados de partidas de futebol. Técnicas e algoritmos de aprendizado de máquina, previamente utilizados no domínio foram pesquisados, estudados e avaliados. Também foi feito o levantamento de características a serem consideradas para auxílio em previsão de resultados de partidas de futebol. O fato de prever resultados em partidas de futebol constitui-se de uma situação de pesquisa motivadora e incentivadora para a busca de soluções computacionais. Por não ser algo trivial e existirem diversos fatores que influenciam o resultado, como o local da partida, a qualidade dos jogadores, o esquema tático, entre outros, este se mostra um problema muito interessante para a aplicação de algoritmos de aprendizado de máquina.

A justificativa científica do trabalho é comprovar a viabilidade de se resolver problemas

complexos, como o resultado de uma partida de futebol, que depende de diversos fatores, com o uso de aprendizado de máquina.

A justificativa social para o desenvolvimento é a possibilidade que o trabalho traria para um melhor tratamento dos dados do futebol e para uso em treinamento de equipes e desenvolvimento de atletas.

O problema de pesquisa a ser respondido é o seguinte: Como é possível, utilizando aprendizado de máquina, auxiliar na previsão do resultado de uma partida de futebol analisando um conjunto de fatores envolvidos?

O objetivo principal foi desenvolver algoritmos baseados em aprendizado de máquina para auxiliar na previsão do resultado de uma partida de futebol a partir de dados estatísticos das duas equipes.

Foram contemplados os seguintes objetivos específicos:

- Obtenção de uma base de dados consistente de equipes e de seu histórico de jogos.
- Análise dos principais dados que afetam o resultado de um jogo de futebol.
- Estudo das técnicas existentes em aprendizado de máquina referentes à previsão e escolha de algumas para aplicação.
- Desenvolvimento de algoritmos de aprendizado de máquina que auxiliem na previsão do resultado de um jogo de futebol.
- Validação dos resultados dos algoritmos com jogos ocorridos na realidade para verificação do percentual de acertos.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo aborda conceitos fundamentais para a realização da pesquisa e a descrição dos algoritmos utilizados nos trabalhos relacionados descritos no capítulo 2.

Há vários algoritmos de aprendizado de máquina na literatura, desta forma, este capítulo apresentará alguns deles com uma breve descrição. Os algoritmos implementados serão detalhados no capítulo 5.

A figura 1 apresenta as etapas de estudo para a fundamentação teórica do trabalho.

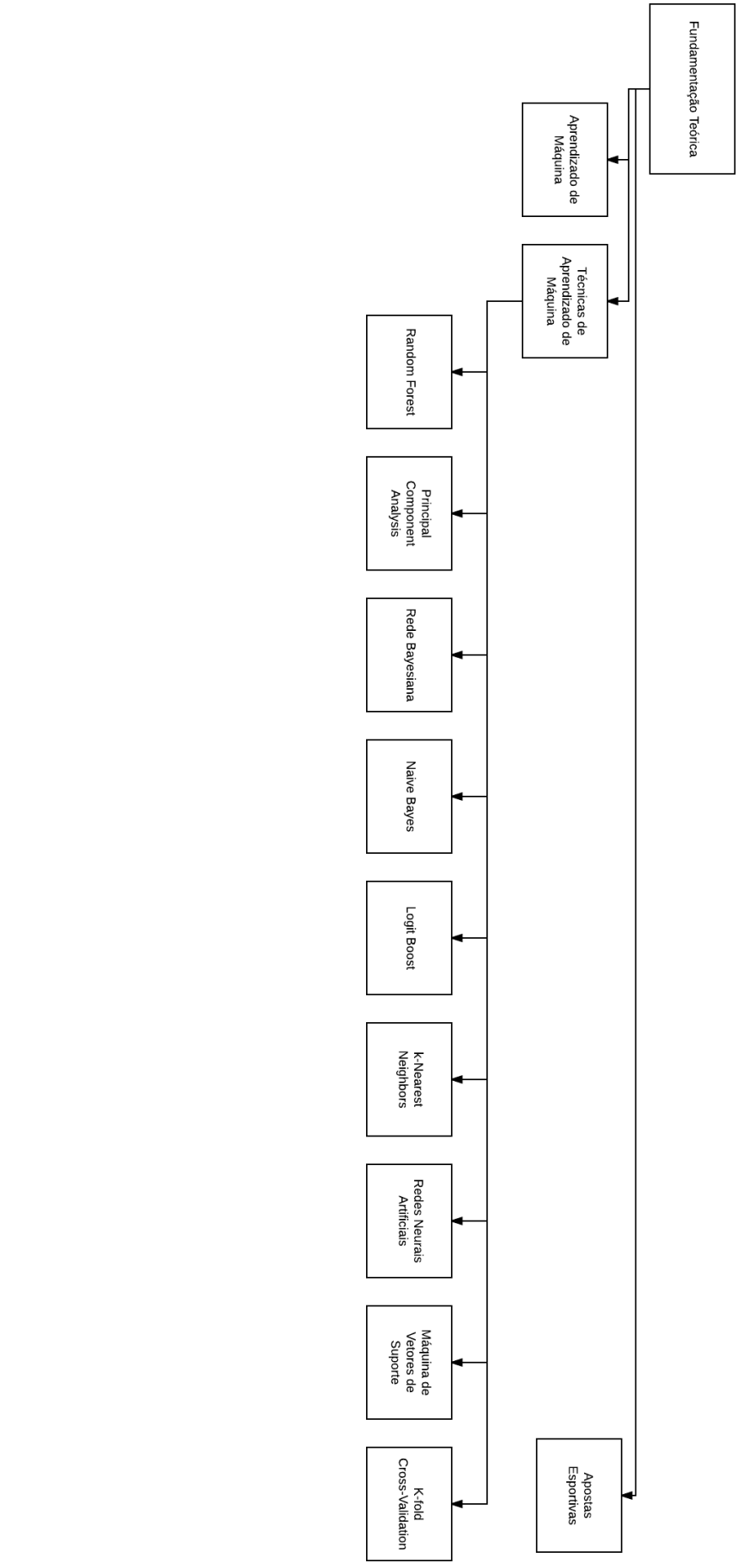


Figura 1 – Etapas de estudo para a fundamentação teórica

Fonte: (AUTORES, 2017)

2.1 Aprendizado de Máquina

Para (MICHALSKI; CARBONELL; MITCHELL, 2013), uma das maiores diferenças entre a forma que pessoas e computadores trabalham é que, enquanto a pessoa faz uma tarefa, ela está pensando em como pode ser mais produtiva na forma de agir.

Segundo (MITCHELL, 1997), o campo de aprendizado de máquina está interessado em resolver o problema de como construir programas de computador que consigam evoluir automaticamente com a sua experiência. Algoritmos de aprendizado de máquina são especialmente úteis nos seguintes casos:

- Problemas de mineração de dados, onde há uma grande base de dados e regularidades ocultas que possam ser descobertas automaticamente.
- Domínios pouco entendidos, onde seres humanos podem não ter conhecimento suficiente para desenvolver algoritmos eficientes.
- Domínios onde o programa deve dinamicamente se adaptar a mudanças no ambiente.

Segundo (MONARD; BARANAUSKAS, 2003), o objetivo da área de aprendizado de máquina é o desenvolvimento de técnicas computacionais sobre o aprendizado e a construção de sistemas de forma automática. Existem dois tipos de aprendizados utilizados em técnicas de aprendizado de máquina, supervisionado e não-supervisionado.

No aprendizado supervisionado, é fornecido um conjunto de exemplos para o programa onde cada exemplo é associado a uma classe. O objetivo do algoritmo é construir um classificador que possa corretamente atribuir uma classe para novos exemplos. Para rótulos de classe discretos, este problema é chamado de classificação e para rótulos de classe contínuos, ele é chamado de regressão (MONARD; BARANAUSKAS, 2003).

Um exemplo de aprendizado supervisionado é a identificação do tipo de animal que está em uma foto e um exemplo de regressão é a predição de valores de ações da bolsa de valores.

No aprendizado não supervisionado, o programa analisa os dados e verifica se alguns deles podem ser agrupados de alguma forma. Após essa análise, o programa forma agrupamentos, que devem ser analisados para verificar o significado de cada um deles no contexto do problema que está inserido (MONARD; BARANAUSKAS, 2003)..

2.2 Random Forest

De acordo com (BREIMAN, 2001), Random Forest é uma combinação de árvores de predição, onde cada árvore depende dos valores de um vetor randômico que tem suas amostras obtidas de forma independente e com a mesma distribuição para todas as árvores da floresta.

O algoritmo irá utilizar os dados disponíveis para criar várias árvores de decisão, cada uma utilizando variáveis diferentes da base de dados. Ao executar o programa, o programa terá a classificação de cada uma das árvores de decisão, onde a classe que for classificada por mais árvores será a classe que o programa irá retornar.

2.3 Principal Component Analysis

Segundo (ABDI; WILLIAMS, 2010), a Análise de Componentes Principais é uma técnica multivariada que analisa dados em uma tabela onde as observações são descritas por variáveis quantitativas, correlacionadas e dependentes uma da outra. O objetivo é extrair informações importantes da tabela, representando estas como um novo conjunto de variáveis ortogonais chamadas de componentes principais.

Os objetivos da técnica são:

- Extrair as informações mais importantes da tabela.
- Comprimir o tamanho dos dados mantendo apenas dados importantes.
- Simplificar a descrição da base de dados.
- Analisar a estrutura das observações e dos dados.

2.4 Rede Bayesiana

Conforme (HECKERMAN; GEIGER; CHICKERING, 1995), uma rede bayesiana é um grafo dirigido que codifica relações probabilísticas entre um número grande de variáveis e tem como objetivo realizar inferências probabilísticas com essas variáveis. Segundo (MARQUES; DUTRA, 2002), uma rede bayesiana é uma representação compacta de uma tabela de probabilidades do universo do problema e consiste do seguinte:

- Conjunto de variáveis e arcos ligando as variáveis.
- Cada variável possui estados mutuamente exclusivos.
- As variáveis e os arcos formam um grafo dirigido sem ciclos.
- Para cada variável existe uma tabela de probabilidade formada pelos pais dessa variável.

Através da tabela de probabilidade entre as variáveis é possível a criação de um modelo preditivo, que calcula o grau de relação dos eventos.

2.5 Naive Bayes

Segundo (BARBOSA et al., 2014), Naive Bayes é um algoritmo de classificação baseado no teorema de Bayes, que encontra as classificações com maiores chances de estarem corretas. O algoritmo calcula as possibilidades de cada atributo em cada umas possíveis classes. A técnica pressupõe que cada uma das possibilidades é independente uma da outra. Com as probabilidades de cada classe e de cada atributo calculadas, é feita a classificação dos dados.

2.6 Logit Boost

Para (HUCALJUK; RAKIPOVIĆ, 2011), o algoritmo do LogitBoost pertence ao grupo de algoritmos chamado "boosting algorithm", que utilizam um grupo de classificadores para obter o resultado final. Quandos estes classificadores são utilizados individualmente, possuem uma precisão em torno de 50%, porém, utilizados em conjunto, os resultados são melhores.

2.7 k-Nearest Neighbors

Segundo (LOPE; MARAVALL et al., 2011), a técnica k-nearest-neighbour é um método para classificação de objetos baseado nos exemplos de treinamento mais próximos na dimensão dos atributos. Os exemplos de treinamento são colocados no espaço de atributos multidimensional. O espaço é dividido em partes onde cada parte recebe o rótulo de uma classe, que representa a classificação do objeto. Um ponto no espaço é atribuído a uma classe se esta é a classe mais frequente atribuída entre os k elementos mais próximos na fase de treinamento.

O parâmetro mais importante desta técnica é o k, que define a quantas unidades de distância os elementos podem estar para serem considerados vizinhos.

2.8 Redes Neurais Artificiais

Conforme (HILL et al., 1994), redes neuras artificiais são modelos matemáticos inspirados pela organização e funcionamento dos neurônios biológicos. Os neurônios podem ser modelados de formas diferentes, dependendo do domínio da aplicação. Estes neurônios possuem entradas, assim como os biológicos possuem, além de dentritos e saídas, como os axiomas dos neurônios biológicos. Assim como na vida real, o aprendizado do sistema ocorre com a al-

teração da força de conexão entre os neurônios.

O método mais comum de aprendizado é o backpropagation, onde os pesos são ajustados para diminuir ao máximo a diferença entre o resultado esperado e o resultado processado pela rede neural. A rede neural consiste de uma camada de entrada, uma de saída e uma ou mais camadas intermediárias. Cada camada consiste de múltiplos neurônios, onde estes estão conectados a neurônios de outras camadas (HILL et al., 1994).

Ao iniciar a rede neural, todos os pesos são inicializados com um valor randômico, depois as observações são enviadas para rede, que calcula a diferença entre o resultado processado e o desejado, ajustando os pesos das conexões, buscando a diminuição do erro.

2.9 Máquina de Vetores de Suporte

Segundo (MARTIN-BARRAGAN; LILLO; ROMO, 2014), as Máquinas de Vetores de Suporte são conhecidas por serem uma técnica de classificação muito poderosa, inclusive para dados com muitas dimensões. A ideia básica de uma Máquina de Vetores de Suporte pode ser explicada geometricamente, se os dados estão em um espaço, o algoritmo encontra o hiperplano que separa os dados com a maior margem possível. Com esse hiperplano, é possível classificar os dados.

2.10 K-fold cross-validation

De acordco com (KOHAVI et al., 1995), o método k-fold cross validation divide a base de dados randomicamente em k partes mutualmente diferentes com aproximadamente o mesmo tamanho. Para cada uma das partes, o classificador é treinado com todo o conjunto de dados menos a parte atual e validado com os dados da parte atual. O resultado da validação é o número de classificações corretas divido pelo número total de classificações efetuadas.

2.11 Apostas esportivas

Segundo (CHAGAS et al., 2016), as apostas esportivas na internet ocorrem no modelo de odds ou cotações. Uma odd ou cotação é a probabilidade de um evento ocorrer, elas são fixadas pelas casas de apostas de acordo com informações relacionadas ao evento, como dados matemáticos, de campanha das equipes, lesão de jogadores, entre outros.

Na seguinte situação: odd para a equipe A igual a 1.50, empate igual a 2.70 e vitória da equipe B igual 3.80. Caso uma pessoa aposte R$ 100,00, irá receber R$ 150,00 caso acertar a

vitória do time A, R$ 270,00 se acertar o empate e R$ 380,00 caso acertar a vitória do time B (CHAGAS et al., 2016).

3 TRABALHOS RELACIONADOS

Este capítulo consiste em apresentar e comparar um grupo de trabalhos que se utilizaram de algoritmos de aprendizado de máquina e mineração de dados para buscar a solução do problema de auxliar na previsão do resultado de uma partida de futebol.

