Relatório Técnico: Modelagem Preditiva da Depressão em Adultos no Brasil utilizando Machine Learning

Pedro Henrique Rodrigues da Silva[1]

[1]Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Av Dom José Gaspar 500, Belo Horizonte, Minas Gerais, 30535-610

Resumo. *Este relatório técnico apresenta os resultados de um estudo exploratório que investigou a utilização de técnicas de aprendizado de máquina para caracterizar a depressão em adultos no Brasil. O estudo utilizou dados da Pesquisa Nacional de Saúde (PNS) de 2019, realizando um pré-processamento rigoroso dos dados e aplicando algoritmos de Árvore de Decisão, Naive Bayes e Random Forest. O desempenho dos modelos foi avaliado por meio de diversas métricas, incluindo precisão, revocação, F1-score, Curva ROC e o coeficiente Phi. Os resultados indicaram que a Floresta Aleatória se destacou como o modelo com melhor desempenho geral, apresentando alta taxa de verdadeiros positivos e a maior associação com a variável alvo. O relatório discute as limitações do estudo, como o desequilíbrio da base de dados, e sugere direções para futuras pesquisas.*

1. Introdução

A depressão é um transtorno mental prevalente globalmente, com impactos significativos na saúde e na qualidade de vida da população [da Saúde 2020]. No Brasil, a Pesquisa Nacional de Saúde (PNS) de 2019, realizada pelo IBGE, revelou que 10,2% dos brasileiros com mais de 18 anos foram diagnosticados com depressão [da Saúde 2020]. A identificação precoce de fatores de risco para o desenvolvimento da depressão é crucial para a implementação de estratégias eficazes de prevenção e tratamento.

Este estudo teve como objetivo explorar o uso de técnicas de aprendizado de máquina para caracterizar a depressão em adultos no Brasil, utilizando dados da PNS 2019. O foco principal foi o desenvolvimento e a avaliação de modelos preditivos de depressão, empregando algoritmos de Árvore de Decisão, Naive Bayes e Random Forest. A análise do desempenho dos modelos se baseou em diversas métricas, permitindo uma avaliação abrangente de suas capacidades preditivas.

2. Metodologia

2.1. Base de Dados

A base de dados utilizada neste estudo foi a Pesquisa Nacional de Saúde (PNS) de 2019, realizada pelo IBGE. A PNS coleta informações sobre saúde, estilo de vida e condições socioeconômicas da população brasileira. A base de dados original contém 293.726 instâncias e 1087 atributos. Após um processo de filtragem e pré-processamento, a base de dados utilizada para o treinamento dos modelos de machine learning foi reduzida para 53.438 instâncias e 40 atributos.

2.2. Pré-Processamento dos Dados

O pré-processamento dos dados foi realizado em oito etapas principais, buscando garantir a qualidade e a confiabilidade da análise, como ilustrado no diagrama da Figura 1:

Etapa 1: Filtragem: Foi utilizada a metodologia CAPTO [Zarate 2023] para extrair os dados relevantes da PNS 2019. Esta etapa foi guiada por um modelo conceitual, disponível no Apêndice B, que relaciona os atributos da base de dados com as dimensões da depressão.

Etapa 2: Combinação de Atributos: Os atributos relacionados à saúde física (peso, altura), renda e trabalho foram combinados em novos atributos, como IMC, renda total, e se a pessoa trabalha ou não. Esta etapa se baseou em estudos que indicam a relevância desses fatores para a depressão [Luppino et al. 2010, Lorant et al. 2003, Stansfeld and Candy 2006, Virtanen et al. 2018, Kessler et al. 1995].

Etapa 3: Tratamento de Valores Ausentes: Foram utilizados métodos de imputação, como a lógica da base de dados, para preencher os valores ausentes. Atributos com mais de 60% de valores ausentes foram excluídos da análise.

Etapa 4: Tratamento de Outliers: Foi utilizada uma combinação de z-score e intervalo interquartil (IQR) para identificar e remover outliers da base de dados.

Etapa 5: Remoção de Atributos: A entropia e a correlação entre os atributos e a variável alvo (depressão) foram utilizadas para remover atributos irrelevantes da base de dados.

Etapa 6: Codificação dos Dados: A codificação de variáveis categóricas nominais (Sexo) e ordinais (renda e IMC) foi realizada para preparar os dados para os algoritmos de machine learning.

