**UNIVERSIDADE PRESBITERIANA MACKENZIE**

**Faculdade de Computação e Informática**

**Uso de machine learning para previsão de casos de depressão em estudantes**

Eduardo Cayres Nishimura - 10400972@mackenzista.com.br

Gustavo Iamamoto Deberdt - 10401167@mackenzista.com.br

Vinicius Nicolini Barros - [10402288@mackenzista.com.br](mailto:10402288@mackenzista.com.br)

**Resumo:**

A depressão entre estudantes universitários é um problema crescente, impactando tanto o desempenho acadêmico quanto a qualidade de vida (Roldan-Espínola, 2024). Este estudo propõe a aplicação de Machine Learning, utilizando a biblioteca Scikit-Learn, para desenvolver um modelo preditivo capaz de identificar alunos com risco de depressão a partir de dados acadêmicos, familiares e comportamentais. A pesquisa utilizará o "Student Depression Dataset", explorando variáveis como histórico escolar, hábitos de vida e satisfação acadêmica. Além dos desafios técnicos, o estudo considera aspectos éticos fundamentais, como privacidade de dados e transparência algorítmica, alinhando-se às diretrizes da LGPD. Com isso, espera-se fornecer uma ferramenta preditiva que auxilie instituições de ensino na identificação precoce de estudantes vulneráveis, permitindo intervenções preventivas mais eficazes.

**Introdução:**

A depressão é uma condição de saúde mental que afeta milhões de pessoas em todo o mundo e tem sido cada vez mais observada entre estudantes universitários (Roldan-Espínola, 2024). O início da vida adulta, somado às pressões acadêmicas e às incertezas sobre o futuro profissional, pode contribuir significativamente para o desenvolvimento de transtornos psicológicos. Segundo estudo realizado na Universidade Estadual de Londrina, 73,9% dos estudantes apresentaram indicativo de depressão, sendo essa condição associada a fatores como insatisfação com o desempenho acadêmico, inatividade física e dependência de mídias sociais (Rufino JV, 2024)​.

Diante desse cenário, métodos eficazes para identificação precoce da depressão entre estudantes tornam-se essenciais para a implementação de estratégias preventivas (Li, S, 2024). O uso de Machine Learning tem se mostrado uma abordagem promissora para a análise de padrões em grandes volumes de dados e a previsão de condições de saúde mental (Arnold, 2024). Ao treinar modelos preditivos com informações sobre histórico escolar, hábitos de vida e fatores socioeconômicos, é possível identificar perfis de risco e auxiliar instituições de ensino na formulação de políticas de apoio psicológico direcionadas(Arnold, 2024).

Diante desse cenário, este projeto será desenvolvido com base na opçãoFramework, empregando a biblioteca Scikit-Learn para construir um modelo de Machine Learning voltado à previsão de casos de depressão em estudantes. O modelo será treinado e testado com um conjunto de dados contendo informações sobre histórico escolar, histórico familiar e estilo de vida, buscando identificar padrões e prever o risco de depressão (Li, S, 2024). Com essa abordagem, espera-se oferecer uma ferramenta preditiva que possa auxiliar instituições de ensino na identificação precoce de estudantes vulneráveis, permitindo intervenções mais eficazes e estratégias de suporte psicológico direcionadas.

**Descrição do Problema:**

O problema da depressão em estudantes universitários tem se tornado uma questão de grande relevância, especialmente entre jovens adultos, grupo etário que apresenta maior susceptibilidade ao estresse psicológico e transtornos psiquiátricos. A transição para a vida adulta envolve desafios significativos, como a adaptação a novas responsabilidades acadêmicas e preocupações com o futuro profissional, fatores que podem contribuir para o desenvolvimento da depressão (Rufino et al.,2024).

A depressão é caracterizada por sintomas como humor deprimido, sensação de vazio ou irritabilidade, acompanhados de alterações cognitivas e somáticas que afetam a capacidade funcional do indivíduo. Estudos apontam que a prevalência de depressão entre universitários varia amplamente, com uma média de 30,6%, podendo chegar a 84,5% em alguns casos. Essa variabilidade sugere a influência de diversos fatores, tanto individuais quanto contextuais (Rufino et al.,2024).

Entre os fatores associados à depressão em estudantes, destacam-se elementos sociodemográficos, como sexo feminino e baixa renda, e aspectos do estilo de vida, incluindo alimentação inadequada, sedentarismo, consumo de álcool e tabaco, além da insatisfação com o tempo de lazer. Adicionalmente, fatores acadêmicos, como a série cursada e a percepção negativa do desempenho acadêmico, também estão correlacionados ao desenvolvimento da doença (Rufino et al.,2024).

