Comparative Evaluation of Class Balancing Strategies for Depression Detection in Reddit Posts

Thallyson G. M. C. Fontenele, Bruno Feres de Souza[1]

[1]Curso de Engenharia da Computação
Universidade Federal do Maranhão (UFMA) – São Luís – MA – Brasil

thallyson.gabriel@discente.ufma.br, bruno.feres@ufma.br

Abstract. *Real-world health datasets are often imbalanced, making it difficult to build effective predictive models. This study evaluates the impact of balancing techniques in detecting signs of depression in Reddit posts using the eRisk 2017 dataset. Four sampling strategies and a no-resampling approach were applied to Random Forest, XGBoost, and GLM models. Evaluation metrics included sensitivity, specificity, and AUC-ROC. Partial undersampling achieved an AUC-ROC of 0.72 with GLM (lambda = 1), while the combined method with oversampling reached 0.68 with RF (mtry = 126). The results highlight the importance of balancing techniques in mental health prediction tasks.*

Resumo. *Conjuntos de dados reais em saúde geralmente são desbalanceados, o que dificulta a construção de modelos preditivos eficazes. Este trabalho avalia o impacto de técnicas de balanceamento na detecção de sinais de depressão em postagens do Reddit, utilizando o conjunto de dados do eRisk 2017. Foram testadas quatro estratégias de amostragem e uma abordagem sem reamostragem, aplicadas aos modelos Random Forest, XGBoost e GLM. As métricas analisadas foram sensibilidade, especificidade e AUC-ROC. A subamostragem parcial atingiu AUC-ROC de 0,72 com GLM (lambda = 1), enquanto a combinação com superamostragem obteve 0,68 com RF (mtry = 126). Os resultados reforçam a importância de técnicas de balanceamento para predição em saúde mental.*

1. Introdução

As redes sociais representam conexões entre indivíduos estabelecidas por meio da comunicação e de interações em ambientes virtuais. Suas principais características são analisadas tanto a partir da estrutura da rede digital quanto das funções sociais que desempenham [Visentini et al. 2018]. A velocidade de interação e a facilidade de acesso à informação nas plataformas digitais têm impulsionado significativamente o rápido crescimento do número de usuários, superando as limitações dos meios tradicionais, como o rádio e a televisão. [Nadaraja and Yazdanifard 2014].

Embora celebradas por sua capacidade de aproximar as pessoas, as redes sociais também são reconhecidas como espaços onde os usuários experienciam uma série de impactos negativos na saúde mental [Sampasa-Kanyinga and Hamilton 2015, Mussio 2019, Junior et al. 2022]. Conforme apontam [Souza and da Cunha 2019], o uso excessivo está associado ao desenvolvimento de dependência, especialmente entre os jovens, uma vez que o ambiente virtual favorece a exposição de relacionamentos e a busca de *status* por meio de curtidas e seguidores [Souza and da Cunha 2019]. Além disso, o ambiente digital

favorece comportamentos hostis, como o bullying virtual, que está associado à manifestação de sintomas depressivos [Sampasa-Kanyinga and Hamilton 2015].

Segundo a [OPAS 2017], a depressão é caracterizada por uma tristeza profunda e persistente, irritabilidade, flutuações de humor e falta de motivação, sendo uma condição de saúde mental que afeta significativamente o bem-estar emocional. Como relata [Junior et al. 2022], em casos mais graves, a depressão pode desencadear pensamentos suicidas, contribuindo para que o suicídio seja a quarta principal causa de morte entre adolescentes [Junior et al. 2022, WHO 2023]. Considerando que os transtornos mentais representam um risco significativo, é fundamental diagnosticar precocemente a depressão e iniciar o acompanhamento em saúde de forma imediata [Junior et al. 2022].

Nesse contexto, cresce o interesse por abordagens computacionais baseadas em dados para apoiar o diagnóstico e a intervenção em saúde mental, especialmente por meio da análise de conteúdos compartilhados em redes sociais [Islam et al. 2018, Rahman et al. 2018, Kim et al. 2021]. No entanto, essas abordagens enfrentam desafios metodológicos importantes. Um dos principais é o desbalanceamento das classes nos dados, o que pode comprometer o desempenho e a imparcialidade dos modelos preditivos [Gentili et al. 2024, Ostojic et al. 2024]. De acordo com [Lin and Chen 2012], a presença de dados desbalanceados é comum em diversas aplicações biomédicas, nas quais o objetivo é identificar com precisão as amostras de maior interesse ou classificar os pacientes em subgrupos clínicos adequados [Lin and Chen 2012]. Diante desse cenário, torna-se essencial aplicar técnicas de balanceamento e estratégias rigorosas de validação, a fim de garantir que os modelos de predição sejam mais robustos e representativos da diversidade populacional. Assim, este trabalho tem por objetivo avaliar o impacto de técnicas de balanceamento de dados na detecção de sinais de depressão em postagens da rede social Reddit.