3.1 pi-football: A Bayesian network model for forecasting Association Football match

O artigo apresentado por (CONSTANTINOU; FENTON; NEIL, 2012), tem como objetivo principal desenvolver um novo modelo de rede bayesiana para predizer os resultados de partidas de futebol, na forma de vitória do time da casa, empate ou vitória do time visitante. Futebol é o esporte mais popular do mundo e o mercado de apostas que mais cresce no mundo, por isto, pesquisadores continuam a criar novos modelos de predição usando técnicas diversas. Enquanto alguns modelos focam na posição das equipes ou no campeão do torneio, este artigo tem como objetivo o resultado individual de cada partida. Uma abordagem comum para este tipo de problema é utilizar a técnica Distribuição de Poisson onde o resultado das partidas é gerado pelos parâmetros de ataque e defesa das duas equipes da partida. Estudos anteriores utilizaram a técnica Rating ELO, iniciamente utilizada para definir a força de um enxadrista, para partidas de futebol, porém, os resultados não foram considerados eficientes. Vários estudos consideraram o impacto de fatores específicos no resultado de uma partida, estes fatores incluem vantagem do time da casa, posse de bola e cartões vermelhos. Pesquisas recentes consideraram redes bayesianas e informações subjetivas para predição de partidas de futebol. Os pesquisadores acreditam que este artigo é importante pelos seguintes motivos: o modelo pode ser utilizado por apostadores para obter uma rentabilidade desejada, os modelos anteriores dependem de estatísticas derivadas de escalas pré-determinadas de qualidade dos times e não de estatísticas particulares de cada uma das equipes, este modelo permite a revisão de predições a partir de dados objetivos, incorporando informações subjetivas de fatores importantes que não foram capturados nos dados históricos, a significância de informações recentes é pesada utilizando graus de incerteza resultando em um procedimento de aprendizagem bayesiana não simétrico, as predições dos resultados foram publicados online antes do início de cada partida e, embora o modelo tenha sido utilizado apenas na liga inglesa, ele pode ser facilmente aplicável em qualquer liga do mundo. Os dados básicos utilizados para construir o modelo foram os resultados de todos as partidas da primeira divisão inglesa desde a temporada 1993/1994 até 2009/2010, esta informação está disponível online no site http://www.football-data.co.uk/englandm.php. Na base de dados, o

nome dos clubes foi alterado para um número que representava o seu número de pontos naquela temporada. Com isso, o mesmo clube poderia receber um número de ranking diferente em cada uma das temporadas. Foram criados 14 níveis de força de time, por exemplo, caso um clube conseguisse mais de 84 pontos, ele estaria no nível 1, caso conseguisse entre 80 e 84, nível 2 e assim por diante. Essa abordagem foi considerada adequada por causa das seguintes vantagens: com isso, se faz o uso máximo do número limitado de dados existentes porque a cada temporada três equipes são rebaixadas e promovidas desta competição, com isso, previsões para equipes com pouco histórico de dados podem ser feitas baseadas no ranking de força e não nos resultados passados, o desafio neste caso não é encontrar jogos passados entre os dois participantes da partida e verificar como foram os resultados, mas sim estimar a força atual das duas equipes e verificar, no passado, como equipes desta mesma força se comportaram jogando em casa ou fora. Não são necessárias atualizações de dados frequentes porque a base já possui milhares de jogos, como não são utilizados nomes de equipes, as mesmas observações podem ser utilizadas para predizer resultados de outras ligas. O modelo, chamado de 'pi-football', gera predições para uma partida baseado em quatro fatores genéricos para cada um dos times: 1- força, 2- forma atual, 3- psicológico, 4- fatiga. O primeiro fator, força, é obtido através da média de pontos nas últimas 5 temporadas, o número de pontos atual e a expectativa de pontos para a temporada, e um dado opcional que é a força subjetiva da equipe, que é feita por um especialista, este dado pode ser importante para casos em que a equipe tem grandes mudanças de uma temporada para a outra. O componente 2, forma atual, é obtida através dos resultados dos jogos recentes, verificando se a equipe está atuando acima ou abaixo do esperado, além de ser afetado por jogadores que não estão disponíveis para a partida ou por jogadores retornando para a equipe. O componente 3, psicológico, é obtido através de indicações subjetivas sobre a motivação, espirito da equipe, problemas de gerência e potenciais tendências para este jogo. O componente 4, fatiga, é determinado pela dificuldade da última partida, o número de dias para descanso, o número de jogadores do time principal que não irão jogar para descansar e participação dos jogadores do time principal em jogos de seleções nacionais. Há muitos modos para avaliar a qualidade de um modelo de predição, podendo considerar precisão ou rentabilidade. Para medir a precisão, foi utilizada a técnica chamada Rank Probability Score. No geral, esta técnica representa a diferença entre o resultado observado e previsto. Os resultados utilizando os dados subjetivos, componentes 2, 3 e 4, foram superiores aos resultados utilizando apenas os dados objetivos. Ao medir a rentabilidade do modelo, foi possível obter lucro utilizando modelo

com todas as odds, que significam a probabilidade de um evento ocorrer, publicadas. O modelo criado é dependente do conhecimento do especialista e caso as entradas dele sejam imprecisas, o modelo não conseguirá bons resultados. O resultados das partidas da temporada 2010/2011 foram publicados online antes delas ocorrerem e os dados subjetivos foram submetidos por um membro do time de pesquisa. Caso estes dados fossem enviados por mais especialistas no futebol inglês, o modelo conseguiria melhores resultados. Não foi possível obter conclusões evidentes sobre a eficiência individual de cada um dos componentes do modelo. Para trabalhos futuros, está a determinação da efetividade dos diferentes componentes adicionando mais evidências de ocorrências relevantes e uma engenharia reversa para ajudar a entender como componentes específicos do modelo ajudam a alcançar as odds dos apostadores. Outra possibilidade de pesquisa é verificar se alterar a força da equipe baseado em dados subjetivos pode melhorar os resultados. Por último, uma melhoria que pode ser implementada é dar mais peso para informações recentes em relação as informações mais antigas.

3.2 Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries

O artigo apresentado por (CONSTANTINOU; FENTON, 2013), tem como objetivo principal desenvolver uma nova abordagem, chamada de pi-rating, para dinamicamente classificar equipes de futebol se baseando apenas nos resultados passados, considerando a discrepância entre eles. Diversos modelos de previsão de resultados foram formulados por pesquisadores. Isso não é uma surpresa pois o futebol é o esporte mais popular do mundo e constitui o mercado de apostas que mais cresce. Determinar a habilidade relativa entre as duas equipes participantes da partida é o elemento mais importante para se prever o resultado de uma partida e a posição atual das equipes na tabela. Porém, utilizar a posição na liga tem algumas desvantagens importantes, a tabela sofre uma alta variação no começo do campeonato e uma variação pequena no final, outro problema é que a tabela está desequilibrada até o final do campeonato, quando todas equipes terão jogado contra todas. Utilizar apenas a posição do time na liga também ignora os resultados de outras competições, além de não permitir a comparação com equipes de outras ligas. Na maior parte da literatura com relação à predição de resultados de futebol, a habilidade de uma equipe de futebol é dependente das estatísticas dos resultados passados. Embora foram feitas inúmeras tentativas para formular modelos de previsão mais precisos, poucos trabalhos fizeram o uso de sistemas de classificação puros. (KNORR-HELD, 2000) foi o pri-

meiro a propor um sistema de classificação mas o desempenho não foi muito bom. No artigo (HVATTUM; ARNTZEN, 2010), foi proposta a utilização do método Rating ELO, que é usado para medir a qualidade de jogadores de xadrez, porém, o sistema não obteve bons resultados ao prever resultados de futebol. (GODDARD, 2005) provou que não há uma diferença grande na capacidade de predição ao utilizar modelos que se baseiam no resultado final ou no número de gols da partida. Este artigo propõe a criação de um novo modelo de classificação que é computacionalmente eficiente e tem uma complexidade baixa. A técnica pode ser utilizada para formular predição baseada no número de gols ou no resultado da partida. O sistema de classificação funciona da seguinte forma: Atribui para cada uma das equipes o valor de desempenho (nota) zero, que significa o valor do time médio da competição em relação as outras equipes. Para calcular a nota de cada uma das equipes, deve-se considerar alguns fatos conhecidos sobre futebol: a) Vantagem do time da casa b) Os resultados mais recentes são mais importantes que os resultados antigos. c) A vitória é mais importante para a equipe do que aumentar a diferença no número de gols. Para se adequar a estes fatos, são propostas as seguintes abordagens para o modelo: a) Diferentes notas quando a equipe jogar em casa e como visitante, em conjunto com uma taxa de aprendizagem definida no sistema que determina o quanto os resultados fora de casa influenciam a nota do time dentro de casa e vice-versa. b) Uma taxa de aprendizagem para definir o quanto os novos resultados irão sobrescrever os antigos. c) Diferença grande de gols entre as duas equipes, a cada partida, é exponencialmente diminuida antes da atualização da nota de cada equipe. Quando uma equipe joga em casa, a sua nova nota em casa será dependente dos seguintes fatores: A nota atual dentro de casa, a nota do time adversário jogando fora de casa e o resultado da partida. A nota de cada uma das equipes é dinâmica e ela aumenta quando o desempenho real é melhor que o previsto para a partida. Para diminuir a importância da grande diferença de gols em uma partida, o sistema utiliza uma fórmula matemática que é executada cada vez que as notas das equipes são atualizadas. A cada partida, a nota das duas equipes é comparada de acordo com o local da partida e é feita a predição do número de gols do time da casa e do time visitante. Para se medir o número de erros do sistema, é feita a comparação entre o número de gols que realmente aconteceram e os gols previstos. O processo de atualização das notas das equipes possui seis passos, o primeiro deles é calcular a diferença de gols prevista contra o adversário médio para cada uma das equipes. O adversário médio é uma equipe que tenha nota zero na base de dados, cada equipe da liga recebe uma nota relativa a este time médio. Por exemplo, uma equipe superior a média das equipes, terá uma nota maior

que zero. O segundo passo é usar a informação obtida para calcular a diferença de gols da partida que irá ocorrer. Depois, deve-se obter o resultado real da partida e calcular o erro do sistema, que é a diferença entre o número de gols previsto e os que realmente ocorreram. Após calcular o erro, deve-se aplicar a fórmula matemática que diminui a importância de um número grande de gols. A partir destes valores calculados e das taxas de aprendizado configuradas no sistema, é possível atualizar as notas das equipes participantes da partida. No futebol, novas informações são sempre mais importantes que as anteriores e, por mais que o desempenho das equipes seja diferente jogando em casa ou fora, ainda é possível obter informações de desempenho fora de casa com dados da equipe jogando em casa, e vice versa. Para a construção do modelo deste artigo, foram testadas diferentes taxas de aprendizagem e as que se comportaram de maneira mais eficiente foram as taxas de 0,035 para o grau de importância de informações antigas e 0,7 para o peso de informação de jogos independente de ser em casa ou fora. Para examinar a capacidade do modelo desenvolvido, foi feita uma comparação de resultados com outros dois modelos que utilizam a técnica Rating ELO. Para fazer a comparação, foram realizadas simulações de apostas para jogos que ocorreram na primeira divisão inglesa entre as temporadas 2007/2008 e 2011/2012. Os dados de apostas e resultados das partidas foram obtidos no site http://www.football-data.co.uk/england.php. Para fazer a simulação das apostas, um procedimento muito simples foi seguido, para cada uma das partidas, foi feita a aposta de uma libra para o resultado com maior discrepância entre a probabilidade que o sistema de previsão calculou e a previsão dos apostadores. Caso nenhuma discrepância fosse encontrada a favor do sistema de predição, nenhuma aposta seria feita. A simulação mostrou um desempenho consistente nas cinco temporadas, com ganhos entre 28% e 37% em apostas com odds entre 2,79 e 3,27. Este resultado mostra que o modelo consegue identificar a habilidade das equipes em qualquer momento da temporada e seria possível gerar lucro nas apostas no longo prazo. Em comparação com os modelos desenvolvidos utilizando a técnica Rating ELO, o modelo desenvolvido neste artigo teve um desempenho melhor. Em comparação com estudos anteriores, que acreditavam que o fator local em partidas de futebol ocasionava o mesmo efeito para todas as equipes, foi possível descobrir que algumas equipes mostravam um desempenho muito melhor em casa, enquanto outras não mostravam diferença ou até exibiam um pior desempenho jogando em casa. Para trabalhos futuros, uma ideia é utilizar as notas das equipes calculadas no método desenvolvido como entrada para outros sistemas, como por exemplo, a rede bayesiana desenvolvida por (Constantinou et al, 2012). Outra ideia é utilizar o método desenvolvido no

artigo utilizando equipes e partidas de diversas competições ao mesmo tempo.

3.3 Discovering the Drivers of Football Match Outcomes with Data Mining

O artigo apresentado por (CARPITA et al., 2015), tem como objetivo principal descobrir os fatores que levam uma equipe de futebol a vencer uma partida. Mais precisamente, o objetivo é, a partir de centenas de variáveis, descobrir aquelas que mais afetam a probabilidade de vitória. Para propósitos de predição, um procedimento comum para modelar a distribuição dos resultados das partidas e se basear no método double poisson distribution. A enorme quantidade de dados brutos disponíveis e a falta de teorias explicando a relação entre estes dados e os resultados fazem do futebol um assunto interessante para o uso de mineração de dados. O artigo segue e estende o trabalho efetuado no artigo Football mining with R, onde os autores aplicaram técnicas de mineração de dados para examinar dados da temporada 2010-2011 da série A do futebol italiano. Neste artigo, os seguintes passos são efetuados, primeiro, são selecionadas as variáveis de informação utilizando a variável de importância Gini, que é computada utilizando o algoritmo Random Forests e corrigida com a abordagem Novel, proposta por Sandri e Zuccolotto. O segundo passo é utilizar as variáveis selecionadas para construir indicadores que podem ser interpretados como fatores que influenciam o resultado de uma partida de futebol. Para concluir, é feita a checagem para descobrir o quanto os fatores obtidos influenciam no resultado das partidas, para isso, é utilizado o Multinomial Logistic Regression model, que foi utilizado no artigo Football mining with R. Neste artigo, são analisados dados de quatro temporadas da série A italiana, disponíveis na base de dados de futebol da empresa Panini Digital, esta empresa é líder em coleta de dados estatísticos sobre futebol, provendo serviços de dados para clubes e mídia. A base de dados contém dados sobre ações que ocorreram na partida, como faltas, roubadas de bola, chutes a gol, entre outros dados. Para cada partida, aproximadamente 1300 eventos são processados e sumarizados em uma base de dados com 482 variáveis. O primeiro passo da análise consiste em selecionar, entre as 482 variáveis, aquelas que têm a maior influência no resultado da partida. Para isso, foi utilizado o algoritmo Random Forest e foram encontradas 17 variáveis que serão utilizadas para construir os indicadores para o resultado da partida. O segundo passo é utilizar estas variáveis e combiná-las em um número menor de indicadores chave que são facilmente interpretáveis, utilizando a técnica Principal Component Analysis (PCA). A principal função da técnica PCA é reduzir a dimensionalidade da base de dados mantendo a variação dentro da base quanto possível. Para facilitar a interpretação dos resultados, foram

feitas duas análises separadamente, uma para os dados referentes a equipe da casa e outra para os dados da equipe visitante. Na análise da equipe da casa, os seguintes três componentes foram encontrados: habilidade do time da casa para criar oportunidades para chutar no gol, habilidades do time casa em jogadas aéreas, para fazer cruzamentos e cabeçadas e, por último, a habilidade defensiva do time da casa. Para o time visitante, os seguintes três componentes de dados foram encontrados: habilidades de defesa, contra-ataque e chutes de longa distância do time visitante, habilidade do time visitante criar oportunidades para chutes a gol e o último é a atitude do time visitante de manter a defesa do time na própria área e a possibilidade de criar efetivas ações de contra ataque. O próximo passo do artigo é verificar qual a real habilidade dos atributos selecionados de predizer os resultados. Para isso, foi utilizado o modelo de classificação Multinomial Logistic Regression, porque ele foi considerado o modelo com melhor balanço entre complexidade e precisão em artigos anteriores. Os resultados do método de classificação confirmaram uma habilidade boa de predição do modelo desenvolvido, porém, foi possível observar que o empate foi o resultado mais difícil de fazer a previsão. Isto pode ser porque o empate é um resultado intrinsecamente caracterizado pela incerteza. Para se resolver este problema, foi feita uma alteração no modelo, aumentando a sensitividade do modelo para o empate, diminuindo a sensitividade da vitória do time da casa. Analisando a matriz de confusão, pode-se perceber que o algoritmo conseguiu prever 71% das vitórias do time da casa, 62% das derrotas do time da casa e 45% dos empates. A partir dos resultados, foi possível observar que a habilidade de criar oportunidades de chutes a gol do time da casa reduzem a possibilidade de empate e derrota para times da casa ou visitante. Um número grande de criação de oportunidades de time visitante aumentam muito a chance de derrota do time da casa. O atributo de criação de jogadas aéreas do time da casa aumentam a possibilidade de derrota ou empate desta equipe, mostrando que esta estratégia não é eficiente, o mesmo acontece para os atributos de defesa do time da casa, mostrando que o time da casa deve ter uma postura o mais agressiva possível para ter chances de vitória. Para o time visitante, também foi visto que a habilidade de gerar contra-ataques aumenta a chance de vitória, já a postura de manter a defesa na própria área, aumenta a chance de derrota. Com o desenvolvimento do artigo, foi possível observar que o time da casa deve ter uma postura agressiva, enquanto o time visitante pode adotar uma postura defensiva, desde de que mantenha a defesa longe da própria área. Também foi descoberto que uma estratégia de bolas aéreas favorece o empate e a derrota. Os resultados podem ser interessantes para treinadores e gerentes de equipes de futebol, que podem utilizar as ideias apresentadas neste estudo para

tomar decisões sobre estratégias de jogo e contratações de jogadores. Os resultados obtidos neste artigo foram estáveis nas quatro temporadas analisadas, isso é importante pois, mesmo com as mudanças nas equipes e jogadores, isto prova que existem algumas regularidades por causa da características de um jogo de futebol.

