

Estudos Comparativos de Técnicas Machine Learning Aplicado em Dados da Covid-19

GONÇALVES, Samuel Pereira¹

LOURENÇO, Ramires de Oliveira¹

SANTOS, Flávia Aparecida Oliveira²

¹Discente – Ciência da Computação – Universidade José do Rosário Vellano

²Docente – Ciência da Computação – Universidade José do Rosário Vellano

RESUMO

O objetivo deste trabalho é realizar um estudo comparativo dos resultados apresentados por diferentes algoritmos de Machine Learning, aplicados em uma base de dados com informações sobre a COVID-19. Neste estudo, são demonstrados os diferentes tipos de comportamentos dos algoritmos utilizados, além de seus respectivos índices de acertos e erros. Para execução dos experimentos, foi utilizado o Weka, que dispõe de diferentes algoritmos para a aplicação dos comparativos. Os algoritmos utilizados nos experimentos são: *Lazy Kstar*, *Naive Bayes Multinomial Text*, *J48 Tree*, *Random Tree* e *Random Forest Tree*.

Palavras-chave

Machine Learning, Inteligência Artificial, Detecção de Falhas, Covid-19, Pandemia.

ABSTRACT

The objective of this work is to carry out a comparative study of the results obtained by different Machine Learning algorithms, scientific in a database with information about a COVID-19. In this study, the different types of behavior of the algorithms used are demonstrated, in addition to their respective success and error criteria. For the execution of the experiments, Weka was used, which has different algorithms for an application of comparatives. The algorithms used in the

experiments are: *Lazy Kstar*, *Naive Bayes Multinomial Text*, *J48 Tree*, *Random Tree* and *Random Forest Tree*.

Keywords

Machine Learning, Artificial Intelligence, Fault Detection, Covid-19, Pandemic.

1 INTRODUÇÃO

Entre o final do ano de 2019 e começo do ano de 2020, foram notificados os primeiros casos de um novo vírus respiratório, que acabou evoluindo para uma pandemia. O vírus em questão, denominado SARS-CoV-2, é o responsável pela doença da Covid-19 (STRABELLI e UIP, 2021).

Durante a pandemia, diversos grupos sociais foram afetados devido às medidas utilizadas para conter o avanço do vírus, como por exemplo, o *lockdown* e o isolamento social. Os estudantes fazem parte de um dos principais grupos afetados pela pandemia. Por esse motivo, o estudante da *University of Engineering & Technology Khulna*, Tamal Joyti Roy, realizou uma pesquisa com um

grupo de estudantes sobre as medidas de segurança adotadas durante o período pandêmico.

A pesquisa realizada por Tamal resultou em uma base de dados completa, que foi utilizada neste trabalho para treino dos algoritmos de *machine learning*. O método de aprendizagem utilizado foi Aprendizado Supervisionado, que, segundo MAINMON e ROKACH (2005), visa encontrar correlações dentre todas as variáveis de entrada com as variáveis de saída. Quando descoberta, essa correlação é representada como uma estrutura chamada de modelo.

Com os testes aplicados na base de dados, espera-se determinar qual dos algoritmos selecionados tem o melhor resultado. Com isso, os algoritmos que demonstraram melhor desempenho podem ser utilizados no desenvolvimento de novos recursos de Inteligência Artificial aplicada aos dados da Covid-19.

2 METODOLOGIA

2.1 Ferramentas e Dados

Para execução dos experimentos, foi utilizado o Weka, um software desenvolvido na linguagem Java. O Weka possui uma biblioteca composta de diversos algoritmos de *Machine Learning*, que podem ser aplicados em diferentes tipos de dados (Witten, et al. 1999). A base de dados utilizada no experimento está disponível no website kaggle (<https://www.kaggle.com/>), que é um repositório de dados gratuito para *data mining*.