A estrutura deste artigo está organizada da seguinte forma: na Seção 2, são apresentados os trabalhos correlacionados; na Seção 3, descrevem-se os métodos e procedimentos do experimento; a Seção 4 apresenta os resultados obtidos e, por fim, a Seção 5 discute as conclusões e perspectivas futuras. As referências utilizadas encontram-se ao final do artigo, na Seção de Referências.

2. Trabalhos Relacionados

Esta seção está dividida em duas partes: primeiro, discutiremos os artigos dos grupos de pesquisa que participaram do Workshop eRisk 2017, promovido pelo CLEF [ERISK 2017] e desenvolveram trabalhos de predição de indicadores de depressão a partir do mesmo conjunto de dados utilizado neste artigo. Em seguida, abordaremos outros trabalhos sobre predição na área da saúde e como lidaram com dados desbalanceados.

2.1. Trabalhos realizados no eRisk 2017

O eRisk 2017, realizado no âmbito do CLEF, foi um workshop pioneiro dedicado à predição precoce de riscos com base em conteúdos publicados na internet [ERISK 2017]. A principal tarefa proposta envolveu a detecção antecipada de sinais de depressão em postagens de usuários da rede social Reddit, desafiando os participantes a desenvolver modelos capazes de emitir alerta a partir da análise sequencial de textos. O objetivo era simular um cenário realista de monitoramento contínuo e avaliação progressiva, exigindo abordagens robustas frente a dados não numéricos e desbalanceados [ERISK 2017].

Embora algumas equipes tenham reconhecido o desbalanceamento dos dados, poucas desenvolveram técnicas de reamostragem ou balanceamento, o que resultou em métricas baixas ou desbalanceadas [Villatoro-Tello et al. 2017, Malam et al. 2017, Farías-Anzaldúa et al. 2017, Losada et al. 2017]. O trabalho de [Almeida et al. 2017] apresentou cinco abordagens, mesclando modelos de AM supervisionados e modelos de Recuperação de Informação (IR). Os autores utilizaram técnicas de Processamento de Linguagem Natural (PLN) que contribuíram para o processo de padronização dos textos, incluindo *tokenização*, conversão para minúsculas, *stemização*, remoção de *stopwords* e pontuação. No entanto, a equipe não realizou intervenções diretas no balanceamento dos dados, obtendo métricas como *F1 Score* (F1) (0,53), precisão (0,48) e *recall* (0,60), que não se destacaram [Losada et al. 2017, Almeida et al. 2017].

Em [Sadeque et al. 2017], foram utilizados bancos léxicos externos relacionados à depressão, combinados com Redes Neurais Recorrentes (RNN) e *Support Vector Machine* (SVM). Embora os autores não mencionem o uso de técnicas de reamostragem, o comportamento dos modelos sugeriu desequilíbrio: a abordagem com RNN obteve o maior recall (0,92) entre todas as equipes participantes. No entanto, o F1 foi baixo (0,34), indicando que o modelo tinha alta sensibilidade, mas ao custo de alta taxa de falsos positivos [Losada et al. 2017, Sadeque et al. 2017].

O trabalho de [Trotzek et al. 2017] abordou diretamente o desbalanceamento das classes para seus classificadores de regressão logística. Os autores utilizaram "pesos de classe modificados" para aumentar o custo dos falsos negativos, uma estratégia que dá maior importância à classe minoritária. Os pesos foram calculados como 1/(1+w) para a classe não deprimida e w/(1+w) para a classe deprimida, com valores específicos de w definidos para diferentes modelos [Trotzek et al. 2017]. Essa abordagem resultou nas melhores métricas de F1 (0,64) e precisão (0,69) entre todos os trabalhos [Losada et al. 2017].