3.4 Predicting football scores using machine learning techniques

O artigo apresentado por (HUCALJUK; RAKIPOVIĆ, 2011), tem como objetivo principal desenvolver um software para predizer os resultados de jogos da Uefa Champions League com uma precisão de 60%. Uma partida de futebol tem apenas três resultados possíveis, vitória do time da casa, empate ou vitória do time visitante. Ainda assim, é muito difícil prever o resultado final, pois a forma de jogar de cada equipe depende de muitos fatores, como as habilidades dos jogadores, as últimas partidas, rivalidades, torcedores, além de vários outros. Além disso, o futebol é um esporte com média baixa de gols, entre dois e três por partida, portanto, um momento de brilhantismo ou um erro de um jogador pode fazer a diferença. Ao desenvolver um sistema de classificação, a escolha dos atributos e do algoritmo de aprendizado afetam o desempenho do sistema. O problema de selecionar os atributos para o sistema pode ser resolvido de duas formas. A primeira é se basear no conhecimento prévio sobre o domínio de partidas de futebol e escolher os atributos que o especialista acredita serem os que mais influenciam no resultado. A segunda forma é utilizar todos os atributos possíveis e, gradualmente, diminuir o número de atributos. Neste artigo, foi utilizada a primeira forma, sendo escolhidas as seguintes propriedades: Os resultados das últimas seis partidas das equipes, o resultado da última partida entre os participantes, a posição atual na tabela do campeonato, número de jogares lesionados do time principal e o número médio de gols feitos e sofridos. Uma outra base de dados, com estas mesmas propriedades e uma adicional, com a avaliação da qualidade das duas equipes por um especialista da área. O objetivo desta base é avaliar o quanto uma avaliação subjetiva impacta o desempenho do sistema. Sobre a escolha do algoritmo de aprendizado, foram utilizadas técnicas diferentes para verificar qual atinge o melhor resultado. As técnicas utilizadas foram as seguintes: Naive Bayes, Bayesian Networks, LogitBoost, The k-nearest algorithm, Random Forest e Artificial neural networks. O software foi implementado utilizando Java com o uso de uma API para o Weka e o software Excel foi utilizado para gerar a base de dados. Na fase de grupos da Uefa Champions League 96 partidas são realizadas, 6 para cada equipe. Cada partida representa um registro com 20 atributos na base de dados. A base de dados é pequena pois não

é possível utilizar partidas dos anos passados porque as equipes participantes são diferentes a cada edição da competição. Os dados foram coletados manualmente de diversos sites da internet. Para testes, os dados foram divididos de três formas, na primeira, três jogos para treino e outros três para testes, na segunda, quatro jogos para treino e dois para testes, e por último, cinco jogos para treino e um para teste. Na análise de resultados, pôde-se observar que a base de dados sem a avaliação do especialista de futebol teve melhores resultados do que a base de dados com a análise subjetiva. Isso pode significar duas coisas, ou os times participantes estavam em um nível parecido e não havia um favoritismo muito grande ou o especialista não possuia muito conhecimento. Sobre os algoritmos de aprendizagem, o algoritmo Naive Bayes teve o pior resultado, seguido por Bayesian Networks. O algoritmo LogitBoost teve um bom desempenho com a base que continha a análise do especialista, este desempenho apenas perdeu para a rede neural artificial, já com a base sem a análise subjetiva, o desempenho não foi muito bom. O algoritmo The k-nearest algorithm teve resultados ruins. O sistema com o algoritmo Random Forest teve um excelente desempenho com a base sem a análise subjetiva e um desempenho mediano com a base com a análise do especialista. Os melhores resultados foram obtidos utilizando rede neural artificial, com 5 camadas e algoritmo Backpropagation, chegando a 68,80% de precisão. Como conclusão, pode-se dizer que foi possível chegar a um resultado satisfatório, com mais de 60% de acertos, utilizando redes neurais artificiais. Para trabalhos futuros, há área para melhoria na seleção de atributos, podendo incluir dados individuais de habilidade e das últimas partidas de cada atleta, além de uma base histórica maior das equipes.

3.5 Prediction of Football Match Results in Turkish Super League Games

O artigo apresentado por (TÜFEKCI, 2016), tem como objetivo principal desenvolver um modelo de predição de resultados para os jogos da liga de futebol turca. Uma quantidade grande de dados históricos, em valores numéricos ou simbólicos, ajudam a resolver o problema. Em estudos recentes com relação à predição de jogos de futebol, algoritmos como redes bayesianas, redes neurais artificiais, máquinas de vetores de suporte (SVM) e árvores de decisão foram aplicados em bases de dados com estatísticas das duas equipes participantes. O propósito do artigo é predizer, para uma partida de futebol, se o time da casa vai vencer, empatar ou perder. Para isso, foram utilizados os métodos de classificação SVMs, Bagging with REP Tree (BREP) e Random Forest (RF). A base de dados incluiu dados de 70 atributos, que são compostos por 69 variáveis de entrada com relação a dados estatísticos do time da casa e visitante, além de uma

variável objetiva, que representou o resultado da partida, em 1222 jogos entre 2009 e 2013. Durante a predição dos resultados, propriedades gerais da partida, propriedades do time da casa e visitante jogando em casa e fora foram levadas em consideração. Para determinar quais atributos eram mais eficientes para tomada de decisão, alguns métodos de seleção foram utilizados e uma base de dados reduzida, com 38 propriedades, foi encontrada. Utilizando a base reduzida, os melhores resultados foram encontrados, chegando a 70.87% de aproveitamento, utilizando a técnica RF. A etapa de pré-processamento dos dados é muito importante e contém os processos de limpeza, integração, transformação e redução dos dados para serem utilizados no algoritmo de aprendizado de máquina. Neste artigo, para fazer a busca de um subconjunto da base de dados, foi aplicada busca genética no software WEKA. Depois da geração dos subconjuntos, um algoritmo de avaliação foi utilizado para definir o melhor deles. Existem algoritmos de avaliação de dois tipos, filtro e wrapper. Neste estudo, um algoritmo de filtro, CFS Subset Evaluator, e três wrapper, que são os avaliadores com SVMS, BREP e RF, foram utilizados. Para comparar os resultados dos algoritmos de tipo SVM, BREP e RF, a técnica de validação cruzada foi empregue, onde a base de dados foi particionada em partes iguais e para cada experimento uma parte foi aplicada para treinamento e outra para validação dos resultados. Os dados originais foram obtidos a partir de várias bases de dados disponíveis na internet. No processo de predição, três métodos de classificação foram aplicados na base de dados original e reduzida, para determinar o modelo mais bem sucedido. Utilizando a técnica SVM, os resultados foram de 67,10% na base original e 67,68% na reduzida, já com classificadores BREP, o resultado foi de 67,43% na base original e 69,48% na base reduzida. O classificador RF conseguiu a maior porcentagem de acerto, 69,97% na base original e 70,41% na base reduzida. De acordo com a base de dados reduzida, as propriedades mais efetivas, com uma taxa de uso de 100% são os dados gerais de uma partida, que especificam se a partida é um clássico e se há alguma penalidade para algumas das equipes, como a ausência de torcida ou troca de estádio. Depois, as propriedades do time visitante jogando fora de casa com 82% de uso e as propriedades do time da casa jogando em casa e fora com 73%. Outras propriedades colocadas na base reduzida foram as estatísticas do time visitante em casa, com 36% de uso, e as estatísticas gerais do time visitante com 27%, além das estatísticas gerais do time da casa, com 18% de uso. Com esta base de dados reduzida e o algoritmo RF, foi possível chegar a uma taxa de acertos 70,87%. O plano de trabalho futuro é predizer os resultados das partidas utilizando redes neurais artificiais.

3.6 Treinamento Supervisionado para Previsão de Partidas de Futebol: Uma Abordagem usando Dados de Videogames

O artigo apresentado por (GOMES et al., 2016), tem como objetivo principal desenvolver um software utilizando técnicas de aprendizado supervisionado para prever resultados de partidas de futebol a partir de dados de caracterização do jogo de videogame Fifa. Um dos principais problemas enfrentados pelos cientistas nas suas pesquisas é a ausência de dados e a dificuldade em obtê-los. Por mais que as técnicas de aprendizado de máquina e mineração de dados tenham avançado muito, é difícil imaginar que algum sistema consiga êxito em aplicações onde uma quantidade de dados relevante não esteja disponível. É grande a complexidade de se prever um evento esportivo e isso se deve ao fato de haver um número grande de variáveis que podem afetar este resultado, como características especificas de cada jogador, problemas psicológicos e fisicos, clima do local do evento, entre outros. Para buscar novas fontes de dados para a previsão de partidas de futebol, o trabalho utilizou dados do jogo eletrônico Fifa. A indústria de jogos eletrônicos investe milhões de dólares a cada ano para fornecer ao usuário uma experiência realista de uma partida de futebol. Para isso, as habilidades dos clubes e jogadores são mensuradas em números e estas informações podem ser utilizadas para aumentar o conhecimento sobre futebol e diminuir a incerteza na previsão de um resultado. Diversos trabalhos já foram realizados para fazer a previsão de eventos esportivos. Alguns trabalham se diferenciam pela forma de predição, onde alguns fazem a previsão do número de gols e outros apenas de qual equipe irá vencer ou empatar. Uma característica comum encontrada em diversos trabalhos é o uso de técnicas de inteligência artificial. Analisando os trabalhos, é possível ver que foram utilizadas técnicas como rede bayesianas, redes neurais artificiais, naive bayes, floresta aleatória, regressão logística e programação genética. Sobre a fonte de dados dos trabalhos, utilizam dados de bookmakers, ou seja, o lucro sobre a possibilidade de cada resultado nos sites de apostas, ou estatísticas sobre os jogos anteriores das equipes. Para a execução do trabalho, foi implementado um crawler web para minerar as informações sobre as características de todos os jogadores e equipes da English Premier League. Também foi buscado no site http://www.football-data.co.uk um conjunto de 380 partidas jogadas por 20 equipes na temporada 2011-2012. Para construir o modelo de previsão de resultados, foi criada uma função de agregação para combinar as características individuais dos jogadores com as médias das equipes. Ao executar a função de agregação para uma equipe, é possível obter uma nota de habilidade utilizada para classificação do resultado das partidas. Na fase de classificação dos resultados, foram utilizados o classificador

Support Vector Machine e a Regressão Logística, que são modelos de treinamento supervisionado. Estes classificadores irão aprender uma função de inferência através de um conjunto de treinamento, onde 80% das partidas serão utilizadas para treino e 20% para validação dos resultados. Nos resultados obtidos, pode-se observar que o classificador Linear SVM obteve um bom desempenho na classificação dos dados, obtendo uma média de 67,88% de previsão. Com isso, é possível concluir que os jogos eletrônicos podem conter poderosas informações capazes de auxiliar na solução de problemas do mundo real. Para trabalhos futuros, os autores pretendem utilizar um conjunto dados estatísticos de partidas de futebol com os dados de jogos eletrônicos e investigar se esta junção pode melhorar os resultados. Também pretendem testar outros métodos probabilísticos. Além disso, também é possível uma melhoria na heurística da construção de variáveis usadas nos classificadores e a construção de um histórico de previsões de um campeonato corrente para ser disponibilizado virtualmente.

3.7 Tabela Comparativa

A tabela 1 mostra um comparativo entre os trabalhos relacionados analisados, levando em consideração os atributos, as técnicas e o tipo de validação realizada.

Tabela 1 – Comparativo entre os trabalhos relacionados analisados

Trabalho Relacionado	Atributos	Algoritmos	Validação
(CONSTANTINOU; FENTON, 2013)	Número de gols, lesões e dados subjetivos.	Algoritmo próprio de classificação de equipes, baseado nos resultados passados.	Simulação de apostas esportivas.
(CARPITA et al., 2015)	482 variáveis, por exemplo: número de faltas e chutes, número de cruzamentos, roubadas de bola, assistências, posse de bola, defesas.	Random Forest, Principal Component Analysis.	Resultados das partidas.
(CONSTANTINOU; FENTON; NEIL, 2012)	Número de gols, lesões e dados subjetivos.	Rede bayesiana.	Simulação de apostas esportivas.
(HUCALJUK; RAKIPOVIĆ, 2011)	Resultados das últimas seis partidas, posição, número médio de gols feitos e sofridos, número de jogadores lesionados, resultado da última partida em que as duas equipes jogaram entre si.	Naive Bayes, Bayesian Networks, LogitBoost, The k-nearest algorithm, Random Forest e Artificial neural networks.	Resultados das partidas.
(TÜFEKCI, 2016)	70 variáveis, por exemplo: atributo indicando se a partida é um clássico, se há punição de estádio, número médio de gols a favor e contra, número de pontos em casa e como visitante.	Support Vector Machine, Random Forest.	Resultados das partidas.
(GOMES et al., 2016)	Dados do jogo de videogame Fifa	Support Vector Machine e Regressão Logística.	Resultados das partidas.
(AUTORES, 2017)	5 conjuntos de atributos, exemplos de dados: Média de gols, posição das equipes na rodada e no ano anterior, chances de vitória segundo as casas de apostas, Dados do jogo de computador Fifa	Random Forest, Support Vector Machine, Redes Neurais Artificais.	Resultados das partidas.

Fonte: (AUTORES, 2017)

3.8 Considerações

Os seis trabalhos apresentados foram escolhidos por apresentarem um objetivo similar ao deste trabalho, o de fazer a previsão do resultado de uma partida de futebol utilizando técnicas de aprendizado de máquina.

Todos os trabalhos seguiram o seguinte fluxo, inicialmente ocorreu a obtenção da base de dados para o sistema, depois um processo de escolha dos atributos utilizados no algoritmo. Após os atributos estarem selecionados, foi feita a aplicação de um algoritmo de aprendizado de máquina para obter o resultado dos jogos a partir da base de dados. Finalmente, foi feita a validação dos resultados gerados pelo sistema desenvolvido.

Sobre os atributos utilizados, cada trabalho selecionou atributos diferentes, dependendo dos dados que estavam disponíveis para o desenvolvimento. Por exemplo, os trabalhos (CONSTANTINOU; FENTON, 2013) e (CONSTANTINOU; FENTON; NEIL, 2012) utilizou dados subjetivos, como moral da equipe e dados sobre lesões dos jogadores, dados que são difíceis de serem obtidos. Já o trabalho realizado por (GOMES et al., 2016) utilizou apenas dados de habilidade no jogo de computador Fifa, que são mais facilmente medidos.

Na fase de aplicação do algoritmo de aprendizado de máquina também foram utilizadas diversas técnicas diferentes, onde a técnica Random Forest se destaca por ter sido utilizada em três trabalhos estudados.

A tabela 2 mostra os três melhores resultados obtidos entre os trabalhos estudados, mostrando as técnicas utilizadas e a sua porcentagem de acerto.

Tabela 2 – Melhores resultados obtidos nos trabalhos relacionados

Algoritmo	Porcentagem de Acertos (Melhor Resultado)	Trabalho Relacionado
Random Forest	70,87	(TÜFEKCI, 2016)
Redes Neurais Artificais	68,80	(HUCALJUK; RAKIPOVIĆ, 2011)
Support Vector Machine	67,88	(GOMES et al., 2016)

Fonte: (AUTORES, 2017)

Para a validação dos resultados, dois trabalhos fizeram simulações de apostas, (CONSTANTINOU; FENTON; NEIL, 2012) e (CONSTANTINOU; FENTON, 2013), e os outros, (TÜFEKCI, 2016), (GOMES et al., 2016), (HUCALJUK; RAKIPOVIĆ, 2011) e (CARPITA et al., 2015), utilizaram os resultados, calculando a porcentagem de acertos do sistema a partir de

partidas reais.

Analisando os estudos correlatos, para este trabalho foram escolhidas as técnicas Random Forest, Support Vector Machine e Redes Neurais Artificiais por terem apresentado bons resultados, conforme é possível ver na tabela 2. Sobre os atributos, foram escolhidos os atributos objetivos, como média de gols marcados, posição das equipes, entre outros, conforme trabalho de (CARPITA et al., 2015). Também foram utilizados os atributos do jogo de computador Fifa, conforme trabalho de (GOMES et al., 2016). Foi escolhido não trabalhar com dados subjetivos, como moral da equipe ou cansaço por exemplo, pela dificuldade de obtenção de dados confiáveis a este respeito.

4 METODOLOGIA

Este capítulo apresenta a caracterização da pesquisa e os procedimentos metodológicos utilizados.

4.1 Caracterização da pesquisa

A pesquisa é exploratória, a partir da análise dos trabalhos relacionados sobre os temas de aprendizado de máquina e do domínio de futebol. O objetivo é se familiarizar e conhecer as técnicas mais utilizadas, a fim de escolher a mais apropriada para o domínio do futebol. É explicativa, pois tem como objetivo descobrir quais os fatores mais interferem no resultado de uma partida de futebol.

Quanto ao ambiente, a pesquisa é bibliográfica, já que foram pesquisados artigos publicados sobre aprendizado de máquina e técnicas utilizadas no domínio de resultados de partidas de futebol e outros esportes.

Com relação aos procedimentos de coleta, a pesquisa é de levantamento visto que foi necessário buscar uma grande base de dados de estatísticas de futebol para o algoritmo processar a fim de encontrar os fatores que mais afetam o resultado de uma partida. Foi feito o download dos arquivos de estatísticas de futebol no site www.football-data.co.uk e criado um programa utilizando a técnica Web Scrapping para extrair os dados de habilidades do jogo de computador Fifa no site futhead.com.

Sobre a natureza da pesquisa, ela é quantitativa pois foram levantados os fatores que envolvem uma partida de futebol e estes foram classificados e analisados, tornando, desta forma, a pesquisa em qualitativa também.

4.2 Bibliometria quantitativa

Para a metodologia utilizada neste trabalho foi desenvolvida uma bibliometria em setembro de 2016 nas bases de dados Web of Science e Scopus. Foram encontrados alguns trabalhos relacionados ao aprendizado de máquina e futebol. Para a análise quantitativa, foram consideradas as publicações realizadas entre 2011 e 2016, sendo realizadas as pesquisas em português e inglês nas áreas de Ciência da Computação, Engenharias e Matemática.

Observando as publicações encontradas, pode-se observar que há muitas pesquisas relacionadas ao tópico de aprendizado de máquina, mostrando que este vem ganhando força nos últimos anos na área de tecnologia. Sobre o domínio futebol, também é possível observar um

número expressivo de trabalhos realizados. Porém, não se encontra um número grande de pesquisas que relacionam o aprendizado de máquina com o tema futebol, o que mostra que há um campo para novas pesquisas.

A tabela 3 mostra o número de trabalhos encontrados em inglês e português nas duas bases científicas.

Tabela 3 – Bibliometria realizada

Critério de pesquisa:"Linha" and "Coluna"	Aprendizado de Máquina		Futebol	
	Scopus	Web of Science	Scopus	Web of Science
Aprendizado de máquina	42377	24314	-	-
Futebol	61	25	7716	7302

Fonte: (AUTORES, 2017)

4.3 Procedimentos Metodológicos

Na figura 2, é possível visualizar os procedimentos metodológicos definidos para o desenvolvimento do trabalho.