Os impactos da depressão nessa população vão além do sofrimento emocional, comprometendo o desempenho acadêmico e podendo gerar um ciclo vicioso de dificuldades. A evolução das gerações e o avanço tecnológico também introduzem novos desafios, como a dependência de redes sociais, que podem intensificar a vulnerabilidade desses jovens (Rufino et al.,2024).

**Discutir a respeito dos aspectos Éticos do uso da IA e a sua Responsabilidade no desenvolvimento da solução:**

A inteligência artificial tem se consolidado como uma ferramenta essencial na análise de dados e previsão de condições de saúde, incluindo a depressão entre estudantes. O uso de machine learning para identificar padrões e prever riscos de transtornos mentais pode proporcionar diagnósticos precoces e intervenções personalizadas. No entanto, essa aplicação levanta questões éticas fundamentais que devem ser rigorosamente consideradas (Santos et al.,2025).

Um dos principais desafios é a confiabilidade dos algoritmos utilizados para a previsão de casos de depressão. A precisão das análises depende da qualidade dos dados e da transparência dos modelos. Algoritmos podem produzir vieses existentes nos dados, levando a falsos positivos ou negativos, o que pode impactar negativamente a vida dos estudantes. Dessa forma, é essencial que os modelos sejam auditáveis, explicáveis e sujeitos a revisão contínua (Santos et al.,2025).

A privacidade e a segurança dos dados também são questões centrais. Informações sensíveis sobre a saúde mental dos estudantes devem ser protegidas contra acessos indevidos, seguindo normativas como o Regulamento Geral de Proteção de Dados e a Lei Geral de Proteção de Dados. O consentimento informado é imprescindível para garantir que os indivíduos tenham conhecimento sobre como seus dados serão utilizados, podendo optar por participar ou não desses sistemas (Santos et al.,2025).

Outro ponto relevante é a responsabilidade das instituições que desenvolvem e aplicam essas soluções. A implementação da IA para prever casos de depressão deve ser acompanhada de protocolos claros para a tomada de decisões baseadas nos resultados obtidos. Profissionais de saúde mental devem estar envolvidos na interpretação dos dados, garantindo que as previsões não sejam usadas de maneira automatizada e descontextualizada (Santos et al.,2025).

Além disso, é fundamental considerar o impacto social da IA na saúde mental dos estudantes. A utilização dessas ferramentas pode gerar estigmatização, ansiedade e preocupações desnecessárias se não forem bem administradas. Assim, abordagens éticas devem assegurar que a tecnologia atue como um complemento à atenção humana e ao suporte psicossocial, e não como um substituto das relações interpessoais e do acompanhamento especializado (Santos et al.,2025).

Diante disso, a adoção da IA na previsão de depressão entre estudantes deve ser realizada de forma responsável e alinhada a princípios éticos bem estabelecidos. Para tanto, é necessária uma abordagem multidisciplinar que envolva profissionais de saúde, especialistas em ética e reguladores para criar diretrizes que assegurem a equidade, a transparência e o respeito à privacidade e à autonomia dos indivíduos. A regulação contínua e a promoção de pesquisas em ética da IA são fundamentais para garantir que essa tecnologia beneficie a sociedade sem comprometer direitos fundamentais (Santos et al.,2025).

**Dataset, análise exploratória preparação**

O dataset selecionado, “Student Depression Dataset”, foi desenvolvido por Adil Shamim e disponibilizado no site “www.kaggle.com”. Ele reúne informações sobre fatores relacionados à saúde mental dos estudantes, com o objetivo de analisar e prever tendências de depressão nesse público. O dataset inclui variáveis demográficas (idade, gênero, cidade), indicadores acadêmicos (CGPA, pressão acadêmica, satisfação com os estudos) e aspectos do estilo de vida (duração do sono, hábitos alimentares, carga de trabalho/estudo). Além disso, considera fatores adicionais, como estresse financeiro, histórico familiar de doenças mentais e relatos de pensamentos suicidas.

**Features:**

id: Identificador único de cada estudante no dataset.

Gender: Gênero do estudante.

Age: Idade do estudante.

City: Cidade onde o estudante reside.

Profession: Ocupação principal do estudante.

Academic Pressure: Nível de pressão acadêmica percebida, variando de 1 a 5.

Work Pressure: Nível de pressão no trabalho.

CGPA: Desempenho acadêmico do estudante em uma escala de 0 a 10.

Study Satisfaction: Grau de satisfação com os estudos, variando de 0 a 5.

Job Satisfaction: Grau de satisfação com o trabalho, onde 0 indica ausência de trabalho.

Sleep Duration: Duração média do sono do estudante.

Dietary Habits: Tipo de alimentação do estudante.

Degree: Curso ou nível acadêmico que o estudante está cursando.

Have you ever had suicidal thoughts: Indica se o estudante já teve pensamentos suicidas.

Work/Study Hours: Média de horas diárias dedicadas ao estudo ou trabalho.

Financial Stress: Grau de estresse financeiro percebido pelo estudante, variando de 1 a 5.

Family History of Mental Illness: Indica se há histórico familiar de doenças mentais.

Depression: Variável-alvo ou rótulo, onde 0 indica que o estudante apresenta sintomas de depressão e 1 indica ausência desses sintomas.

A análise exploratória dos dados foi realizada utilizando a biblioteca de pandas para leitura e manipulação do dataset. O objetivo dessa etapa foi compreender a estrutura dos dados, identificar possíveis inconsistências e preparar o conjunto para o treinamento do modelo.

Inicialmente, algumas features foram removidas da base, as colunas “Job Satisfaction”, “Work Pressure”, “Profession” e “id” foram descartadas, pois continham valores pouco informativos ou que não agregam valor ao modelo preditivo, como é caso do identificador único (id), que não possui influência sobre a variável-alvo.

Para garantir que o modelo pudesse processar todas as informações corretamente, os dados categóricos foram convertidos para valores numéricos utilizando o LabelEncoder da biblioteca sklearn. Esse processo foi essencial para transformar variáveis qualitativas, como gênero e hábitos alimentares, em representações numéricas adequadas ao modelo de machine learning. Além disso, para entender melhor as relações entre as variáveis, foi gerada uma matriz de correlação utilizando as bibliotecas seaborn e matplotlib. Esse gráfico permite visualizar as interdependências entre os atributos do dataset e avaliar a influência de cada um sobre a variável-alvo (Depression). Após a análise da matriz, constatou-se que nenhuma feature precisaria ser removida, pois todas apresentavam alguma relevância para o estudo.

Com esses procedimentos, garantimos que os dados estejam preparados de forma otimizada para a etapa de modelagem, garantindo que as informações mais relevantes fossem utilizadas na construção do modelo preditivo.

**Metodologia**

A pesquisa tem como objetivo explorar o potencial do Machine Learning na identificação e classificação de possíveis casos de depressão entre estudantes, com base em variáveis relacionadas ao seu estilo de vida, hábitos e condições socioemocionais. Espera-se que essa abordagem possibilite um avanço significativo na maneira como a saúde mental dos estudantes é avaliada, fornecendo subsídios para intervenções mais eficazes e personalizadas. Ao identificar os principais determinantes que contribuem para o desenvolvimento da depressão, o estudo pode servir de base para ações preventivas e políticas educacionais voltadas ao bem-estar estudantil.

Para a resolução do problema proposto, utilizaremos a abordagem de Machine Learning baseada no algoritmo Random Forest, uma técnica de aprendizado supervisionado amplamente empregada para problemas de classificação (Breiman, L. 2001). O processo será dividido nas seguintes etapas:

Pré-processamento dos dados: O dataset "Student Depression Dataset" será tratado para garantir a qualidade dos dados. Serão removidas colunas irrelevantes e tratados valores ausentes ou inconsistentes. Além disso, variáveis categóricas serão codificadas numericamente utilizando o LabelEncoder.

Divisão dos dados: Os dados serão divididos para treino e teste

Normalização dos dados: As variáveis numéricas serão normalizadas para garantir um melhor desempenho do modelo.

Treinamento do modelo: Será utilizada a técnica Random Forest para a construção do modelo preditivo. O algoritmo consiste em múltiplas árvores de decisão, onde cada árvore contribui para a decisão final por meio de um processo de votação. Esse método reduz o risco de overfitting e melhora a precisão do modelo.

Avaliação do modelo: O desempenho será avaliado por métricas como precisão, recall, F1-score e matriz de confusão. Será utilizada validação cruzada K-fold para garantir a robustez do modelo.

**Resultados:**

Os resultados obtidos durante a avaliação do modelo ao longo das cinco pastas de validação cruzada indicam um desempenho consistente e satisfatório na tarefa de classificação de URLs. O tempo total de treinamento somou em média 18 segundos, demonstrando boa eficiência computacional para o processo.

Em termos de desempenho, tanto nos dados de treino quanto nos de teste, as métricas apresentaram valores estáveis. A acurácia variou entre aproximadamente 81% e 83%, com médias acima de 82%, evidenciando uma boa capacidade do modelo em acertar a maioria das previsões. O recall, que mede a capacidade de identificar corretamente as instâncias positivas (como URLs maliciosas), também se manteve estável, com valores médios por volta de 81%. Já o F1-score, que equilibra precisão e recall, apresentou resultados próximos a 81% nos testes, refletindo um bom equilíbrio entre os erros de falso positivo e falso negativo.

Esses resultados indicam que o modelo está bem ajustado, com baixa variação entre treino e teste, sugerindo que ele se generaliza bem para dados não vistos. O uso da validação cruzada reforça a confiabilidade desses indicadores, comprovando a robustez do modelo desenvolvido para a detecção de phishing ou classificação de URLs.

#### **Conclusão:**

Em conclusão, este estudo demonstrou que o uso de algoritmos de Machine Learning, especificamente o Random Forest, pode ser uma estratégia eficaz na identificação precoce de casos de depressão entre estudantes universitários. A partir de um conjunto de dados bem estruturado e um processo metodológico foi possível alcançar resultados consistentes, com alta acurácia e estabilidade nas métricas de avaliação. Além disso, o modelo mostrou-se promissor como uma ferramenta de apoio à tomada de decisão em ambientes educacionais, contribuindo para a formulação de ações preventivas mais direcionadas. Apesar dos avanços, destaca-se a importância da consideração contínua dos aspectos éticos, garantindo que a tecnologia seja aplicada de forma responsável, segura e sempre em complemento ao cuidado humano.

Os resultados obtidos indicam que os objetivos esperados foram, em grande parte, atingidos. O modelo Random Forest alcançou precisão superior a 80% em todas as execuções, demonstrando boa capacidade preditiva na identificação de estudantes com risco de depressão. Além disso, a consistência das métricas como acurácia, recall e F1-score reforça a eficácia do modelo na generalização dos padrões aprendidos. Embora a identificação dos fatores mais influentes não tenha sido detalhada nos resultados apresentados, a boa performance sugere que o modelo conseguiu captar relações significativas nos dados. Assim, a abordagem proposta mostra-se promissora como uma ferramenta de apoio para monitoramento e prevenção, alinhando-se ao objetivo de integrar tecnologia à promoção da saúde mental no ambiente acadêmico.

**Referências:**

Roldan-Espínola, L., Riera-Serra, P., Roca, M., García-Toro, M., Coronado-Simsic, V., Castro, A., ... & Gili, M. (2024). Depression and lifestyle among university students: A one-year follow-up study. *The European Journal of Psychiatry*, *38*(3), 100250.

Rufino, J. V., Sirtoli, R., Rodrigues, R., Girotto, E., Andrade, S. M. de ., & Guidoni, C. M.. (2024). Indicativo de depressão e fatores associados em estudantes de graduação de uma universidade pública. Cadernos Saúde Coletiva, 32(3), e32030335. <https://doi.org/10.1590/1414-462X202432030335>

Arnold, P. I. M., Janzing, J., & Hommersom, A. (2024). Machine learning for antidepressant treatment selection in depression. *Drug Discovery Today*, 104068.

Li, S., Shi, J., Shao, C., Sznajder, K. K., Wu, H., & Yang, X. (2024). Predicting Depression, Anxiety, and Their Comorbidity among Patients with Breast Cancer in China Using Machine Learning: A Multisite Cross‐Sectional Study. *Depression and Anxiety*, *2024*(1), 3923160.

Santos, Bernardina, et al. “INTELIGÊNCIA ARTIFICIAL NA MEDICINA: ASPECTOS ÉTICOS E REGULATÓRIOS.” Revista Interdisciplinar Científica Aplicada, vol. 18, no. 4, 2024, pp. 36–49, portaldeperiodicos.animaeducacao.com.br/index.php/rica/article/view/25906. Accessed 27 Mar. 2025.

SHAMIM, Adil**.** Student Depression Dataset. Kaggle, 2023. Disponível em:<https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset>. Acesso em: Acesso em: 26 mar. 2025.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.