

Utilizando Modelos de Machine Learning para a Caracterização da Depressão em Adultos no Brasil

Pedro Henrique Rodrigues da Silva¹, Luiz Enrique Zárte Galvez¹

¹Curso Ciência de Dados, Pontifícia Universidade Católica de Minas Gerais
Av Dom José Gaspar 500, Belo Horizonte, Minas Gerais, 30535-610

Abstract. *The objective of the study is to characterize the profile of adult individuals with depression based on the most recent National Health Survey (PNS 2019). As a methodology, a method for knowledge discovery through the application of white-box and black-box algorithms is proposed. The Random Forest algorithm stood out for its best overall performance, reaching an average F1-score measurement of 82% for the test set. The results also corroborate the need to consider socioeconomic factors, lifestyle and physical health conditions in the prevention and treatment of depression.*

Resumo. *O objetivo do estudo é a caracterização do perfil de indivíduos adultos com depressão a partir da mais recente Pesquisa Nacional de Saúde (PNS 2019). Como metodologia, é proposto um método para descoberta de conhecimento por meio da aplicação de algoritmos caixa-branca e caixa-preta. O algoritmo Floresta Aleatória se destacou pelo melhor desempenho geral, atingindo uma média da medida F1-score de 82% para o conjunto de teste. Os resultados também corroboram a necessidade de considerar fatores socioeconômicos, estilo de vida e condições de saúde física na prevenção e tratamento da depressão.*

1. Introdução

A depressão, transtorno mental que afeta milhões de pessoas no mundo, é caracterizada por uma tristeza profunda e persistente, perda de interesse e prazer nas atividades, alterações no apetite e no sono, fadiga e sentimentos de inutilidade e desesperança. No Brasil, a situação é particularmente preocupante, com a depressão representando um problema crescente de saúde pública, impactando significativamente a qualidade de vida dos indivíduos e gerando relevantes consequências socioeconômicas às famílias e sociedade [MS-BRASIL 2020].

De acordo com a Pesquisa Nacional de Saúde (PNS) de 2019 (<https://www.pns.icict.fiocruz.br/>), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), 10,2% dos brasileiros com mais de 18 anos foram diagnosticados com depressão. Esses dados evidenciam a necessidade urgente de pesquisas que aprofundem no conhecimento sobre os fatores que contribuem para o desenvolvimento da doença no contexto brasileiro.

Apesar dos avanços de pesquisas na medicina sobre a depressão, a maioria dos estudos se concentra em aspectos químicos e biológicos do cérebro, com pouca atenção aos fatores socioambientais e comportamentais que podem influenciar o desenvolvimento e o diagnóstico do transtorno. O presente estudo busca preencher essa lacuna de conhecimento, utilizando técnicas de aprendizado de máquina para caracterizar o perfil de indivíduos com depressão a partir da PNS 2019. O objetivo principal é desenvolver e avaliar modelos preditivos de depressão, utilizando algoritmos como Árvore de Decisão, Naïve Bayes e Floresta Aleatória. Com isso, pretende-se auxiliar na identificação precoce da doença e no desenvolvimento de estratégias de intervenção mais eficazes.

Como estratégia principal da metodologia adotada, para o entendimento do domínio de problema e construção de modelos mais representativos, foi utilizado o Método CAPTO [Zarate et al. 2023]. O método permite construir um modelo conceitual que integra diferentes dimensões da depressão, incluindo fatores socioeconômicos, estilo de vida, condições de saúde física e aspectos comportamentais. A partir da análise do modelo conceitual e dos dados da PNS 2019, foi selecionado um conjunto de dados específico para a aplicação de um processo de descoberta de conhecimento e aplicação dos algoritmos de aprendizado de máquina de forma a caracterizar o indivíduo com depressão.

Este trabalho é estruturado em quatro seções principais. Na segunda seção trabalhos relacionados ao tema são apresentados. Na terceira seção, a metodologia adotada é apresentada. Na quarta seção, os experimentos e análise dos resultados são discutidos. A seção Conclusões e Trabalhos Futuros analisa criticamente os resultados obtidos, explorando suas implicações para a compreensão da depressão no contexto brasileiro. Discutimos também as limitações do estudo e os próximos passos para pesquisas futuras.