Etapa 7: Separação dos Dados de Treinamento e Teste: A base de dados foi dividida em conjuntos de treinamento (80%) e teste (20%).

Etapa 8: Balanceamento dos Dados de Treinamento: Foi aplicada a técnica de undersampling para equilibrar a distribuição das classes da variável alvo, corrigindo o desequilíbrio da base de dados.

2.3. Modelos de Machine Learning

Foram utilizados três algoritmos de aprendizado de máquina para construir os modelos preditivos de depressão: Árvore de Decisão, Naive Bayes e Random Forest.

Árvore de Decisão: O algoritmo da Árvore de Decisão foi utilizado para identificar padrões e regras de decisão complexas a partir dos dados, utilizando o critério de entropia para a divisão dos nós da árvore.

Naive Bayes: O algoritmo Naive Bayes é uma técnica de classificação probabilística baseada no teorema de Bayes. A versão utilizada foi o Gaussian Naive Bayes, que assume uma distribuição normal para os atributos numéricos.

Random Forest: O algoritmo Random Forest combina múltiplas árvores de decisão para melhorar a precisão e robustez do modelo.

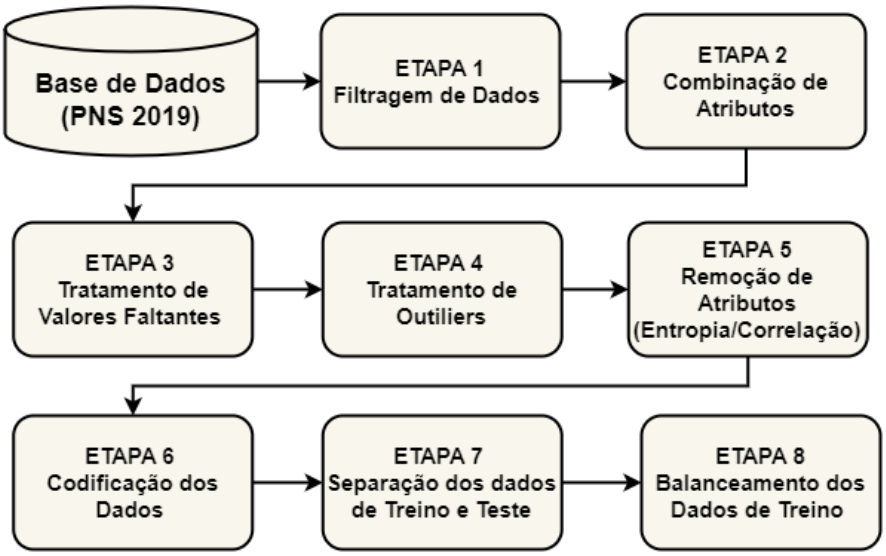

Figura 1. Diagrama das etapas de pré-processamento de dados aplicadas neste estudo.

2.4. Ajuste de Hiperparâmetros

A técnica de RandomizedSearchCV foi utilizada para determinar os melhores hiperparâmetros para cada modelo de Machine Learning, avaliando um subconjunto aleatório das combinações possíveis.

2.5. Métricas de Avaliação

O desempenho dos modelos foi avaliado por meio das seguintes métricas:

Precisão: Indica a proporção de previsões positivas corretas (pessoas com depressão) em relação ao total de previsões positivas. **Revocação (Recall):** Indica a proporção de exemplos positivos (pessoas com depressão) corretamente classificados como positivos em relação ao total de exemplos positivos. **F1-Score:** Combina precisão e revocação, representando a média harmônica entre as duas métricas. **Curva ROC:** Representa o desempenho do modelo em diferentes thresholds de classificação, mostrando a relação entre a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR). **AUC:** Representa a área sob a curva ROC, sendo um indicador do poder de discriminação entre as classes. **Coeficiente Phi:** Mede a força da associação entre as previsões dos modelos e a variável alvo.

3. Resultados e Discussões

Os resultados obtidos foram analisados por meio de diversas métricas, como precisão, revocação, F1-score e a Curva ROC. A Figura 2 apresenta um resumo do desempenho dos modelos para a classe majoritária (sem depressão) e

a classe minoritária (depressão) em relação a essas métricas.