Por fim, [Errecalde et al. 2017] reconheceu o desbalanceamento do conjunto de dados e adotou inicialmente a abordagem denominada *Temporal Variation of Terms* (TVT), dividindo o conjunto de dados em blocos de tempo iguais e analisando a variação dos termos. Posteriormente, complementaram seus métodos com modelos clássicos de AM, como *Random Forest* (RF), *Naive Bayes* (NB) e Árvores de Decisão. Como resultado, além de obterem o melhor ERDE (*Early Risk Detection Error*), métrica do eRisk que penaliza classificações tardias, também apresentaram resultados moderados em F1 (0,59), precisão (0,48) e *recall* (0,79) [Losada et al. 2017].

2.2. Análise preditiva na área da saúde

De acordo com [Mena et al. 2012], e reforçado por [dos Santos 2018], na realização de análises preditivas em saúde, uma particularidade dos conjuntos de dados é o desbalanceamento entre as classes de resposta. Ou seja, ao coletarmos dados do mundo real, observa-se uma maior frequência de registros da classe de indivíduos saudáveis em comparação à classe de indivíduos não saudáveis, sendo esta última geralmente a mais relevante nas análises preditivas [Mena et al. 2012].

O desbalanceamento de classes é reconhecido por [Ostojic et al. 2024] como um dos principais desafios na aplicação de modelos de Aprendizado de Máquina (AM) na pesquisa psiquiátrica e na prática clínica. As melhores estratégias para mitigar esse de-

sequilíbrio em modelos de AM dividem-se em duas categorias principais: (a) abordagens baseadas em dados, que atuam por meio da reamostragem, aumentando ou diminuindo o número de amostras; e (b) abordagens baseadas em algoritmos, que modificam a regra de aprendizado do modelo de AM para considerar a razão de desequilíbrio [Lin and Chen 2012].

Além disso, [Mena et al. 2012] ressalta que, em tarefas de classificação na área da saúde, a sensibilidade (*recall*), a especificidade e a Área Sob a Curva ROC (AUC ROC) são métricas determinantes na avaliação de modelos de AM. A análise conjunta da sensibilidade e da especificidade é especialmente relevante, uma vez que a classe minoritária, frequentemente a de maior interesse clínico, tende a ser mais afetada por erros de classificação. Em busca de maior acurácia geral, os modelos muitas vezes favorecem a especificidade, o que pode reduzir a capacidade de detectar verdadeiros positivos e comprometer a generalização em novos dados [Mena et al. 2012, dos Santos 2018]. Já a AUC ROC avalia a capacidade discriminativa do modelo ao distinguir entre pacientes com e sem o desfecho de interesse, permitindo observar o *trade-off* entre sensibilidade e especificidade ao longo de diferentes limiares de decisão [Ostojic et al. 2024, dos Santos 2018].

Estudos demonstram que, embora o balanceamento de dados represente um desafio na mineração de dados, ele melhora significativamente as métricas preditivas [Abdulsadig and Rodriguez-Villegas 2024, Araf et al. 2024]. O estudo de caso apresentado por [Gentili et al. 2024] demonstra que a aplicação de técnicas de balanceamento de dados, como *Oversampling*, *Undersampling*, *Synthetic Minority Oversampling Technique* (SMOTE) e *Cost-Sensitive Learning*, resultou em uma melhora substancial no *recall*, além de melhorias significativas na acurácia [Gentili et al. 2024].

3. Metodologia da pesquisa

Esta seção apresenta a metodologia de pesquisa adotada neste estudo. A abordagem segue o processo de *Knowledge Discovery from Data* (KDD), conforme proposto por [Morais and Ambrósio 2007], aplicado à mineração de texto. O KDD é estruturado em quatro etapas principais: identificação do problema, preparação dos dados (pré-processamento), mineração de dados e pós-processamento. Essa abordagem permite a extração de conhecimento útil a partir do conjunto de dados, alinhando-se aos objetivos deste trabalho. O notebook com os códigos está disponível publicamente[1].

3.1. Identificação do problema

Nesta etapa, realizamos a identificação e definição dos objetivos a serem alcançados pelo experimento [Morais and Ambrósio 2007]. Uma forma de compreender o problema é estudar a base de dados e responder a perguntas como: de onde foi extraída? Qual é o formato? Qual é o conteúdo?