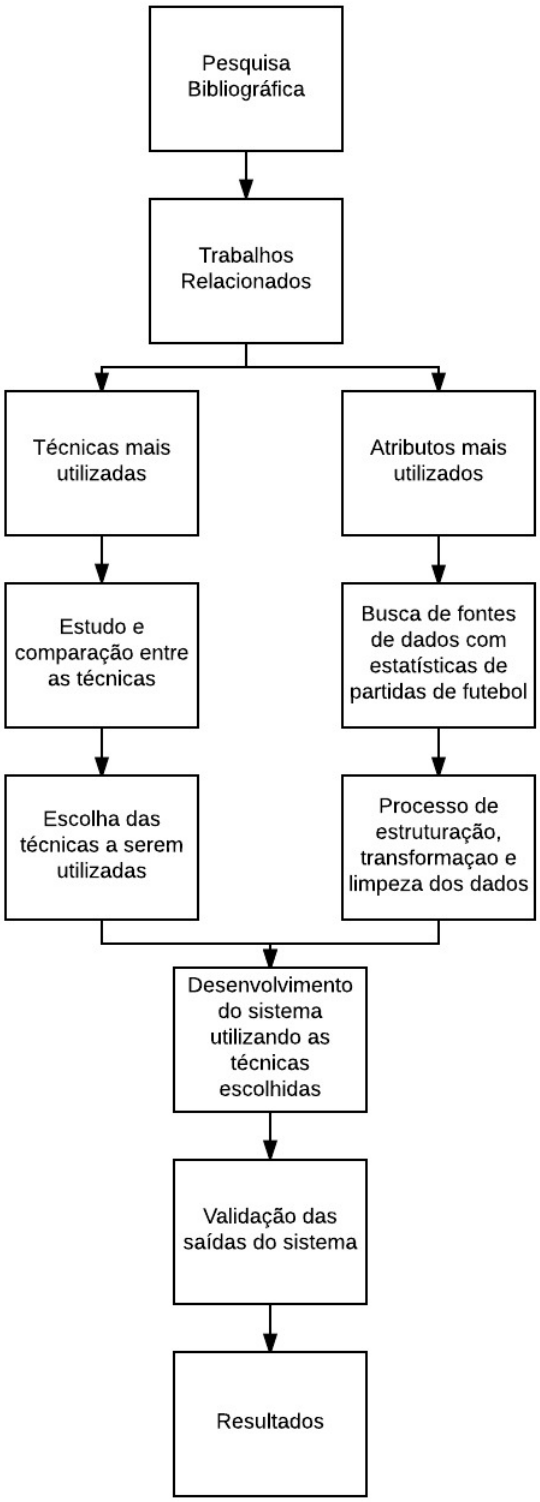


Figura 2 – Procedimentos metodológicos

Fonte: (AUTORES, 2017)

Inicialmente, foram obtidos dados estatísticos sobre o domínio futebol. Após a obtenção dos dados, foi feita uma estruturação e limpeza destes dados. Neste fase, foram corrigidos e padronizados os campos com relação ao tamanho, erros de digitação, presença de caracteres estranhos e formatação. Um exemplo de limpeza aconteceu com os nomes dos clubes nos arquivos de histórico de resultados. Os nomes dos clubes nos arquivos de dados estavam diferentes do que no site com as habilidade do jogo de computador Fifa. Por isso, foi necessário mapear o nome no site com os dados do Fifa para cada equipe do arquivo de resultados.

Com os dados estruturados e limpos, foi feita a seleção dos atributos mais importantes no auxílio à previsão dos resultados de partidas de futebol.

Durante o processo de seleção dos atributos, a base de dados foi dividida em alguns grupos, onde cada grupo possuía alguma característica. Alguns grupos possuíam apenas dados relacionado ao jogo de computador Fifa, outros grupos apenas dados estatísticos sobre as partidas realizadas e outros com a junção destes tipos de atributos. A partir da criação dos grupos, foram realizadas diversas execuções dos programas de auxílio a previsão e mantidos os grupos de atributos com melhores resultados.

Com a base populada e os atributos escolhidos, foi desenvolvido um sistema, utilizando as técnicas de aprendizado de máquina Random Forest, Support Vector Machine e Redes Neurais, que obtiveram uma melhor taxa de aproveitamento em trabalhos relacionados, para auxiliar na previsão de resultados de futebol.

O próximo capítulo descreverá as características e funcionalidades do sistema desenvolvido.

5 SISTEMA DE APRENDIZADO DE MÁQUINA NO AUXÍLIO EM PREVISÃO DE RESULTADOS DE PARTIDAS DE FUTEBOL

O desenvolvimento do trabalho foi baseado na figura 3.

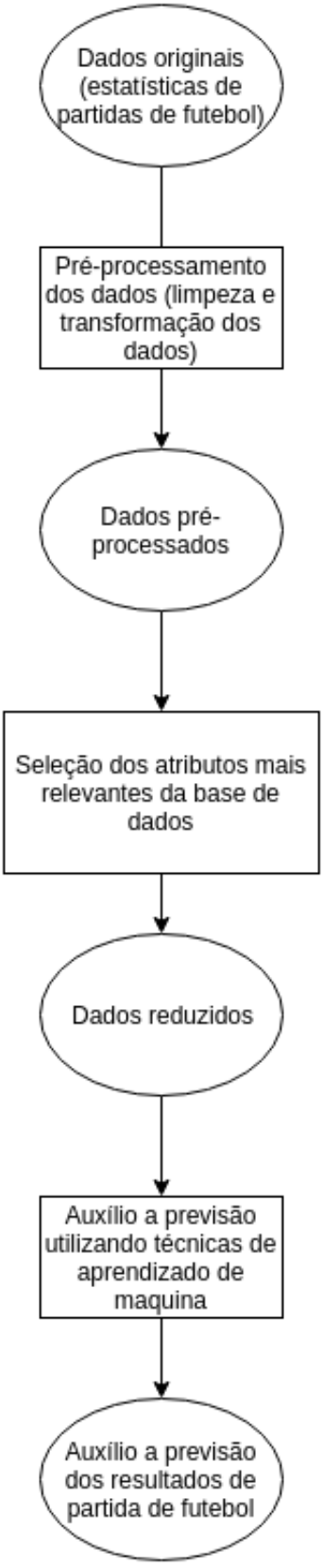


Figura 3 – Fluxograma do trabalho desenvolvido

Fonte: (AUTORES, 2017)

A primeira fase do trabalho consistiu na realização da importação de dados estatísticos sobre partidas de futebol. Após, foi realizada a limpeza e a transformação dos dados, salvando os dados pré-processados em um banco de dados relacional. Na terceira fase, foi realizada a seleção dos atributos a serem utilizados no programa de auxílio a previsão e então implementados programas utilizando as técnicas de aprendizado de máquina escolhidas a partir do estudo dos trabalhos relacionados. Com os programas desenvolvidos, a última etapa foi a validação dos resultados do programa com os resultados reais das partidas.

5.1 Técnicas Escolhidas

Analisando os trabalhos relacionados, as técnicas Random Forest, Support Vector Machine e Redes Neurais Artificiais foram as técnicas que obtiveram os melhores resultados, conforme explicado na seção 3.8. Por isso, foram utilizadas estas três técnicas, a fim de comparar os resultados de cada um dos algoritmos, utilizando a mesma base de dados.

Para validação dos resultados foi utilizada a técnica k-fold cross validation, onde o sistema divide a base em algumas partes, e para cada uma destas partes, o sistema faz o treinamento com todos os dados da base menos a parte que está sendo validada e valida o classificador com a parte que está sendo atualmente validada.

Para cada técnica e conjunto de atributos utilizado, o sistema retornou uma porcentagem de acertos dos resultados das partidas do conjunto de validação. Esta porcentagem foi utilizada na validação do trabalho realizado.

O desenvolvimento do sistema foi na linguagem Python, pois ela possui bibliotecas de código livre que implementam diversos tipos de algoritmos de aprendizado de máquina, incluindo aqueles que foram escolhidos para o desenvolvimento deste trabalho. O hardware utilizado para o desenvolvimento deste trabalho foi um computador pessoal com 16GB de memória RAM, um processador Intel(R) Core(TM) i7-4710MQ CPU @ 2.50GHz e um disco rígido de estado sólido de 128GB.

5.2 Base de Dados

Os sites que foram utilizados para obtenção de dados estão listados a seguir:

- www.football-data.co.uk: Dados estatísticos do campeonato inglês, com os resultados e dados estatísticos de todas as partidas, desde a temporada 1993/1994.
- futhead.com: Dados de habilidades de todos os clubes e jogadores do campeonato inglês

no jogo de computador Fifa.

O trabalho combinou dados de partidas reais com as habilidades dos clubes e jogadores no jogo Fifa para auxiliar na previsão de resultados de futebol.

5.3 Modelagem do Banco de Dados

A figura 4 mostra o modelo relacional do banco de dados que foi utilizado no trabalho.

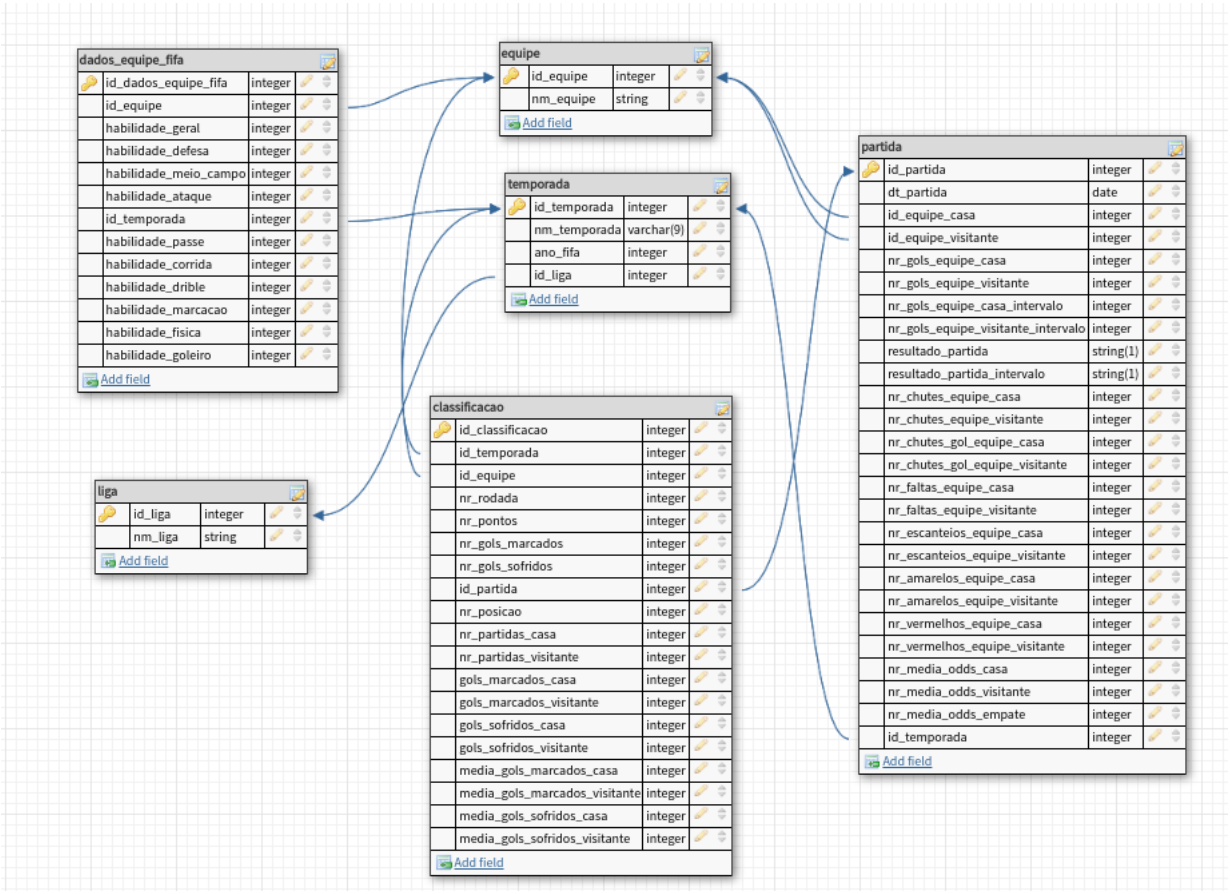


Figura 4 – Modelo relacional do banco de dados a ser utilizado

Fonte: (AUTORES, 2017)

O modelo contém os dados de partidas reais de futebol, que estão na tabela partida, além dos dados de habilidade de cada um dos clubes e seus jogadores, que estão nas tabelas relacionadas a jogador e equipe. Estes dados de habilidade são referentes ao jogo de computador Fifa.

Com a utilização desta base relacional, foi possível ligar os dados estatísticos com as habilidades das equipes e dos jogadores do jogo de computador. Estas informações de habilidade são atribuídas através de análises de especialistas de futebol.

As próximas seções explicam cada tabela do modelo relacional criado.

5.3.1 Tabela Partida

A figura 5 mostra os atributos da tabela partida.

partida	
id_partida	integer
dt_partida	integer
id_equipe_casa	integer
id_equipe_visitante	integer
nr_gols_equipe_casa	integer
nr_gols_equipe_visitante	integer
nr_gols_equipe_casa_intervalo	integer
nr_gols_equipe_visitante_intervalo	integer
resultado_partida	integer
resultado_partida_intervalo	integer
nr_chutes_equipe_casa	integer
nr_chutes_equipe_visitante	integer
nr_chutes_gol_equipe_casa	integer
nr_chutes_gol_equipe_visitante	integer
nr_faltas_equipe_casa	integer
nr_faltas_equipe_visitante	integer
nr_escanteios_equipe_casa	integer
nr_escanteios_equipe_visitante	integer
nr_amarelos_equipe_casa	integer
nr_amarelos_equipe_visitante	integer
nr_vermelhos_equipe_casa	integer
nr_vermelhos_equipe_visitante	integer
nr_media_odds_casa	integer
nr_media_odds_visitante	integer
nr_media_odds_empate	integer
id_liga	integer

Figura 5 – Modelagem da tabela partida

Fonte: (AUTORES, 2017)

Esta é a principal tabela do modelo, pois contém o resultado da partida e os dados estatísticos, como número de gols, faltas, escanteios, cartões, entre outros.

Estes dados foram obtidos do site football-data.co.uk, que disponibiliza os mesmos dados da tabela no formato de CSV de todas as partidas do campeonato inglês desde a temporada 1993/1994.

5.3.2 Tabela Equipe

A figura 6 mostra os atributos da tabela equipe.

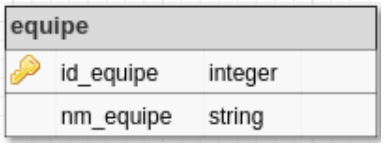


Figura 6 – Modelagem da tabela Equipe

Fonte: (AUTORES, 2017)

Esta tabela do banco de dados contém apenas o nome da equipe e é utilizada para armazenar as equipes das partidas e jogadores.

5.3.3 Tabela Dados_equipe_fifa

A figura 7 mostra os atributos da tabela dados_equipe_fifa.

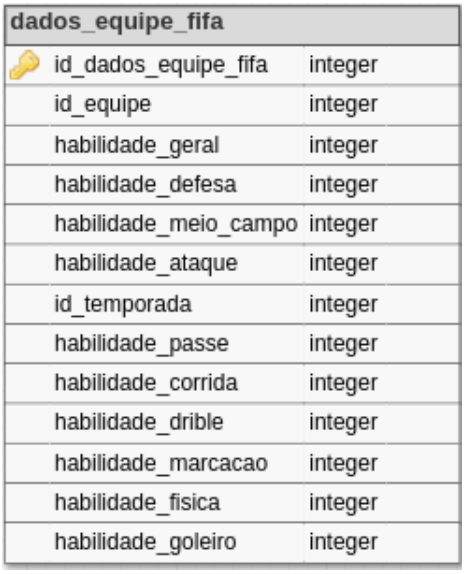


Figura 7 – Modelagem da tabela Dados_equipe_fifa

Fonte: (AUTORES, 2017)

Esta tabela armazena a média de habilidades dos jogadores dos clubes no jogo Fifa 2017.

No jogo, cada jogador possui um valor de habilidade para cada fundamento, como passe, corrida, entre outros. Cada atributo tem um valor entre 0 e 100, onde o valor 0 representa falta de habilidade e 100 muita habilidade no fundamento. Por exemplo, um jogador com valor de passe 0 é um passador ruim, já com valor 100 é um excelente passador. Além disso, o jogador possui um atributo chamado habilidade_geral, que é calculado a partir da média de habilidade de todos os fundamentos.

Os atributos desta tabela guardam a média de habilidade dos jogadores em cada um dos setores do campo e em cada um dos fundamentos. Por exemplo, o atributo habilidade_defesa armazena a média de habilidade geral de todos os jogadores da defesa.

Estes dados foram obtidos no site futhead.com, que contém a habilidade de todos os clubes do campeonato inglês no jogo.

5.3.4 Tabela Liga

A figura 8 mostra os atributos da tabela liga.

