*MACHINE LEARNING* APLICADO NA PREVISÃO DE PARTIDAS DE FUTEBOL DA PREMIER LEAGUE

Felipe de Oliveira

*Resumo* — A previsão de resultados de uma partida de futebol, através da utilização de algoritmos de *Machine Learning* é uma área de pesquisa que vem ganhando destaque, em parte devido à grande popularidade do esporte, aumento da quantidade de dados disponíveis e a imprevisibilidade existe no esporte. O presente trabalho apresenta um estudo do desempenho de diferentes classificadores para prever o resultado de partidas de futebol do campeonato Premier League. A partir de dados que retratam partidas realizadas no período 2005-2017, comparou-se o desempenho de três classificadores.

Palavras-Chave—Machine Learning, Premier League, futebol, previsão, python.

Abstract — The prediction of the results of a football match using *Machine Learning* algorithms is a research area that has been gaining prominence, in part due to the great popularity of the sport, increasing the amount of data available and the unpredictability in sports. The present work presents a study of the performance of different classifiers to predict the result of football matches of the Premier League championship. Based on data from 2005-2017 games, the performance of three classifiers was compared.

Keywords — Machine Learning, Premier League, soccer, prevision, python.

# Introdução

O futebol é o esporte mais praticado e assistido no mundo (DOBSON; 2001), sendo a Premier League (EPL, English Premier League) a liga futebolística com maior audiência, sendo que nos últimos anos, observou-se um aumento na utilização de métodos estatísticos para identificação de fatores determinantes para a previsão de resultados esportivos (HOPKINS, 2010), sendo a Premier League um dos campeonatos mais analisados, devido em parte à sua popularidade, imprevisibilidade, e audiência. Assim, observa-se que um dos problemas relacionados a previsões esportivas é a previsão do resultado final de partidas de futebol, sabendo que atualmente a acurácia média dos melhores modelos de previsão encontrados na literatura não ultrapassa a figura de 60% .Através da comparação de diversos classificadores, observa-se que a previsão do resultado de partidas de futebol através da utilização de classificadores de *Machine Learning* esbarra em um problema relacionado com a previsão de empates observaram acurácia inferior na previsão de instâncias como sendo pertencentes a classe empate quando comparada com a acurácia da classificação de instâncias como vitória do mandante ou visitante. Assim, é possível confirmar a necessidade de mais estudos de forma a possibilitar a replicação ou a solução deste problema ao se utilizar diferentes abordagens. Desta forma, neste trabalho busca-se discutir alternativas para previsão do resultado final de partidas de futebol utilizando-se métodos de aprendizado de máquina O objetivo principal deste trabalho é realizar uma avaliação do desempenho de três algoritmos de classificação quando os mesmos são utilizados para prever resultados de partidas de futebol contidas entre as temporadas 2005e 2017 da Premier League.

## Sobre o Simpósio

Este simpósio ocorre anualmente promovendo o encontro de mestrandos sobre na área de aprendizado de máquina em que são discutidos temas de abordagem aprendizado de máquina cotidiano.

# Fundamentação teórica

*Machine Learning* é uma importante área da [inteligência artificial](https://www.cetax.com.br/blog/inteligencia-artificial-beneficios-riscos/) onde é possível criar algoritmos para ensinar uma determinada máquina a desempenhar tarefas. Um algoritmo de *Machine Learning* possibilita pegar um conjunto de dados de entrada e com base em determinados padrões encontrados gerar as saídas. Cada entrada desse conjunto de dados possui suas próprias *features*, e ter um conjunto delas é o ponto inicial fundamental para qualquer algoritmo de Machine Learning.

Algumas áreas de aplicação de métodos de *Machine Learning* são:

* Detecção de fraudes.
* Resultados de pesquisa na Web.
* Anúncios em tempo real, tanto em páginas da web como em dispositivos móveis.
* Análise de sentimento baseada em texto.
* Pontuação de crédito e melhores ofertas.
* Previsão de falhas em equipamento.
* Novos modelos de precificação.
* Detecção de invasão em uma determinada rede.
* Reconhecimento de determinados padrões e imagens.
* Filtragem de spams em e-mail.

