PREMIER-LEAGUE: PREVENDO VENCEDORES

Murilo V. Pinheiro, Guilherme de M. Furtado & João Paulo A. Lima
Graduandos de Ciência da Computação
Universidade Federal do Ceará (UFC)
CEP 60.440-900 – Fortaleza – CE – Brasil

ABSTRACT

Trabalho Final da Disciplina de Aprendizagem de Máquina onde nós desenvolvemos um projeto utilizando dados por nós escolhidos aplicando todo o conhecimento desenvolvido na disciplina. Nosso projeto procura prever o resultados de partidas da Liga Inglesa, Premier-League utilizando uma série de modelos de Aprendizagem de Máquina e de técnicas e conhecimentos adquiridos no decorrer da disciplina.

1 INTRODUÇÃO

O futebol, considerado o esporte mais famoso do mundo, tem conquistado cada vez mais popularidade nos últimos anos, especialmente no contexto das apostas esportivas. A predição de partidas de futebol tem atraído diversas pessoas que tem paixão pelo futebol, e com a popularização das apostas, os fans do esporte estão cada vez buscando melhores ferramentas que proporcionem predições e probabilidades mais confiáveis. A predição de partidas de futebol é um problema desafiador por pode envolver uma enorme variedade de fatores, como desempenho do time nos últimos jogos, horário dos jogos, time da casa e time visitante, entre muitos outros.

Nesse contexto, nosso trabalho tem como objetivo desenvolver modelos de aprendizado de máquina capazes de classificar com eficiência os resultados das partidas da Premier League, uma das ligas mais prestigiadas e competitivas do mundo, situada na Inglaterra. Pretendemos realizar uma análise exploratória de dados, utilizando informações recentes e acumuladas sobre os times ao longo das temporadas, a fim de identificar padrões que possam ser utilizados na previsão dos resultados. A escolha da Premier League se deve a sua popularidade e a grande disponibilidade de informações para a criação do conjunto de dados. Ao abranger um período que vai de 2001 a 2022 e ter diversas estatísticas dos jogos, a liga proporciona uma ampla quantidade de instâncias e atributos para a nossa análise.

É importante destacar que, embora o aprendizado de máquina seja uma abordagem promissora, não podemos garantir com certeza absoluta a precisão das previsões obtidas. O futebol é um esporte complexo, influenciado por diversos fatores imprevisíveis, como lesões, decisões arbitrais e individualismo. No entanto, acreditamos que a aplicação de técnicas de aprendizado de máquina pode nos ajudar a ter uma ideia de fatores que influenciam significativamente em resultado de partidas.

Ao longo deste trabalho, apresentaremos a análise dos dados, a metodologia utilizada, os resultados obtidos e as considerações finais sobre a aplicação de modelos de aprendizado de máquina na previsão de resultados nas partidas da Premier League.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 PROBLEMA

O problema de prever resultados de partidas de esportes ou jogos, sempre foi altamente requisitado no mundo todo, seja pelas apostas ou apenas pela diversão e pela torcida. Em um país movido pelo futebol, buscar unir nosso mundo acadêmico com o mundo dos esportes é sempre uma tarefa complicada. Nosso objetivo é compreender como funcionam as partidas de futebol de uma das maiores ligas do mundo e tentar desenvolver modelos que, até certo nível, posso prever os resultados de um dos esportes mais imprevísíveis que já existiu.

Definir um vencedor nunca é fácil por isso nosso problema pode ser considerado de certa forma até complexo, mas utilizando as técnicas e algoritmos de aprendizagem de máquina, talvez sejamos capazes de encontrar resultados satisfatórios. A primeira parte do nosso trabalho buscou entender um pouco melhor do problema e por isso, a análise exploratória dos dados é crucial pro desenvolvimento do nosso projeto, como também na compreensão dos nossos futuros resultados.

Algumas das informações obtidas na análise dos dados que foram importantes para nosso desenvolvimento:

- O time da casa possui uma larga vantagem em relação ao time visitante em relação ao resultado das partidas, o time da casa vence 46.02% enquanto que o time visitante ganha 28.94% das partidas, sendo o resto das partidas os empates.
- Nos últimos anos, o time visitante tem tido um aumento na porcentagem de vitórias enquanto que o time da casa tem tido uma diminuição, o seguinte gráfico ilustra esse comportamento, pode ser observado em 1.