2. Trabalhos Relacionados

Diversos estudos exploram a aplicação de *Machine Learning* (ML) na área da saúde mental, com o objetivo de prever a depressão. [Zulfiqar et al. 2020] utilizaram *Deep Learning* e *Transfer Learning* para prever a gravidade da depressão utilizando dados de mídias sociais, concluindo que a combinação desses métodos supera abordagens tradicionais. [Losada-Barrios et al. 2021] buscaram detectar depressão em estudantes universitários usando algoritmos como *Random Forest* e *Support Vector Machines* (SVM), destacando a *Random Forest* como mais eficaz. [Acharya et al. 2022] utilizaram dados de eletrocardiograma (ECG) para prever sintomas depressivos com *SVM* e *Random Forest*, e concluíram que dados fisiológicos têm grande potencial na identificação da depressão. Já [Li et al. 2023] aplicaram ML a dados de *smartphones* para prever alterações na gravidade dos sintomas, evidenciando o valor da coleta passiva de dados para monitoramento da depressão.

3. Materiais e Métodos

3.1. Descrição da base de dados

Os dados utilizados neste estudo foram coletados pela Pesquisa Nacional de Saúde (PNS, 2019), realizada pelo IBGE. A PNS é uma pesquisa domiciliar de caráter nacional que coletou informações sobre a saúde da população brasileira, incluindo dados sociodemográficos, estilo de vida, acesso aos serviços de saúde e indicadores de saúde física e mental [Batista et al. 2021]. A base de dados possui originalmente 293.726 registros e 1087 atributos.

3.2. Preparação do conjunto de dados

A metodologia para preparação do conjunto de dados e construção dos modelos de aprendizado envolveu as etapas descritas na Figura 1, as quais são sucintamente descritas a seguir.

Etapa 1 - Entendimento do domínio e seleção conceitual de atributos: O trabalho utiliza o Método CAPTO, que busca a captura e transformação do conhecimento tácito em conhecimento explícito. O Método consiste de cinco etapas principais: 1) Socialização: formação de grupos de trabalho, e identificação de dimensões e aspectos; 2) Mapeamento: explicitação do conhecimento tácito; 3) Combinação: troca de experiências, e combinação dos modelos/mapas explicitados; 4) Focalização: indicar possíveis atributos em relação a sua relevância; 5) Congruência: harmonização entre a expectativa e a disponibilidade dos dados. O modelo conceitual

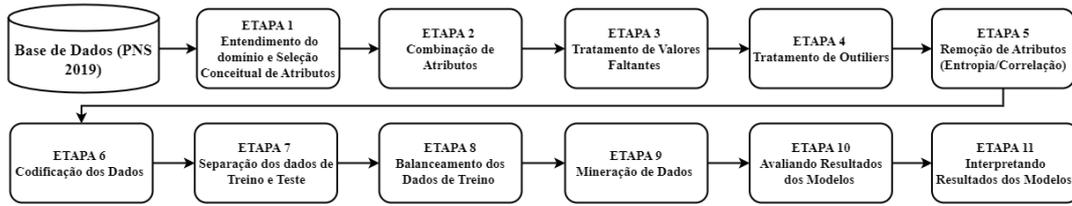


Figura 1. Diagrama das etapas de pré-processamento de dados aplicadas neste estudo.

resultante é mostrado na Figura 2. O modelo mostra a relação entre os atributos identificados e as dimensões da depressão em adultos. Como resultado, os seguintes módulos da PNS e a quantidade de atributos selecionados em cada um são mostrados: Módulo C - Características gerais dos moradores (10 atributos), Módulo D - Características de educação dos moradores (1 atributo), Módulo E - Ocupação (13 atributos), Módulo F - Renda (7 atributos), Módulo J - Estado de saúde e utilização de serviços de saúde (5 atributos), Módulo M - Características do trabalho e apoio social (2 atributos), Módulo N - Percepção do estado de saúde (2 atributos), Módulo P - Estilos de Vida (42 atributos), Módulo Q - Doenças Crônicas (18 atributos) e Módulo W - Antropometria (6 atributos). A seleção teve como referência o modelo conceitual construído.