APREDINZADO SUPERVISIONADO

Nos é dado um conjunto de dados rotulados que já sabemos qual é a nossa saída correta e que deve ser semelhante ao conjunto, tendo a ideia de que existe uma relação entre a entrada e a saída.

Problemas de aprendizagem supervisionados são classificados em problemas de “regressão” e “classificação”. Em um problema de regressão, estamos tentando prever os resultados em uma saída contínua, o que significa que estamos a tentando mapear variáveis ​​de entrada para alguma função continua.

REGRESSÃO GAUSSIANA (RG)

processo gaussiano é um modelo estatístico em que as observações ocorrem em um domínio contínuo, por exemplo, tempo ou espaço. Em um processo gaussiano, cada ponto em algum espaço de entrada contínua está associada com uma [variável aleatória](https://pt.wikipedia.org/wiki/Vari%C3%A1vel_aleat%C3%B3ria) com [distribuição normal](https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_normal). Além disso, cada conjunto finito dessas variáveis ​​aleatórias tem uma distribuição normal multivariada. A distribuição de um processo gaussiano é a distribuição conjunta de todas as infinitas variáveis aleatórias, e, como tal, é uma distribuição de funções com um domínio contínuo.

Visto como um algoritmo de aprendizado de máquina, um processo gaussiano utiliza "[aprendizagem preguiçosa](https://pt.wikipedia.org/wiki/Lazy_Learning)" e uma medida da similaridade entre os pontos (a função kernel) para prever o valor de um ponto invisível a partir de dados de treinamento. A previsão não é apenas uma estimativa para esse ponto, mas tem também a informação da incerteza. É uma distribuição gaussiana unidimensional (que é uma distribuição marginal nesse ponto).

SUPPORT VECTOR MACHINE (SVR)

O SVR padrão toma como entrada um conjunto de dados e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVR um classificador linear binário não probabilístico. Dados um conjunto de exemplos de treinamento, cada um marcado como pertencente a uma de duas categorias, um [algoritmo](https://pt.wikipedia.org/wiki/Algoritmo" \o "Algoritmo) de treinamento do SVR constrói um modelo que atribui novos exemplos a uma categoria ou outra. Um modelo SVR é uma representação de exemplos como pontos no espaço, mapeados de maneira que os exemplos de cada categoria sejam divididos por um espaço claro que seja tão amplo quanto possível. Os novos exemplos são então mapeados no mesmo espaço e preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles são colocados.

Em outras palavras, o que uma SVR faz é encontrar uma linha de separação, mais comumente chamada de hiperplano entre dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes.

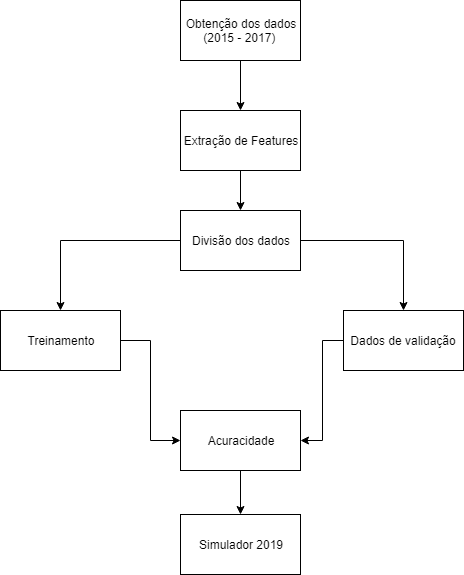
REGRESSÃO LINEAR MULTIPLA (RL)

A regressão, em geral, tem como objetivo tratar de um valor que não se consegue estimar inicialmente.