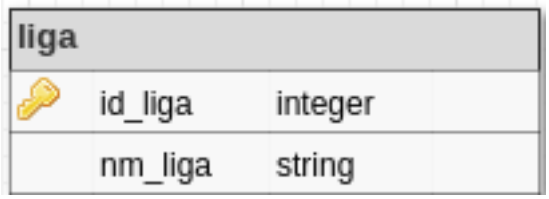


Figura 8 – Modelagem da tabela Liga

Fonte: (AUTORES, 2017)

Esta tabela contém apenas o nome da liga e é utilizada para armazenar a liga de cada uma das partidas.

5.3.5 Tabela Temporada

A figura 9 mostra os atributos da tabela temporada.



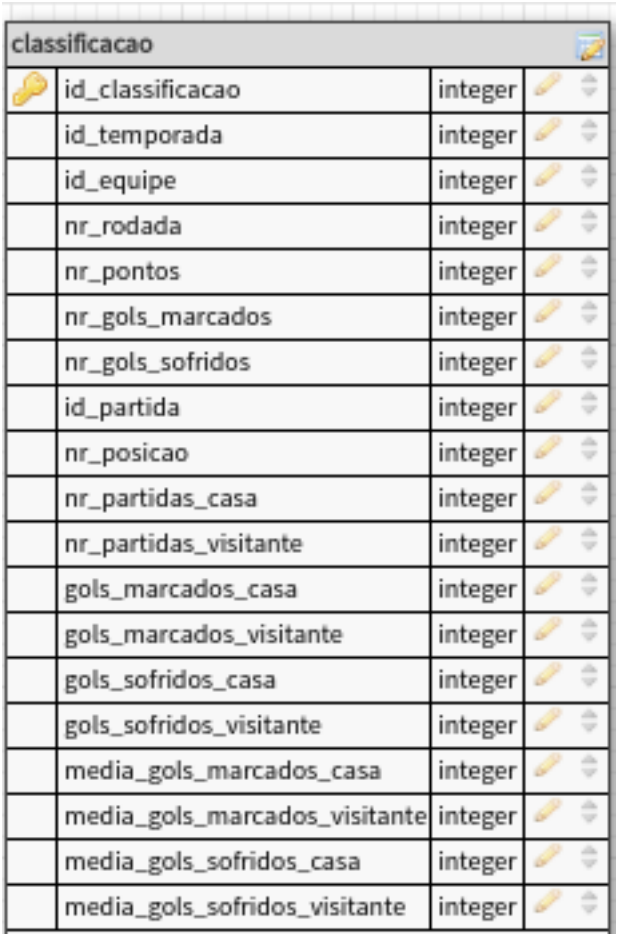
Figura 9 – Modelagem da tabela Temporada

Fonte: (AUTORES, 2017)

Esta tabela contém apenas o ano da temporada e é utilizada para armazenar qual a temporada de cada uma das partidas da base de dados. Esta informação é importante no cálculo de classificação.

5.3.6 Tabela Classificação

A figura 10 mostra os atributos da tabela classificação.



classificacao		
id_classificacao	integer	
id_temporada	integer	
id_equipe	integer	
nr_rodada	integer	
nr_pontos	integer	
nr_gols_marcados	integer	
nr_gols_sofridos	integer	
id_partida	integer	
nr_posicao	integer	
nr_partidas_casa	integer	
nr_partidas_visitante	integer	
gols_marcados_casa	integer	
gols_marcados_visitante	integer	
gols_sofridos_casa	integer	
gols_sofridos_visitante	integer	
media_gols_marcados_casa	integer	
media_gols_marcados_visitante	integer	
media_gols_sofridos_casa	integer	
media_gols_sofridos_visitante	integer	

Figura 10 – Modelagem da tabela Classificação

Fonte: (AUTORES, 2017)

Esta tabela contém a posição, o número de pontos, o número e a média de gols marcados como mandante e como visitante de cada uma das equipes em cada uma das temporadas.

5.4 Importação dos dados

O objetivo da fase de importação é popular a base de dados modelada e apresentada na figura 4.

A importação foi realizada importando arquivos textos que continham dados estatísticos sobre partidas de futebol para uma base de dados e com a utilização da técnica web scraping para percorrer várias páginas de um site importando dados relacionados ao jogo de computador Fifa. As próximas seções detalham o processo de importação.

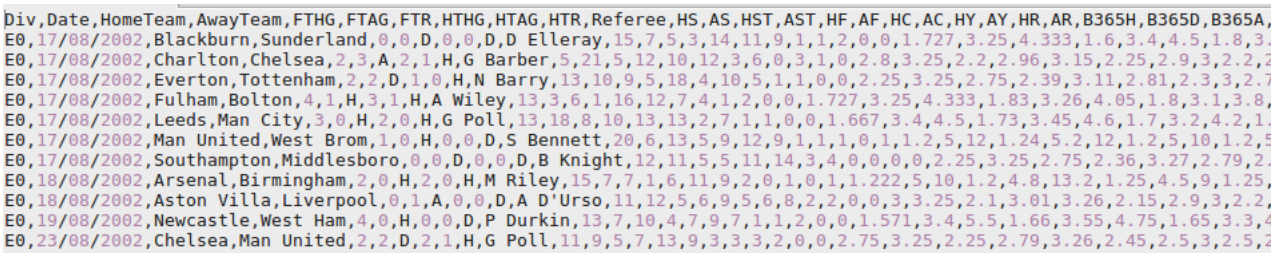
5.4.1 Dados de resultados das partidas

Para popular a tabela partida, foi feito o download de 16 arquivos no formato TXT com dados de partidas que ocorreram no campeonato inglês entre os anos de 2001 e 2017. Cada arquivo contava com os dados de uma temporada. Os arquivos estão disponíveis no link http://www.football-data.co.uk/englandm.php.

A figura 11 mostra os dados de um dos arquivos. Nela é possível verificar os seguintes atributos:

- 2ª coluna: Data da partida
- 3ª coluna: Equipe mandante
- 4ª coluna: Equipe visitante
- 5ª coluna: Gols da equipe mandante
- 5ª coluna: Gols da equipe visitante
- 6ª coluna: Resultado da partida (Vitória da equipe mandante, visitante ou empate)
- 7ª coluna: Gols da equipe mandante no intervalo
- 8ª coluna: Gols da equipe visitante no intervalo
- 9ª coluna: Resultado da partida no intervalo (Vitória da equipe mandante, visitante ou empate)
- 10ª coluna: Resultado da partida (Vitória da equipe mandante, visitante ou empate)
- 11ª coluna: Chutes da equipe mandante
- 12ª coluna: Chutes da equipe visitante

- 13ª coluna: Chutes a gol da equipe mandante
- 14ª coluna: Chutes a gol da equipe visitante
- 15ª coluna: Faltas da equipe mandante
- 16ª coluna: Faltas da equipe visitante
- 17ª coluna: Escanteios da equipe mandante
- 18ª coluna: Escanteios da equipe visitante
- 19ª coluna: Cartões amarelos da equipe mandante
- 20ª coluna: Cartões amarelos visitante
- 21ª coluna: Cartões vermelhos da equipe mandante
- 22ª coluna: Cartões vermelhos da equipe visitante
- A partir da 23ª coluna: Chances de vitória de cada equipe para 7 empresas de apostas diferentes



```
Div,Date,HomeTeam,AwayTeam,FTHG,FTAG,FTR,HTHG,HTAG,HTR,Referee,HS,AS,HST,AST,HF,AF,HC,AC,HY,AY,HR,AR,B365H,B365D,B365A,
E0,17/08/2002,Blackburn,Sunderland,0,0,D,0,0,D,D Elleray,15,7,5,3,14,11,9,1,1,2,0,0,1.727,3.25,4.333,1.6,3.4,4.5,1.8,3.
E0,17/08/2002,Charlton,Chelsea,2,3,A,2,1,H,G Barber,5,21,5,12,10,12,3,6,0,3,1,0,2.8,3.25,2.2,2.96,3.15,2.25,2.9,3,2.2,2
E0,17/08/2002,Everton,Tottenham,2,2,D,1,0,H,N Barry,13,10,9,5,18,4,10,5,1,1,0,0,2.25,3.25,2.75,2.39,3.11,2.81,2.3,3,2.7
E0,17/08/2002,Fulham,Bolton,4,1,H,3,1,H,A Wiley,13,3,6,1,16,12,7,4,1,2,0,0,1.727,3.25,4.333,1.83,3.26,4.05,1.8,3.1,3.8,
E0,17/08/2002,Leeds,Man City,3,0,H,2,0,H,G Poll,13,18,8,10,13,13,2,7,1,1,0,0,1.667,3.4,4.5,1.73,3.45,4.6,1.7,3.2,4.2,1.
E0,17/08/2002,Man United,West Brom,1,0,H,0,0,D,S Bennett,20,6,13,5,9,12,9,1,1,1,0,1,1.2,5,12,1.24,5.2,12,1.2,5,10,1.2,5
E0,17/08/2002,Southampton,Middlesboro,0,0,D,0,0,D,B Knight,12,11,5,5,11,14,3,4,0,0,0,0,2.25,3.25,2.75,2.36,3.27,2.79,2.
E0,18/08/2002,Arsenal,Birmingham,2,0,H,2,0,H,M Riley,15,7,7,1,6,11,9,2,0,1,0,1,1.222,5,10,1.2,4.8,13.2,1.25,4.5,9,1.25,
E0,18/08/2002,Aston Villa,Liverpool,0,1,A,0,0,D,A D'Urso,11,12,5,6,9,5,6,8,2,2,0,0,3,3.25,2.1,3.01,3.26,2.15,2.9,3,2.2,
E0,19/08/2002,Newcastle,West Ham,4,0,H,0,0,D,P Durkin,13,7,10,4,7,9,7,1,1,2,0,0,1.571,3.4,5.5,1.66,3.55,4.75,1.65,3.3,4
E0,23/08/2002,Chelsea,Man United,2,2,D,2,1,H,G Poll,11,9,5,7,13,9,3,3,3,2,0,0,2.75,3.25,2.25,2.79,3.26,2.45,2.5,3,2.5,2
```

Figura 11 – Arquivos com dados estatísticos de partidas de futebol

Fonte: (AUTORES, 2017)

Após o processamento de todos os arquivos, a base de dados contava com 5570 partidas de 39 equipes.

5.4.2 Dados de habilidade no Fifa

Para a importação dos dados de habilidade dos jogadores no Fifa foi utilizada a técnica chamada Web Scraping, onde um programa de computador faz a extração de dados de uma página da internet. Para cada página de cada jogador de cada uma das equipes do sistema para cada uma das temporadas entre 2010 e 2017 foi realizada uma requisição para obter os dados do site futhead.com. Com os dados obtidos, foi feito o processamento e transformados os dados para a base de dados relacional.

A figura 12 mostra parte dos dados da equipe Manchester City no site.



Figura 12 – Site mostrando dados de habilidade da equipe Manchester City

Fonte: (AUTORES, 2017)

Após a importação, a base de dados contava com os seguintes dados de habilidades de 39 equipes em 7 temporadas, que são medidos em um número inteiro, entre 0 e 100:

- Habilidade geral: Média da habilidade geral de todos os jogadores de uma equipe
- Habilidade de goleiros: Média da habilidade geral de todos os goleiros de uma equipe
- Habilidade de defensores: Média da habilidade geral de todos os defensores de uma equipe
- Habilidade de meio-campo: Média da habilidade geral de todos os meio-campos de uma equipe
- Habilidade de atacantes: Média da habilidade geral de todos os atacantes de uma equipe
- Habilidade de passe: Média da habilidade de passe de todos os jogadores de uma equipe
- Habilidade de corrida: Média da habilidade de corrida de todos os jogadores de uma equipe
- Habilidade de drible: Média da habilidade de drible de todos os jogadores de uma equipe
- Habilidade de marcação: Média da habilidade de marcação de todos os jogadores de uma equipe
- Habilidade física: Média da habilidade física de todos os jogadores de uma equipe

5.5 Pré-Processamento dos Dados

Com a base de dados populada foi necessário fazer o cálculo de alguns atributos que não foram encontrados na fonte original e que são importantes para a previsão do resultado de uma partida de futebol.

5.5.1 Cálculo da média de odds de cada uma das equipes

O arquivo original com dados das partidas possuía um valor de odd para 7 empresas de apostas diferentes. Para o sistema, não era interessante ter vários valores diferentes para cada uma das empresas, mas sim um valor que representasse as chances da equipe sair vencedora, de acordo com os apostadores.

Para isso, no processo de transformação dos dados, foi realizada a média de chances para cada um dos resultados possíveis entre as 7 empresas de apostas diferentes.

Na figura 13, é possível visualizar o arquivo com vários valores de odds para cada partida. A seguir estão listados os seis primeiros atributos e o que eles representam.

- B365H: Chances de vitória do time da casa segundo casa de apostas bet365
- B365D: Chances de empate segundo casa de apostas bet365
- B365A: Chances de vitória do time visitante segundo casa de apostas bet365
- BWH: Chances de vitória do time da casa segundo casa de apostas bwin
- BWD: Chances de empate segundo casa de apostas bwin
- BWA: Chances de vitória do time visitante segundo casa de apostas bwin

```
,B365H,B365D,B365A,BWH,BWD,BWA,IWH,IWD,IWA,LBH,LBD,LBA,PSH,PSD,PSA,WHH,WHD,WHA,VCH,VCD,VCA
9,2.5,3.2,2.9,2.5,3.2,3.25,55,2.55,2.43,3.35,3.21,3.3,3.1,40,2.4,2.3,1.68,1.61,32,-0.25,2.
```

Figura 13 – Arquivo mostrando odds de diferentes empresas para uma partida

Fonte: (AUTORES, 2017)

5.5.2 Cálculo da classificação e médias de atributos das equipes em cada uma das rodadas

No arquivo de importação dos dados, não é possível visualizar a classificação de cada uma das equipes ou informações relevantes como o número médio de gols marcados na temporada, número médio de gols sofridos, entre outros.

Os seguintes atributos foram calculados para cada uma das partidas do banco de dados:

- Posição das equipes na classificação no momento da partida.
- Posição final das equipes na temporada anterior.
- Média de gols marcados em casa e fora de casa na temporada atual e na anterior.
- Média de gols sofridos em casa e fora de casa na temporada atual e na anterior.

5.5.2.1 Cálculo da posição das equipes na classificação no momento da partida

Para o cálculo da posição das equipes, foi utilizado o seguinte algoritmo:

Para cada uma das equipe em cada uma das temporadas, o sistema buscou cada uma das partidas e calculou quantos pontos a equipe possuía em cada uma das rodadas.

Após calcular o número de pontos de todas as equipes, o sistema atualizou a posição de cada equipe em cada uma das rodadas na tabela de classificação fazendo a ordenação das equipes por pontos, saldo de gols, gols marcados e gols sofridos.

No momento de buscar uma partida, foi possível encontrar na tabela de classificação a posição atual da equipe.

5.5.2.2 Cálculo da posição final das equipes na temporada anterior

Como a posição das equipes em cada uma das rodadas de cada uma das temporadas já estava calculada, para conhecer a posição na temporada anterior o sistema buscou a posição da equipe na rodada de número 38, que é a última, na temporada anterior.

5.5.2.3 Cálculo da média de gols marcados e sofridos

Para o cálculo da média de gols marcados e sofridos, foi utilizado o seguinte algoritmo:

Para cada uma das equipes em cada uma das temporadas, o sistema buscou cada uma das partidas e calculou a média de gols marcados e sofridos por cada uma das equipes como mandante e como visitante.

No momento de fazer a previsão do resultado, foi possível visualizar na tabela de classificação a média de gols marcados em casa e fora por cada uma das equipes. Para buscar a média de gols da temporada anterior, foi possível utilizar os dados gravados na última rodada, de número 38, desta temporada.

5.5.3 Cálculo de habilidades por setor para cada uma das equipes

No site onde foi realizada a importação dos dados de habilidades de jogadores no Fifa estão disponíveis apenas informações sobre cada um dos jogadores. Por isso, foi necessário calcular e armazenar a média de habilidades de cada uma das equipes por setor e habilidade.

Cada atributo de habilidade para cada jogador é medido em um número inteiro, entre 0 e 100.

As médias a seguir foram calculadas na fase de pré-processamento dos dados:

- Habilidade dos goleiros.
- Habilidade do setor defensivo.
- Habilidade do setor de meio-campo.
- Habilidade do setor de ataque.
- Habilidade de drible.
- Habilidade de marcação.
- Habilidade do chute.
- Habilidade física.
- Habilidade de velocidade.

5.6 Técnicas desenvolvidas

Esta seção lista as técnicas e o método de validação desenvolvidos neste trabalho.

5.6.1 Rede Neural Artificial

O sistema desenvolvido utilizou a rede neural Multilayer Perceptron utilizando Backpropagation para treinamento.

Diferentes modelos de RNAs foram desenvolvidos (WILDE, 2013) (FAUSETT, 1994) (HAYKIN, 2004), abordando a definição e o funcionamento de suas arquiteturas. Entre os modelos, destaca-se o MLP (Multi-Layer Perceptron) (PARKS; LEVINE; LONG, 1998), que foi utilizado neste trabalho, por ser um modelo de aprendizado supervisionado, que, neste caso, relacionou os atributos que representam as estatísticas das equipes de futebol com o resultado das partidas.