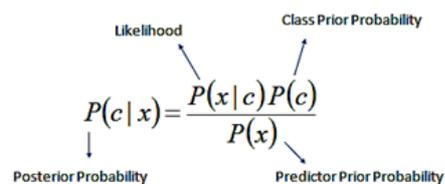
2.2 Criação do Ambiente

Antes da realização dos experimentos, foram selecionadas quatro perguntas da base de dados para serem submetidas ao Weka. As perguntas utilizadas foram:

- Se a vacina não estiver disponível, você poderá se mover livremente no futuro?
- Você acredita que a flexibilização do lockdown pode aumentar a propagação de COVID-19 no futuro?
- Você manteve um relacionamento saudável com sua família durante o lockdown?
- Você usa máscara quando sai?

Após a seleção dos dados para análise, foi feita a escolha de cinco algoritmos distintos para serem aplicados em cada uma das perguntas. Os algoritmos selecionados possuem diferentes características, sendo um algoritmo probabilístico (*bayes*), um algoritmo classificador baseado em exemplos (*lazy*) e três algoritmos de árvore de decisão (*tree*).

O algoritmo probabilístico, chamado *Multinomial Naive Bayes*, utiliza a fórmula ilustrada na Figura 1 (Fernando, 2019):

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$


$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

FIGURA 1 – Fórmula Naive Bayes

O algoritmo classificador baseado em exemplos, conhecido como Kstar, faz parte dos algoritmos de aprendizagem lenta, também conhecido como algoritmo preguiçoso. Esse algoritmo utiliza de famosas funções de distância, como a Distância Euclidiana, para encontro das medidas (CHELLAM; L; S, 2018).

Um dos algoritmos de árvore de decisão selecionados é chamado de Random Tree. Esse algoritmo baseia-se na elaboração de uma árvore considerando uma quantidade N de atributos, escolhidos de modo aleatório para cada nó (MARIN; LOPES, 2021).

De acordo com Oshiro (2021), o Random Forest também é um algoritmo de árvore de decisão. Conforme ilustrado na Figura 2, nesse algoritmo, são geradas várias árvores de decisão onde cada uma tem suas próprias singularidades. Após isso, é realizado um cálculo de média com o resultado de todas as árvores, gerando um resultado final, comumente mais precisa que a Random Tree.

Por fim, o algoritmo J48 é uma variação implementada na linguagem Java, baseada no algoritmo C4.5. Conforme ilustrado na Figura 3, esse algoritmo também realiza a elaboração de uma árvore de decisão com dados pré-montados. Após a montagem da árvore, o algoritmo seleciona em cada nó o atributo mais competente que realiza a divisão das amostras em subamostras homogêneas.

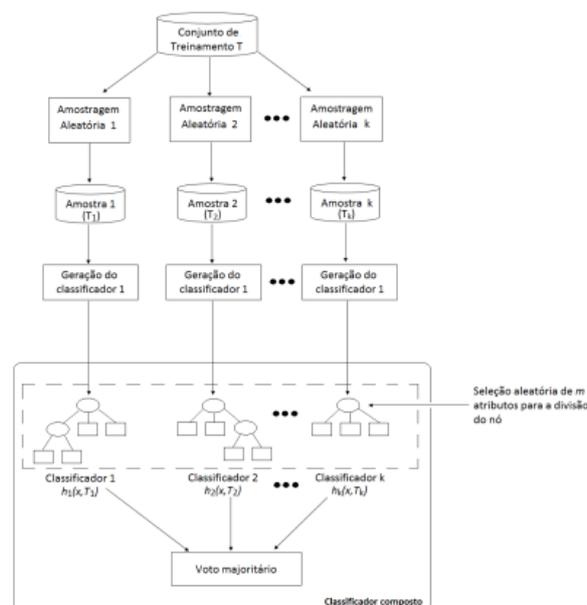


FIGURA 2 - Diagrama Random Forest

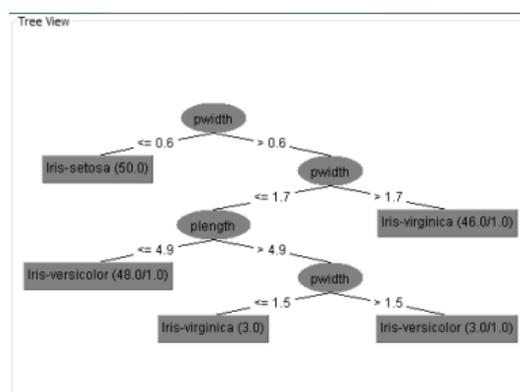


FIGURA 3 – Esquema de Árvore J48

Com os dados preparados e algoritmos selecionados, através do software Weka (Figura 4), foi configurado o ambiente para início dos experimentos. Todos os algoritmos foram aplicados utilizando o método de *Cross-Validation* (Validação Cruzada), que é uma técnica onde os dados são divididos em dois conjuntos, sendo um conjunto para teste e outro

para avaliação de desempenho do modelo (BERRAR, 2018).

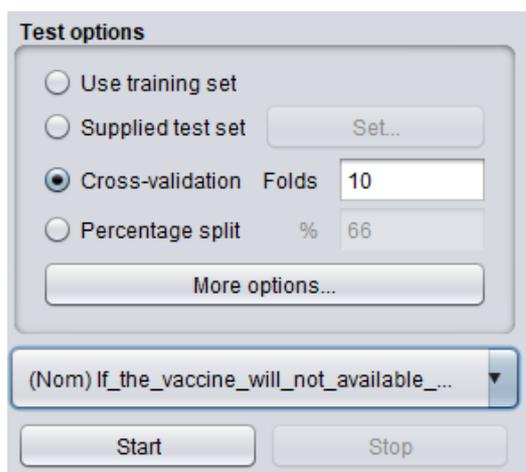


FIGURA 4 – Ambiente de Teste (WEKA)

3 RESULTADOS E DISCUSSÃO

Após a aplicação dos algoritmos, foi feito um balanço do desempenho individual de cada um nos dados selecionados. Os resultados foram condensados em gráficos para análise. A seguir, são listados os resultados obtidos para cada pergunta, considerando os diferentes tipos de algoritmos utilizados.

Pergunta 1. Se a vacina não estiver disponível, você poderá se mover livremente no futuro?

Algoritmo NaiveBayesMultinomialText

Correctly Classified Instances 402 - 72.6944 %

Incorrectly Classified Instances 151 - 27.3056 %

Algoritmo KStar

Correctly Classified Instances 411 - 74.3219 %

Incorrectly Classified Instances 142 - 25.6781 %

Algoritmo RandomForest

Correctly Classified Instances 367 - 66.3653 %

Incorrectly Classified Instances 186 - 33.6347 %

Algoritmo RandomTree

Correctly Classified Instances 414 - 74.8644 %

Incorrectly Classified Instances 139 - 25.1356 %

Algoritmo J48

Correctly Classified Instances 406 - 73.4177 %

Incorrectly Classified Instances 147 - 26.5823 %

Na Figura 5, é apresentado um comparativo dos algoritmos resultados obtidos pelos algoritmos na análise dos dados da Pergunta 1.

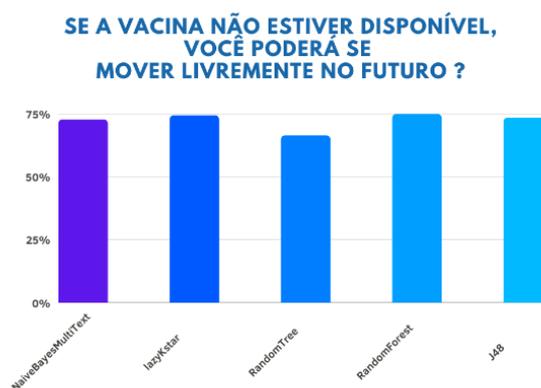


FIGURA 5 – Análise Pergunta 1

Pergunta 2. Você acredita que a flexibilização do lockdown pode aumentar a propagação de COVID-19 no futuro?

Algoritmo NaiveBayesMultinomialText

Correctly Classified Instances 441 - 79.7468 %
 Incorrectly Classified Instances 112 - 20.2532 %

Algoritmo KStar

Correctly Classified Instances 422 - 76.311 %
 Incorrectly Classified Instances 131 - 23.689 %

Algoritmo RandomTree

Correctly Classified Instances 392 - 70.8861 %
 Incorrectly Classified Instances 161 - 29.1139 %

Algoritmo RandomForest

Correctly Classified Instances 445 - 80.4702 %
 Incorrectly Classified Instances 108 - 19.5298 %

Algoritmo J48

Correctly Classified Instances 439 - 79.3852 %
 Incorrectly Classified Instances 114 - 20.6148 %

Na Figura 6, é apresentado um comparativo dos algoritmos resultados obtidos pelos algoritmos na análise dos dados da Pergunta 2.

Pergunta 3. Você manteve um relacionamento saudável com sua família durante o lockdown?

Algoritmo NaiveBayesMultinomialText

Correctly Classified Instances 501 - 90.5967%
 Incorrectly Classified Instances 52 - 9.4033 %

VOCÊ ACREDITA QUE A FLEXIBILIZAÇÃO DO LOCKDOWN PODE AUMENTAR A PROPAGAÇÃO DE COVID-19 NO FUTURO ?

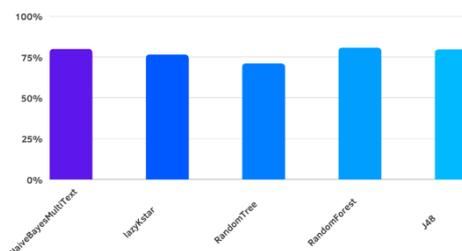


FIGURA 6 – Análise Pergunta 2

Algoritmo KStar

Correctly Classified Instances 484 - 87.5226 %
 Incorrectly Classified Instances 69 - 12.4774 %

Algoritmo RandomTree

Correctly Classified Instances 468 - 84.6293 %
 Incorrectly Classified Instances 85 - 15.3707 %

Algoritmo RandomForest

Correctly Classified Instances 500 - 90.4159 %
 Incorrectly Classified Instances 53 - 9.5841 %

Algoritmo J48

Correctly Classified Instances 501 - 90.5967 %
 Incorrectly Classified Instances 52 - 9.4033 %

Pergunta 4. Você usa máscara quando sai?

Algoritmo NaiveBayesMultinomialText

Correctly Classified Instances 513 - 96.0674 %
 Incorrectly Classified Instances 21 - 3.9326 %

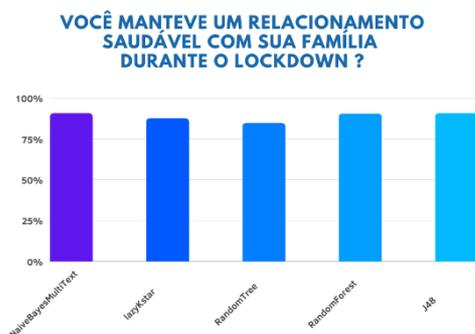


FIGURA 7 – Análise Pergunta 3

Algoritmo KStar

Correctly Classified Instances 513 - 96.0674 %

Incorrectly Classified Instances 21 - 3.9326 %

Algoritmo RandomTree

Correctly Classified Instances 499 - 93.4457 %

Incorrectly Classified Instances 35 - 6.5543 %

Algoritmo RandomForest

Correctly Classified Instances 513 - 96.0674 %

Incorrectly Classified Instances 21 - 3.9326 %

Algoritmo J48

Correctly Classified Instances 509 - 95.3184 %

Incorrectly Classified Instances 25 - 4.6816 %

4 CONCLUSÃO

De modo geral, a maioria dos algoritmos teve um bom desempenho no quesito índice de acertos, com destaque para o *Random Forest* e *NaiveBayesMultinomialText*. O *Random Tree*, por sua vez, foi o que apresentou o pior

desempenho comparado aos demais, porém, ainda sim foi satisfatório, sempre com índice acima de 50% de acertos.