A regressão linear é chamada "linear" porque se considera que a relação da resposta às variáveis é uma [função linear](https://pt.wikipedia.org/wiki/Fun%C3%A7%C3%A3o_linear" \o "Função linear) de alguns parâmetros. Os modelos de regressão que não são uma função linear dos parâmetros se chamam modelos de [regressão não-linear](https://pt.wikipedia.org/wiki/Regress%C3%A3o_n%C3%A3o-linear" \o "Regressão não-linear). Sendo uma das primeiras formas de análise regressiva a ser estudada rigorosamente, e usada extensamente em aplicações práticas. Isso acontece porque modelos que dependem de forma linear dos seus parâmetros desconhecidos, são mais fáceis de ajustar que os modelos não-lineares aos seus parâmetros, e porque as propriedades estatísticas dos estimadores resultantes são fáceis de determinar.

# Metodologia

Este capítulo apresenta a metodologia utilizada no presente trabalho. A Figura 1 apresenta um fluxograma que resume as etapas a seguir.



1. Estrutura da metodologia empregado ao longo do projeto.

O conjunto de dados utilizado compreende todas as partidas realizadas no período de 2005 a 2017 no campeonato britânico Premier League. Estes dados foram obtidos gratuitamente no website football-data.co.uk. Toda a codificação foi implementada em Python, utilizando bibliotecas adicionais como a biblioteca pandas (leitura de arquivos do tipo CSV), numpy (cálculos em geral), e scikit-learn (machine learning) entre outras.

A partir dos dados coletados foram extraídas *features* a serem utilizadas no treinamento dos classificadores selecionados. Utilizaram-se 21 *features*, descritas a seguir:

1. Gols marcados
2. Gols tomados
3. Número de cartões amarelos
4. Número de cartões Vermelhos
5. Número de faltas ocorridas
6. Número de escanteios no jogo
7. Chance de ganhar o jogo
8. Número de defesas
9. Porcentagem de defesas
10. Razão de defesas e total de chutes
11. Razão de gols e chutes no gol
12. Razão de gols por chutes dados
13. Média de chutes ao gol por partida
14. Média de chutes leva por partida

Estruturou-se por tanto duas matrizes, apresentadas abaixo, sendo que a matriz A representas as *features* apresentas e matriz B se o time em casa venceu (1), empatou (0.5), perdeu (0).

.

Com o output é se classificado o resultado por meio das regras apresenta abaixo, tabela 1.

1. Regra de classificação de output

|  |  |
| --- | --- |
| Regra | Classificação |
| Se y > =0.55 | Vitória |
| Se 0.55>y>0.45 | Empate |
| Se y<=0.45 | Derrota |

Classificado o resultado dos jogos será calculado a acurácia por meios meio da equação 1:

(1)

Sendo que classificação correta significa a quantidade de partidas que o sistema que a classificação do programa foi igual ao resultado real e classificação errada representa a quantidade de vezes que houve divergência entre a classificação do programa e os resultados reais.

# Apresentação dos resultados

Primeiramente, analisou-se o desempenho dos classificadores variando os testes de modo aleatório, compilou-se os dados e registrou se os dados na Tabela II.

1. Acuracidade dos modelos testados

|  |  |  |  |
| --- | --- | --- | --- |
|  | Regressão Linear múltipla | SVR | Gaussiana |
| Acuracidade | 0.5144 | 0.4903 | 0.3841 |

Definido o método mais eficiente foi se utilizado o programa foi criado uma simulação de um campeonato completo, utilizando o método regressão linear múltipla. O resultado obtido pode ser visto na Tabela III:

1. Campeonato simulado por meio do programa

|  |  |  |
| --- | --- | --- |
| Posição | Time | Pontos |
| 1 | Burnley | 60 |
| 2 | Leicester | 60 |
| 3 | Crystal Palace | 59 |
| 4 | Southampton | 58 |
| 5 | West Ham | 58 |
| 6 | Watford | 58 |
| 7 | Bournemouth | 58 |
| 8 | West Brom | 56 |
| 9 | Arsenal | 54 |
| 10 | Everton | 54 |
| 11 | Man United | 54 |
| 12 | Liverpool | 54 |
| 13 | Swansea | 53 |
| 14 | Man City | 50 |
| 15 | Newcastle | 45 |
| 16 | Tottenham | 45 |

∎

# Conclusões

Pode-se perceber ao decorrer do projeto que o processo de estimativa de um resultado de um jogo tem muito a evoluir, visto que a tabela do nosso campeonato simulado diferiu muito da classificação real da temporada 2018-2019 na figura 2.