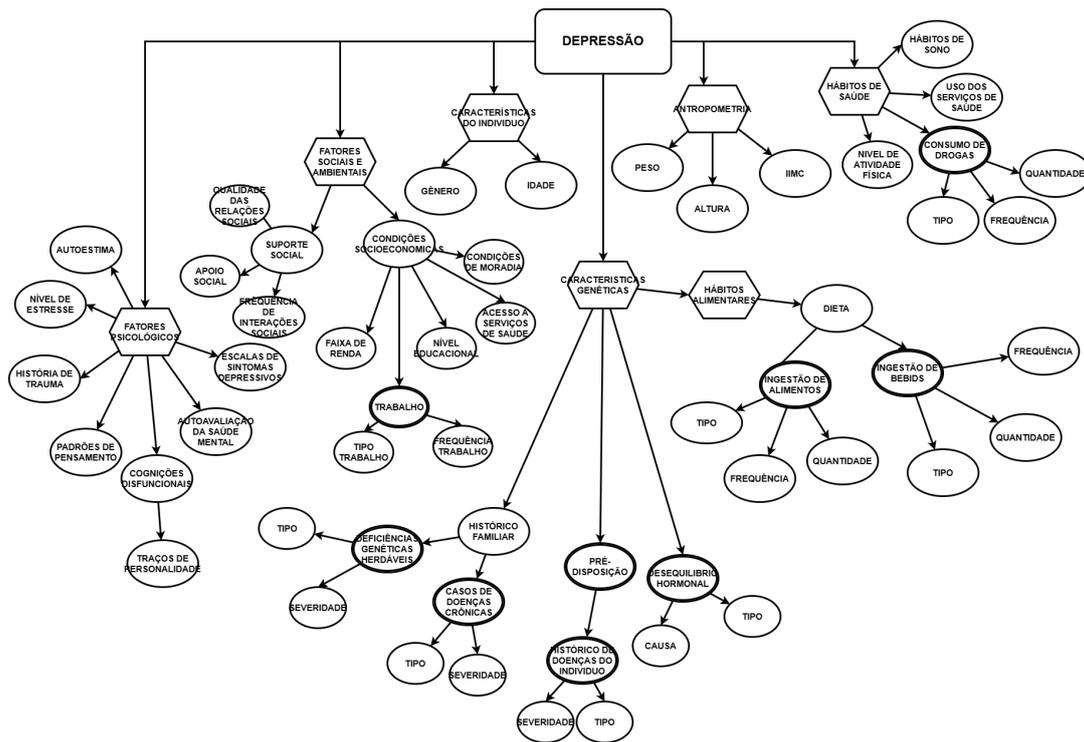


Figura 2. Modelo Conceitual para a Depressão em Adultos

Etapa 2 - Combinação de Atributos: *a) Combinação de atributos relacionados à saúde física:* Para a combinação de atributos relacionados à saúde física, foi consultada a literatura e a própria estrutura da base de dados. Atributos como peso e altura, que possuíam mais de uma medição, foram consolidados considerando a medição final ou a medição que estava preenchida em caso das outras constarem valores ausentes. Além da consolidação, foi calculado o IMC por ser considerado um fator importante na relação com a depressão [Luppino et al. 2010].

b) *Combinação de atributos relacionados à renda e trabalho:* Para analisar a renda como um fator socioeconômico que pode influenciar a saúde mental, todos os atributos relacionados à renda foram consolidados em um único atributo denominado "Renda total". Estudos como [Lorant et al. 2003], [Stansfeld and Candy 2006], e [Virtanen et al. 2018] destacam a correlação entre baixos níveis de renda e o aumento do risco de depressão.

Os atributos relacionados a trabalho foram agrupados, considerando apenas se a pessoa trabalha ou não. No caso das horas trabalhadas, os valores foram preenchidos com a soma das colunas correspondentes, e o valor zero foi atribuído aos indivíduos que não trabalhavam.

c) *Combinação de atributos relacionados às doenças crônicas:* Quanto às doenças crônicas, todos os atributos relacionados foram consolidados em um único atributo para identificar se a pessoa possui ou não uma doença crônica. Estudos como [Kessler et al. 1995], [Beck et al. 1979], [Kendler et al. 2006], [Smith et al. 2006], [Beck et al. 1961], [Kuehner 2017], e [Fiske et al. 2009] destacam a relevância de doenças crônicas como fatores de risco para a depressão.