O processo de treinamento e classificação utilizando esta técnica está descrito na figura 14. Inicialmente, são inicializados os neurônios e as conexões entre eles com valores padrões. Após

a inicialização é feito o treinamento da rede neural, onde cada um dos registros de treinamento é enviado para a rede neural e os pesos das conexões são atualizados de acordo com a saída esperada e a saída gerada pela rede neural para o registro de treinamento. Após o número de iterações atingir um valor máximo ou o erro da rede neural estiver abaixo do valor desejado, o treinamento da rede neural é concluído.

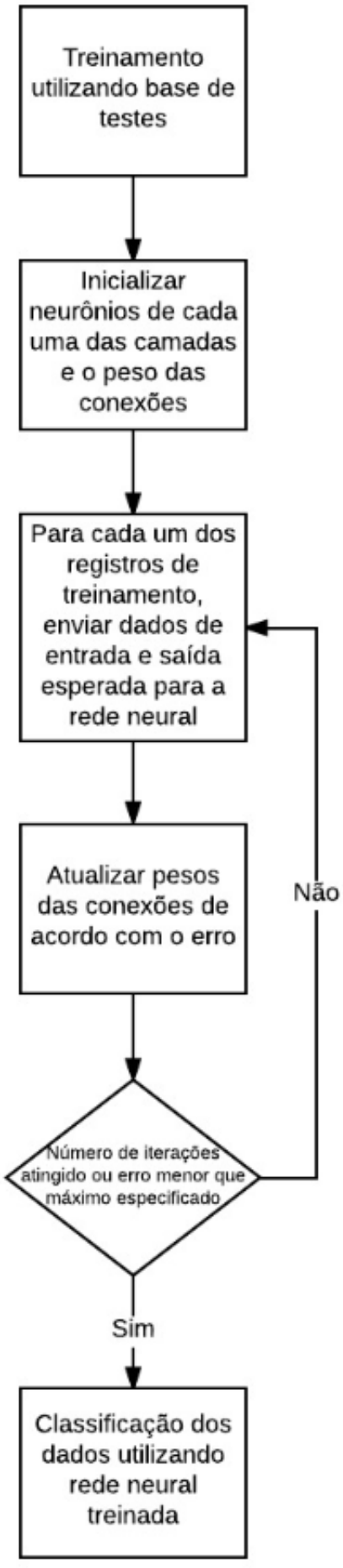


Figura 14 – Fluxo de execução da técnica Rede Neural Artificial

Fonte: (AUTORES, 2017)

5.6.2 Random Forest

O processo de previsão utilizado na técnica Random Forest é descrito na figura 15. O algoritmo desta técnica consiste na criação de um número definido de árvores de decisão. Ao fazer a classificação do registro, o programa encontra a classificação que foi prevista pelo maior número de árvores de decisão.

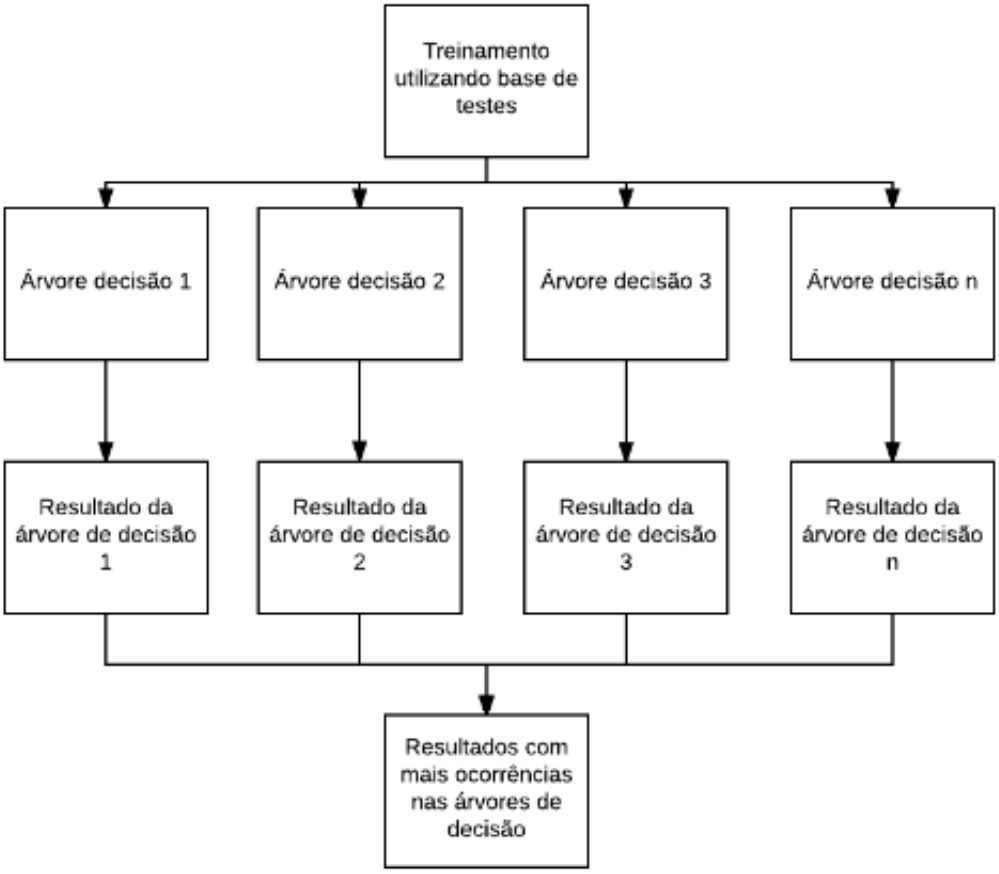


Figura 15 – Fluxo de execução da técnica Random Forest

Fonte: (AUTORES, 2017)

5.6.3 Support Vector Machine

O objetivo da técnica Support Vector Machine é obter um hiperplano que classifique todos os registros em suas classes. A figura 16 mostra dados classificados por um algoritmo utilizando a técnica.

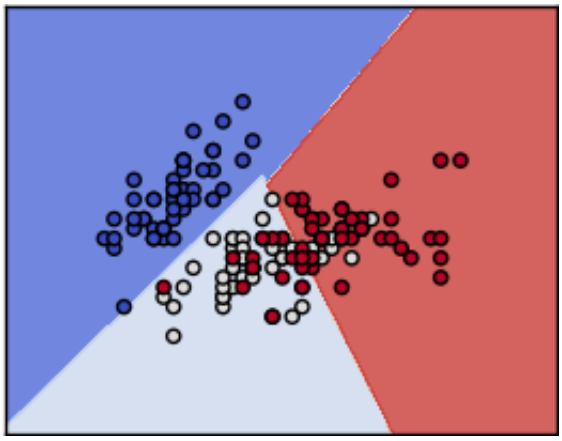


Figura 16 – Exemplo de classificação utilizando SVM

Disponível em: scikit-learn.org/stable/modules/svm.html. Acessado em 17/06/2017

O processo para chegar no hiperplano é descrito na figura 17. Inicialmente, o sistema faz o cálculo de hiperplanos que dividem as classes dos registros da base de treinamento. Após encontrar os hiperplanos, o sistema utiliza aquele que possuir a maior margem entre as classes, pois este tem maior possibilidade de classificar corretamente novos registros.



Figura 17 – Fluxo de execução da técnica Support Vector Machine

Fonte: (AUTORES, 2017)

5.6.4 K-fold cross validation

Segundo (KOHAVI et al., 1995), o método k-fold cross-validation é uma técnica para calcular a capacidade de generalização de um sistema de predição. Os dados originais são

divididos em k partes, no caso do sistema desenvolvido, eles são divididos em 10 partes, onde cada parte possui um conjuntos de dados com aproximadamente o mesmo tamanho das outros. Para cada parte é feito o treinamento do programa com todos os dados da base menos os dados da parte atual e a validação dos resultados do programa é feita com o conjunto de dados da parte atual. Para cada parte de dados é gerada uma porcentagem de acertos do programa. O valor a ser utilizado para validação do sistema será a média de acertos entre todas as partes geradas.

Quanto maior o valor de k da técnica k-fold cross validation, mais custoso o processamento, maior a variação dos dados, aumentando a confiabilidade dos testes, já quando menor o k, menos custoso o processamento, menor a variação dos dados, diminuindo a confiabilidade dos testes. De uma forma geral, muitas aplicações utilizam k igual a 10.

O processor de execução da técnica está descrito na figura 18.

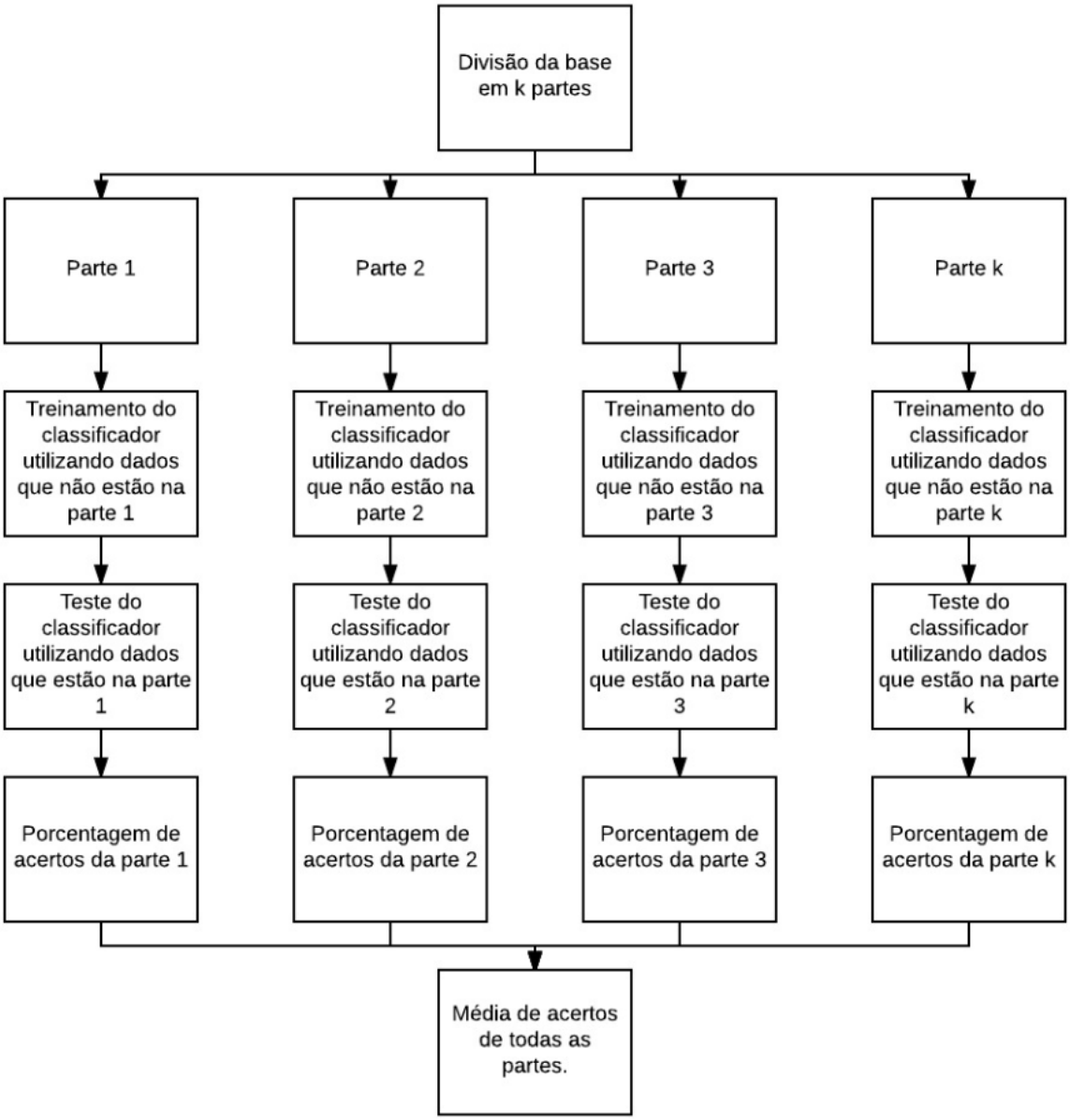


Figura 18 – Fluxo de execução da técnica k-fold cross validation

Fonte: (AUTORES, 2017)

5.6.5 Scikit-learn

Segundo (PEDREGOSA et al., 2011), o Scikit-learn é um módulo do python que faz a integração entre a linguagem e uma série de algoritmos de aprendizado de máquina para problemas supervisionados e não supervisionados. O foco do projeto é fornecer algoritmos de aprendizados de máquinas para usuários não especialistas em uma linguagem de alto nível.

As técnicas implementadas no trabalho utilizaram este projeto.

5.6.6 Pandas

Segundo (MCKINNEY et al., 2010), pandas é um biblioteca do python com estruturas de dados e ferramentas de estatística. A biblioteca pandas fornece uma classe chamada Dataframe, que tem a função de armazenar os dados a serem utilizados no treinamento e validação dos programas.

No sistema desenvolvido, os dados do banco de dados foram transformados em Dataframes e enviados para os métodos que fazem a previsão e a validação dos resultados.

5.7 Interface de utilização do sistema

Para possibilitar o auxílio à previsão de resultados para usuários foi desenvolvido um sistema WEB, onde é possível visualizar a porcentagem de acertos de cada conjunto de atributos e técnica, além da previsão de resultados de um conjunto de partidas.

O sistema foi desenvolvido utilizando HTML, Javascript e o framework Flask.

O Flask é um microframework desenvolvido em Python e foi desenvolvido com a ideia de ser muito rápida a criação de um projeto. O termo microframework significa que o Flask tem por objetivo manter o core do projeto simples, mas flexível.

5.7.1 Interface Inicial

Na figura 19 é possível visualizar as duas opções do sistema, que estão listadas a seguir:

- Calcular porcentagem de acertos do sistema: Interface para o sistema retornar a porcentagem de acertos de um conjunto de atributos e uma técnica. A precisão é calculada utilizando k-fold cross validation.
- Prever o resultado de partidas de futebol: Interface para o sistema exibir ao usuário a previsão do resultado de um conjunto de partidas de futebol. O usuário pode escolher quais partidas, atributos e técnicas deseja utilizar na previsão.

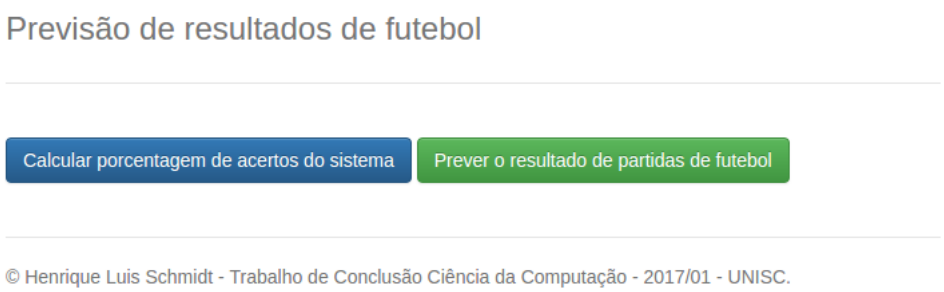


Figura 19 – Interface inicial do sistema desenvolvido

Fonte: (AUTORES, 2017)

Nas próximas seções será explicado o desenvolvimento e o uso das opções.

5.7.2 Cálculo de porcentagem

Esta interface oferece ao usuário a possibilidade de escolher quais atributos e técnicas ele deseja utilizar para verificar a porcentagem de acertos do sistema.

Na figura 20, é possível visualizar as escolhas do usuário, o resultado e tempo de processamento do sistema.



Figura 20 – Interface para cálculo de porcentagem de acertos

Fonte: (AUTORES, 2017)

A lista de execuções disponíveis, que representam os atributos utilizados, é a mesma apresentada na seção 5.8. Para cada uma das técnicas é exibido um botão com o texto configuração avançada que será explicado na seção 5.7.3.

Sempre que o usuário clica em Executar, a porcentagem de acertos e o tempo exibido são atualizados.

5.7.3 Configuração avançada de técnicas

Para cada uma das técnicas, o sistema possibilita a alteração dos parâmetros de execução dos algoritmos. Nas seções a seguir, a configuração avançada para cada uma das técnicas é explicada.

5.7.3.1 Random Forest

Na figura 21 são exibidos os parâmetros de configuração da técnica Random Forest, que são explicados na lista a seguir.

- **Número de árvores na floresta:** Número de árvores a serem utilizadas na técnica.
- **Número mínimo de amostras para um nodo interno:** Número mínimo de registros que cada nodo interno deve possuir.
- **Profundidade máxima da árvore:** Número máximo de níveis para cada uma das árvores
- **Número de atributos para utilizar em cada árvore:** Número máximo de atributos que a técnica pode utilizar em cada árvore.