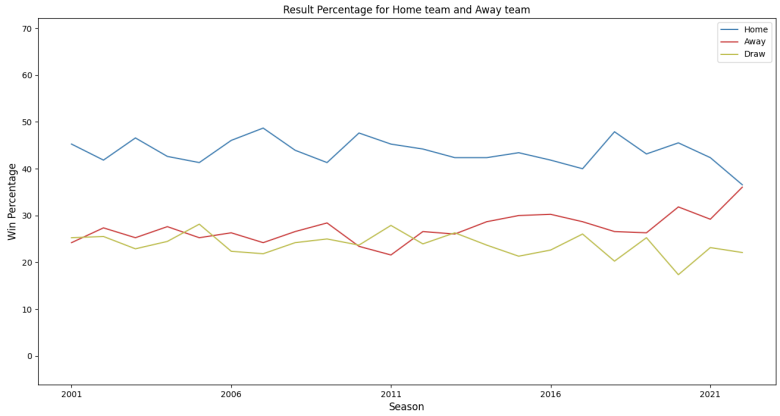

Figura 1: Vitórias de casa, vitórias de fora e empates ao longo das temporadas.

- Observamos a também que a taxa de vitória de um time em casa cai bastante quando ele está em várias derrotas seguidas, o inverso também acontece quando está vencendo muito, podemos observar em 2.

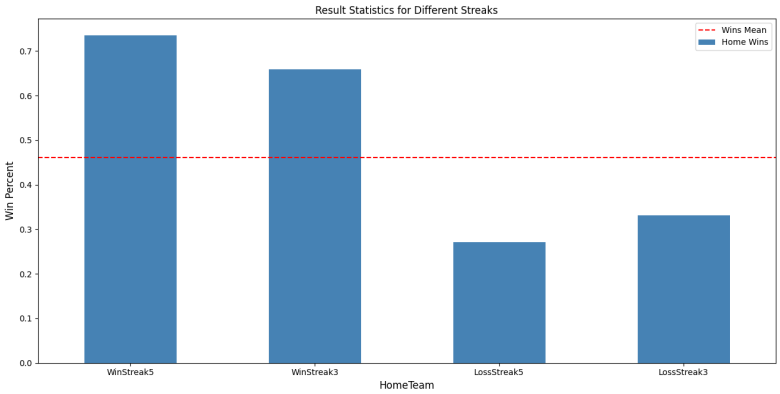

Figura 2: Taxas de Vitórias em casa quando consideramos uma sequência de vitórias ou derrotas.

2.2 DADOS OBTIDOS

A predição será feita com base nos dados de jogos passados. O dataset pré-processado possui 30 atributos comuns a todas as temporadas e consistem em estatísticas de cada partida das temporadas de 2000 a 2021 da Premier League Inglesa. Todos os dados foram obtidos do site football-data.co.uk.

Para que os modelos levem em consideração o desempenho dos times em partidas anteriores, agrupamos os dados por time e criamos atributos acumulativos normalizados e atributos "móveis", ou seja, que são obtidos das últimas três ou cinco partidas. Depois disso, não desejando inviezar os modelos, descartamos os atributos de nome dos times e também os inutilizáveis, como aqueles exclusivos a temporadas mais recentes, e os irrelevantes como, por exemplo, o nome do juíz da partida. Todos os dados utilizados nos modelos foram construídos com base nos históricos do time naquela temporada, e, principalmente no desempenho nos seus últimos três jogos. Com isso, o modelo é capaz de entender um pouco como está a situação daquele time na temporada e no momento atual, facilitando o entedimento de como ele está no presente e futuro, que é aquilo que se busca prever.

Vale lembrar também, que qualquer dado referente a partida atual foi completamente retirado dos dados, nenhum registro de um jogo tem dado sobre o jogo atual, pois seria vazar dados e informações que, em teoria, não teríamos no momento da predição.