d) *Combinação de atributos relacionados ao saneamento:* Atributos referentes à moradia e ao saneamento básico também foram consolidadas, utilizando métricas do IBGE. O tempo total de exercício em horas foi preenchido, consolidando atributos de total de horas e total de minutos que a pessoa praticava exercícios por semana.

e) *Combinação de Atributos Relacionados à Alimentação:* Para consolidar as informações sobre alimentação, foram agrupados os atributos que representam o consumo de diferentes grupos alimentares em novos atributos, como: *Consumos de Grãos e Tubérculos, Leguminosas, Carnes, Ovos, Verduras e Legumes, Frutas, Laticínios, Oleaginosas, Consumo de Refrigerantes, Sucos Industrializados, Sucos Naturais, Bebidas Lácteas, Consumo de Salgadinhos e Biscoitos Salgados, Sobremesas Industrializadas, Embutidos e Alimentos Processados, Pães Industrializados, Molhos Industrializados, Refeições por Lanches Rápidos, e Consumo de Sal.* Por fim, os atributos relacionados à quantidade de bebidas que a pessoa consumia por semana foram consolidadas em um único atributo.

A análise do consumo alimentar é relevante para a compreensão da relação entre dieta e depressão, conforme evidenciado por estudos como [Barros and et al. 2021]. O estudo aborda a importância da terapia nutricional para a saúde mental, justificando nossa decisão de analisar o consumo alimentar da PNS. Outros estudos, como [Lai et al. 2015] e [Skogen et al. 2014], também reforçam a relação entre alimentação e a depressão.

Etapa 3 - Tratamento de Valores Ausentes: Após a etapa de combinação de atributos, a presença de valores ausentes em alguns atributos foi constatada. Atributos com mais de 60% de valores ausentes foram excluídos da análise.

Etapa 4 - Tratamento de Outliers: *Outliers* são dados discrepantes que podem prejudicar a precisão dos modelos de aprendizado, levando a resultados menos confiáveis [Dondena et al. 2017].

Para identificar e remover os *outliers*, primeiramente, foi utilizado o método do z-score para identificar e remover *outliers* que se desviam drasticamente da média.

Valores com z-score acima do limite de valor 4 foram considerados *outliers* e removidos da base de dados. Após a remoção dos *outliers* mais extremos com o z-score, aplicamos o método do intervalo interquartil (IQR) para identificar e remover *outliers* restantes. O IQR é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) da distribuição dos dados. Os *outliers* foram definidos como os valores que estavam além de 1,5 vezes o IQR, a partir do

limite inferior ($Q1 - 1,5 * IQR$) e do limite superior ($Q3 + 1,5 * IQR$) [Dondena et al. 2017].

Etapa 5 - Remoção de Atributos: Após o tratamento de *outliers*, atributos irrelevantes foram removidos da base de dados, utilizando o cálculo da entropia e a correlação entre os atributos e a variável alvo (presença ou ausência de depressão). Os atributos "Saneamento Básico", "Total de horas trabalhadas" e "Última consulta médica" apresentaram baixa entropia, indicando classificação direta sobre a presença ou ausência de depressão e, portanto, foram removidos [Witten et al. 2016]. O cálculo da correlação também foi realizado, mas nenhum atributo com correlação relevante para ser removido da base de dados foi identificado.

Etapa 6 - Codificação dos Dados: A base de dados da PNS encontra-se formatada preferencialmente para o tipo de variáveis categóricas ordinais. No entanto, foi necessário realizar a codificação de variáveis categóricas nominais e de variáveis numéricas por meio de discretização. A variável "Sexo" foi codificada utilizando *one-hot encoding*, separando-a em atributos binários ("Masculino" e "Feminino") para evitar que o modelo atribuísse uma ordem arbitrária às categorias. As variáveis de renda e IMC, transformadas em categóricas ordinais, receberam valores numéricos de 1 a 6, seguindo padrões da OMS para IMC e do IBGE para renda, para que o modelo compreendesse a relação hierárquica entre as categorias.

Após a preparação, o conjunto de dados possui 53.438 instâncias e 40 atributos. A descrição do conjunto de dados encontra-se na Tabela 1.