A Floresta Aleatória se destacou como o modelo com melhor desempenho geral, apresentando taxas significativas de verdadeiros positivos (TP) e um bom equilíbrio entre

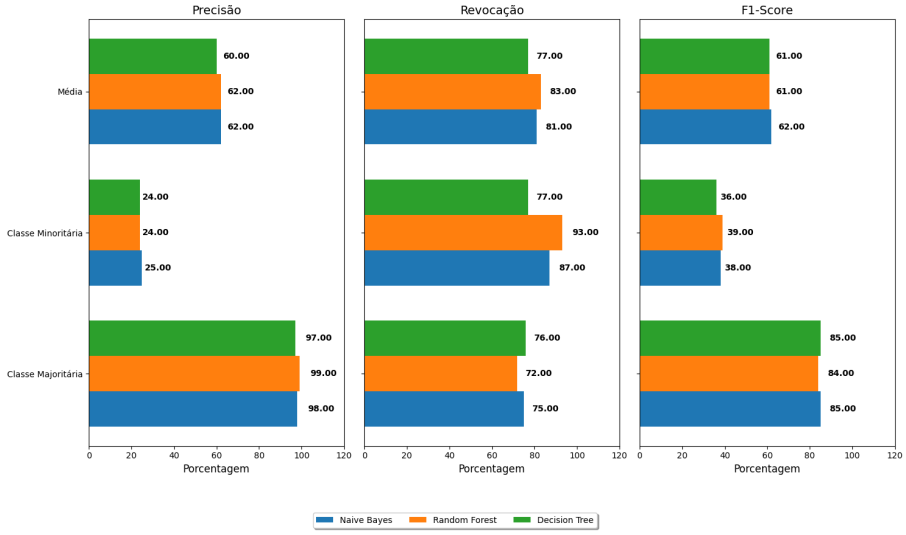

Figura 2. Comparação das métricas de precisão, revocação e F1-score para os modelos de Machine Learning.

precisão e revocação (F1-score). No entanto, é importante notar que a precisão dos modelos para a classe minoritária (depressão) ainda se mostrou baixa, possivelmente devido ao desequilíbrio da base de dados. A análise da curva ROC, ilustrada na Figura 3, corroborou a superioridade da Floresta Aleatória em relação aos outros modelos, demonstrando uma maior capacidade de discriminar entre as classes.

O coeficiente Phi, que mede a força da associação entre as previsões dos modelos e a variável alvo, também indicou a Floresta Aleatória como o modelo com a maior associação (Phi = 0.3893), seguido pelo modelo Naive Bayes (Phi = 0.3769). A Árvore de Decisão apresentou o menor coeficiente Phi (Phi = 0.3124), sugerindo uma associação mais fraca entre suas previsões e a classe real. A Figura 4 mostra a comparação dos coeficientes Phi entre os modelos.

4. Limitações do Estudo

O principal desafio encontrado neste estudo foi o desequilíbrio da base de dados, com uma predominância de casos sem depressão. Essa limitação afetou a precisão dos modelos para a classe minoritária (depressão).

5. Conclusões

Este estudo demonstrou o potencial das técnicas de aprendizado de máquina para a caracterização da depressão em adultos no Brasil, utilizando dados da PNS 2019. A Floresta Aleatória se destacou como o modelo com melhor desempenho, demonstrando alta capacidade de detecção de casos de depressão e a maior associação com a variável alvo. No entanto, o desequilíbrio da base de dados é uma limitação importante, que precisa ser considerada em futuras pesquisas. As informações geradas por este estudo podem auxiliar na elaboração de estratégias de prevenção e tratamento mais eficazes, direcionando os esforços para os grupos mais vulneráveis e promovendo a saúde mental da população brasileira. Investigar a aplicação de técnicas de balanceamento de dados mais avançadas para melhorar a precisão dos modelos para a classe minoritária,

