Analisando tendências e preditores de saúde mental entre estudantes

Matheus Luciano de Caldas Figueiredo, Juan Vila Nova Rojas Moreno, João Victor da Paz Nascimento, Davi Gleristone Alves Gomes, Luciano de Souza Cabral

Curso Médio integrado a Desenvolvimento de Sistemas – Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE - Campus jaboatão dos Guararapes)

Caixa Postal 54080-000 – Jaboatão dos Guararapes – PE – Brazil

Mlcf@discente.ifpe.edu.br,

Jvrnm@discente.ifpe.edu.br,

Jvpn1@discente.ifpe.edu.br,

dgag@discente.ifpe.edu.br,

luciano.cabral@jaboatao.ifpe.edu.br

Abstract

Analyzing Trends and Predictors of Mental Health Among Students

Student mental health is a growing concern within the academic environment. This paper analyzes factors associated with depression in students. Using the public dataset "student_depression_dataset.csv," an Exploratory Data Analysis (EDA) was performed to understand the variables and their distributions. Subsequently, an interactive web application was developed using the Streamlit platform, which employs a pre-trained Artificial Intelligence (AI) model to predict the probability of a student exhibiting depressive symptoms based on their answers. The objective is to demonstrate the practical application of AI models as accessible tools for screening and awareness. The data analysis and comparison with reference studies reinforce that factors such as academic pressure and financial stress are relevant predictors.

Resumo.

A saúde mental dos estudantes é uma preocupação crescente no ambiente acadêmico. Este trabalho analisa os fatores associados à depressão em estudantes. Utilizando o dataset público "student_depression_dataset.csv" , foi realizada uma análise exploratória dos dados (EDA) para compreender as variáveis e suas distribuições. Em seguida, foi implementada uma aplicação web interativa utilizando a plataforma Streamlit , que utiliza um modelo de inteligência artificial (IA) pré-treinado para prever a probabilidade de um estudante apresentar sintomas depressivos com base em suas respostas. O objetivo é demonstrar a aplicação prática de modelos de IA como ferramentas acessíveis para triagem e conscientização. A análise dos dados e a comparação com estudos de referência reforçam que fatores como pressão acadêmica e estresse financeiro são preditores relevantes.

1. Introdução

Nós nos baseamos em artigos que utilizaram o mesmo dataset, o student_depression_dataset.csv do criador Adil Shamim no site Kaggle. O motivo dessa escolha veio do fato de nós lidarmos com esse assunto de depressão e doenças mentais anteriormente e diariamente, todos nós já nos envolvemos em trabalhos que foram relacionados a esse tópico, o tornando algo que já possuímos certo conhecimento.

2. Metodologia

O estudo foi conduzido em três etapas principais: (1) Análise Exploratória dos Dados (EDA), (2) Teste e Seleção de Modelos de Inteligência Artificial, e (3) Desenvolvimento de uma aplicação web preditiva.

2.1. Conjunto de Dados

	id	Gender	Age	City	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
1	8	Female	24.0	Bangalore	Student	2.0	0.0	5.90	5.0	0.0	'5-6 hours'	Moderate	BSc	No	3.0	2.0	Yes	0
2	26	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5.0	0.0	'Less than 5 hours'	Healthy	BA	No	9.0	1.0	Yes	0
4	32	Female	25.0	Jaipur	Student	4.0	0.0	8.13	3.0	0.0	'5-6 hours'	Moderate	M.Tech	Yes	1.0	1.0	No	0
5	33	Male	29.0	Pune	Student	2.0	0.0	5.70	3.0	0.0	'Less than 5 hours'	Healthy	PhD	No	4.0	1.0	No	0
6	52	Male	30.0	Thane	Student	3.0	0.0	9.54	4.0	0.0	'7-8 hours'	Healthy	BSc	No	1.0	2.0	No	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
27896	140685	Female	27.0	Surat	Student	5.0	0.0	5.75	5.0	0.0	'5-6 hours'	Unhealthy	'Class 12'	Yes	7.0	1.0	Yes	0
27897	140686	Male	27.0	Ludhiana	Student	2.0	0.0	9.40	3.0	0.0	'Less than 5 hours'	Healthy	MSc	No	0.0	3.0	Yes	0
27898	140689	Male	31.0	Faridabad	Student	3.0	0.0	6.61	4.0	0.0	'5-6 hours'	Unhealthy	MD	No	12.0	2.0	No	0
27899	140690	Female	18.0	Ludhiana	Student	5.0	0.0	6.88	2.0	0.0	'Less than 5 hours'	Healthy	'Class 12'	Yes	10.0	5.0	No	1
27900	140699	Male	27.0	Patna	Student	4.0	0.0	9.24	1.0	0.0	'Less than 5 hours'	Healthy	BCA	Yes	2.0	3.0	Yes	1

27881 rows × 18 columns

Figura 1. Tabela com os dados originais do dataset vindo do Kaggle.

O conjunto de dados utilizado foi o "student_depression_dataset.csv", obtido na plataforma Kaggle. Este dataset é composto por 27.901 entradas e 18 colunas (variáveis). As variáveis incluem:

Dados Demográficos: Gender (Gênero), Age (Idade), City (Cidade), Degree (Nível de Escolaridade).