Atributos	Descrição	Valores Únicos
Apoio_Familiar	Número de familiares que pode contar em momentos com dificuldades	4
Apoio_de_Amigos	Número de amigos que pode contar em momentos com dificuldades	4
Tipo_de_Trabalho	Tipo de Trabalho	8
Curso_Mais_Elevado	Nível educacional mais alto que frequentou	15
Diagnostico_Depressao (Target)	Foi diagnosticado ou não com depressão	2
Estado_de_Saude	Estado atual de saúde	3
Pratica_Exercicio	Pratica ou não exercício físico	2
Exercicio_Mais_Frequente	Qual exercício mais frequente que pratica	18
Procura_Atendimento_Saude	Procura atendimento de saúde	2
Problemas_Sono	Possui problemas de sono	2
idade	Idade	40
trabalhou	Trabalhou ou não no período de referencia	2
doencas_cronicas	Possui ou não doenças crônicas	2
moradia_vulneravel	Está ou não em condições de moradia vulneráveis	2
frequencia_exercicio	Frequência de exercícios	8
freq_bebida_alcolica	Frequência de bebidas alcoólicas	3
Consumo de Graos e Tuberculos	Consumo de Graos e Tuberculos	3
Consumo de Leguminosas	Consumo de Leguminosas	9

Atributos	Descrição	Valores Únicos
Consumo de Carnes	Consumo de Carnes	23
Consumo de Ovos	Consumo de Ovos	2
Consumo de Verduras e Legumes	Consumo de Verduras e Legumes	14
Consumo de Frutas	Consumo de Frutas	13
Consumo de Laticínios	Consumo de Laticínios	12
Consumo de Oleaginosas	Consumo de Oleaginosas	2
Consumo de Refrigerantes	Consumo de Refrigerantes	12
Consumo de Sucos Industrializados	Consumo de Sucos Industrializados	12
Consumo de Sucos Naturais	Consumo de Sucos Naturais	8
Consumo de Bebidas Lacteas	Consumo de Bebidas Lacteas	2
Consumo de Salgadinhos e Biscoitos	Consumo de Salgadinhos e Biscoitos	2
Consumo de Doces e Sobremesas Industrializadas	Consumo de Doces e Sobremesas Industrializadas	10
Consumo de Embutidos e Alimentos Processado	Consumo de Embutidos e Alimentos Processados	3
Consumo de Paes Industrializados	Consumo de Paes Industrializados	3
Consumo de Molhos Industrializados	Consumo de Molhos Industrializados	2
Substituição de Refeições por Lanches Rápidos	Substituição de Refeições por Lanches Rápidos	2
Consumo de Sal	Consumo de Sal	5
sexo_masculino	Sexo Masculino	2
sexo_feminino	Sexo Feminino	2
renda_discretizada	Faixa de Renda Total	6
imc_discretizado	Faixa de IMC	6

Tabela 1: Atributos selecionados da PNS 2019

Etapa 7 - Divisão da Base de Dados em Conjuntos de Treinamento e Teste: Com o objetivo de avaliar o desempenho dos modelos de aprendizado de forma robusta, a base de dados foi dividida em dois conjuntos distintos: um conjunto de treinamento e um conjunto de teste. A divisão dos dados foi realizada com a proporção de 80% para o conjunto de treinamento e 20% para o conjunto de teste. Para garantir a reprodutibilidade dos resultados em diferentes execuções, o parâmetro `random_state=42` foi definido, garantindo que a mesma divisão dos dados fosse obtida.

Etapa 8 - Balanceamento da Base de Dados: Uma análise prévia da base de dados revelou um desbalanceamento significativo na distribuição das classes da variável alvo, com predominância de casos de Não-Depressão. Para corrigir essa disparidade e evitar vieses nos modelos, foi aplicada a técnica de undersampling utilizando a classe `RandomUnderSampler` da biblioteca `imblearn.under_sampling`. Essa técnica consiste em remover aleatoria-

mente exemplos da classe majoritária (Não-Depressão) até que a distribuição das classes seja equilibrada. No caso específico deste estudo, foi aplicada tanto ao conjunto de treinamento quanto ao conjunto de teste, garantindo que o modelo fosse treinado em um conjunto de dados mais balanceado e reduzindo a probabilidade de vieses em favor da classe majoritária. A Tabela 2 demonstra a divisão dos conjuntos de dados para treino e para teste.

Tabela 2. Divisão da Base de Dados para Treino e Teste

Classe	Base para Treino	Base para Teste
SIM: Com Depressão	3383	1445
NÃO: Sem Depressão	3383	1445
TOTAL	6766	2890

Etapa 9 - Mineração de dados: Neste estudo, foram utilizados três algoritmos de aprendizado de máquina: Árvore de Decisão, Naïve Bayes, como algoritmos caixa-branca e Floresta Aleatória, como algoritmo caixa-preta, *ensemble*. A escolha desses algoritmos se baseou em sua capacidade interpretativa e de desempenho de ambas categorias [Liu et al. 2022], [Loyola-González 2019].

Para determinar os melhores hiper-parâmetros para cada modelo, foi utilizada a técnica de `RandomizedSearchCV`. A técnica avalia o desempenho do modelo para diferentes combinações de hiper-parâmetros dentro de um intervalo pré-definido. Essa técnica é particularmente útil quando o espaço de busca de hiper-parâmetros é grande [Witten et al. 2016]. O processo de treinamento foi realizado com validação-cruzada de 10 dobras.

4. Experimentos e Análise dos Resultados

Após o processo de treinamento, os modelos apresentaram resultados satisfatórios considerando a medida F1-Score, a qual corresponde a uma medida que combina precisão e revocação, representando a média harmônica global entre as duas métricas. Um F1-score alto indica um bom equilíbrio entre precisão e revocação.

Os modelos foram testados com instâncias não vistas durante o treinamento (ver Tabela 2). A Figura 3 ilustra a comparação das métricas de precisão, revocação e F1-score para cada modelo. O valor médio para a medida F1-score foi de 79% para árvore de decisão, de 82% para Floresta Aleatória, e 80% para o classificador Bayesiano. Apesar dos resultados satisfatórios em relação à revocação e ao F1-score, a precisão dos modelos se manteve levemente menor para o diagnóstico Com-Depressão. Isto pode indicar, um ligeiro viés do modelo para tender a diagnosticar indivíduos com depressão.

Para uma análise mais precisa do desempenho, é importante considerar as taxas de classificação, como a taxa de verdadeiros positivos (TP) e verdadeiros negativos (TN). Os modelos, demonstraram um bom desempenho na detecção de casos de depressão (altas taxas de TP). A Árvore de Decisão, por exemplo, classificou corretamente 83,38% dos casos de depressão, enquanto a Floresta Aleatória atingiu 92,66% de TP, demonstrando um desempenho superior. A análise da curva ROC, presente na Figura 4 corroboram essa observação, demonstrando que a Floresta Aleatória possui um desempenho superior em relação aos demais modelos (AUC = 0,83).

No entanto, os modelos Árvore de Decisão e Naive Bayes também oferecem uma performance satisfatória, podendo ser considerados opções viáveis por sua simplicidade de interpretabilidade dos resultados. A Árvore de Decisão, por exemplo, demonstrou uma taxa de verdadeiros negativos (TN) de 73,80%, indicando que conseguiu identificar corretamente 73,80%

