Using Machine Learning to identify profiles of individuals with depression

Carlos D. Maia, Cristiane N. Nobre, Marco Paulo S. Gomes, Luis E. Zárate

Pontifícia Universidade Católica de Minas Gerais, Curso de Ciência de Dados
carlosdiasmaia@gmail.com, nobre@pucminas.br, marcopaulo@pucminas.br, zarate@pucminas.br

Abstract. Depression is a major public health problem in Brazil, affecting millions of individuals each year. While the prevalence of depression in Brazil has been well-documented, there is still a need for more accurate and timely predictions of depression trends to improve treatment and prevention strategies. In this study, we explored the potential of machine learning algorithms to forecast depression trends in Brazil using data from the National Health Survey conducted by the Brazilian Institute of Geography and Statistics. We compared the performance of various machine learning models in depression trends, including decision trees, random forests, support vector machines, and neural networks. Additionally, we aimed to identify key risk factors for depression trends in Brazil, including age, gender, income, education, and marital status. These findings have important implications for public health policies and mental healthcare in Brazil. Our study provides insights into the use of machine learning algorithms to predict and prevent depression trends and highlights the potential of data-driven approaches to improve mental health outcomes in Brazil.

CCS Concepts: • **Applied computing** → **Health informatics**; • **Computing methodologies** → **Machine learning algorithms**; **Supervised learning by classification**.

Keywords: Machine Learning, Health, Depression

1. INTRODUÇÃO

A depressão é uma das principais causas de incapacidade em todo o mundo, afetando mais de 300 milhões de pessoas de todas as idades [Woody et al. 2017]. No Brasil, a depressão é um grande problema de saúde pública, afetando cerca de 5,8% da população [BRASIL 2023]. Apesar do significativo impacto da depressão no Brasil, previsões precisas e oportunas das tendências de depressão ainda são escassas, o que pode prejudicar os esforços de tratamento e prevenção. Para abordar essa lacuna, pesquisadores têm recorrido a métodos, técnicas e algoritmos de aprendizado de máquina para identificar padrões a partir de bases de dados públicas ou privadas que poderiam ser usadas para prever tendências de depressão. A aprendizagem de máquina é um subcampo da inteligência artificial que utiliza modelos estatísticos e algoritmos para analisar conjuntos de dados e identificar padrões que podem ser usados para fazer previsões [Alpaydin 2014]. Estudos recentes têm demonstrado o potencial de algoritmos de aprendizado de máquina para prever tendências de depressão em outras populações [Na et al. 2020] [Sharma e Verbeke 2020].

Neste artigo, exploramos o potencial de algoritmos de aprendizado de máquina para identificar perfis e fatores-chave de depressão no Brasil, utilizando dados do mais recente estudo da Pesquisa Nacional de Saúde (PNS) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em 2019 [Instituto Brasileiro de Geografia e Estatística] . Propomos uma metodologia baseada no entendimento do domínio de problema e, após etapas de preparação de dados, comparamos o desempenho de vários modelos de aprendizado de máquina na caracterização da depressão, incluindo Árvores de decisão, Florestas aleatórias, Máquinas de vetores de suporte e Redes neurais artificiais. Também identificamos fatores de risco-chave para tendências de depressão no Brasil, o que poderia induzir políticas de saúde pública e abordagens personalizadas de cuidados. Nosso estudo contribui para o crescente corpo de literatura sobre o uso de aprendizado de máquina para melhorar os resultados de saúde mental e destaca o potencial de abordagens baseadas em dados para enfrentar o impacto global da depressão.

Este trabalho está estruturado em três partes principais: os trabalhos relacionados, que apresentam as abordagens utilizadas por outros pesquisadores para abordar o problema da depressão com aprendizado de máquina; a metodologia, que é dividida em materiais e métodos; e, os materiais, que buscam explicar e contextualizar a base de dados da PNS-2019, enquanto os métodos abrangem todo o pré-processamento realizado na base de dados e na modelagem, utilizando diversos modelos e extraindo o máximo de conhecimento do modelo que apresentou melhor desempenho; as considerações finais . O objetivo final é traçar perfis e identificar fatores-chave para a depressão.

2. TRABALHOS RELACIONADOS

A depressão é um transtorno mental comum e grave que afeta significativamente a qualidade de vida das pessoas. Considerando a importância de se prevenir tal doença, diversos trabalhos buscam, por meio de diversas técnicas de aprendizado de máquina, criar modelos estatísiticos para indentificar padrões relevantes e fazer previsões.