2.3 TRABALHOS RELACIONADOS

Quanto à trabalhos relacionados tivemos como base para a ideia inicial baseando-se no artigo Ajgaonkar (2021), no entanto o autor utiliza-se de outras features, e faz uma análise diferente da nossa, os dados também são desatualizados em relação aos nossos, mas trazem ótimos resultados, ainda melhores, talvez devido às features empregadas, como não temos acesso à elas, como foram criadas, não utilizamos da mesma e resolvemos selecionar e criar outras features que trouxessem informações relevantes.

3 METODOLOGIA

3.1 DIVISÃO EM DOIS SUBPROBLEMAS

Nosso trabalho problema em geral foi dividido em dois subproblemas. Ao analisar todos os dados e como eles se comportavam decidimos criar esses dois subproblemas para entender melhor como poderíamos desenvolver nosso trabalho. O primeiro problema e o original busca fazer uma classificação multiclasse, sendo as classes em vitória de Casa (*HomeWin*), vitória de Fora (*AwayWin*) e Empate (*Draw*), esse problema apesar de aparentar simples pode ser muito complexo devido à natureza muitas vezes imprevísivel do esporte, devido a isso buscamos mudar um pouco nosso objetivo para reduzirmos nossa busca a uma classificação binária.

Como dito, o segundo problema foi a simplificação do primeiro, basicamente procuramos resolver o problema de: "O time de casa vai vencer?"; transformando as três classes anteriores em apenas duas, sendo a primeira vitória de casa e a segunda derrota de casa. Dessa forma, fomos capazes de entender melhor nosso ambiente de estudo e foi mais fácil de desenvolver técnicas para classificação binária que é um local de estudo mais conhecido por nosso grupo.

3.2 MODELOS DE APRENDIZAGEM DE MÁQUINA

Ao início, buscamos escolher modelos famosos como MLP (*MultiLayer Perceptron*), uma rede neural simples e com poucas camadas, *Random Forest*, um ensemble muito utilizado e Regressão Logística, um modelo linear muito utilizado em classificação, para resolver o problema da classificação multiclasse, assim utilizamos cada um deles com os mesmos dados, visando comparar o desempnho cada um em relação ao problema inicial.

Para o segundo caso reutilizamos dois dos modelos já citados, a *Random Forest* e Regressão Logística, mas optamos por não reutilizamos a rede neural, substituindo-a por uma *Extreme Gradient Boosting*, ou melhor *XGBoost*, um algoritmo baseado em árvores de decisões altamente poderoso, utilizado tanto para regressão quanto para classificação. Um detalhe importante é que a *Random Forest* utilizada foi feita com uma técnica diferente das demais. Tanto para a classificação binária

quanto multiclasse, buscamos prever o valor de gols de cada time, através de um problema de regressão, modificando o problema original, após isso tornamos a utilizar um limiar para determinar a vitória de um time ou outro, e no caso da multiclasse o empate.

3.3 Treinamento e Avaliação

No treinamento de todos os modelos foram feitas Validações Cruzadas para garantir uma melhor hiperparametrização e buscar melhores resultados no teste. A Validação Cruzada foi feito com 20% do conjunto de validação em relação ao treino, e foram feitas 5 vezes para cada modelo. Após o treinamento e validação dos hiperparâmetros é efetuado o teste do modelo, calculando as métricas, procuramos fazer uma breve análise dos resultados buscando trazer informações e um melhor entedimento de como o resultados de futebol e algoritmos de aprendizado de máquina funcionam.

Para a avaliação resolvemos utilizar três formas de avaliação dos resultados, sendo elas as métricas Acurácia, a Precisão e a F1-Score, e a quarta sendo a Matriz de Confusão Normalizada dos resultados do teste. Como temos uma espécie de série temporal, nossos dados de treino e teste foram divididos da seguinte forma: os dados de treino são os jogos do ano de 2001 até o ano de de 2015, o treino é composto dos jogos de 2016 até 2022. Ressaltamos que a Precisão e a F1-Score foram utilizados as variações *macro* no problema de multiclasse.

De acordo da natureza dos nossos dados, que dependem dos últimos três jogos de um time, nós retiramos dos dados de treino e teste todos os três primeiros jogos de cada time de cada temporada, dessa forma nosso treino e teste não pega dados ”incompletos”. Além disso, foi feito todo o tratamento padrão, os dados foram todos codificados em valores numéricos, e todos os valores foram normalizados para ficarem no intervalo de $[0, 1]$.