Fatores Acadêmicos: Academic Pressure (Pressão Acadêmica), CGPA (Coeficiente de Rendimento Acumulado), Study Satisfaction (Satisfação com os Estudos).

Fatores de Estilo de Vida e Pessoais: Sleep Duration (Duração do Sono), Dietary Habits (Hábitos Alimentares), Work/Study Hours (Horas de Trabalho/Estudo), Financial Stress (Estresse Financeiro), Family History of Mental Illness (Histórico Familiar de Doença Mental).

Variável Alvo: Depression (Depressão), um indicador binário (0 ou 1) que representa a ausência (0) ou presença (1) de sintomas depressivos.

2.2. Análise Exploratória dos Dados (EDA)

Os dados foram carregados e processados em um ambiente Google Colab, utilizando a biblioteca Pandas da linguagem Python. A análise estatística descritiva inicial foi realizada com a função data.describe() para obter médias, desvios padrão e quartis das variáveis numéricas.

Para uma análise mais aprofundada, foi empregada a biblioteca ydata-profiling (anteriormente pandas-profiling). Esta ferramenta gerou um relatório de perfilamento que detalha a distribuição de cada variável (histogramas), identifica valores ausentes, calcula estatísticas e mapeia correlações entre as variáveis, facilitando a compreensão da estrutura dos dados.

2.3. Teste e Seleção de Modelos de IA

Uma vez que o objetivo era prever uma categoria (Depressão = 0 ou 1), o problema foi definido como uma tarefa de classificação binária. Para selecionar o algoritmo de machine learning (Aprendizado de Máquina) mais adequado, foi realizada uma etapa de experimentação comparativa.

Foram testados dois modelos de classificação distintos:

Decision Tree Classifier (Árvore de Decisão): Um modelo que cria uma estrutura de "fluxograma" para tomar decisões.

Random Forest Classifier (Floresta Aleatória): Um modelo de conjunto (Ensemble) que constrói múltiplas árvores de decisão e combina seus resultados para obter uma predição mais robusta.

O critério de avaliação utilizado para comparar os modelos foi a acurácia (Accuracy), que mede a proporção de previsões corretas. O modelo RandomForestClassifier apresentou o desempenho superior, alcançando uma acurácia de 83%. Em comparação, o DecisionTreeClassifier obteve 76% de acurácia.

Devido ao seu maior desempenho, o RandomForestClassifier foi o modelo selecionado para a implementação final.

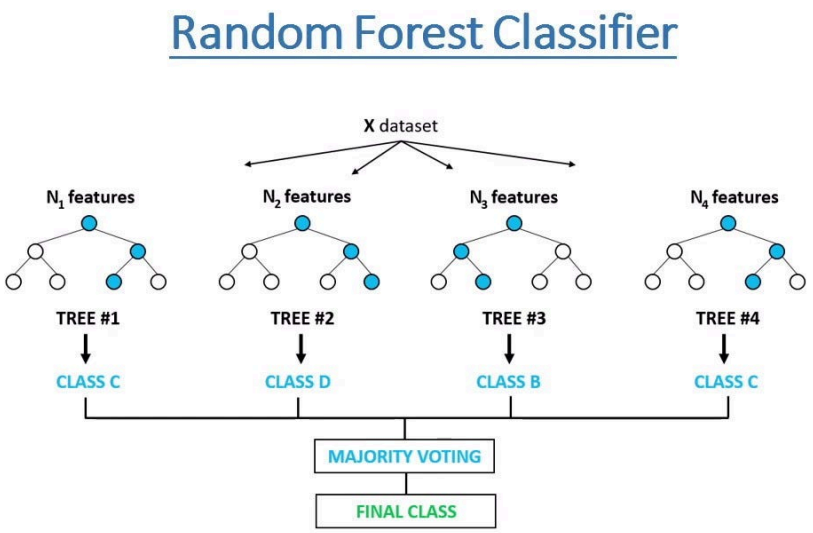

Figura 2. Representação visual do processamento do Random Forest Classifier.

2.4. Desenvolvimento da Aplicação Preditiva

A etapa final consistiu na implementação de uma ferramenta prática de software. A aplicação web foi construída com a biblioteca Streamlit.

A arquitetura da aplicação foi desenvolvida para treinar o modelo de inteligência artificial no momento de sua execução. Ao ser iniciada, a aplicação primeiro realiza o carregamento e processamento do conjunto de dados "student_depression_dataset.csv" diretamente de seu repositório online.

Imediatamente após o preparo dos dados, o aplicativo executa uma função (train_model) que instancia e treina o modelo RandomForestClassifier (utilizando os parâmetros otimizados nos testes, como n_estimators=100, max_depth=10 e class_weight='balanced').

Este modelo, recém-treinado pela própria aplicação, é então mantido em memória para realizar as predições. Quando o usuário insere seus dados na interface (idade, pressão acadêmica, etc.) e solicita a análise, essas informações são processadas e submetidas ao modelo, que executa a predição (model.predict) e calcula a probabilidade (model.predict_proba) do diagnóstico. O resultado é então exibido de forma clara na tela.

Referências

CHAUKHAN, Ankit. Random Forest Classifier and its Hyperparameters. *Analytics Vidhya (Medium)*, 23 fev. 2021. Disponível em: https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

SHAMIM, A. "Student Depression Dataset". *Kaggle*, [s. l.], atual. há 7 meses. Disponível em: https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset