Avaliação de Estratégias de Reamostragem na Predição de Desfechos em Dados de Saúde Ocupacional

Daniel Bortot de Salles[1], **Fernanda Sumika Hojo de Souza**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de Ouro Preto (UFOP)
Ouro Preto – MG – Brazil

daniel.bortot@aluno.ufop.edu.br, fsumika@ufop.edu.br

Abstract. *Research on human psychological conditions has grown rapidly lately. Mental disorders are multifactorial and have no single explanation. To contribute to this topic, this study aims to classify, using supervised machine learning techniques, the outcomes of patients with work-related mental disorders. The dataset was obtained from SUS/SINAN. Random Forest, XGBoost, and several resampling techniques were applied to predict patient outcomes. The results show that severe class imbalance affects model performance, and the three best distinct models achieved an F1-Score of 0.472, a balanced accuracy of 0.77, and an ROC-AUC of 0.858. This study concludes that the models must be aligned with clinical priorities when defining optimization criteria.*

Resumo. *O estudo sobre as condições psicológicas do ser humano vem crescendo rapidamente nos últimos anos. Os transtornos mentais são multifatoriais e não possuem uma explicação única. Para contribuir com esse tema, este trabalho busca classificar, por meio de técnicas de aprendizado de máquina supervisionado, a evolução dos pacientes com transtornos mentais relacionados ao trabalho. A base de dados utilizada foi obtida a partir do SUS/SINAN. Foram utilizados o Random Forest, o XGBoost e diversas técnicas de reamostragem para predizer a evolução do paciente. Os resultados mostram que o alto desbalanceamento impacta o desempenho dos modelos, e os três melhores modelos distintos obtidos tiveram 0,472 de F1-Score, 0,77 de acurácia balanceada e 0,858 de ROC-AUC. Conclui-se que os modelos devem estar alinhados às prioridades clínicas na definição do critério de otimização.*

1. Introdução

Segundo levantamento recente do *International Labour Organization* (ILO), quase 60% da população mundial encontra-se inserida ativamente no mercado de trabalho [ILO 2022]. Por outro lado, dados do *World Health Organization* (WHO) estimam que 15% dos trabalhadores adultos apresentavam transtornos mentais em 2019 [WHO 2024]. No Brasil, dados de 2024 mostram que o país registrou mais de 470 mil afastamentos do trabalho por transtornos mentais [G1 2025], sendo o maior número alcançado em dez anos. Trabalhos da área da saúde, humanitária e emergencial geralmente carregam um risco elevado de exposição a eventos adversos, impactando na saúde mental dos trabalhadores [Brasil. Ministério da Saúde and Fundação Oswaldo Cruz 2024].

Os transtornos mentais são multifatoriais e não há uma explicação única [Organização Pan-Americana da Saúde (OPAS) 2025]. Algumas causas do aumento

dos problemas de saúde mental no trabalho são tarefas excessivas, horas inflexíveis, locais de trabalho em situação precária, cultura corporativa prejudicial, instabilidade empregatícia e o conflito de demanda entre as tarefas domésticas e profissionais. No Brasil, a alta dos casos pode-se alinhar ao luto pós-pandemia [Organização Pan-Americana da Saúde (OPAS) 2022], anos de isolamento e insegurança financeira [Campêlo 2023] com o aumento do custo de vida.

Existem ações efetivas para prevenir, proteger e promover a estabilidade mental no trabalho, bem como dar suporte àqueles que sofrem dessas condições. Alguns programas adotados pelas empresas incluem: treinamento psicológico, redução do estigma em relação à saúde mental, acomodações confortáveis, apoio a iniciativas dos colaboradores e a construção de um ambiente comunicativo, ativo e aberto [WHO 2024].

No Brasil, se diagnosticado com um transtorno mental, o paciente pode atestar a necessidade de se ausentar do trabalho ou realizar um tratamento [Instituto Nacional do Seguro Social - INSS 2025]. Se o período do afastamento ultrapassar 15 dias, na condição de segurado do Instituto Nacional do Seguro Social (INSS) e com pelo menos 12 meses de contribuição, o trabalhador pode solicitar o benefício por incapacidade temporária. Caso o motivo do afastamento seja decorrente do próprio trabalho, como em casos de Síndrome de Burnout[1], o trabalhador pode ser beneficiado por incapacidade temporária acidentária, que não exige contribuição previdenciária. Esse benefício garante que, após seu retorno às atividades, o trabalhador não seja demitido sem justa causa pelos próximos 12 meses.

O Ministério da Saúde incluiu novas doenças e agravos relacionados ao trabalho (DART) na lista nacional de notificação compulsória de doenças, agravos e eventos de saúde pública [Ministério da Saúde 2025]. Isso significa que essas enfermidades deverão ser notificadas obrigatoriamente por profissionais de saúde na Rede de Atenção à Saúde (RAS) em suas atividades profissionais. O objetivo é ampliar a vigilância das doenças relacionadas à saúde do trabalhador, por serem doenças evitáveis e passíveis de prevenção.

O aprendizado de máquina (AM) vem sendo utilizado em diversas áreas, como saúde [Morsoleto et al. 2025b], ações judiciais [Gomes et al. 2019], entre outras, a fim de encontrar padrões implícitos nos dados disponíveis atualmente. No contexto dos transtornos mentais associados ao trabalho, o uso do aprendizado de máquina supervisionado pode contribuir para a predição e prevenção de possíveis dificuldades de pacientes diagnosticados e para a maior compreensão das variáveis. Assim, tornam-se disponíveis ferramentas capazes de auxiliar profissionais da saúde, fornecendo indicadores de risco e apoio no acompanhamento da evolução dos casos, contribuindo para um tratamento mais ágil, direcionado e eficaz.

Este estudo busca avaliar modelos de aprendizado supervisionado capazes de prever se um paciente notificado com transtorno mental relacionado ao trabalho terá como evolução a cura ou a incapacidade temporária, isto é, a necessidade de afastamento laboral por tempo indeterminado. Para isso, foram utilizados algoritmos clássicos na literatura, além de técnicas voltadas ao desbalanceamento dos dados, tendo os resultados comparados por meio de diversas métricas de desempenho. Apesar de atuar sobre indivíduos já

[1] https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/s/sindrome-de-burnout

diagnosticados, a identificação de padrões pode contribuir indiretamente para estratégias preventivas, ao permitir a detecção precoce de perfis de risco em populações semelhantes.

A principal contribuição deste trabalho não consiste na proposição de um novo algoritmo, mas na análise do comportamento de modelos clássicos quando submetidos a diferentes estratégias de reamostragem e critérios de otimização em uma base nacional desbalanceada. A escolha da métrica a ser otimizada impacta a seleção do modelo, afetando decisões em saúde ocupacional. Embora modelos computacionais não substituam uma avaliação clínica individual, podem auxiliar na identificação de padrões associados a desfechos mais graves, contribuindo para a priorização de pacientes.

2. Trabalhos Relacionados

Diversos trabalhos da literatura combinam o uso de aprendizado de máquina e bases de dados públicas da área da saúde com o objetivo de gerar modelos preditivos para auxiliar a predição e prevenção de desfechos indesejáveis. A seguir, alguns trabalhos que focaram em bases de dados do DATASUS[2] (Departamento de Informação e Informática do SUS) são apresentados.

O estudo descrito em [Jesus et al. 2020] utiliza dados do SUS para desenvolver modelos de aprendizado de máquina para predizer a mortalidade de gêmeos até um ano nascidos no Brasil. Diversas abordagens de aprendizado de máquina são avaliadas, incluindo técnicas de balanceamento de dados e ajustes nos modelos para melhorar a detecção da classe minoritária que representa 3,91% das instâncias. O melhor modelo encontrado foi o XGBoost, com um F1-Score de 57,30% e ROC-AUC de 95%. Os autores observam que métodos de reamostragem podem melhorar significativamente métricas como recall e F1-Score, porém frequentemente à custa do aumento de falsos positivos.

O trabalho apresentado em [Moreira et al. 2021] investiga a aplicação de técnicas de aprendizado de máquina em dados de saúde provenientes do SUS, abordando um problema de classificação no contexto da predição de diabetes tipo 1 na gestação com 1,63% de instâncias na classe minoritária. O estudo compara diferentes algoritmos, incluindo métodos baseados em árvores e modelos ensemble, avaliando seu desempenho por meio da métrica F1-Score. Os resultados indicam que a proposta pode ser um recurso relevante, apresentando sensibilidade e precisão superiores a 90%.

O estudo descrito em [Rodrigues et al. 2024] avalia algoritmos de aprendizado de máquina, utilizando dados de registros nacionais, para prever a perda de seguimento durante o tratamento da tuberculose. A base de dados inclui todos os casos de tuberculose notificados ao SINAN entre 2015 e 2022, excluindo menores de 18 anos, grupos vulneráveis e casos de tuberculose resistente a medicamentos. As amostras apresentam um desbalanceamento de 33,33% para a classe minoritária. Os modelos de predição apresentaram uma área sob a curva entre 0,71 e 0,72. A técnica *Light Gradient Boosting* apresentou o melhor desempenho preditivo, equilibrando especificidade e sensibilidade.

O trabalho apresentado em [Morsoleto et al. 2025a] utilizou modelos de aprendizado de máquina para predizer a mortalidade infantil, ciente de desbalanceamento. Dessa forma, múltiplas técnicas de amostragem foram utilizadas: subamostragem, sobreamostragem e métodos híbridos. Os resultados apontam que o *Random Under-Sampling*

[2] `https://datasus.saude.gov.br/`

apresenta o melhor *recall*, crítico para identificar a classe positiva, enquanto o *Random Over-Sampling* apresenta a melhor precisão, minimizando falsos positivos. Segundo os resultados, *Edited Nearest Neighbours* apresentou o melhor balanceamento entre *recall* e precisão, com valores próximos de 0,44.

O estudo descrito em [Barros et al. 2025] busca classificar, por meio do aprendizado de máquina, os resultados favoráveis ou desfavoráveis do tratamento da tuberculose utilizando dados clínicos e sociodemográficos. Os dados apresentavam desbalanceamento de 28,08% para a classe minoritária, portanto, os modelos foram treinados em conjuntos de dados pré-processados com técnicas de balanceamento (subamostragem, sobreamostragem e SMOTE) e avaliados usando métricas como F1-Macro e AUC-ROC. Os melhores resultados foram obtidos no cenário que combinou um intervalo temporal mais amplo com atributos derivados adicionais, no qual o modelo Random Forest alcançou uma acurácia de 85,7%, F1-Macro de 85,7% e MCC de 71,6%.

3. Metodologia

Para a realização deste trabalho, extraiu-se do Sistema de Informação de Agravos de Notificação (SINAN) dados sobre transtornos mentais relacionados ao trabalho agrupados na base de dados DRT Transtorno Mental[3]. Nela é registrado todo caso de sofrimento emocional em suas diversas formas de manifestação, tais como: choro fácil, tristeza, medo excessivo, doenças psicossomáticas, agitação, irritação, nervosismo, ansiedade, taquicardia, sudorese, insegurança, entre outros sintomas que podem indicar o desenvolvimento ou agravamento de transtornos mentais utilizando os CID10[4](Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde). Todo o processo metodológico é representado pelo fluxograma da Figura 1.

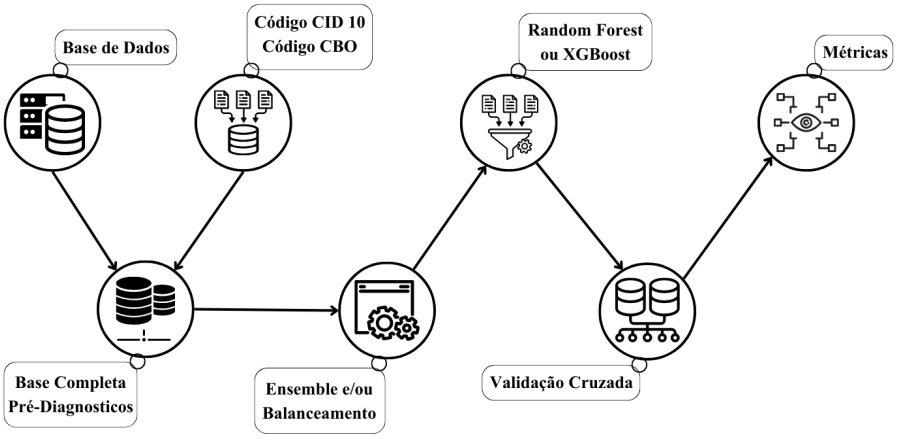

Figura 1. Fluxograma da metodologia experimental, incluindo pré-processamento dos dados, treinamento dos modelos e avaliação.

Este estudo engloba os anos de 2015 a 2025, representando 11 anos de abrangência. Foram incluídos pacientes com idade entre 18 e 60 anos, visando a população em faixa etária ativa no mercado de trabalho. Inicialmente, a base possui 25.287 instâncias e 62 atributos, que passaram por um tratamento de dados para obter um

[3]https://portalsinan.saude.gov.br/drt-transtorno-mental
[4]https://cid10.com.br/

melhor estudo com os algoritmos. Para isso, a base de dados foi dividida em duas partes: variáveis pré-diagnóstico e pós-diagnóstico do paciente. Dessa forma, é possível utilizar os algoritmos para classificar os dados pré-diagnóstico e comparar com os modelos com os atributos pós-diagnóstico.

Atributos pré-diagnóstico incluíram os seguintes dados dos pacientes: sexo, gestante, raça, idade, escolaridade, situação de trabalho, terceirizado, código CBO, tempo de ocupação, tempo de exposição, hábitos (álcool, tabaco, drogas, psicofármacos), região do país e diagnóstico CID-10. Analisando os tipos dos atributos, verifica-se a presença de dados mistos, sendo 4 contínuos e 12 categóricos. Os atributos contínuos são: "Idade", "Escolaridade", "Tempo de Ocupação" e "Tempo de Exposição". Entre os atributos categóricos restantes, "Sexo" é binário assimétrico, "Gestante" e "Terceirizado" são binários simétricos, e os demais são categóricos nominais. "Sit trabalho" tem diversas categorias, por fins de simplificação, foi dividido em: emprego formal, outros/ignorado, desempregado/aposentado e trabalho informal. Esses rótulos foram obtidos agrupando os diversos subgrupos presentes nesse atributo.

Antes da aplicação dos modelos, todos os atributos categóricos nominais foram transformados por meio da técnica de *One-Hot Encoding*, de modo a evitar que os algoritmos interpretassem indevidamente qualquer relação ordinal entre suas categorias. Os dados sobre o diagnóstico específico (CID10) e a classificação brasileira de ocupações (CBO) do indivíduo que sofreu o agravo estavam no formato de sigla. Foi utilizada a integração do código CID10 ao agrupamento obtido pelo DataSUS[5] e com o código CID10 da tabela dos grandes grupos disponibilizada pelo Ministério do Trabalho e Emprego[6].

Atributos pós-diagnóstico incluíram informações temporais (ano, mês, semana, dia), regime de tratamento e condutas adotadas. Analisando os tipos de atributos, observa-se que os dados são mistos: quatro contínuos ("Ano", "Semana Not", "Mês" e "Qtd Condutas") e dez categóricos. Dentre os categóricos, "Dia da Semana" corresponde aos sete dias da semana, "Regime" pode assumir os valores Ambulatorial, Hospitalar ou conter dados ausentes, e "Individual", "Mudança", "Nenhum", "Coletiva", "Afast Desgaste", "Afast Trab", "Conduta", "CAPS" e "CAT" são variáveis binárias assimétricas. Para aplicação nos modelos de aprendizado de máquina, somente os atributos "Dia da Semana" e "Regime" foram submetidos à técnica de *One-Hot Encoding*.

Por fim, tem-se a variável "Evolução", que representa o atributo alvo do estudo. Para classificar a evolução do paciente, foi necessário filtrar todas as possibilidades existentes: "cura", "cura não confirmada", "incapacidade temporária", "incapacidade permanente parcial", "incapacidade permanente total" e "óbito por doença relacionada ao trabalho". Neste trabalho, optou-se por utilizar as classes "cura" e "incapacidade temporária", as quais são as mais frequentes na base de dados. É importante ressaltar que as variáveis de condutas adotadas não possuem relação direta com o desfecho final, ou seja, em muitos casos há conduta e ainda assim a evolução é incapacidade.

A base de dados disponibilizada no portal DATASUS é gerada a partir de um formulário do SUS preenchido pelo próprio paciente ou por um funcionário de saúde.

[5] `http://www2.datasus.gov.br/cid10/V2008/descrcsv.htm`
[6] `https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/cbo`

Vale ressaltar que a maioria das questões do formulário possui a opção de ser ignorada. Dessa forma, considera-se essa possibilidade equivalente ao valor nulo. Entretanto, como em alguns casos o valor nulo pode ter significado relevante, como na variável referente ao uso de drogas, optou-se por tratar todas as entradas nulas como uma categoria distinta.

Uma das principais características deste conjunto de dados é o desbalanceamento entre as classes. A "incapacidade temporária" possui aproximadamente 16 vezes mais amostras do que a classe "cura". Para lidar com esse problema, são utilizadas técnicas de sobreamostragem, subamostragem, híbridas e de *ensemble*. Os métodos de reamostragem utilizados pertencem à biblioteca Imbalanced-learn[7]. Os algoritmos de subamostragem empregados foram *Random Under-Sampling* e *Edited Nearest Neighbours*. Os métodos de sobreamostragem aplicados foram *Random Over-Sampling*, SMOTE e ADASYN. Por fim, os métodos híbridos, que combinam subamostragem e sobreamostragem, foram *SMOTE-ENN* e *SMOTE-Tomek Links*.

É importante ressaltar que este estudo apresenta uma característica incomum se tratando de base de dados de saúde desbalanceada: a classe minoritária é a "cura", ou seja, menos relevante do ponto de vista clínico. Dessa forma, um modelo que apresentasse os melhores resultados, nesse contexto, seria aquele que simplesmente previsse todos os casos como "incapacidade temporária". Por esse motivo, optou-se por inverter a importância das classes, buscando um modelo mais equilibrado e capaz de representar ambas as classes imparcialmente.

Devido ao desbalanceamento das classes, é proposta uma estratégia de *ensemble*, ilustrada pela Figura 2, dividindo as amostras da classe majoritária até equilibrar a quantidade de amostras da classe minoritária, nesse caso, 16 vezes. Dessa forma, tem-se 16 bases de dados com classes equilibradas, nas quais se pode aplicar a técnica de *ensemble* com 16 modelos. Além disso, considera-se a possibilidade de dividir a classe majoritária em menor número de partes, como 8 e 4 vezes, resultando em divisões menos desequilibradas dos dados para posterior aplicação das técnicas de balanceamento.

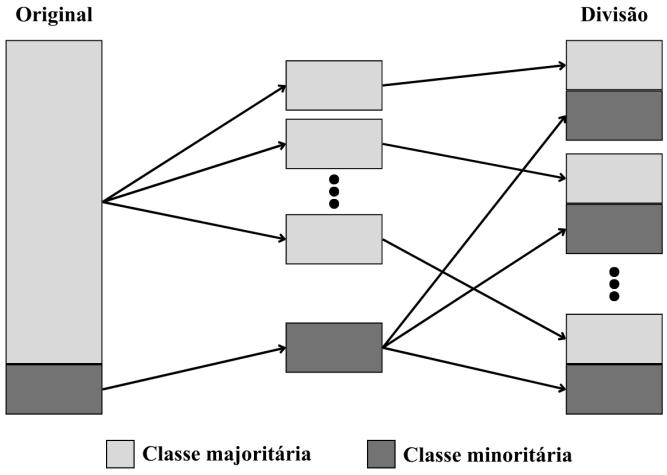

Figura 2. Exemplo ilustrativo do funcionamento de um modelo em ensemble que constrói classificadores balanceando amostras da classe alvo.