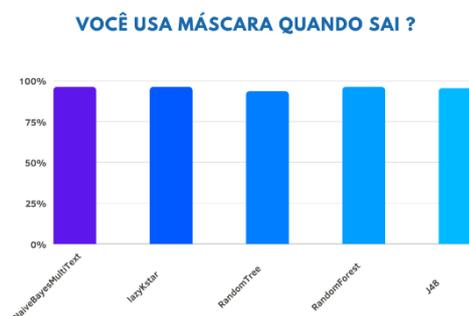


FIGURA 8 – Análise Pergunta 4

Com base nos resultados obtidos, é possível observar que o desenvolvimento do Machine Learning se faz muito necessária no cenário atual, onde uma boa análise de dados pode extrair informações importantes e úteis para tomadas de decisões. Automatizar este processo com algoritmos bem desenvolvidos pode trazer maior produtividade a depender do contexto em que estiver aplicado.

REFERÊNCIAS

UIP, David Everson; STRABELLI, Tânia Mara Varejão. COVID-19 e o Coração. Disponível em: <<https://www.scielo.br/j/abc/a/NWKKJDxLthW/Sb53XFV9Nhvn/?lang=pt>>. Acesso em: 23 set. 2021

VIEIRA, Mateus Vitor. UTILIZAÇÃO DE APRENDIZADO SUPERVISIONADO NA PREDIÇÃO DA DEMANDA DE ENERGIA NO PROCESSO DE PRODUÇÃO DE CUMENO. Disponível em: <[https://repositorio.ufscar.br/bitstream/handle/ufscar/14630/TCC-](https://repositorio.ufscar.br/bitstream/handle/ufscar/14630/TCC-MATEUS%20VITOR%20VIEIRA.pdf?sequence=1&isAllowed=y)

MATEUS%20VITOR%20VIEIRA.pdf?sequence=1&isAllowed=y>. Acesso em: 23 set. 2021.

WITTEN, Ian H et al. Weka: ferramentas e técnicas práticas de aprendizado de máquina com implementações Java. Disponível em: <<https://researchcommons.waikato.ac.nz/handle/10289/1040>>. Acesso em: 23 set. 2021.

FERNANDO, Jonathan Radot. Multinomial Naive Bayes. Disponível em: <https://github.com/JonathanRadotski/multinomial_naivebayes#multinomial-naive-bayes>. Acesso em: 23 set. 2021.

CHELLAM, Aditya; L, Ramanathan; S, Ramani. Intrusion Detection in Computer Networks using Lazy Learning Algorithm. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1877050918308408>>. Acesso em: 23 set. 2021.

MARIN, Maikon Aloan; LOPES, Fabrício Martins. Indução de Árvores de Decisão para a Inferência de Redes Gênicas. Disponível em: <<http://paginapessoal.utfpr.edu.br/fabricio/fabricio-martins-lobes/pesquisa/orientacoes/relatorio-pibic-2013-maikon-marin.pdf>>. Acesso em: 23 set. 2021.

OSHIRO, Thais Mayumi. Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica. Disponível em: <<https://www.teses.usp.br/teses/disponiveis/95/95131/tde-15102013-183234/publico/monografia.pdf>>. Acesso em: 23 set. 2021.

BERRAR, Daniel. Cross-validation. Disponível em: <https://www.researchgate.net/profile/Daniel-Berrar/publication/324701535_Cross-Validation/links/5cb4209c92851c8d22ec4349/Cross-Validation.pdf>. Acesso em: 23 set. 2021

LIMA, Isafas; PINHEIRO, Carlos AM; SANTOS, Flávia A. Oliveira. Inteligência artificial. Elsevier Brasil, 2016.