Em Na et al. [2020], os autores utilizaram a pesquisa nacional de bem-estar coreana, o *Korea Welfare Panel Study* (KoWePS), para desenvolver um modelo de aprendizado de máquina. Os autores utilizaram diversas técnincas de pré-processamento para melhorar a qualidade dos dados e geraram um modelo baseado em Floresta Aleatória, alcançando uma acurácia de 0,862, mostrando que é viável a utilização de bases de dados de censos para criação de modelos preditivos que avaliam a depressão.

Já Sharma e Verbeke [2020] propuseram um modelo de predição. Porém, ao invés de utilizarem dados pessoais e demográficos, eles utilizaram uma base de dados de biomarcadores extraídos de amostra de sangue e urina de 11.081 pessoas na Holanda. Considerando que o dataset se encontrava altamente desbalanceado, foram testados diversos métodos de balanceamento. O algoritmo escolhido foi o *Extreme Gradient Boosting* (XGBoost) e, em conjunto com os dados que foram balanceados com *oversampling*, obtiveram uma acurácia de 0,9729.

Em Richter et al. [2020], os autores submeteram 125 pessoas a uma bateria de testes que quantificou diversos vieses cognitivo-emocionais nesse grupo. Após pré-processamento, os dados foram aplicados para o algoritmo Floresta Aleatória, utilizando métricas de tamanho de amostragem, para que cada vez mais o modelo pudesse representar o grupo. Com isso, o modelo de previsão foi utilizado para diferenciar os participantes sintomáticos (ou seja, sintomas elevados de depressão, ansiedade ou ambos) do grupo de controle não sintomático, que revelou uma precisão de 71,44%.

3. METODOLOGIA

3.1 Material

Para este trabalho, foi utilizada a base de dados da Pesquisa Nacional de Saúde (PNS), conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE), em 2019. A PNS é uma pesquisa nacionalmente representativa que coleta informações sobre vários aspectos da saúde, incluindo a saúde mental, de indivíduos com 18 anos ou mais. A pesquisa usou uma estratégia de amostragem multietapas complexa para selecionar uma amostra representativa da população brasileira. A PNS coletou dados sobre várias características sociodemográficas, incluindo idade, gênero, renda, educação e estado civil, além de dados sobre comportamentos e resultados de saúde, incluindo a depressão. A base de dados original da PNS-2019 possui 1.088 atributos e 293.726 instâncias. A pesquisa agrupa os dados coletados em 26 módulos, como mostrados, de forma sintética, na tabela I.

3.2 Métodos

Considerando a grande quantidade de atributos disponíveis, foi aplicado, primeiramente, uma pré-seleção conceitual de atributos baseada em conhecimento explícito (literatura) e de especialistas de

Tabela I: Separação da base de dados, por módulos

Módulo	Nome	Descrição Detalhada
A	Informações do domicílio	Coleta informações sobre características do domicílio, como tipo de moradia e abastecimento de água.
B	Visitas domiciliares de Equipes de Saúde da Família e de Agentes de Endemias	Avalia a frequência das visitas realizadas por equipes de saúde e de agentes de endemias aos domicílios.
C	Características gerais dos moradores	Coleta informações demográficas e socioeconômicas dos moradores, como idade, sexo, raça e nível de escolaridade.
D	Características de educação das pessoas de 5 anos ou mais de idade	Investigação sobre o nível de escolaridade e frequência escolar das pessoas com 5 anos ou mais.
E	Trabalho dos moradores do domicílio	Coleta informações sobre a situação de trabalho dos moradores, como ocupação, setor de atividade e rendimentos.
F	Rendimentos domiciliares	Avalia a renda total e a composição dos rendimentos do domicílio, incluindo trabalho, aposentadorias e benefícios sociais.
G	Pessoas com deficiências	Identifica a presença de deficiências físicas, mentais, intelectuais ou sensoriais nos moradores.
I	Cobertura de plano de saúde	Analisa a proporção de pessoas cobertas por plano de saúde, seja público ou privado.
J	Utilização de serviços de saúde	Coleta informações sobre o acesso e a utilização dos serviços de saúde, incluindo consultas médicas e exames.
K	Saúde dos indivíduos com 60 anos ou mais e cobertura de mamografia entre mulheres de 50 anos ou mais	Avalia a saúde e a cobertura de exames de rotina em indivíduos com 60 anos ou mais, e em mulheres de 50 anos ou mais.
L	Crianças com menos de 2 anos	Analisa a saúde e os cuidados recebidos por crianças com menos de 2 anos, incluindo amamentação e imunização.
M	Características do trabalho e apoio social	Coleta informações sobre condições de trabalho, como carga horária e exposição a riscos, e sobre apoio social no ambiente de trabalho.
N	Percepção do estado de saúde	Avalia a percepção dos indivíduos sobre seu próprio estado de saúde e qualidade de vida.
O	Acidentes	Investigação sobre a ocorrência de acidentes domésticos, de trabalho e de transporte nos domicílios.
P	Estilos de vida	Analisa comportamentos relacionados à saúde, como alimentação, atividade física, tabagismo, consumo de álcool e uso de drogas.
Q	Doenças crônicas	Identifica a presença de doenças crônicas, como diabetes, hipertensão e asma, nos moradores.
R	Saúde da mulher	Avalia a saúde das mulheres, incluindo acesso a serviços de saúde específicos, planejamento familiar e exames preventivos.
S	Atendimento pré-natal	Coleta informações sobre o atendimento pré-natal recebido pelas mulheres durante a gravidez.
U	Saúde bucal	Analisa a saúde bucal da população, incluindo cuidados, problemas dentários e utilização de serviços odontológicos.
Z	Paternidade e pré-natal do parceiro	Investigação sobre a participação do parceiro na paternidade e no pré-natal.
V	Violência	Avalia a exposição a situações de violência e seus impactos na saúde dos moradores.
T	Doenças transmissíveis	Coleta informações sobre a ocorrência de doenças transmissíveis, como HIV/AIDS e tuberculose, nos domicílios.
Y	Atividade sexual	Investigação sobre a atividade sexual e o uso de métodos contraceptivos.
AA	Relações e condições de trabalho	Avalia a relação entre saúde e condições de trabalho, incluindo carga horária, jornada noturna e exposição a agentes nocivos.
H	Atendimento médico	Coleta informações sobre o atendimento médico recebido pelos moradores, incluindo consultas, exames e hospitalização.
W	Antropometria	Avalia medidas antropométricas, como altura, peso e índice de massa corporal (IMC), nos moradores.