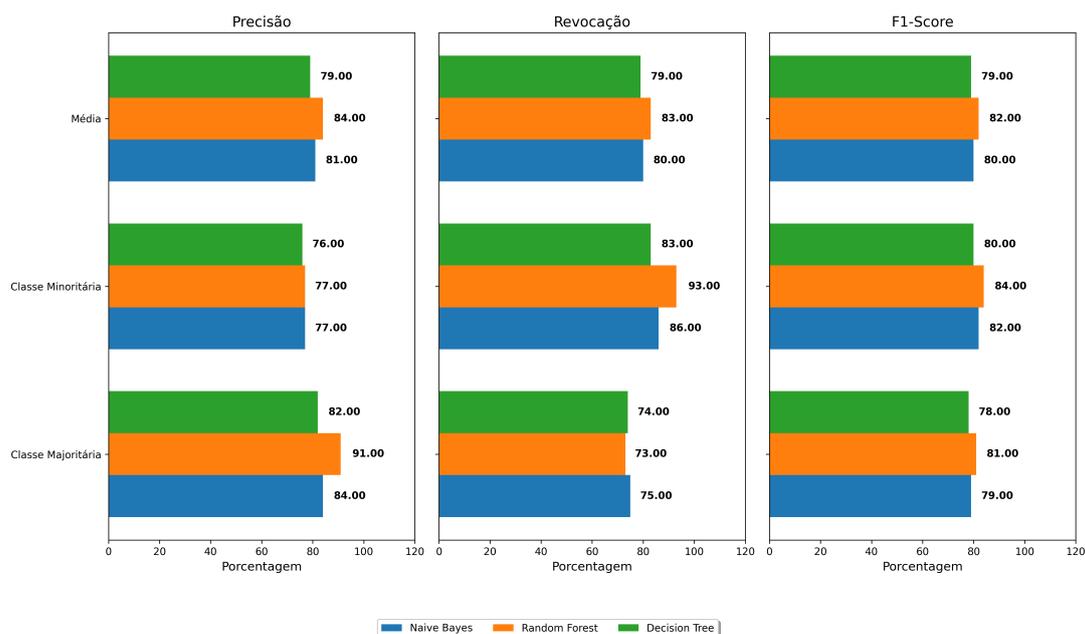


Figura 3. Comparando Modelos de Machine Learning

dos casos sem depressão. A Floresta Aleatória, apesar de ter obtido a melhor taxa de TP, teve uma TN de 72,50%, o que sugere que classificou incorretamente uma porcentagem maior de indivíduos como positivos para a depressão. Já o modelo Naïve Bayes apresentou uma TN de 74,70%, uma taxa intermediária em relação aos outros modelos.

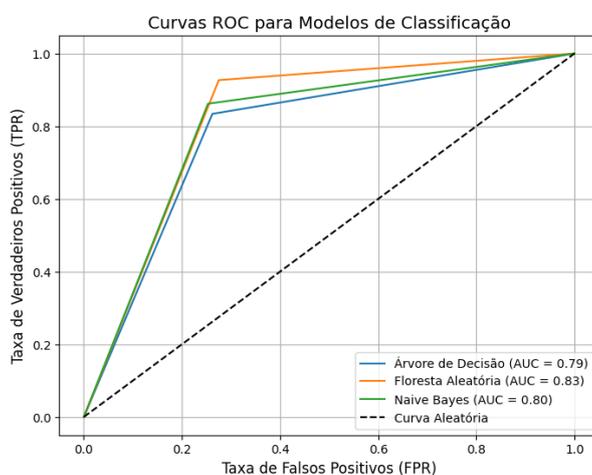


Figura 4. Comparando Modelos de Machine Learning com curvas ROC

4.1. Análise das Regras da Árvore de Decisão

A Árvore de Decisão gerou um conjunto de regras para a predição de depressão, envolvendo diversos fatores de risco, tais como: problemas de sono, consumo de alimentos, falta exercício físico, doenças crônicas, baixa renda e baixo apoio social. A Tabela 3 apresenta um resumo das regras mais relevantes geradas pelo modelo.

As regras indicam, por exemplo, que a presença de problemas de sono, em conjunto com outros fatores como baixo apoio familiar e consumo de frutas abaixo do recomendado, aumenta o risco de depressão em homens. Esses resultados corroboram estudos como o de

Tabela 3. Regras da Árvore de Decisão para Predição de Depressão

Nó	Atributo	Valor de Corte	Decisão
Raiz	Problemas_Sono	1.5	Ramo 1
Ramo 1	sexo_masculino	0.5	Ramo 2
Ramo 2	Consumo de Oleaginosas	1.5	Ramo 3
Ramo 3	Consumo de Frutas	11.5	Ramo 4
...

[Luppino et al. 2010], que demonstraram uma associação significativa entre o IMC e o risco de depressão, e o de [Lorant et al. 2003], que destacam a correlação entre baixos níveis de renda e o aumento do risco de depressão. A análise das regras da Árvore de Decisão apresentada na Tabela 3 demonstra a complexa interação entre diversos fatores que podem influenciar o risco de depressão. A identificação desses fatores, especialmente a relação entre problemas de sono, consumo alimentar, doenças crônicas e rede social, pode ser útil para a elaboração de estratégias de prevenção e tratamento mais eficazes.