Uma base de dados é uma coleção de informações organizadas de maneira estruturada ou não estruturada, sendo um recurso essencial para treinar algoritmos de AM [Zhang et al. 2020]. No contexto da mineração de dados, um conjunto de dados de treinamento precisa atender a características fundamentais para garantir um bom desempenho dos algoritmos de classificação: conter um grande volume representativo de informações, além de apresentar conteúdo de alta qualidade [de Oliveira Melo and Cortes 2021].

[1] `https://github.com/thallyson1997/Text_Mining_Depression`

Em nosso experimento, utilizamos o *banco de dados* do Workshop CLEF eRisk 2017 – Previsão Antecipada de Risco na Internet: Fundamentos Experimentais. Este foi um estudo exploratório voltado para a detecção precoce do risco de depressão. O desafio envolve processar de forma sequencial as evidências e identificar os primeiros sinais de depressão o mais cedo possível. A abordagem foca na avaliação de soluções de mineração de texto e, portanto, concentra-se em analisar publicações provenientes de redes sociais. A base de dados é composta por dois conjuntos de dados: treinamento e teste; dividida cronologicamente em vinte conjunto de dados (dez conjuntos de dados de treinamento e dez conjuntos de dados de teste) [ERISK 2017]. Na Tabela 1 e 2, constam as principais estatísticas do conjunto de dados de treinamento e teste.

Tabela 1. Estatística dos dados de treinamento.

	Treinamento	
	Depressivo	Controlado
Número de usuários	83	403
Número de postagens	30851	264172
Média de postagens por usuários	371.7	655.5
Média de palavras por postagens	27.6	21.3

Tabela 2. Estatística dos dados de teste.

	Teste	
	Depressivo	Controlado
Número de usuários	52	349
Número de postagens	18706	217665
Média de postagens por usuários	359.7	623.7
Média de palavras por postagens	26.9	22.5

3.2. Preparação dos dados

Antes do treinamento, os conjuntos de dados passaram por um processo de preparação envolvendo técnicas de limpeza e transformação para adequar os dados aos modelos de AM [Morais and Ambrósio 2007]. Foram utilizadas funções e pacotes especializados em Processamento de Linguagem Natural (PLN), como os pacotes *stringr* (versão 1.5.1) e *tm* (versão 0.7.16), disponíveis no ambiente de desenvolvimento integrado RStudio.

A primeira etapa consistiu na remoção de expressões regulares por meio da função *str_replace_all*. Essas expressões atuavam como marcadores textuais padronizados, tais como identificadores de usuário, data e hora da postagem, além da origem do texto, presentes em todas as amostras, mas que não apresentavam valor informativo para o treinamento ou teste do modelo.

Na sequência, foram eliminados números e sinais de pontuação utilizando a função *gsub*. Posteriormente, aplicou-se a função removeWords para retirar *stopwords*, que são palavras comuns na língua (como pronomes, preposições, conjunções e outras classes gramaticais de função) e geralmente irrelevantes em tarefas de aprendizado de máquina por não carregarem significado contextual útil.

Concluída a limpeza textual, foram realizadas as transformações. Todos os caracteres foram convertidos para minúsculas com a função *tolower*, e as palavras foram reduzidas ao radical por meio da função *stemDocument*. Em seguida, aplicaram-se cinco abordagens distintas de balanceamento amostral.

A vetorização com *Term Frequency–Inverse Document Frequency* (TF-IDF), utilizando as funções *DocumentTermMatrix* e *weightTfIdf*, foi aplicada após cada reamostragem, exceto nos métodos D e E, nos quais o SMOTE requer amostras já vetorizadas.

3.2.1. Método A — Sem modificação no número de amostras (linha de base)

Neste método, mantivemos a quantidade original de 486 amostras no conjunto de treinamento, com vetorização por TF-IDF.

3.2.2. Método B — Subamostragem com classes balanceadas

Equalizamos o número de amostras entre as duas classes, reduzindo a classe majoritária [Gentili et al. 2024, Abdulsadig and Rodriguez-Villegas 2024]. Embora [Gentili et al. 2024] utilize a técnica de subamostragem aleatória, priorizamos a exclusão de textos mais curtos (em número de caracteres), visando preservar o conteúdo informativo.

3.2.3. Método C — Subamostragem parcial

Aplicamos o mesmo procedimento de redução da classe majoritária, como no método B, porém com proporção diferente [Gentili et al. 2024]. Nesta abordagem, a quantidade da classe majoritária foi ajustada para ser duas vezes e meia maior que a da classe minoritária, totalizando 290 amostras.