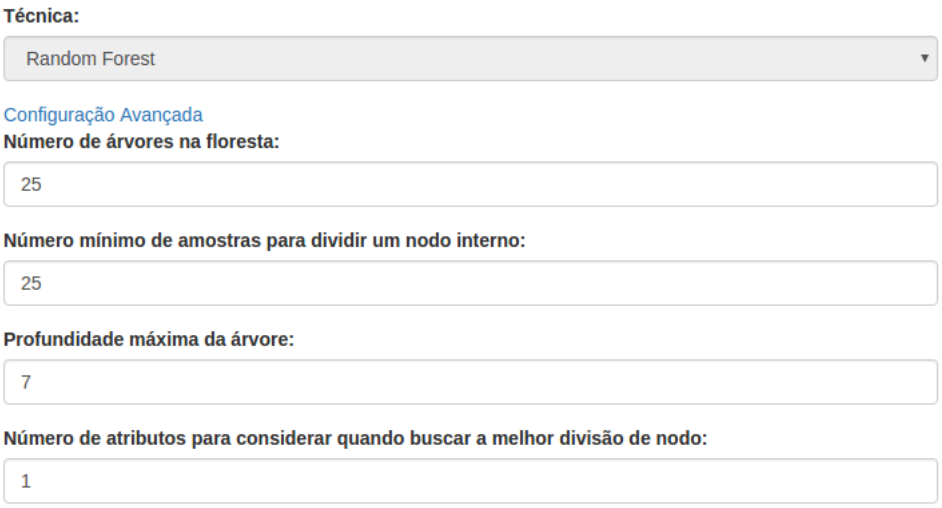


Figura 21 – Configuração avançada da técnica Random Forest

Fonte: (AUTORES, 2017)

5.7.3.2 Support Vector Machine

Na figura 22 são exibidos os parâmetros de configuração da técnica Support Vector Machine, que são explicados na lista a seguir.

- **Kernel:** Tipo do kernel utilizado no algoritmo.
- **Coeficiente do kernel:** Segundo (PEDREGOSA et al., 2011), o coeficiente do kernel influencia no quanto cada registro da base de treinamento influencia no classificador.
- **Parâmetro de penalidade do erro:** Segundo (PEDREGOSA et al., 2011), o parâmetro representa o quanto se quer evitar classificar errado cada registro da base de treinamento. Para valores altos, o classificador irá buscar um hiperplano de separação com margens pequenas, caso os valores sejam baixos, será buscado um classificador com margens maiores entre as possíveis classificações.



Figura 22 – Configuração avançada da técnica Random Forest

Fonte: (AUTORES, 2017)

5.7.3.3 Redes Neurais Artificais

Na figura 23 são exibidos os parâmetros de configuração da técnica Redes Neurais Artificiais, que são explicados na lista a seguir.

- **Número de camadas ocultas:** Número de camadas ocultas na rede neural entre a camada de entrada e de saída.
- **Número de neurônios nas camadas ocultas:** Número de neurônios de cada uma das camadas ocultas.
- **Número máximo de iterações:** Número de iterações a serem realizadas no treinamento da rede neural. A cada iteração, serão atualizados os pesos de cada uma das conexões.



Figura 23 – Configuração avançada da técnica Redes Neurais Artificiais

Fonte: (AUTORES, 2017)

5.7.4 Previsão de resultados de partidas de futebol

Esta interface permite ao usuário carregar as partidas de uma determinada equipe em uma temporada ou partidas em um período especificado.

Após carregar as partidas, o usuário pode escolher o conjunto de atributos, a técnica e a configuração avançada das técnicas para visualizar o resultado de cada uma das partidas carregadas.

Na figura 24, é possível visualizar a escolha da liga, temporada e equipe para busca das partidas.

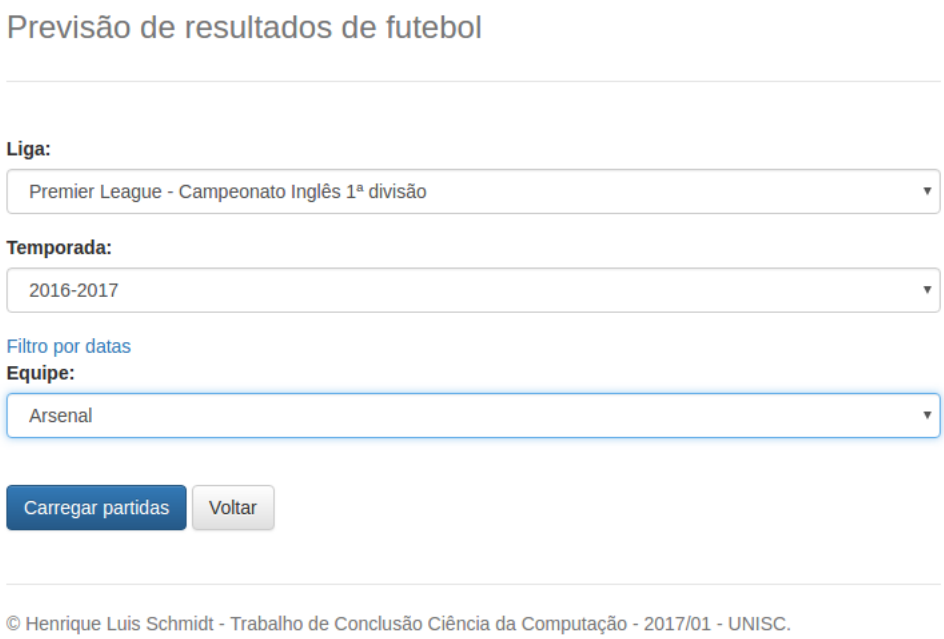


Figura 24 – Seleção de partidas por equipe

Fonte: (AUTORES, 2017)

O sistema também permite a busca de partidas em um período de datas, conforme exibido na figura 25.



Figura 25 – Seleção de partidas por período de datas

Fonte: (AUTORES, 2017)

Após a seleção da configuração de busca, o sistema exibe todas as partidas do banco de dados que atendem os requisitos da busca, conforme exibido na figura 26.

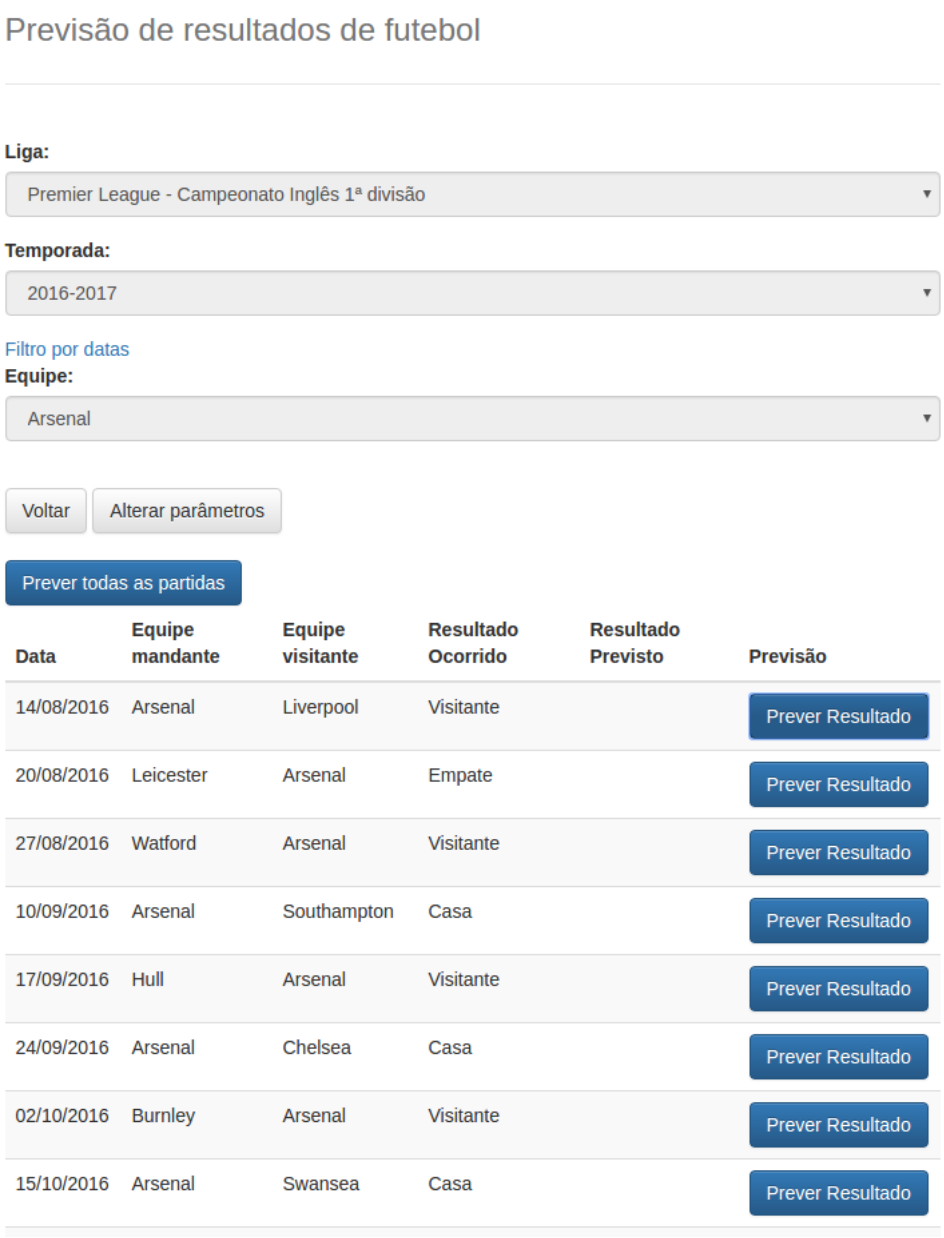


Figura 26 – Exibição de partidas fitradas

Fonte: (AUTORES, 2017)

A partir da exibição da lista de partidas, o sistema permite duas opções, que são listadas a seguir.

- Prever resultado de uma partida: Ao clicar no botão Prever Resultado de uma linha da lista de partidas, o sistema irá utilizar todas as outras partidas do banco de dados para treinar o algoritmo e utilizará esta linha para teste, exibindo o resultado calculado na coluna Resultado Previsto.
- Prever resultados da lista de partidas exibida: Ao clicar no botão Prever todas as partidas,

que está acima da lista de partidas, o sistema irá fazer a previsão de todas as partidas que estão na lista, utilizando todas as outras partidas do banco de dados como treino para o algoritmo. O resultado previsto das partidas será exibido na coluna Resultado Previsto e a porcentagem de acertos será exibida acima da lista.

Na figura 27, é possível visualizar a configuração da previsão de resultados das partidas. O usuário pode escolher o conjunto de atributos, a técnica e fazer a configuração avançada, conforme explicado na seção 5.7.3.

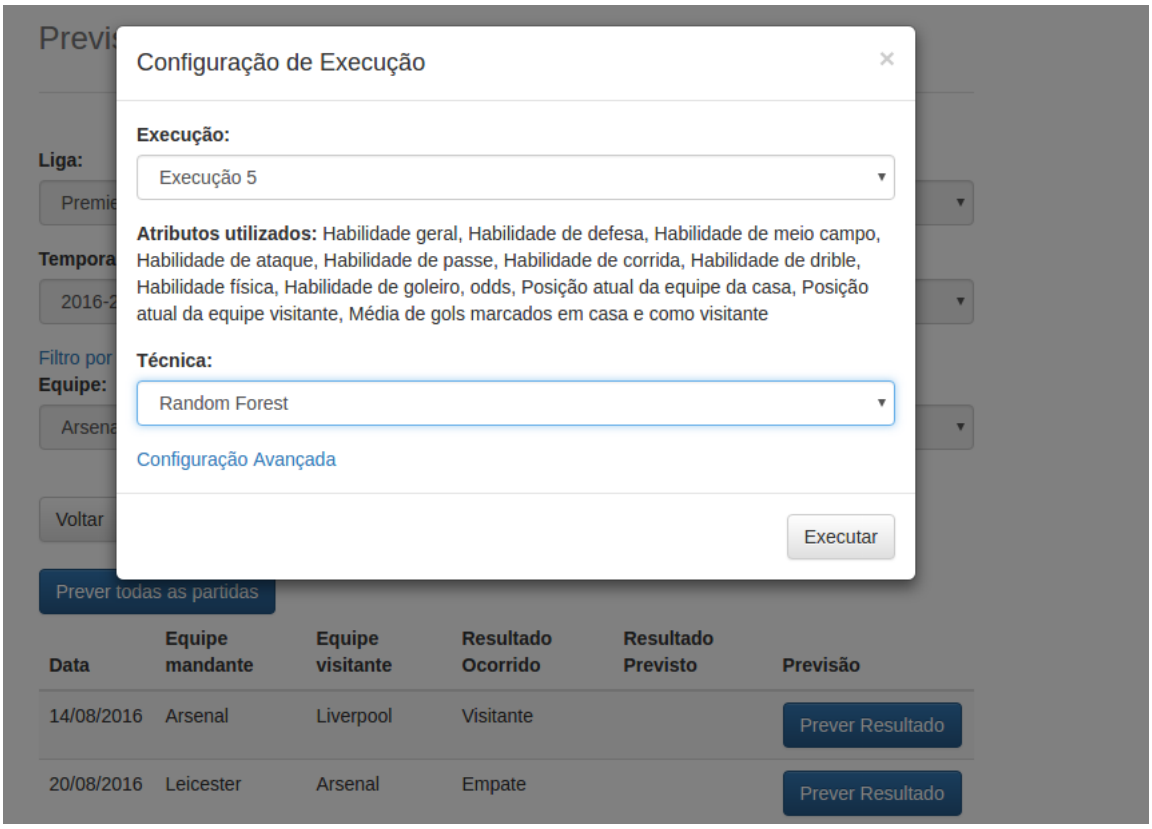


Figura 27 – Configuração da predição de partidas de futebol

Fonte: (AUTORES, 2017)

Após o usuário escolher a configuração de execução, o sistema exibe o resultado para cada uma das partidas e a porcentagem de acertos acima da lista, como é possível visualizar na figura 28.

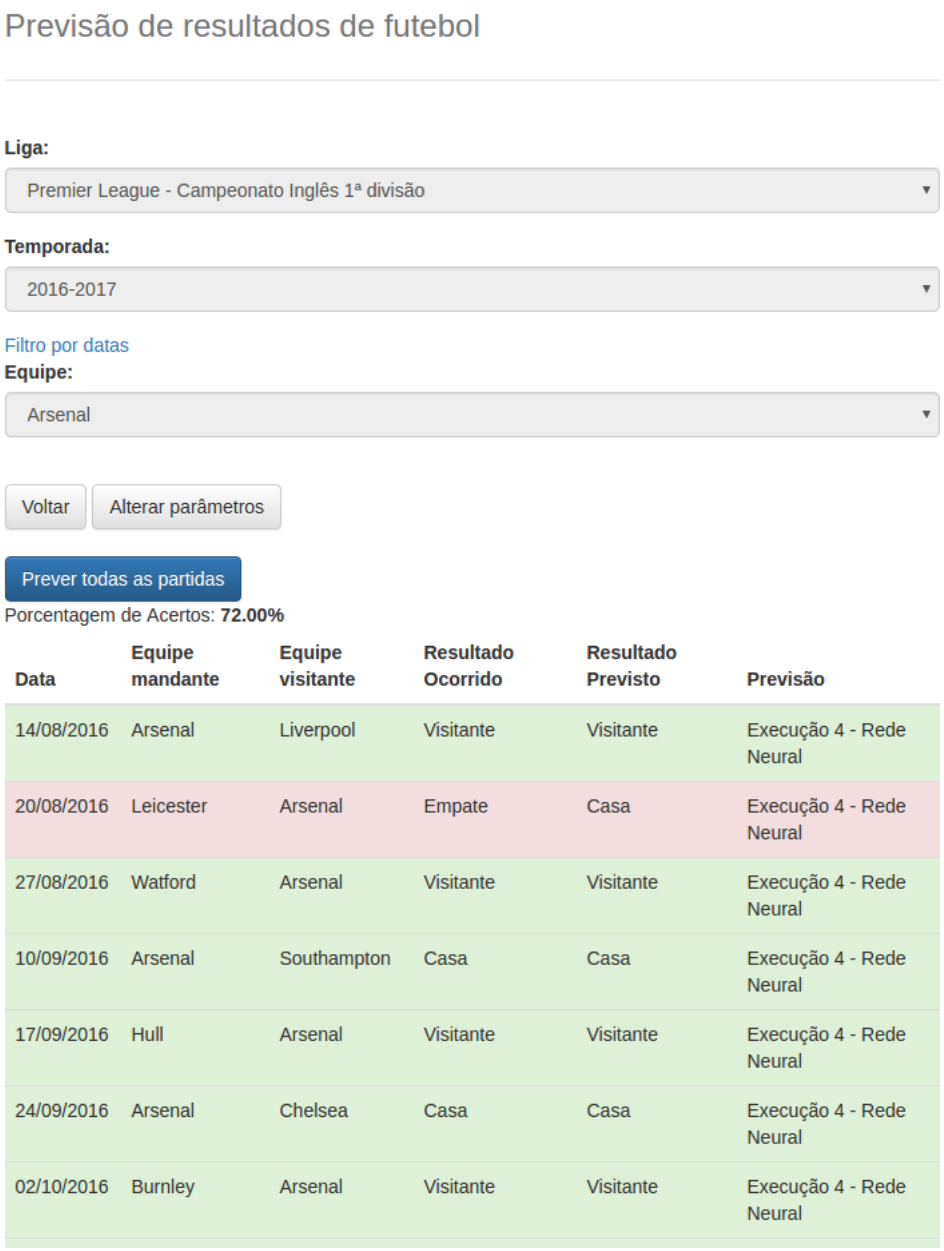


Figura 28 – Exibição da previsão de resultados

Fonte: (AUTORES, 2017)

5.7.5 Melhoria de desempenho

Para cada execução do programa é necessário fazer a busca de todos os dados de treinamento e validação na base de dados. Para isso é executado um comando SELECT que junta os dados de várias tabelas em uma lista de registros.