[7] https://imbalanced-learn.org/stable/

Os algoritmos utilizados foram a Floresta Aleatória[8] e o XGBoost[9]. Ambos empregam múltiplas árvores de decisão em sua construção, porém cada um apresenta características próprias. Esses modelos foram escolhidos por serem técnicas amplamente utilizadas na literatura, facilitando a comparação com outros estudos, além de oferecerem maior interpretabilidade quando comparados a arquiteturas profundas. Resultados preliminares com outros algoritmos também se mostraram inferiores aos propostos.

As métricas analisadas foram: Acurácia, *Recall*, Precisão, F1-Score, ROC-AUC e Acurácia balanceada. Dessa forma, têm-se múltiplas formas de avaliar os modelos com métricas da literatura e relacionadas a bases desbalanceadas. Outra variação implementada nos modelos consiste em aplicar um limiar na probabilidade das classes na predição da base de teste. Para isso, será utilizada a curva de Precisão-*Recall* para definir o limiar que maximize o F1-Score. O modelo resultante será novamente avaliado em todas as métricas, a fim de verificar se houve uma melhora significativa nos demais indicadores de desempenho.

Para realizar os testes, foi realizada uma validação cruzada de 10 partições. Dessa forma, têm-se 10 modelos com 90% para treino e 10% para teste, com estratificação, a fim de garantir que a base de teste contenha uma quantidade satisfatória de amostras da classe minoritária. O procedimento descrito na Figura 2 é aplicado somente aos dados de treinamento da partição em questão. Vale ressaltar que, ao ter a divisão do *ensemble* por 16 vezes, não são aplicadas as técnicas de balanceamento, pois as bases já se encontram balanceadas pela própria divisão, o que não ocorre nas divisões em 8 e 4 partes. A partir desses testes, é analisada a média dos resultados da validação cruzada em todas as métricas descritas e serão apresentados os modelos que se destacarem. Para fins de replicabilidade, tanto o treinamento quanto os testes foram realizados com a semente de aleatoriedade fixada em 42.

4. Resultados

De acordo com a análise descritiva dos dados, a quantidade total de transtornos mentais relacionados ao trabalho notificados no período investigado foi de 14.031 casos, sendo o pico de notificações no ano de 2024 (2.804), com faixa etária predominante de 35 a 49 anos, maioria do sexo feminino em todo Brasil, sendo a cor/raça branca a mais afetada em quase todas as regiões brasileiras, nível de escolaridade predominantemente de nível superior completo e o agente comunitário de saúde mostrou-se a ocupação mais afetada. A utilização de drogas psicoativas foi ausente na maior parte dos registros, evolução dos casos majoritariamente com incapacidade temporária e a emissão de CAT não ocorreu na maioria dos casos.

Os resultados da avaliação experimental realizada são apresentados na Tabela 1, com ênfase na métrica F1-Score. Como se pode observar, o melhor modelo foi a *Random Forest*, apresentando um F1-Score de aproximadamente 0,467, utilizando todos os atributos da base de dados, sem aplicar o método de *ensemble*, e sem recorrer ao tratamento do desbalanceamento de amostras, mas ajustando o limiar a fim de maximizar essa métrica.

Vale ressaltar que, apesar de apresentar o maior valor de F1-Score, esse modelo

[8] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

[9] `https://xgboost.readthedocs.io/`

obteve um desempenho relativamente mais baixo na métrica de acurácia balanceada, com valor aproximado de 0,69. Mantendo o F1-Score como referência, observa-se que a utilização de todos os atributos se mostrou importante, e que a alteração do limiar teve impacto significativo nos resultados.

Tabela 1. Os 10 melhores modelos na métrica F1-Score.