3.2.4. Método D — Superamostragem com SMOTE

Neste método, aplicamos a técnica SMOTE para gerar novas amostras sintéticas da classe minoritária. O SMOTE cria amostras artificiais com base nos K vizinhos mais próximos de cada ponto minoritário [Gentili et al. 2024, Abdulsadig and Rodriguez-Villegas 2024]. Com o uso do pacote *smotefamily* (versão 1.4.0), o conjunto final resultou em 735 amostras, com K=81.

3.2.5. Método E — Combinação de subamostragem e superamostragem

Por fim, neste método, combinaram-se as estratégias dos métodos C e D [Gentili et al. 2024]. Primeiramente, a classe majoritária foi reduzida para manter uma proporção de duas vezes e meia em relação à classe minoritária. Em seguida, realizou-se a vetorização textual com TF-IDF e, posteriormente, aplicou-se o SMOTE para gerar novas amostras da classe minoritária. O conjunto final ficou composto por 373 amostras (K=81).

3.3. Treinando e testando os dados

Após a preparação, realiza-se o treinamento dos modelos de AM com base nos dados processados [Morais and Ambrósio 2007]. O objetivo é ajustar o modelo aos padrões existentes no conjunto de treinamento e, em seguida, avaliar sua capacidade de generalização com base no conjunto de teste.

Os conjunto de dados de treinamento foram utilizados para treinar cada modelo individualmente de RF, *eXtreme Gradient Boosting* (XGBoost) e *Lasso and Elastic-Net Regularized Generalized Linear Models* (GLM), em seus respectivos conjunto de dados de teste. Cada conjunto de treinamento foi pareado com o respectivo conjunto de teste da mesma ordem cronológica (mais antigo com mais antigo, mais recente com mais recente). Essa metodologia de divisão temporal permite uma avaliação robusta do desempenho dos modelos ao longo do tempo, considerando a dinâmica temporal dos dados. Isso proporciona uma compreensão mais abrangente da capacidade dos modelos de generalizar e se adaptar a novos cenários.

Para criar modelos de AM para classificação, utilizamos o pacote *caret* (versão 7.0.1) para treinar os modelos e extrair métricas como acurácia, F1, especificidade, *recall* e a Área Sob a Curva (AUC-ROC). Treinamos os modelos através da função *train* usando o pacote *randomForest* (versão 4.7.1.2) para RF, o pacote *xgboost* (versão 1.7.11.1) para XGBoost e o pacote *glmnet* (versão 4.1.8) para GLM. Após o teste, extraímos métricas de desempenho de cada modelo usando as funções *confusionMatrix* e *roc* do pacote *pROC* (versão 1.18.5).

3.4. Pós-processamento

Na fase de pós-processamento, foram conduzidos diversos ciclos de retreinamento e reavaliações dos modelos, visando ajustar os hiperparâmetros e otimizar o desempenho preditivo. Esse processo foi realizado por meio de uma abordagem empírica baseada em tentativa e erro, na qual diferentes combinações de valores foram testadas iterativamente, com base nas métricas de avaliação obtidas. Os experimentos foram repetidos até que os modelos atingissem um desempenho estável e satisfatório em relação às métricas de acurácia, AUC-ROC, F1, sensibilidade e especificidade.

Tabela 3. Parâmetros utilizados no modelo RF.

Random Forest - "rf"	
mtry	2, 126, 250

Tabela 4. Parâmetros utilizados no modelo XGBoost.

XGBoost - "xgbTree"	
nrounds	250, 300, 350
max_depth	4
eta	0.1
gamma	3
colsample_bytree	0.6
min_child_weight	1
subsample	0.5

Tabela 5. Parâmetros utilizados no modelo GLM.

GLM - "glmnet"	
lambda	0.1, 0.5, 1
alpha	0.01

As Tabelas 3, 4 e 5 apresentam, respectivamente, os hiperparâmetros finais definidos para os modelos RF, XGBoost e GLM, resultantes do processo de refinamento. Os parâmetros 'mtry', 'lambda' e 'nrounds' referem-se, respectivamente, ao número de variáveis consideradas em cada divisão (RF), ao coeficiente de regularização do tipo Elastic Net (GLM com alpha = 0,01) e ao número de iterações (XGBoost).