5. Conclusões e Trabalhos Futuros

Este estudo demonstra o potencial do aprendizado de máquina para a caracterização do perfil de indivíduos adultos com depressão no Brasil, utilizando dados da Pesquisa Nacional de Saúde (PNS) de 2019. O algoritmo Floresta Aleatória se destacou como o modelo com maior desempenho, evidenciando sua capacidade de identificar padrões e fatores de risco associados à depressão. A análise também reforça a importância de considerar fatores socioeconômicos, estilo de vida e condições de saúde física na compreensão da depressão. No entanto, o estudo apresenta algumas limitações. A base de dados da PNS 2019 possui algumas lacunas em relação aos dados sobre condições de saúde mental, o que pode ter influenciado a precisão dos resultados. Além disso, a análise se concentrou em uma amostra de indivíduos adultos, limitando a generalização dos resultados para outros grupos populacionais. Outros fatores, como aspectos psicológicos e genéticos, que também podem influenciar a depressão, não foram considerados neste estudo.

Agradecimentos

Os autores agradecem o apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo No 303133/2021-0, e do Fundo de Incentivo à Pesquisa (FIP) da PUC Minas, Processo No 30914-1S/2024

Referências

- Barros, M. B. d. A. and et al. (2021). Association between health behaviors and depression: findings from the 2019 brazilian national health survey. *Revista Brasileira de Epidemiologia*, 24(suppl 2):e210010.
- Batista, H. M. C. d., Paim, A. B., Siqueira, B. S., Ebecken, N. F. F., and Dias, A. C. (2021). Fatores que podem desencadear depressão: uma aplicação do aprendizado de máquina aos dados da pesquisa nacional de saúde no brasil. *P2P E INOVAÇÃO*, 7:164–185.
- Beck, A. T., Rush, A. J., Shaw, B. F., and Emery, G. (1979). *Cognitive therapy of depression*. Guilford press.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571.

- Dondena, L. M., Ferretti, E., Maragoudakis, M., Sapino, M., and Errecalde, M. L. (2017). Predicting depression: a comparative study of machine learning approaches based on language usage. *Cuadernos de Neuropsicologia*, 11:42–54.
- Fiske, A., Wetherell, J. L., and Gatz, M. (2009). Depression in older adults. *Annual Review of Clinical Psychology*, 5:363–389.
- Kendler, K. S., Gatz, M., Gardner, C. O., and Pedersen, N. L. (2006). A swedish national twin study of lifetime major depression. *American Journal of Psychiatry*, 163(1):109–114.
- 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.
- Kuehner, C. (2017). Gender differences in unipolar depression: an update of epidemiological findings and possible explanations. *Acta Psychiatrica Scandinavica*, 95(3):163–174.
- Lai, H. M. X., Cleary, M., Sitharthan, T., and Hunt, G. E. (2015). Prevalence of comorbid substance use, anxiety and mood disorders in epidemiological surveys, 1990–2014: A systematic review and meta-analysis. *Drug and alcohol dependence*, 154:1–13.
- Liu, Y., Pu, C., Xia, S., Deng, D., Wang, X., and Li, M. (2022). Machine learning approaches for diagnosing depression using eeg: A review. *Transl Neurosci*, 13(1):224–235.
- Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., and Anseau, M. (2003). Socio-economic inequalities in depression: a meta-analysis. *American Journal of Epidemiology*, 157(2):98–112.
- Loyola-González, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.
- 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.
- MS-BRASIL (2020). Ministério da saúde: A saúde mental no brasil: Indicadores de morbidade.
- Skogen, J. C., Harvey, S. B., Henderson, M., Stordal, E., Mykletun, A., and Øverland, S. (2014). Anxiety and depression among abstainers and low-level alcohol consumers: The nord-trøndelag health study. *Addiction*, 109(2):269–277.
- Smith, K. J., Victor, C., and Bartholomew, J. (2006). Factors associated with the self-reported health status of older people in the united kingdom. *Ageing society*, 26(4):607–627.
- 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.
- Witten, I. H., Frank, E., and Hall, M. A. (2016). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann.
- Zarate, L., Petrocchi, B., Dias Maia, C., Felix, C., and Gomes, M. P. (2023). Capto - a method for understanding problem domains for data science projects: Capto - um método para entendimento de domínio de problema para projetos em ciência de dados. *Concilium*, 23(15):922–941.