domínio, que buscou selecionar apenas atributos que poderiam estar associados ao problema da depressão, dividindo o problema em diversas dimensões, selecionando, assim, um total de 126 atributos, como é resumido na Tabela II.

Esses atributos foram filtrados e transformados, a fim de diminuir a dimensionalidade da base de dados e de aumentar a capacidade interpretativa dos resultados, além de diminuir possíveis vieses e ruídos. Após esse processo, foram definidos 30 atributos, sendo eles mostrados na Tabela III.

a) Pré-processamento: Após segmentação dos dados, foram realizadas fusões e transformações gerando novos atributos a partir de outras variáveis apresentadas na PNS, sendo estes:

- *Segmentação dos dados*: a partir de uma análise exploratória inicial foi decidido segmentar a base de dados considerando apenas pessoas adultas (de 20 a 59 anos), restando, assim, 63.782 instâncias, uma vez que os padrões de depressão variam muito entre as diferentes faixas etárias. Este procedimento foi adotado para evitar caracterização de populações diferentes.

- *Alimentação*: Baseado nos atributos de alimentação do módulo P, foram criadas 3 categorias

Tabela II: Dimensões consideradas

Atributo	Justificativa de escolha	Localização dos atributos na base
	Dimensão Hábitos Alimentares	
Ingestão de alimentos	A alimentação está intrinsecamente ligada à produção de alguns hormônios, como a serotonina, que, em falta, é um dos causadores da depressão.	Atributos P6 a P27 da BD-PNS
Ingestão de álcool	O consumo regular de álcool altera a química do cérebro, afetando a produção de serotonina.	Atributos P50 e P52 da BD-PNS
Ingestão de tabaco e derivados	Os efeitos gerados pelo tabaco (ex: euforia) contribuem para a dificuldade de alcançar a manutenção da abstinência dessa droga, assim como os sintomas desagradáveis pela falta da substância, como compulsão aumentada, irritabilidade, ansiedade, dificuldade de concentração, agitação, sensação de sonolência ou embotamento, bem como reações de hostilidade, que podem levar a um possível quadro de depressão.	Atributos P50 e P52 da BD-PNS
	Dimensão Condições-socioeconômicas	
Renda	Estudos indicam que populações com baixa renda tendem a desenvolver depressão com mais frequência.	E18 da BD-PNS
Plano de saúde	Planos de saúde ajudam na questão do acesso mais frequente a médicos e psicólogos, que podem favorecer a prevenção de possíveis casos de depressão.	I00102 da BD-PNS
Trabalho	Cargas excessivas de trabalho são extremamente nocivas ao corpo humano e podem afetar diretamente o humor e, principalmente, o psicológico dos funcionários, que, ao se sentirem sobrecarregados, podem desenvolver depressão, improdutividade e outros fatores.	E12/E14A/E17, E19,M5D da BD-PNS
Acesso aos serviços de saúde	Acesso frequente a médicos e psicólogos pode favorecer a prevenção de possíveis casos de depressão.	Módulo J - Utilização dos serviços de saúde: J001...J054; Fonte: BD-PNS
	Dimensão Condições Físicas e mentais	
Gravidez	As grandes mudanças físicas, hormonais, psíquicas e de interação social pelas quais as grávidas passam deixam esse grupo mais propenso a desenvolver depressão.	Q03002 e P5 da BD-PNS
Deficiências	Tendo em vista que deficiências podem causar afastamentos sociais, sentimentos de incapacidade e de não pertencimento a grupos, deficiências se mostram fortes causadores da depressão.	Módulo G da BD-PNS
Saúde Mental	O estado de saúde mental reflete diretamente na depressão, uma vez que ela tem como quadro inicial diagnósticos de ansiedade e outros transtornos.	Módulo Q da BD-PNS
Doenças crônicas	Pessoas com doenças crônicas podem apresentar limitações, como de mobilidade, de alimentação, de atividade física e de realização de atividades cotidianas na vida pessoal, social ou no trabalho. Os problemas e implicações inerentes dessas restrições podem levar a transtornos de humor e de depressão.	
	Características do indivíduo	
Sexo	O índice de mulheres com depressão é duas vezes maior do que em homens.	C6 da BD-PNS
Estado Civil	A relação entre os níveis de ansiedade e/ou de depressão com o estado civil obteve significância estatística ($p < 0{,}05$), em que se verificou que a maioria dos indivíduos que não apresentaram níveis de ansiedade e de depressão era casada (75,9%), seguida dos viúvos (20,7%).	C011
Frequenta/frequentou escola	A baixa escolaridade está associada a pior autoavaliação da memória, maior incidência de demências, piora dos sintomas depressivos, maior comparecimento aos serviços médicos, aumento do consumo de medicamentos e elevada prevalência de queixas inespecíficas.	D001, D00201, D008, D00901
Idade	De acordo com dados publicados pelo IBGE, em 2019, a população entre 60 e 64 anos é a mais afetada pela depressão.	C7 da BD-PNS
IMC	A presença de uma elevada taxa de adiposidade é um fator de risco estabelecido para muitas doenças psiquiátricas, incluindo depressão e ansiedade. Há ampla evidência que liga o índice de massa corporal (IMC) mais elevado a maiores chances de depressão e de ansiedade.	P00104 e P00404

Tabela III: Descrição dos atributos considerados

Atributo	Descrição atributo	Valores
		Categórico
Sexo	Gênero do entrevistado	Masculino = 1 Feminino = 2
		Numérico
Idade	Idade do entrevistado no momento da entrevista	000 a 130
Estado Civil	Estado civil do entrevistado	Categórico - Casado(a) = 1, Divorciado(a) = 2, Viúvo(a) = 3, Solteiro(a) = 4
Saber ler e escrever	Se o entrevistado sabe ler e escrever	Dicotômica - Sim = 1, Não = 2
Frequenta escola/creche	Se o entrevistado frequenta escola ou creche	Dicotômica - Sim = 1, Não = 2
Curso Que Frequenta	Curso que o Entrevistado frequenta, atualmente	Categórico - Creche = 1, Pré-escola = 2, Classe de alfabetização - CA = 3, Alfabetização de jovens e adultos = 4, Antigo primário (elementar) = 5, Antigo ginasial (médio 1^{o} ciclo) = 6, Regular do ensino fundamental ou do 1^{o} grau = 7, Educação de jovens e adultos (EJA) ou supletivo do ensino fundamental = 8, Antigo científico, clássico etc. (médio 2^{o} ciclo) = 9, Regular do ensino médio ou do 2^{o} grau = 10, Educação de jovens e adultos (EJA) ou supletivo do ensino médio = 11, Superior - graduação = 12, Especialização de nível superior (duração mínima de 360 horas) = 13, Mestrado = 14, Doutorado = 15.
Já frequentou escola/creche	Se o entrevistado já frequentou escola ou creche	Dicotômica - Sim = 1, Não = 2
Grau de ensino	Grau de ensino mais elevado que o entrevistado já alcançou	Categórico - Creche = 1, Pré-escola = 2, Classe de alfabetização - CA = 3, Alfabetização de jovens e adultos = 4, Antigo primário (elementar) = 5, Antigo ginasial (médio 1^{o} ciclo) = 6, Regular do ensino fundamental ou do 1^{o} grau = 7, Educação de jovens e adultos (EJA) ou supletivo do ensino fundamental = 8, Antigo científico, clássico etc. (médio 2^{o} ciclo) = 9, Regular do ensino médio ou do 2^{o} grau = 10, Educação de jovens e adultos (EJA) ou supletivo do ensino médio = 11, Superior - graduação = 12, Especialização de nível superior (duração mínima de 360 horas) = 13, Mestrado = 14, Doutorado = 15.
Tipo de trabalho	Qual o tipo de trabalho do entrevistado	Categórico - Área de administração, Agropecuárioas/Pesaca/Florestal, Ciências Intelectuais, Diretor/Gerente, Estudante, Muntenção/Reparos, Militares, Serviços Industriais, Trabalhos Manuais/Artesões, Nível Técnico, Vendedores, Não informado/Desempregado
Trabalho	Se o entrevistado trabalhoU na semana de referência	Dicotômica - Sim = 1, Não = 2
Rendimento	Rendimento bruto mensal do entrevistado	Numérico
Horas Trabalho	Horas que trabalhava por semana	Numérico - De 0 a 120
Atividade Física Regular	Se o entrevistado prática ou não ativade física regular	Dicotômica - Sim = 1, Não = 2
Tem Deficiência	Se o entrevistado tem qualquer tipo de limitação	Dicotômica - Sim = 1, Não = 0
Trabalho noturno	Número de horas de trabalho noturno	Numérico - De 0 a 10
Acolhimento Familiar	Número de familiares que o entrevistado pode contar em momentos ruins	Numérica - De 0 a 3
Má Alimentação	Alimentos ultraprocessados ou industrilaizados consumidos	Categórico - não consume alimentos ultraprocessados/industrializados = 0, consume poucos alimentos ultraprocessados/industrializados, consome quantia aceitável de alimentos ultraprocessados/industrializados = 1, consume muitos alimentos ultraprocessados/industrializados = 2
Alimentação Balanceada	Qualidade da alimentação do entrevistado	Dicotômica - Alimentação ruim = 3, Alimentação aceitável = 2, Alimentação boa = 1
Ingere Álcool	Se a entrevistado ingere álcool	Dicotômica - Sim = 1, Não = 0
Raça	Raça do entrevistado	Categórico - 1 = Branca, 2 = Preta, 3 = Amarela, 4 = Parda, 5 = Indígena
IMC	Indíce de massa corporal do entrevistado	Numérico
Gravidez	Se a entrevistada está grávida, naquele momento	Dicotômica - Sim = 1, Não = 2
Tem Outra Doença	Se o entrevistado tem outra doença crônica	Dicotômica - Sim = 1, Não = 0
Tem Depressão	Se o entrevistado tem depressão	Dicotômica - Sim = 1, Não = 2
Fuma Atualmente	Se o entrevistado fuma atualmente	Dicotômica - Sim = 1, Não = 2
Fumou no Passado	Se o entrevistado já fumou algum produto de tabaco	Dicotômica - Sim = 1, Não = 2
Quantidade de Fumo Atual	Quantos cigarros o entrevistado fuma por dia	Numérico

de alimentação, relacionando a frequência do consumo de tipos de alimentos que constituem uma alimentação básica (contendo proteínas, carboidratos, fibras, frutas e verduras). Alimentação ruim = 3, Alimentação aceitável = 2, Alimentação boa = 1, uma vez que a alimentação é fundamental para a produção de alguns hormônios relacionados a depressão. [Ljungberg et al. 2020].

- *Deficiências*: Baseado no módulo G, que trata deficiências presentes nos indivíduos, foram criadas as categorias: Possui deficiência = 1, Não possui deficiências = 0, uma vez que pessoas com deficiências, por conta de suas limitações, tendem a ter mais depressão [Noh et al. 2016].

- *Alimentação inadequada*: Baseado nas informações acerca da alimentação apresentadas no módulo P, foram criadas 3 valores de categorias vinculadas à quantidade de alimentos ultraprocessados e industrializados que o entrevistado consumia, sendo elas: Não consume alimentos ultraprocessados/industrializados = 0, Consume poucos alimentos ultraprocessados/industrializados = 1, Consome quantia aceitável de alimentos ultraprocessados/industrializados = 2, Consume muitos alimentos ultraprocessados/industrializados = 3, uma vez que a alta ingestão de alimentos ultraprocessados pode trazer sintomas como ansiedade e depressão [Lane et al. 2022].

- *Presença de outras doenças*: Se o entrevistado em questão possui qualquer outra doença crônica apontada no módulo Q, um novo atributo foi definido: Não-possui = 0, e Possui = 1.

- *Consumo de álcool*: Foram divididos em 2 grupos, Não-consume álcool = 0 e Consume-álcool = 1, uma vez que o consumo de álcool influencia na não produção de hormônios que podem ser ligados a depressão [McHugh 2019].

- *Grandes Grupos de Trabalho*: Com as informações da variável E01201, foi definido um novo atributo, que visa trazer os 10 grandes grupos de profissões abordadas na Classificação Brasileira de Ocupações para pesquisas domiciliares de 2010. Foi adicionada, também, a categoria estudante, caso o entrevistado em questão não trabalhasse, mas estudasse. Estudos compravam que a depressão está mais presente em determinadas atividades específicas de trabalho [Schonfeld e Bianchi 2021].

Atributos originais foram combinados para gerar novo atributo, por exemplo:

IMC: Utilizando o peso e altura reportada pelos entrevistados no módulo P, foi calculado o IMC dos indivíduos, uma vez que indivíduos obesos tendem a se sentir insatisfeitos com seus corpos, gerando problemas relacionados a ansiedade. Também existem estudos que relacionam a baixa produção de hormônios que previnem a depressão com a obesidade [Blasco et al. 2020].

Codificação da base de dados: Alguns atributos correspondem a variáveis categóricas nominais (por exemplo, Estado Civil) que precisam ser adequadamente codificadas, e outras trazem valores quantitativos contínuos (por exemplo, Salário). Depois, as variáveis Raça e Estado Civil foram codificadas utilizando o One-Hot-Encoding. Após esse processo, a base de dados passou a ter 58 atributos.

- *Imputação de dados ausentes*: Para isso, foi utilizado o MissForest, do pacote MissingPy, para imputar dados ausentes em variáveis categóricas em seu conjunto de dados. O MissForest é uma técnica baseada em florestas aleatórias adaptada especificamente para tratar variáveis categóricas e dados ausentes.

- *Balanceamento*: Como se trata de uma base de dados desbalanceada (5.792 pessoas com depressão e 57.990 pessoas sem depressão) foi necessário fazer o balanceamento da base de dados. Foram testados diversos métodos para realizar esse passo e o que gerou melhores resultados foi o Random Under Sampler, da biblioteca imblearn, no python,deixando assim o dataset com 10.546 entradas.

- *Tratamento de outliers*: Utilizando uma função que calculava o limite inferior e superior das colunas quantitativas salário e idade, foram retiradas as instâncias que possuíam um desses valores abaixo de 200 e maiores que 4.000, no atributo salário, e o atributo idade que não possuía outliers.

Então, a base foi dividida no conjunto de treino e teste, sendo que o conjunto de teste tinha 20% do total da base de dados.

b) Modelagem: O problema tratado é de classificação para descrever o perfil de pessoas com depressão. Para encontrar os melhores padrões entre os indivíduos com depressão, foram aplicados 4 diferentes modelos de aprendizado no conjunto de dados resultante, a fim de comparar seu desempe-

nho. Foram escolhidos como algoritmos:

1) Floresta aleatória (FA): com os seguintes parâmetros: profundidade máxima = 10, número máximo de características = raiz quadrada, número mínimo de amostras em uma folha = 1, número de estimadores = 400, critério = gini, utilizando reposição.

2) Rede Neural Artificial (RNA): para encontrar os melhores parâmetros para a rede neural, foi aplicado o teorema de Kolmogorov [Schmidt-Hieber 2021], que estabelece que uma RNA é um aproximador universal, se a rede possui uma camada de entrada, com a quantidade de neurônios iguais a quantidade de atributos (N), uma camada escondida com o $2 \times N + 1$ e uma camada de saída, e função de ativação sigmoide. Sendo assim, a rede neural possui uma camada de entrada de 58 neurônios, uma camada escondida de 117 neurônios e uma camada de saída.

3) Support Vector Machines (SVM): foi utilizado com os seguinte paramêtros: $C = 0,1, kernel =' linear', degree = 1$.

4) Modelo Árvore de Decisão (AD): com os seguintes parâmetros: profundidade máxima = 10, número máximo de características = raiz quadrada, número mínimo de amostras em uma folha = 5, número de estimadores = 300, critério = gini.

Em todos os modelos, excetuando a RNA, foi utilizado o *RandomizedSearchCV*, do *scikit-learn*, para encontrar os melhores parâmetros. Foi, também, utilizada a biblioteca *KFold*, do *scikit-learn*, com $k = 10$, para se encontrar o conjunto de treinamento onde o modelo performava melhor.

4. RESULTADOS E DISCUSSÕES

A Figura 1 apresenta os resultados dos testes para as métricas recall, precisão, e F1-score, para os quatro algoritmos avaliados.

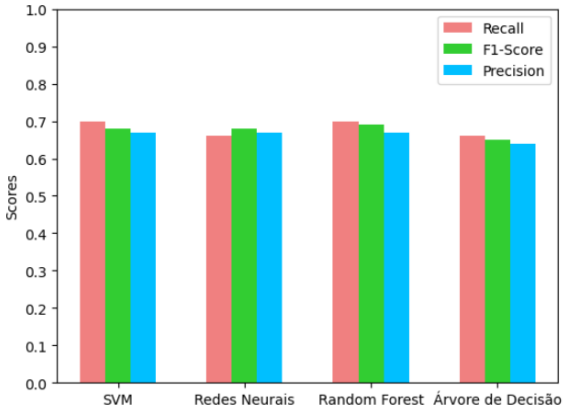

Fig. 1: Resultados dos testes

É possível observar que o melhor desempenho foi com Floresta Aleatória, e por ser considerado um método não interpretável (caixa-preta), utilizamos o algoritmo SHapley Additive exPlanations (SHAP), para trazer interpretabilidade ao modelo.

Pela Figura 2, pode-se observar que o atributo "Sexo" está relacionado com o fato de uma pessoa ter depressão. Já é conhecido que mulheres tendem a ter mais depressão que homens, por questões relacionadas a fatores biológios, socioeconômicos e psicossociais [Albert 2015].

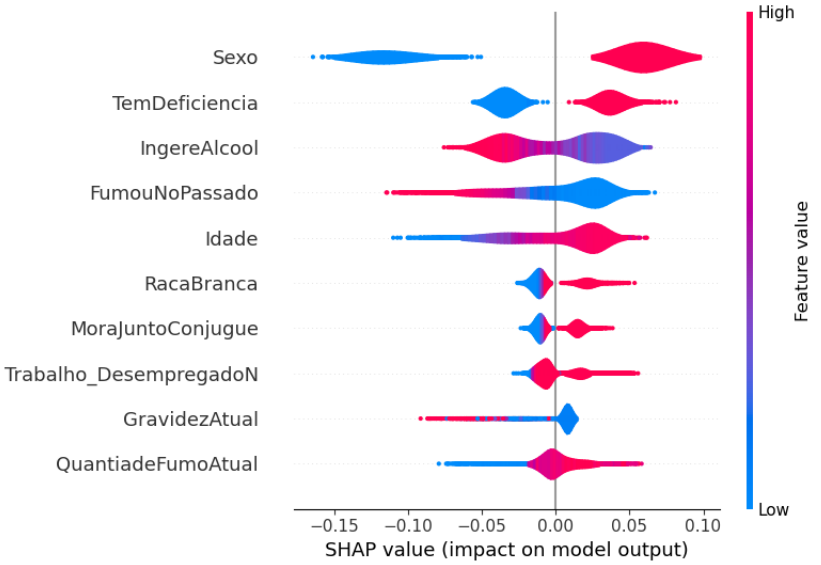

Fig. 2: Atributos mais importantes com SHAP

Também, é possível observar que o fato de a pessoa possuir algum tipo de limitação é relacionado com a depressão. Tal relação pode ser justificada, principalmente, pela estigmatização de pessoas que se enquadram nesses grupos e por conta de desafios físicos e funcionais [Noh et al. 2016].

Outra relação que pode ser observada é o fato de que pessoas que ingerem álcool tendem a ter mais depressão. Estudos mostram como o álcool afeta os neurotransmissores do cérebro, como a serotonina, que desempenha um papel importante na regulação do humor [Kim et al. 2021].

Além disso, foi utilizada a biblioteca te2rules, do Python [Lal et al. 2022], com os parâmetros $num\_stages = 300$, $min\_precision = 0{,}95$, para extrair regras desse modelo. Foram extraídas 208 possíveis regras entre as 300 árvores geradas, sendo que as mesmas representam 100% das previsões positivas e 85,61% das previsões negativas. Duas das regras geradas foram:

Regra 1: Se é divorciada E está gravida atualmente E tem o IMC maior que 25,5 E ingere álcool E fuma mais que 9 cigarros semanalmente ENTÃO tem depressão.

Regra 2: Se não realiza atividade física regular E fumou no passado E tem idade superior a 28 anos E ingere álcool E não mora com o cônjuge E não é branco E é desempregado ENTÃO tem depressão.

Ao analisar as regras, os resultados demonstram associações já conhecidas da depressão, como, por exemplo, a presença maior de depressão em pessoas do sexo feminino [Albert 2015], que pessoas que já têm algum tipo de problema físico estão associadas a uma maior depressão [Noh et al. 2016], além do IMC mais elevado.

Além disso, percebe-se que os atributos mais importantes na Figura 2 também aparecem com mais frequência nas regras. Isto corrobora com a importância de combinar métodos para extração do conhecimento adquirido pelos modelos de aprendizado, a fim de elucidar melhor o problema analisado.

5. CONSIDERAÇÕES FINAIS

Ao explorar o potencial dos algoritmos de aprendizado de máquina, como aumento do gradiente, florestas aleatórias, máquinas de vetores de suporte e redes neurais, com base nos dados da Pesquisa Nacional de Saúde, do IBGE, em 2019, pudemos analisar e comparar diferentes modelos para prever as tendências de depressão. Esses modelos podem fornecer informações valiosas sobre fatores de risco-

chave para a depressão, no Brasil, permitindo a identificação de políticas de saúde pública mais eficazes e a implementação de abordagens personalizadas de cuidados.