4. Resultados e Discussões

Os resultados foram obtidos por meio da avaliação da acurácia, especificidade, sensibilidade, *F-Score* e valor AUC-ROC. Ao final, calculou-se a média das métricas correspondentes a cada parâmetro avaliado. Esse procedimento permitiu uma análise abrangente do desempenho dos modelos sob diferentes configurações dos algoritmos utilizados.

Nas Tabelas 5 a 7, estão as médias das métricas para três valores de *mtry* no algoritmo RF: 2, 126 e 250. Essas médias foram calculadas com base nos resultados dos cinco métodos de reamostragem empregados.

Tabela 6. Métrica RF (*mtry* = 2)

mtry=2	A	B	C	D	E
Acurácia	0.87	0.27	0.55	0.87	0.49
Sensib.	0.00	0.99	0.75	0.18	0.89
Especif.	1.00	0.17	0.52	0.97	0.43
F-Score	NA	0.26	0.30	0.26	0.31
AUC-ROC	0.50	0.58	0.64	0.58	**0.66**

Tabela 7. Métrica RF (*mtry* = 126)

mtry=126	A	B	C	D	E
Acurácia	0.88	0.32	0.49	0.86	0.54
Sensib.	0.19	0.98	0.90	0.37	0.86
Especif.	0.98	0.22	0.43	0.93	0.50
F-Score	0.29	0.27	0.31	0.40	0.33
AUC-ROC	0.59	0.60	0.66	0.65	**0.68**

Tabela 8. Métrica RF (*mtry* = 250)

mtry=250	A	B	C	D	E
Acurácia	0.88	0.33	0.51	0.85	0.54
Sensib.	0.21	0.97	0.90	0.39	0.85
Especif.	0.98	0.24	0.45	0.91	0.49
F-Score	0.30	0.27	0.32	0.39	0.32
AUC-ROC	0.59	0.60	**0.68**	0.65	0.67

Observa-se que, no caso de *mtry* = 2, o valor do F-Score aparece como "NA"(*Not Available*). Isso ocorreu devido à ausência de exemplos positivos reais na classe prevista durante a validação, impossibilitando o cálculo do F-Score, que depende tanto da precisão quanto da sensibilidade.

Para o modelo XGBoost, as médias das métricas foram calculadas para os três valores dos parâmetros *nrounds* utilizados: 250, 300 e 350, em relação aos cinco métodos de reamostragem. Esses resultados estão detalhados nas Tabelas 9 a 11.

Tabela 9. Métrica XGBoost (*nrounds* = 250)

nrounds=250	A	B	C	D	E
Acurácia	0.88	0.32	0.52	0.85	0.52
Sensib.	0.22	0.97	0.86	0.38	0.86
Especif.	0.97	0.22	0.47	0.92	0.47
F-Score	0.31	0.27	0.32	0.40	0.32
AUC-ROC	0.60	0.60	0.66	0.65	**0.67**

Tabela 10. Métrica XGBoost (*nrounds* = 300)

nrounds=300	A	B	C	D	E
Acurácia	0.88	0.31	0.52	0.85	0.53
Sensib.	0.23	0.98	0.87	0.38	0.86
Especif.	0.97	0.21	0.46	0.92	0.48
F-Score	0.33	0.27	0.32	0.40	0.32
AUC-ROC	0.60	0.59	0.67	0.65	**0.67**

Tabela 11. Métrica XGBoost (*nrounds* = 350)

nrounds=350	A	B	C	D	E
Acurácia	0.88	0.31	0.52	0.85	0.52
Sensib.	0.23	0.98	0.86	0.38	0.86
Especif.	0.97	0.21	0.47	0.92	0.48
F-Score	0.33	0.27	0.32	0.40	0.32
AUC-ROC	0.60	0.59	0.66	0.65	**0.67**

E, por fim, para o modelo GLM, as médias das métricas foram calculadas para os três valores dos parâmetros *lambda* utilizados: 0.1, 0.5 e 1, em relação aos cinco métodos de reamostragem. Esses resultados estão detalhados nas Tabelas 12 a 14.