4 Experimentos

Após o treino, validação e teste, sob as condições de experimentos já apresentados, os resultados estão descritos e resumidos nos gráficos 3 e 4.

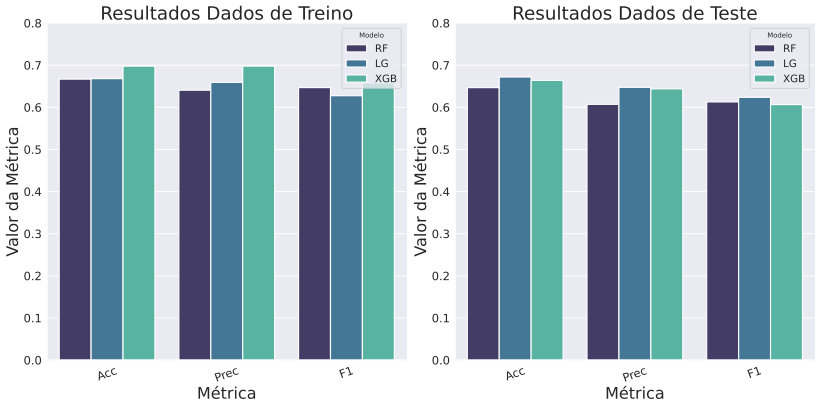

Figura 3: Resultados da Classificação Binária.

Nossos Resultados mostram como apesar de seguirem padrões demonstrados na nossa análise dos dados, ainda assim pode ser muito difícil prever o resultado de uma partida com uma alta certeza. O que mostra que o primeiro problema pode ser até complexo, mesmo com diversos jogos do passado ainda é complicado encontrar esses padrões.

O segundo retorna melhores resultados, devido a natureza do problema. Para observar melhor comportamento dos classificadores, as Matrizes de Confusão com melhores resultados foram as das figuras 5 e 6. As outras Matrizes de Confusão tiveram resultados semelhantes, mas um pouco piores.

As Matrizes de Confusão mostram a principal problemática da multiclassificação, os empates são muito difíceis de prever, isso pode se dever a uma não linearidade entre o empate e os dados que

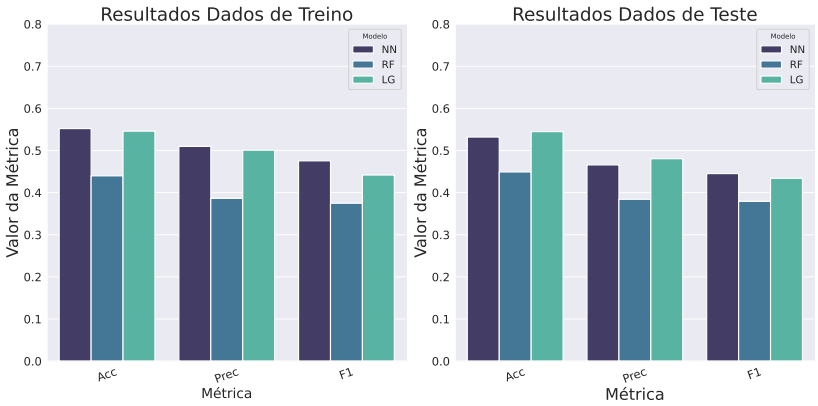

Figura 4: Resultados da Classificação Multiclasse.

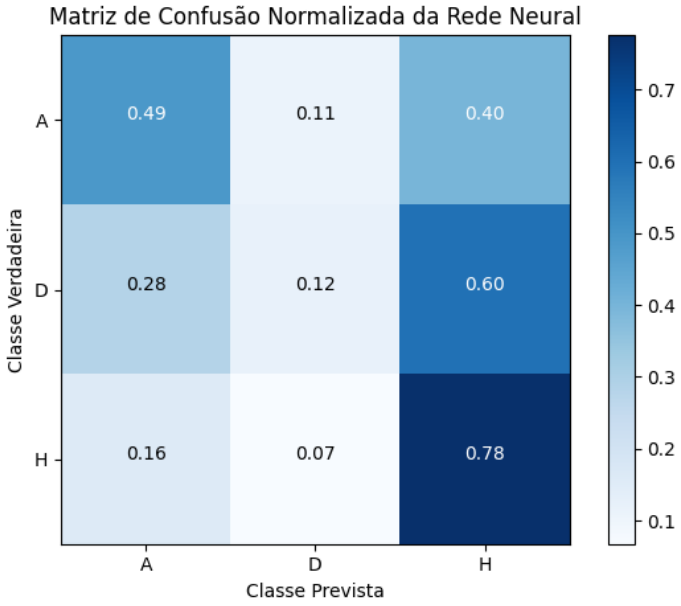

Figura 5: Matriz de Confusão Rede Neural.