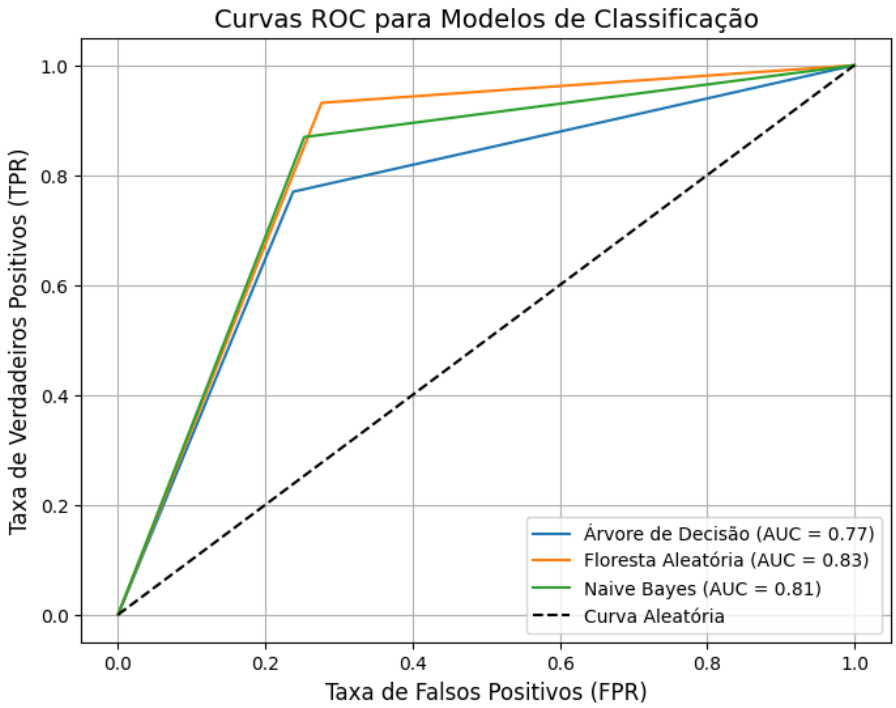

Figura 3. Curvas ROC para os modelos de classificação.

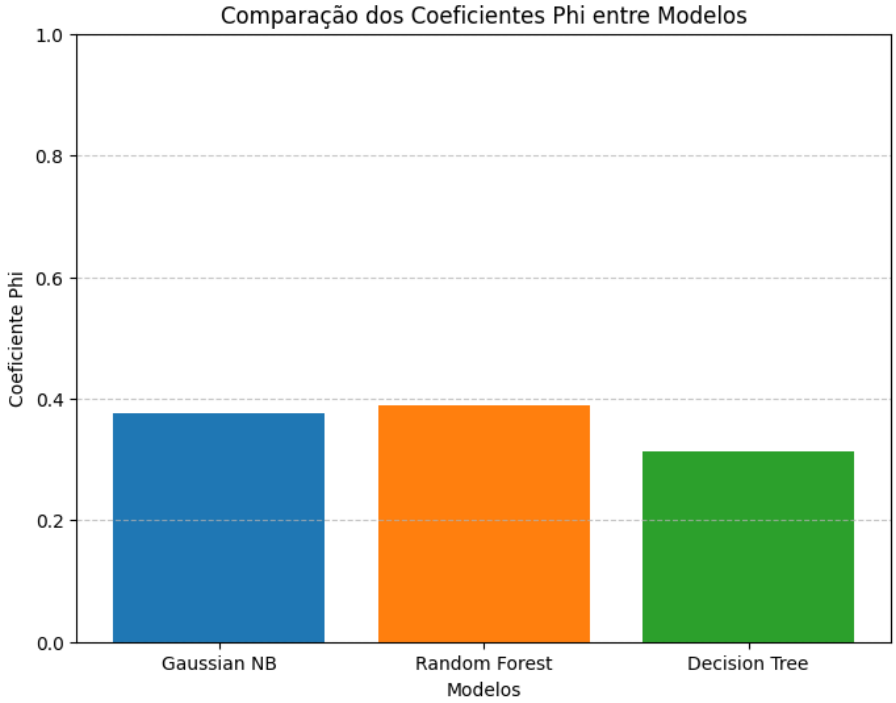

Figura 4. Comparação dos Coeficientes Phi entre os modelos.

Apêndice B: Modelo Conceitual

Referências

da Saúde, M. (2020). A saúde mental no brasil: Indicadores de morbidade.