Tabela 12. Métrica GLM (*lambda* = 0,1)

lambda=0.1	A	B	C	D	E
Acurácia	0.86	0.43	0.64	0.81	0.60
Sensib.	0.15	0.92	0.77	0.48	0.81
Especif.	0.96	0.36	0.62	0.86	0.57
F-Score	0.21	0.29	0.36	0.39	0.34
AUC-ROC	0.56	0.64	**0.70**	0.67	0.69

Tabela 13. Métrica GLM (*lambda* = 0,5)

lambda=0.5	A	B	C	D	E
Acurácia	0.87	0.42	0.70	0.83	0.60
Sensib.	0.04	0.92	0.72	0.41	0.84
Especif.	0.99	0.35	0.70	0.89	0.56
F-Score	0.10	0.29	0.39	0.39	0.35
AUC-ROC	0.52	0.63	**0.71**	0.65	0.70

Tabela 14. Métrica GLM (*lambda* = 1)

lambda=1	A	B	C	D	E
Acurácia	0.87	0.43	0.77	0.85	0.61
Sensib.	0.00	0.92	0.64	0.37	0.83
Especif.	1.00	0.36	0.79	0.92	0.58
F-Score	0.04	0.30	0.42	0.38	0.36
AUC-ROC	0.50	0.64	**0.72**	0.64	0.71

As estratégias de subamostragem parcial (C) e a combinação com superamostragem (E) apresentaram os melhores desempenhos. A técnica C alcançou AUC-ROC de 0,72 com GLM (*lambda* = 1), enquanto a técnica E obteve 0,68 com RF (*mtry* = 126), indicando maior capacidade de discriminação dos casos positivos. A linha de base (A) obteve os piores resultados em sensibilidade e AUC, enquanto B e D apresentaram boa sensibilidade, mas baixa especificidade, resultando em muitos falsos positivos.

Em relação a referências comparativas do próprio eRisk 2017, o modelo de [Trotzek et al. 2017] com regressão logística ponderada alcançou F1 de 0,64, enquanto [Errecalde et al. 2017] obteve F1 de 0,59 utilizando técnicas como TVT e Random Forest. Embora nosso melhor F1 (0,42) seja inferior aos valores reportados na literatura, destaca-se que o modelo GLM com subamostragem parcial atingiu AUC-ROC de 0,72, indicando uma boa capacidade discriminativa entre classes, métrica que não foi explicitamente relatada nos trabalhos analisados.

5. Conclusão

Os resultados demonstraram variações relevantes no desempenho dos modelos conforme a estratégia de reamostragem aplicada. O objetivo central deste estudo foi avaliar o impacto de técnicas de balanceamento na predição de sinais de depressão, com ênfase nas métricas sensibilidade, especificidade e AUC-ROC, essenciais em cenários com dados desbalanceados e alta sensibilidade social.

Contudo, reconhece-se a ausência de testes estatísticos formais que validem as diferenças entre os métodos avaliados, aspecto que será abordado em estudos futuros por meio de técnicas como ANOVA ou Wilcoxon. Também se propõe a incorporação de *embeddings* semânticos mais avançados e uma comparação sistemática com referências comparativas da literatura. Por fim, destaca-se a importância de considerar princípios éticos no desenvolvimento de soluções automatizadas aplicadas à saúde mental, especialmente em relação à privacidade, consentimento e possíveis estigmatizações em ambientes sensíveis como redes sociais.

Referências

Abdulsadig, R. S. and Rodriguez-Villegas, E. (2024). A comparative study in class imbalance mitigation when working with physiological signals. *Front. Digit. Health*, pages 1–11.

Almeida, H., Briand, A., and Meurs, M.-J. (2017). Detecting Early Risk of Depression from Social Media User-generated Content . pages 1–12.

Araf, I., Idri1, A., and Chairi3, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. *Artifcial Intelligence Review*, 57(80):1–72.

de Oliveira Melo, W. E. and Cortes, O. A. C. (2021). Utilizando Análise de Sentimentos e SVM na Classificação de Tweets Depressivos. *XI Computer on the Beach*.

dos Santos, H. G. (2018). *Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina.* PhD thesis, Universidade de São Paulo (USP), São Paulo.

ERISK (2017). eRisk 2017: Early risk prediction on the Internet: experimental foundations. https://erisk.irlab.org/2017/index.html.

Errecalde, M. L., Villegas, P., Funez, D. G., Uelay, J. G., and Cagnina, L. C. (2017). Temporal Variation of Terms as concept space for early risk predi tion. pages 1–12.

Farías-Anzaldúa, A. A., y Gómez, M. M., López-Monroy, A. P., and González-Gurrola, L. C. (2017). UACH-INAOE participation at eRisk2017. pages 1–8.