Na figura 29, é exibido o comando que busca os dados da execução 5. Como são utilizadas expressões complexas e diversas tabelas, a execução estava demorando em torno de 4 minutos.

```
SELECT
    CASE WHEN (def_casa.habilidade_geral)-def_visitante.habilidade_geral IS NOT NULL THEN (def_casa.habilidade_geral+50)-def_v
    CASE WHEN (def_casa.habilidade_defesa) - def_visitante.habilidade_defesa IS NOT NULL THEN (def_casa.habilidade_defesa+50)
    CASE WHEN (def_casa.habilidade_meio_campo) - def_visitante.habilidade_meio_campo IS NOT NULL THEN (def_casa.habilidade_mei
    CASE WHEN (def_casa.habilidade_ataque) - def_visitante.habilidade_defesa IS NOT NULL THEN (def_casa.habilidade_ataque+50)
    CASE WHEN (def_casa.habilidade_passe) - def_visitante.habilidade_passe IS NOT NULL THEN (def_casa.habilidade_passe+50) - d
    CASE WHEN (def_casa.habilidade_corrida) - def_visitante.habilidade_corrida IS NOT NULL THEN (def_casa.habilidade_corrida+5
    CASE WHEN (def_casa.habilidade_drible) - def_visitante.habilidade_drible IS NOT NULL THEN (def_casa.habilidade_drible+50)
    CASE WHEN (def_casa.habilidade_fisica) - def_visitante.habilidade_fisica IS NOT NULL THEN (def_casa.habilidade_fisica+50)
    CASE WHEN (def_casa.habilidade_goleiro) - def_visitante.habilidade_goleiro IS NOT NULL THEN (def_casa.habilidade_goleiro+5
    (SELECT nr_posicao FROM classificacao WHERE id_partida = p.id_partida AND id_equipe = p.id_equipe_casa) AS nr_posicao_casa
    (SELECT nr_posicao FROM classificacao WHERE id_partida = p.id_partida AND id_equipe = p.id_equipe_visitante) AS nr_posicao
    (SELECT CASE WHEN (SELECT nr_posicao FROM classificacao WHERE id_temporada = c.id_temporada-1 AND nr_rodada = 38 AND
    id_equipe = c.id_equipe) IS NULL THEN 10 ELSE (SELECT nr_posicao FROM classificacao WHERE id_temporada = c.id_temporada-1
    id_equipe = c.id_equipe) END FROM classificacao c WHERE id_partida = p.id_partida AND id_equipe = p.id_equipe_casa) AS nr_
    (SELECT CASE WHEN (SELECT nr_posicao FROM classificacao WHERE id_temporada = c.id_temporada-1 AND nr_rodada = 38 AND
    id_equipe = c.id_equipe) IS NULL THEN 10 ELSE (SELECT nr_posicao FROM classificacao WHERE id_temporada = c.id_temporada-1
    id_equipe = c.id_equipe) END FROM classificacao c WHERE id_partida = p.id_partida AND id_equipe = p.id_equipe_visitante) A
    p.nr_media_odds_casa - nr_media_odds_visitante AS "odds",
    (SELECT media_gols_marcados_casa FROM classificacao WHERE id_partida = p.id_partida AND id_equipe = p.id_equipe_visitante)
    (SELECT media_gols_marcados_visitante FROM classificacao WHERE id_partida = p.id_partida AND id_equipe = p.id_equipe_visit
    p.resultado_partida AS "resultado_partida",
    p.dt_partida,
        p.id_temporada,
        p.id_equipe_casa,
        p.id_equipe_visitante,
        p.id_partida
     FROM partida p
JOIN equipe e_casa ON e_casa.id_equipe = p.id_equipe_casa
LEFT JOIN dados_equipe_fifa def_casa ON def_casa.id_equipe = e_casa.id_equipe AND def_casa.id_temporada = p.id_temporada
JOIN equipe e_visitante ON e_visitante.id_equipe = p.id_equipe_visitante
LEFT JOIN dados_equipe_fifa def_visitante ON def_visitante.id_equipe = e_visitante.id_equipe AND def_visitante.id_temporada =
```

Figura 29 – Comando para busca dos dados de execução

Fonte: (AUTORES, 2017)

Para a melhora do desempenho e para deixar a previsão dos resultados mais rápida, foi criada uma tabela para armazenar o resultado deste comando. Após a execução da importação dos dados, foi executado um programa que populou esta tabela com os resultados. A tabela criada para a execução 5 é exibida na figura 30.



Figura 30 – Tabela para armazenar dados da execução 5

Fonte: (AUTORES, 2017)

Após a criação da tabela, o sistema passou a fazer uma consulta simples na tabela execucao_5, apenas filtrando as partidas que estão na base de dados de treinamento ou de validação. Com a nova tabela, as consultas diminuíram de 4 minutos para 50 milissegundos. Foram criadas tabelas para todos os conjuntos de atributos descritos na seção 5.8.

5.8 Resultados

Foram desenvolvidos algoritmos para fazer a previsão de resultados utilizando as técnicas Random Forest (RF), Support Vector Machine (SVM) e Redes Neurais Aritificais (RNA). Para a validação do programa foi utilizada a teécnica k-cross-validation, com o valor de k igual a 10.

Atributos diferentes para cada uma das execuções foram selecionados para verificar quais têm maior relevância na previsão dos resultados. Em algumas das execuções foram utilizados apenas atributos relacionados ao jogo de computador Fifa, em outras apenas dados históricos e também feitas execuções com a combinação de ambos os tipos de dados.

Na tabela 4 estão listadas a porcentagem de acertos de cada técnica com cada grupo de atributos, para uma das execuções realizadas.

Tabela 4 – Resultados das execuções

Conjunto de atributos	Número de registros	RF	SVM	RNA
1	2530	49,24%	46,08%	42,33%
2	2530	50,87%	46,48%	46,16%
3	5570	57,36%	57,03%	58,09%
4	5570	58,07%	56,44%	58,77%
5	5570	56,28%	51,83%	56,46%

Fonte: (AUTORES, 2017)

Os dados de cada um dos conjuntos de atributos estão listados a seguir:

- 1: Habilidade geral, habilidade do setor defensivo, habilidade do setor de meio campo, habilidade do setor de ataque.
- 2: Habilidade geral, habilidade do setor defensivo, habilidade do setor de meio campo, habilidade do setor de ataque, habilidade de passe, habilidade de corrida, habilidade de goleiro, habilidade física, odds
- 3: Posição atual do time da casa, posição atual do time visitante
- 4: Posição atual do time da casa, posição atual do time visitante, posição das equipes no ano anterior, média de gols das equipes em casa e como visitante
- 5: Habilidade geral, habilidade do setor defensivo, habilidade do setor de meio campo,

habilidade do setor de ataque, habilidade de passe, habilidade de corrida, habilidade de goleiro, habilidade física, posição atual do time da casa, posição atual do time visitante, odds, posição das equipes no ano anterior, média de gols das equipes em casa e como visitante

Como é possível visualizar nos conjuntos, o conjunto 1 possui apenas dados de habilidade no jogo de computador Fifa. O conjunto 2 possui os dados de habilidade no jogo em conjunto com as chances de vitória de acordo com os apostadores. O conjunto três utiliza apenas as posições das equipes na partida atual e o conjunto 4 faz a união da posição atual com a posição das equipes no ano anterior e a média de gols. O conjunto 5 é a união de todos os conjuntos. O critério para separação dos conjuntos foi o teste com diferentes tipos de atributos, alguns utilizando dados estatísticos, outros utilizando dados do jogo Fifa e outros combinando os dados.

O melhor resultado atingido pelo trabalho foi utilizando a técnica Rede Neural Artificial com o conjunto de atributos 4, que utiliza as posições das equipes na temporada atual, temporada passada e a média de gols das equipes, sendo possível atingir 58,77% de acertos no resultado de uma partida de futebol.

O conjunto 5, que possui todos os atributos, não teve um resultado melhor que os outros e isso mostra a importância da fase de seleção de atributos, pois uma quantidade muito grande de atributos pode ajudar na redução do percentual de acertos.

Analisando a tabela 2, é possível visualizar que a porcentagem de acertos deste trabalho foi menor que a dos trabalhos relacionados, podendo isso ser resultados das causas listadas a seguir:

- Dados mais detalhados sobre as partidas. Alguns trabalhos possuíam dados estatísticos como posse de bola, assistências e outros dados subjetivos como moral da equipe. Exemplos destes são encontrados nos trabalhos (CARPITA et al., 2015), (CONSTANTINOU; FENTON, 2013) e (TÜFEKCI, 2016).
- Execução das técnicas com diferentes parâmetros. Por exemplo, número de camadas intermediárias na Rede Neural Artificial ou número de árvores na técnica Random Forest.

6 CONCLUSÃO

Este trabalho de conclusão propôs, através do uso de técnicas de aprendizado de máquina, auxiliar na previsão de resultados de futebol.

Para atingir este objetivo foram estudados trabalhos relacionados que faziam a predição de eventos esportivos, foram levantadas as técnicas utilizadas por cada um dos trabalhos e feita uma pesquisa para buscar uma compreensão do funcionamento e objetivo de cada um dos algoritmos. Também foi realizada uma comparação entre os melhores resultados destes trabalhos com o objetivo de selecionar uma lista de técnicas a serem desenvolvidas.

Para a realização deste trabalho foi necessária a busca de dados estatísticos sobre o domínio futebol, analisadas as fontes de dados dos trabalhos relacionados e encontrados dois sites que possuem uma boa quantidade de dados para as entradas do sistema.

Foi desenvolvido um sistema para importar os dados dos dois sites e com os dados estruturados realizou-se uma limpeza nos dados e a seleção dos atributos mais importantes. Com os dados limpos e uma base de dados menor, foi feito o desenvolvimento de um sistema para auxiliar a previsão dos resultados de partidas de futebol utilizando as três técnicas escolhidas, Random Forest, Support Vector Machine e Redes Neurais Artificiais. Após o desenvolvimento, fez-se uma comparação entre os resultados de cada uma das técnicas para definir qual a mais eficiente no domínio deste trabalho.

Com o desenvolvimento deste trabalho, verificou-se que é possível obter um bom aproveitamento no auxílio à previsão de resultados de futebol utilizando técnicas de aprendizado de máquina. Com as técnicas Random Forest, Redes Neurais Artificiais e Support Vector Machine e com as informações de posição das equipes na rodada da partida, na temporada anterior e a média de gols marcados e sofridos foi possível chegar a um aproveitamento de 58,77%, que é muito bom levando em consideração que há três possibilidades de resultado e um programa que gerasse resultados aleatórios levaria a uma porcentagem de acertos de 33,33%.

Na comparação deste trabalho com os trabalhos relacionados, a porcentagem de acertos foi um pouco menor, podendo isso ser resultado de atributos diferentes na base de dados ou configurações de execuções diferentes para as técnicas de aprendizado de máquina. Os trabalhos relacionados não mostraram todo o conjunto de dados, por isso, há uma dificuldade para avaliar este ponto.

O trabalho também comprovou a viabilidade da utilização de técnicas de aprendizado de máquina para problemas complexos, onde programas tradicionais não têm êxito na solução.

Além disso, foi possível identificar pontos importantes de favoritismo de uma equipe em uma partida, o que pode auxiliar no treinamento de atletas em clubes de futebol.

Para trabalhos futuros, sugere-se a utilização de uma base de dados maior, com partidas de mais ligas de diversos países, o que poderia deixar o treinamento do classificador desenvolvido mais eficiente. Outro ponto a ser estudado é a utilização de atributos diferentes no auxílio à previsão dos resultados. A validação deste trabalho utilizou a técnica k-fold cross validation e trabalhos futuros podem utilizar outras técnicas de validação, como leave-one-out, monte carlo e ROC, aumentando a confiabilidade do percentual de acertos retornado pelo sistema.

Este trabalho teve como foco o auxílio à previsão de partidas em ligas de longa duração e há a possibilidade de auxílio na previsão de torneios curtos, como a Copa do Mundo.

Existem diversas técnicas na área de aprendizado de máquina e este trabalho abordou três delas. Trabalhos futuros podem implementar configurações de execução diferentes nas técnicas escolhidas neste trabalho e selecionar outras para fazer a comparação de desempenho.

Outra oportunidade para trabalhos futuros é o desenvolvimento de uma forma automática de obtenção de resultados de futebol, fazendo com que o programa esteja sempre treinado com os últimos resultados ocorridos na realidade, possibilitando a melhoria do percentual de acertos do sistema sem a intervenção de nenhum usuário.

REFERÊNCIAS

ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010.

BARBOSA, R. M. et al. The use of decision trees and naïve bayes algorithms and trace element patterns for controlling the authenticity of free-range-pastured hens' eggs. *Journal of food science*, Wiley Online Library, v. 79, n. 9, p. C1672–C1677, 2014.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

CARPITA, M. et al. Discovering the drivers of football match outcomes with data mining. *Quality Technology & Quantitative Management*, Taylor & Francis, v. 12, n. 4, p. 561–577, 2015.

CHAGAS, J. M. et al. A (im) possibilidade de regulamentação das apostas esportivas no ordenamento jurídico brasileiro. Florianópolis, SC, 2016.

CONSTANTINOU, A. C.; FENTON, N. E. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, v. 9, n. 1, p. 37–50, 2013.

CONSTANTINOU, A. C.; FENTON, N. E.; NEIL, M. pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, Elsevier, v. 36, p. 322–339, 2012.

CONSTANTINOU, A. C.; FENTON, N. E.; NEIL, M. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. *Knowledge-Based Systems*, Elsevier, v. 50, p. 60–86, 2013.

FAUSETT, L. *Fundamentals of neural networks: architectures, algorithms, and applications*. [S.l.]: Prentice-Hall,, 1994.

GODDARD, J. Regression models for forecasting goals and match results in association football. *International Journal of forecasting*, Elsevier, v. 21, n. 2, p. 331–340, 2005.

GOMES, B. G. et al. Treinamento supervisionado para previsao de partidas de futebol: Uma abordagem usando dados de videogames. 2016.

HAYKIN, S. A comprehensive foundation. *Neural Networks*, v. 2, n. 2004, p. 41, 2004.

HECKERMAN, D.; GEIGER, D.; CHICKERING, D. M. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, Springer, v. 20, n. 3, p. 197–243, 1995.

HILL, T. et al. Artificial neural network models for forecasting and decision making. *International Journal of Forecasting*, Elsevier, v. 10, n. 1, p. 5–15, 1994.

HUCALJUK, J.; RAKIPOVIĆ, A. Predicting football scores using machine learning techniques. In: IEEE. *MIPRO, 2011 Proceedings of the 34th International Convention*. [S.l.], 2011. p. 1623–1627.

HVATTUM, L. M.; ARNTZEN, H. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, Elsevier, v. 26, n. 3, p. 460–470, 2010.

KNORR-HELD, L. Dynamic rating of sports teams. *Journal of the Royal Statistical Society: Series D (The Statistician)*, Wiley Online Library, v. 49, n. 2, p. 261–276, 2000.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

LI, Z. et al. Water pipe condition assessment: a hierarchical beta process approach for sparse incident data. *Machine learning*, Springer, v. 95, n. 1, p. 11–26, 2014.

LOPE, J. de; MARAVALL, D. et al. Robust high performance reinforcement learning through weighted k-nearest neighbors. *Neurocomputing*, Elsevier, v. 74, n. 8, p. 1251–1259, 2011.

MARQUES, R. L.; DUTRA, I. Redes bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. *Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil*, 2002.

MARTIN-BARRAGAN, B.; LILLO, R.; ROMO, J. Interpretable support vector machines for functional data. *European Journal of Operational Research*, Elsevier, v. 232, n. 1, p. 146–155, 2014.

MCGOVERN, A. et al. Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Machine learning*, Springer, v. 95, n. 1, p. 27–50, 2014.

MCKINNEY, W. et al. Data structures for statistical computing in python. In: SCIPY AUSTIN, TX. *Proceedings of the 9th Python in Science Conference*. [S.l.], 2010. v. 445, p. 51–56.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013.

MITCHELL, T. M. Machine learning. *Artificial Intelligence*, 1997.

MJOLSNESS, E.; DECOSTE, D. Machine learning for science: state of the art and future prospects. *Science*, American Association for the Advancement of Science, v. 293, n. 5537, p. 2051–2055, 2001.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, v. 1, n. 1, 2003.

PARKS, R. W.; LEVINE, D. S.; LONG, D. L. *Fundamentals of neural network modeling: Neuropsychology and cognitive neuroscience*. [S.l.]: Mit Press, 1998.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

RUDIN, C.; WAGSTAFF, K. L. Machine learning for science and society. *Machine Learning*, Springer, v. 95, n. 1, p. 1–9, 2014.

TÜFEKCI, P. Prediction of football match results in turkish super league games. In: SPRINGER. *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015*. [S.l.], 2016. p. 515–526.

WILDE, P. D. *Neural network models: theory and projects*. [S.l.]: Springer Science & Business Media, 2013.