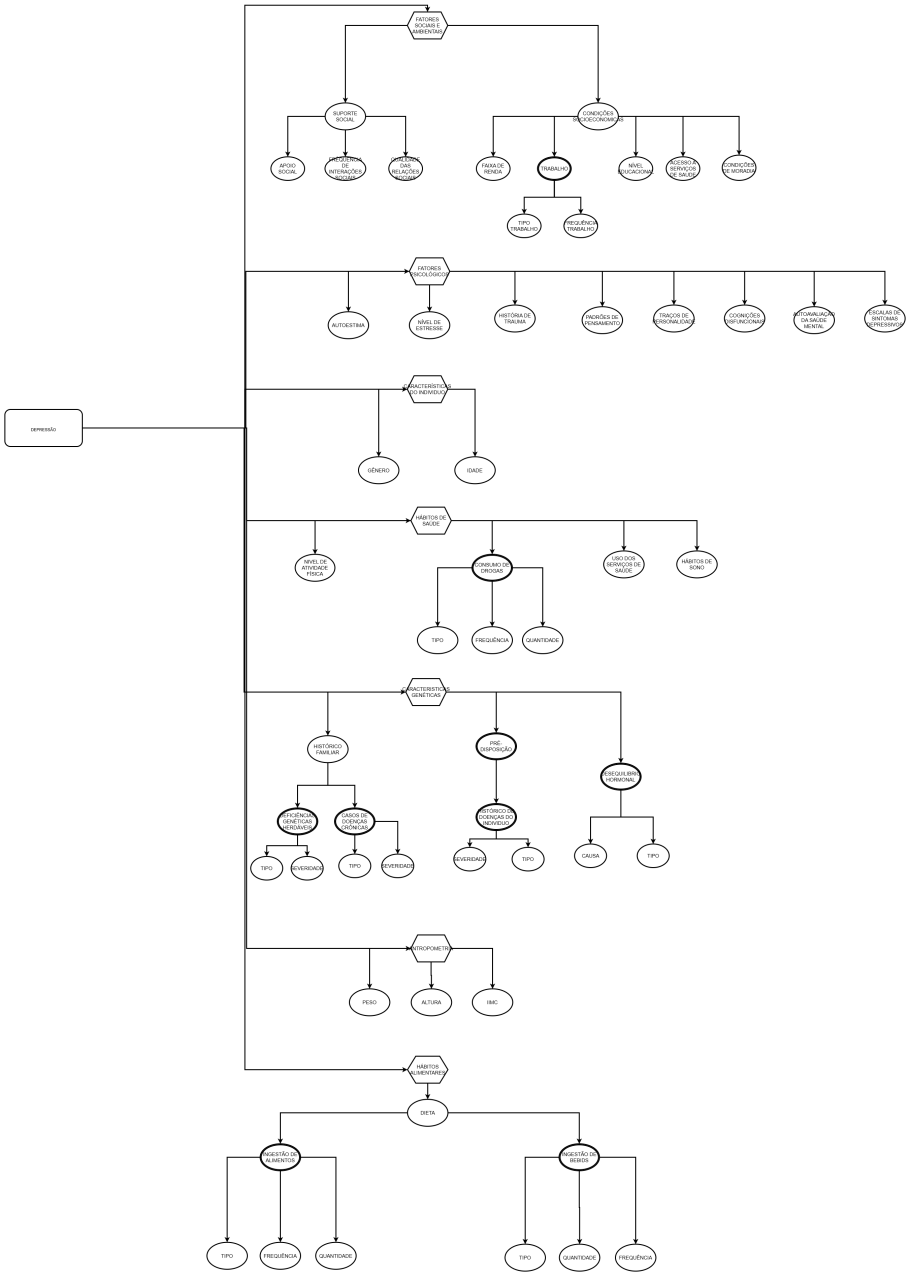

Figura 5. Modelo Conseitual - Método CAPTO.

Kessler, R. C., Davis, C. G., and Kendler, K. S. (1995). Childhood adversity and adult psychiatric disorder in the us national comorbidity survey. *Psychological medicine*, 25(1):51–67.

Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., and Ansseau, M. (2003). Socioeconomic inequalities in depression: a meta-analysis. *American Journal of Epidemiology*, 157(2):98–112.

Luppino, F. S., de Wit, L. M., Bouvy, P. F., Stijnen, T., Cuijpers, P., Penninx, B. W., and Zitman, F. G. (2010). Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. *Archives of general psychiatry*, 67(3):220–229.

Stansfeld, S. and Candy, B. (2006). Psychosocial work environment and mental health—a meta-analytic review. *Scandinavian journal of work, environment health*, 32(6):443–462.

Virtanen, M., Stansfeld, S. A., Fuhrer, R., and Ferrie, J. E. (2018). Overtime work as a predictor of major depressive episode: a 5-year follow-up of the whitehall ii study. *PLoS One*, 13(8):e0202224.

Zarate, F. e. a. (2023). The capto method for transforming tacit knowledge into explicit knowledge in the context of knowledge management systems. *Journal of Knowledge Management*.