Gentili, E., Franchini, G., Zese, R., Alberti, M., Ferrara, M., Domenicano, I., and Grassi, L. (2024). Machine learning from real data: A mental health registry case study. *Computer Methods and Programs in Biomedicine Update*, 5(100132):1–10.

Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., and Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. *Health Information Science and Systems*, 6(8):1–12.

Junior, E. S. S., de Melo, J. A. B., da Silva, A. P., de A. Silva, T., de C. Chaves, A. P., de Souza, A. F., de S. G. júnior, J., and do N. Santana, S. (2022). Depression among adolescents who frequently use social networks: a literature review. 8(3):18838–18851.

Kim, J., Lee, D., and Park, E. (2021). Machine Learning for Mental Health in Social Media: Bibliometric Study. *Journal of Medical Internet Research*, 23(3):1–17.

Lin, W.-J. and Chen, J. J. (2012). Class-imbalanced classifiers for high-dimensional data. *Oxford University Press*, 14(1):13–26.

Losada, D. E., Crestani, F., and Parapar, J. (2017). CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. pages 1–18.

Malam, I. A., Arziki, M., Bellazrak, M. N., Benamara, F., Kaidi, A. E., Es-Saghir, B., He, Z., Housni, M., Moriceau, V., Mothe, J., and Ramiandrisoa, F. (2017). IRIT at e-Risk. pages 1–7.

Mena, L. J., Orozco, E. E., Felix, V. G., Ostos, R., Melgarejo, J., and Maestre, G. E. (2012). Machine learning approach to extract diagnostic and prognostic thresholds:

application in prognosis of cardiovascular mortality. *Computational and Mathematical Methods in Medicine*, pages 1–6.

Morais, E. A. M. and Ambrósio, A. P. L. (2007). Mineração de Textos. https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF$_0$05 − 07.*pdf*.

Mussio, R. A. P. (2019). A geração Z e suas respostas comportamental e emotiva nas redes sociais virtuais. *European psychiatry*, 3(3):204–217.

Nadaraja, R. and Yazdanifard, R. (2014). Social Media Marketing: Advantages and Disadvantages. *Social Media Marketing*, pages 1–10.

OPAS (2017). Com depressão no topo da lista de causas de problemas de saúde, OMS lança a campanha "Vamos conversar". https://www.paho.org/pt/noticias/30-3-2017-com-depressao-no-topo-da-lista-causas-problemas-saude-oms-lanca-campanha-vamos.

Ostojic, D., Lalousis, P. A., Donohoe, G., and Morris, D. W. (2024). The challenges of using machine learning models in psychiatric research and clinical practice. *European Neuropsychopharmacology*, 88:53–65.

Rahman, R. A., Omar, K., Noah, S. A. M., and Danuri, M. S. N. M. (2018). A Survey on Mental Health Detection in Online Social Network. *International Journal on Advanced Science Engineering Information Technology*, 8(4-2):1431–1436.

Sadeque, F., Xu, D., and Bethard, S. (2017). UArizona at the CLEF eRisk 2017 Pilot Task: Linear and Recurrent Models for Early Depression Detection . pages 1–9.

Sampasa-Kanyinga, H. and Hamilton, H. (2015). Social networking sites and mental health problems in adolescents: The mediating role of cyberbullying victimization. *European psychiatry*, 30(8):1021–1027.

Souza, K. and da Cunha, M. X. C. (2019). Impacts of the use of virtual social networks on adolescents' mental health: A systematic review of literature. *Revista Educação, Psicologia e Interfaces*, 3(3):204–217.

Trotzek, M., Koitka, S., and Friedrich, C. M. (2017). Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression - FHDO Biomedical Computer Science Group (BCSG) . pages 1–17.

Villatoro-Tello, E., de-la Rosa, G. R., and Jiménez-Salazar, H. (2017). UAM's participation at CLEF eRisk 2017 task: Towards modelling depressed bloggers. pages 1–9.

Visentini, C., Cassidy, M., Bird, V. J., and Priebe, S. (2018). Social networks of patients with chronic depression: A systematic review . *Journal of Affective Disorders*, 241:571–578.

WHO (2023). Depressive disorder (depression). https://www.who.int/news-room/fact-sheets/detail/depression.

Zhang, D., Yin, C., Zeng, J., Yuan, X., and Zhang, P. (2020). Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Medical Informatics and Decision Making*, 20(280):1–11.