1. Classificação premier League 2018-2019

##### Agradecimentos

Ao professor Leonardo X por auxiliar na realização deste projeto desde da idealização aos métodos interessantes serem realizados.

##### Referências

1. L. Lamport, *A Document Preparation System: LATEX, User’s Guide and Reference Manual.* Addison Wesley Publishing Company, 1986.
2. F. C. Silva e J. J. Sousa, “Esta referência é apenas um exemplo”, *Revista de Exemplos*, v. 5, pp. 52–55, Maio 1999.
3. DOBSON, S.; GODDARD, J. The Economics of Football. 2001. Disponível em: <http://www.cambridge.org/>..
4. *HOPKINS, W. G. Impact Factors of Journals in Sport and Exercise Science and Medicine for 2010. Sportscience, 2010. v. 14, p. 60–62. Disponível em:* [*http://www.sportsci.org/2010/wghif.htm*](http://www.sportsci.org/2010/wghif.htm)*.*
5. Rasmussen, C. E. (2004). «Gaussian Processes in Machine Learning». Advanced Lectures on Machine Learning. Col: Lecture Notes in Computer Science. 3176. [S.l.: s.n.] pp. 63–71. [ISBN](https://pt.wikipedia.org/wiki/International_Standard_Book_Number) [978-3-540-23122-6](https://pt.wikipedia.org/wiki/Especial:Fontes_de_livros/978-3-540-23122-6). [doi](https://pt.wikipedia.org/wiki/Digital_object_identifier):[10.1007/978-3-540-28650-9\_4](https://dx.doi.org/10.1007%2F978-3-540-28650-9_4)
6. *Grimmett, Geoffrey; David Stirzaker (2001). Probability and Random Processes. [S.l.]:*[*Oxford University Press*](https://pt.wikipedia.org/wiki/Oxford_University_Press)*.*[*ISBN*](https://pt.wikipedia.org/wiki/International_Standard_Book_Number)[*0198572220*](https://pt.wikipedia.org/wiki/Especial:Fontes_de_livros/0198572220)
7. Green, William H. (2003). Econometric Analysis, fifth edition. [S.l.]: Prentice Hall. [ISBN](https://pt.wikipedia.org/wiki/International_Standard_Book_Number) [0-13-066189-9](https://pt.wikipedia.org/wiki/Especial:Fontes_de_livros/0-13-066189-9)
8. Hosmer, David W.; Stanley Lemeshow (2000). Applied Logistic Regression, 2nd ed. [S.l.]: New York; Chichester, Wiley. [ISBN](https://pt.wikipedia.org/wiki/International_Standard_Book_Number) [0-471-35632-8](https://pt.wikipedia.org/wiki/Especial:Fontes_de_livros/0-471-35632-8)
9. REIS, E., *Estatistica Descritiva* (2ª ed.). Lisboa: Edições Sílabo, 1994
10. Liu, W.; Principe, J.C.; Haykin, S. (2010). [Kernel Adaptive Filtering: A Comprehensive Introduction](https://web.archive.org/web/20160304042652/http:/www.cnel.ufl.edu/~weifeng/publication.htm). [S.l.]: [John Wiley](https://pt.wikipedia.org/wiki/John_Wiley_%26_Sons). [ISBN](https://pt.wikipedia.org/wiki/International_Standard_Book_Number) [0-470-44753-2](https://pt.wikipedia.org/wiki/Especial:Fontes_de_livros/0-470-44753-2). Consultado em 20 de setembro de 2016. Arquivado do [original](http://www.cnel.ufl.edu/~weifeng/publication.htm) em 4 de março de 2016

Maria da Silva e José da Silva¸ Faculdade de Ciências Aplicadas, Universidade Estadual de Campinas (UNICAMP), Limeira-SP, Brasil, E-mails: maria@unicamp.br, jose@unicamp.br. Este trabalho foi parcialmente financiado pelo CNPq (XX/XXXXX-X).