Modelo de previsão de doenças com base em sintomas usando algoritmos de aprendizado de máquina

Bruno Correa[1], Gabriel Azevedo[1], Pedro Henrique Moreira[1], João Teramatsu[1]
[1]PUC - MG
{brunoinnecco@hotail.com,gabri10.fernandes23@gmail.com,pedrohenriqmoreira@gmail.com,joao.teramatsu@sga.pucminas.br}

Abstract

Esta pesquisa foi realizada utilizando três algoritmos distintos de *Machine Learning* para prever qual doença o paciente possui após ser treinado com uma base de dados. Os algoritmos foram treinados e testados em um conjunto de dados de registros de pacientes, e seu desempenho foi avaliado usando precisão, *recall* e *F1 score*. Os resultados mostraram que os três algoritmos alcançaram altos níveis de precisão na previsão da doença correta, sendo que um algoritmo se destacou dos demais. Este estudo demonstra o potencial do *Machine Learning* no diagnóstico médico e destaca a importância de usar vários algoritmos para garantir previsões precisas.

Keywords

Machine Learning, Diagnóstico Médico, Previsão de Doenças, Classificação, *Support Vector Machine (SVM)*, Random Forest, *Gradient Boosting Machine (GBM)*, Inteligência Artificial, Saúde Digital, Análise de Dados Médicos

ACM Reference Format:
Bruno Correa[1], Gabriel Azevedo[1], Pedro Henrique Moreira[1], João Teramatsu[1]. 2024. Modelo de previsão de doenças com base em sintomas usando algoritmos de aprendizado de máquina. In *Proceedings of ConferenceName'Year: ACM Conference (ConferenceName'Year).* ACM, Belo Horizonte, BH, BR, 3 pages.

1 Introdução

A aplicação de técnicas de *Machine Learning* na área da saúde tem sido um foco de pesquisa intensiva, dada a sua capacidade de analisar grandes volumes de dados e identificar padrões complexos. Neste contexto, o presente trabalho se concentra no uso do *Machine Learning* para auxiliar no diagnóstico de doenças com base em sintomas apresentados pelos pacientes. O processo de diagnóstico é uma tarefa complexa que exige dos profissionais da saúde um alto nível de conhecimento e experiência. No entanto, com o avanço das tecnologias de Inteligência Artificial, tornou-se possível desenvolver sistemas que aprendem com os dados e são capazes de fazer previsões precisas. O objetivo deste estudo é treinar um modelo de *Machine Learning* utilizando uma base de dados contendo diversos tipos de sintomas, visando predizer qual doença um paciente pode ter. A relevância deste problema reside na possibilidade de melhorar a eficiência e a precisão dos diagnósticos médicos. Ao fornecer uma ferramenta que auxilia no processo de diagnóstico, podemos contribuir para um tratamento mais rápido e eficaz, melhorando assim a qualidade do atendimento aos pacientes.

ConferenceName'Year, June, 2024, Belo Horizonte, Brazil
2024.

2 Materiais e métodos

2.1 Descrição da base de dados

A base de dados utilizada neste estudo foi obtida na plataforma Kaggle e destina-se à aplicação de métodos de aprendizado de máquina no campo médico. Ela é composta por dois arquivos: um para treinamento e outro para testes, ambos contendo 132 colunas após a remoção da última coluna. As primeiras 132 colunas representam sintomas binários (0 ou 1), indicando a ausência ou presença do sintoma, enquanto a última coluna, que foi removida, correspondia ao prognóstico, mapeando esses sintomas para uma das 42 possíveis doenças. Os sintomas incluem uma ampla variedade de condições, desde coceira e erupções cutâneas até dor articular e estomacal, proporcionando uma base rica e diversificada para o treinamento de modelos preditivos. Cada registro no conjunto de dados representa um conjunto específico de sintomas associados a um diagnóstico particular, permitindo que os modelos de aprendizado de máquina aprendam a classificar efetivamente essas combinações. Essa base de dados destaca a importância da aplicação de técnicas avançadas de ciência de dados na medicina, permitindo melhorias na precisão dos diagnósticos médicos e, consequentemente, no tratamento de pacientes. Através do uso eficiente de métodos de aprendizado de máquina, espera-se que esta pesquisa contribua para avanços significativos na prática médica e na saúde pública.

2.2 Etapas de pré-processamento

Para preparar a base de dados para a modelagem, foram realizadas diversas etapas de pré-processamento essenciais. Primeiramente, a última coluna dos arquivos de treinamento e teste, que continha o prognóstico, foi removida para assegurar que o algoritmo pudesse fazer previsões sem acesso às respostas durante o treinamento. Em seguida, os dados foram carregados utilizando a biblioteca *pandas* e as colunas de sintomas foram separadas da coluna de prognóstico nos dados de treinamento. O conjunto de dados de treinamento foi dividido em subconjuntos de treino e validação, com uma proporção de 80/20, utilizando a função *train_test_split*, permitindo uma avaliação robusta do modelo.