No entanto, é fundamental reconhecer que o uso de algoritmos de aprendizado de máquina na previsão de tendências de depressão possui limitações e desafios. É necessário um cuidadoso tratamento e interpretação dos dados, consideração ética e responsabilidade no uso dessas abordagens. Além disso, é importante garantir que as intervenções baseadas em dados sejam complementadas por uma abordagem holística que envolva profissionais de saúde e bem-estar mental.

Além disso, conclui-se que é possível utilizar a base de dados da Pesquisa Nacional de Saúde para descobrir padrões entre as pessoas que possuem depressão, no Brasil. Porém, o modelo fica limitado, uma vez que a base de dados não possuí dados temporais para suportar um modelo de previsão. Ou seja, descobrir se certo indivíduo terá ou não depressão, de acordo com seus hábitos atuais.

Em resumo, este estudo destaca a promessa do aprendizado de máquina na previsão de tendências de depressão, fornecendo insights valiosos para informar políticas de saúde pública e melhorar os cuidados personalizados. Esperamos que essa pesquisa incentive futuros estudos e abra caminho para avanços significativos no campo da saúde mental, especialmente no combate à depressão e no seu impacto na sociedade.

REFERÊNCIAS

Albert, P. R. Why is depression more prevalent in women? *Journal of psychiatry & neuroscience: JPN* 40 (4): 219, 2015.

Alpaydin, E. *Introduction to Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2014.

Blasco, B., García-Jiménez, J., Bodoano, I., and Gutiérrez-Rojas, L. Obesity and depression: Its prevalence and influence as a prognostic factor: A systematic review. *Psychiatry investigation* vol. 17, 08, 2020.

BRASIL. Depressão, 2023.

Instituto Brasileiro de Geografia e Estatística. Pesuisa Nacional de Saúde. `https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=29270`.

Kim, H., Yoo, J., Han, K., Fava, M., Mischoulon, D., Park, M. J., and Jeon, H. J. Associations between smoking, alcohol consumption, physical activity and depression in middle-aged premenopausal and postmenopausal women. *Frontiers in Psychiatry* vol. 12, pp. 2437, 2021.

Lal, G. R., Chen, X., and Mithal, V. Te2rules: Extracting rule lists from tree ensembles. *arXiv preprint arXiv:2206.14359*, 2022.

Lane, M. M., Gamage, E., O'Neil, A., Jacka, F., Marx, W., Dissanayaka, T., Ashtree, D., Travica, N., Gauci, S., and Lotfalian, M. Ultra-processed food consumption and mental health: A systematic review and meta-analysis of observational studies. *Nutrients* vol. 14, pp. 2568, 06, 2022.

Ljungberg, T., Bondza, E., and Lethin, C. Evidence of the importance of dietary habits regarding depressive symptoms and depression. *International Journal of Environmental Research and Public Health* vol. 17, pp. 1616, 03, 2020.

McHugh, R. Alcohol use disorder and depressive disorders. *Alcohol Research: Current Reviews* vol. 40, 10, 2019.

Na, K.-S., Cho, S.-E., Geem, Z. W., and Kim, Y.-K. Predicting future onset of depression among community dwelling adults in the republic of korea using a machine learning algorithm. *Neuroscience Letters* vol. 721, pp. 134804, 01, 2020.

Noh, J.-W., Kwon, Y. D., Park, J., Oh, I.-H., and Kim, J. Relationship between physical disability and depression by gender: a panel regression model. *PloS one* 11 (11): e0166238, 2016.

Richter, T., Fishbain, B., Markus, A., Richter-Levin, G., and Okon-Singer, H. Using machine learning-based analysis for behavioral differentiation between anxiety and depression. *Scientific Reports* vol. 10, 10, 2020.

Schmidt-Hieber, J. The kolmogorov–arnold representation theorem revisited. *Neural Networks* vol. 137, pp. 119–126, 2021.

Schonfeld, I. and Bianchi, R. From burnout to occupational depression: Recent developments in research on job-related distress and occupational health. *Frontiers in Public Health* vol. 9, pp. 1–6, 12, 2021.

Sharma, A. and Verbeke, W. J. M. I. Improving diagnosis of depression with xgboost machine learning model and a large biomarkers dutch dataset (n = 11,081). *Frontiers in Big Data* vol. 3, 2020.

Woody, C., Ferrari, A., Siskind, D., Whiteford, H., and Harris, M. A systematic review and meta-regression of the prevalence and incidence of perinatal depression. *Journal of Affective Disorders* vol. 219, pp. 86–92, 2017.