utilizamos, mas isso pode se dever também à natureza do esporte, como já citado antes. Dessa forma, reduzir o problema trouxe diversos benefícios na resolução do nosso problema, além de trazer novas comparações e possíveis interpretações.

5 Conclusão

Sob o contexto da popularização de apostas esportivas e uma maior necessidade para análise de partidas e previsão de resultados, buscamos desenvolver modelos de aprendizado de máquina para prever os resultados das partidas da Premier League. Utilizamos principalmente de estatísticas acumulativas da temporada e das partidas mais recentes, na esperança de identificar padrões que pudessem contribuir para a classificação precisa dos resultados das partidas.

Entretanto, foi visto que a tarefa de prever o resultado de partidas não é tão fácil, diversos fatores imprevisíveis podem afetar uma partida, além de não ser muito raro a occorência de ”zebras”no futebol, onde um time bem menos prestigiado ganha de um time bem prestigiado. Embora tenhamos observado correlações interessantes entre certas variáveis e os resultados das partidas, essas impre-

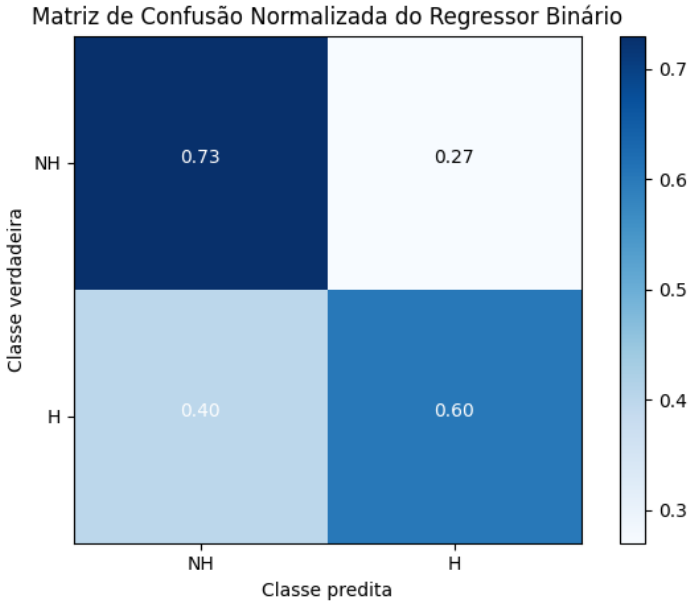

Figura 6: Matriz de Confusão da Regressão Logística para Classe binária.

visibilidades inerentes ao esporte se mostraram um obstáculo significativo. Os modelos tiveram uma acurácia por volta de 67% para classificação binária e por volta de 50% em classificação multiclasse, com o Random Forest tendo um pior desempenho nesse tipo de classificação.

Algumas propostas de investigação futuras são:

- Incorporar ao dataset dados de bets, que podem ser um ótimo atributo para identificar a diferença de desempenho dos times nos últimos jogos.
- Incluir o repertório de jogos entre equipes, algumas equipes podem ter mais vantagem sobre outras equipes no contexto histórico.
- Incorporação de novos atributos ou criação de atributos a partir de outros já existentes.
- Utilização de modelos mais complexos de aprendizagem de máquina.

Aplicando todas essas ideias e trabalhar nas construção de novas pode fazer com que nosso resultados elevem-se de diversas formas diferentes.

REFERÊNCIAS

Ajgaonkar, Bhoyar, P. S. (2021). Prediction of winning team using machine learning. *International Journal of Engineering Research & Technology (IJERT)*. Special Issue.

Bengio, Y. and LeCun, Y. (2007). Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT Press.

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.