Para lidar com valores ausentes, foi aplicada a imputação com a mediana das colunas numéricas, uma escolha que ajuda a evitar distorções causadas por valores extremos, utilizando a biblioteca *SimpleImputer*. Posteriormente, um modelo de *Gradient Boosting* foi treinado com hiperparâmetros padrão e seu desempenho foi avaliado no conjunto de validação utilizando métricas como acurácia, precisão, recall e F1-score. As colunas do conjunto de teste foram alinhadas com as do conjunto de treino para garantir consistência, e as previsões no conjunto de teste foram realizadas. Por fim, as previsões foram salvas em um novo arquivo *CSV*, garantindo que todas

as etapas do pré-processamento contribuíssem para a integridade e precisão do modelo preditivo.

2.3 Descrição dos métodos utilizados

2.3.1 ***Support Vector Machine SVM.*** O algoritmo Support Vector Machine (SVM) é um modelo de aprendizado de máquina supervisionado que é usado para classificação e regressão. O SVM funciona encontrando o hiperplano que melhor separa os dados em classes distintas, maximizando a margem entre as classes. Ele é eficaz em lidar com dados complexos e de alta dimensionalidade e é amplamente utilizado em problemas de classificação, como reconhecimento de imagem, detecção de spam e diagnóstico médico. Nesta pesquisa foi utilizado o *SVM* com um *kernel* linear. Os principais hiperparâmetros ajustados foram:

- `Parâmetro de regularização que controla a margem de erro. Avaliamos os valores`: [0.1, 1, 10]
- `Tipo de kernel a ser usado. Utilizamos 'linear'.`
- `gamma`: Coeficiente do *kernel*. Avaliamos os valores ['scale', 'auto'].

A técnica de *Grid Search* com validação cruzada (*GridSearchCV*) foi utilizada para ajustar os hiperparâmetros. A melhor combinação encontrada foi *C=1* e *gamma='scale'*.

2.3.2 ***Random Forest.*** O algoritmo Random Forest é um modelo de aprendizado de máquina supervisionado que se destaca por sua capacidade de lidar com dados complexos e de alta dimensionalidade. Ele utiliza uma técnica de *ensemble learning*, onde várias árvores de decisão são criadas durante o treinamento e suas previsões são combinadas para produzir uma previsão final mais precisa. Cada árvore de decisão é construída usando uma amostra aleatória dos dados de treinamento e um subconjunto aleatório das variáveis de entrada, tornando o modelo menos suscetível a *overfitting*. Foi utilizado o *RandomForestClassifier*. Os principais hiperparâmetros ajustados foram:

- `n_estimators`: Número de árvores na floresta. Foram avaliados os valores [100, 200, 300].
- `max_depth` Profundidade máxima das árvores. Foram avaliados os valores [None, 10, 20].
- `max_features`: Número de recursos a serem considerados para a melhor divisão. Foram avaliados os valores ['auto', 'sqrt']

.

2.3.3 ***Gradient Boosting Machine GBM.*** O *Gradient Boosting Machine (GBM)* é um algoritmo de *machine learning* que cria um modelo preditivo forte combinando múltiplos modelos fracos, tipicamente árvores de decisão, de forma iterativa. Cada nova árvore é ajustada para corrigir os erros das árvores anteriores, melhorando a precisão do modelo a cada iteração. Para ajustar os hiperparâmetros, utilizamos a técnica de Grid Search com validação cruzada (GridSearchCV). O ajuste foi feito avaliando uma faixa de valores para cada hiperparâmetro chave:

O modelo **GBM (Gradient Boosting Machine)** apresentou desempenho perfeito nos dados de teste, com precisão, *recall* e *F1 score* de 100%. Tanto o **SVM** quanto o **RandomForest** tiveram desempenhos muito próximos, mas um pouco inferiores ao do **GBM**.

Table 1: Valores dos Hiperparâmetros Avaliados e os Melhores Valores Encontrados

Hiperparâmetro	Valores Avaliados	Melhor Valor
n_estimators	[50, 100, 150]	150
learning_rate	[0.01, 0.1, 0.2]	0.1
max_depth	[3, 4, 5]	4

3 Resultados e Discussões

A presente pesquisa analisou o desempenho de três algoritmos de aprendizado de máquina - *Support Vector Machine (SVM)*, *RandomForest* e *Gradient Boosting Machine (GBM)* - em uma tarefa de classificação. As métricas utilizadas para avaliar o desempenho foram precisão, *recall* e *F1 score*. O algoritmo GBM apresentou um desempenho excepcionalmente alto, alcançando 100% em todas as métricas avaliadas. A estratégia de *boosting* empregada pelo GBM, que combina múltiplos modelos fracos para formar um modelo forte, pode ser a responsável por tal performance. Entretanto, é crucial ressaltar a necessidade de cautela ao interpretar esses resultados, pois um desempenho perfeito pode indicar um possível sobreajuste do modelo aos dados de treinamento. Os algoritmos SVM e *RandomForest* também demonstraram desempenhos notáveis. O SVM, que busca encontrar o hiperplano que maximiza a margem entre as classes, provou ser eficaz na tarefa de classificação binária. Similarmente, o *RandomForest*, que emprega um método de *ensemble* ao combinar várias árvores de decisão, mostrou-se eficiente na redução do *overfitting* e na melhoria da generalização do modelo. A tabela a seguir demonstra a comparação entre o resultado de cada algoritimo e como o GBM teve uma performace superior se comparada com os outros

Table 2: Métricas de Desempenho dos Modelos de Aprendizado de Máquina

Modelo	Precisão (*Precision*)	*Recall*	*F1 Score*
SVM	98.81%	97.62%	97.62%
Random Forest	98.81%	97.62%	97.62%
GBM	100%	100%	100%

A seleção do algoritmo mais adequado é fortemente dependente dos objetivos específicos do projeto. Se a minimização de falsos negativos é priorizada, todos os três algoritmos são apropriados, com o GBM exibindo um *recall* perfeito. Analogamente, se a redução de falsos positivos é o foco, o GBM novamente se destaca com uma precisão impecável. Contudo, é imprescindível validar esses resultados com um conjunto de dados de teste independente ou aplicar técnicas de validação cruzada para assegurar a robustez do modelo selecionado. Futuras pesquisas poderiam explorar ainda mais a comparação desses algoritmos em diferentes cenários e conjuntos de dados.

4 Considerações finais

Este estudo demonstrou a eficácia de três algoritmos de aprendizado de máquina - *Support Vector Machine (SVM)*, *Random Forest* e *Gradient Boosting Machine (GBM)* - na previsão de doenças com

base em sintomas apresentados por pacientes. Entre os três, o *GBM* apresentou desempenho superior, com precisão, *recall* e *F1 score* perfeitos. A análise das matrizes de confusão corroborou esses resultados, evidenciando que o *GBM* classificou todas as instâncias corretamente, sem erros de classificação nos dados de validação. As matrizes de confusão dos modelos *SVM* e *Random Forest* mostraram uma precisão de 98.81%, *recall* de 97.62% e *F1-score* de 97.62%, destacando ainda sua eficácia, embora com um ligeiro aumento nos erros de classificação comparado ao *GBM*. Estes resultados destacam o potencial do *Machine Learning* no diagnóstico médico e a importância de utilizar diferentes algoritmos para assegurar previsões precisas. Para trabalhos futuros, recomenda-se a otimização dos hiperparâmetros através de técnicas como busca em grade (*Grid Search*) ou busca aleatória (*Random Search*), além da utilização de conjuntos de dados maiores e mais variados para validar ainda mais a robustez dos modelos.

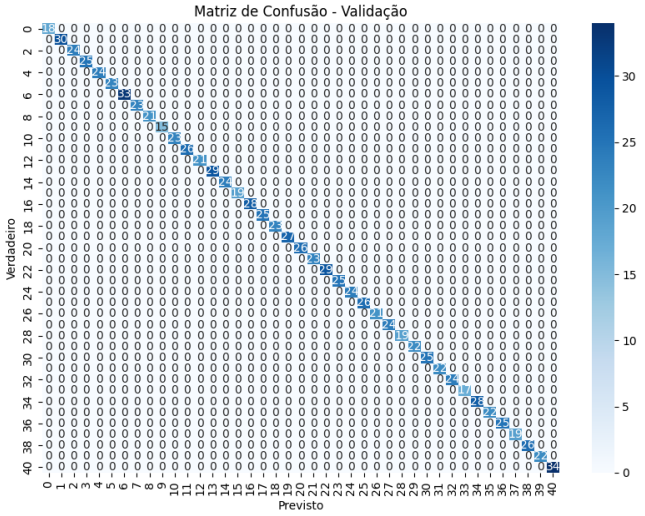

Figure 1: Matriz de confusão.

5 Utilização do GPT

Neste estudo, utilizamos o GPT para correção de erros ortográficos e para auxiliar na correção de erros e testes dos algoritmos. O GPT foi utilizado como uma ferramenta de suporte para garantir que o texto estivesse livre de erros de digitação e ortografia, melhorando assim a clareza e a qualidade geral do manuscrito. Além disso, o GPT foi empregado para validar e testar os algoritmos de aprendizado de máquina implementados, assegurando que os métodos fossem executados corretamente e produzissem resultados precisos.

6 Código desenvolvido

https://github.com/TrabalhoPraticoIA2024/TP2IA/tree/main

7 References

- https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning
- https://seer.ufrgs.br/index.php/rita/article/download/rita_v14_n2_p43-67/3543/18885
- https://seer.ufrgs.br/index.php/rita/article/download/rita_v14_n2_p43-67/3543/18885