Modelo	Base	Ensamble	Imblearn	Limiar	Acurácia	Recall	Precisão	**F1-Score**	ROC-AUC	BalancedAcc
RF	Todas	1	Default	Sim	0.943257	0.420732	0.538409	**0.467885**	0.847067	0.698323
RF	Todas	4	Default	Sim	0.946341	0.392683	0.581896	**0.463453**	0.858248	0.686814
RF	Todas	4	RandomOverSampler	Sim	0.940459	0.425610	0.506303	**0.456790**	0.857617	0.699123
RF	Todas	1	RandomOverSampler	Sim	0.945696	0.385366	0.585852	**0.455829**	0.851748	0.683041
RF	Todas	1	EditedNearestNeighbours	Sim	0.944548	0.390244	0.549479	**0.453128**	0.848315	0.684718
XGB	Todas	4	Default	Sim	0.941894	0.406098	0.517720	**0.451605**	0.851538	0.690739
RF	Todas	8	Default	Sim	0.944620	0.382927	0.586397	**0.449630**	0.857844	0.681326
RF	Todas	8	RandomOverSampler	Sim	0.943472	0.393902	0.541919	**0.449397**	0.858571	0.685861
XGB	Todas	8	Default	Sim	0.941966	0.402439	0.517000	**0.449208**	0.852599	0.689063
XGB	Todas	4	RandomOverSampler	Sim	0.942826	0.396341	0.526788	**0.449128**	0.851317	0.686662

Focando-se na métrica de acurácia balanceada, os resultados são dos dez melhores modelos apresentados na Tabela 2. Observa-se que o melhor resultado difere daquele obtido pela métrica de F1-Score, sendo o *Random Forest* o melhor modelo, utilizando toda a base de dados, o método de *ensemble* com 16 divisões e sem alterações no limiar de decisão. É interessante destacar que esse modelo apresentou um valor de F1-Score baixo (aproximadamente 0,28), principalmente devido à baixa precisão (aproximadamente 0,17), o inverso do modelo anterior, que possuía um equilíbrio dessas métricas.

Dessa forma, conclui-se que o modelo com melhor acurácia balanceada consegue predizer de forma mais eficiente a classe positiva e minoritária. Mantendo a acurácia balanceada como referência, observa-se que a utilização de todos os atributos continuou sendo importante, e que a alteração do limiar não teve impacto significativo nos resultados, reforçando a assimetria entre a acurácia balanceada e o F1-Score neste estudo.

Tabela 2. Os 10 melhores modelos na métrica de acurácia balanceada.

Modelo	Base	Ensamble	Imblearn	Limiar	Acurácia	Recall	Precisão	F1-Score	ROC-AUC	**BalancedAcc**
RF	Todas	16	Default	Nao	0.768436	0.771951	0.172935	0.282462	0.853942	**0.770084**
RF	Todas	8	RandomUnderSampler	Nao	0.768508	0.769512	0.172525	0.281778	0.853915	**0.768979**
XGB	Todas	16	Default	Nao	0.767432	0.769512	0.171506	0.280443	0.854673	**0.768407**
XGB	Todas	4	RandomUnderSampler	Nao	0.763486	0.771951	0.169289	0.277637	0.847650	**0.767454**
XGB	Todas	8	RandomUnderSampler	Nao	0.764778	0.762195	0.168843	0.276391	0.851677	**0.763567**
RF	Todas	4	RandomUnderSampler	Nao	0.768077	0.753659	0.169731	0.276995	0.851740	**0.761319**
XGB	Todas	8	RandomOverSampler	Nao	0.857747	0.637805	0.237825	0.346276	0.853856	**0.754649**
XGB	Todas	8	SMOTE	Nao	0.869082	0.620732	0.253155	0.359374	0.853356	**0.752668**
XGB	Todas	8	ADASYN	Nao	0.872310	0.614634	0.257242	0.362491	0.852322	**0.751524**
XGB	Todas	4	SMOTEENN	Nao	0.810258	0.681707	0.190876	0.298050	0.832637	**0.750000**

Considerando a área sob a curva ROC, os resultados são apresentados na Tabela 3. O melhor modelo foi o *Random Forest*, utilizando todos os atributos disponíveis da base de dados, com um *ensemble* de 8 divisões, aplicando a técnica de reamostragem *RandomOverSampler* e sem necessidade de ajuste no limiar de decisão, uma vez que a métrica ROC e sua área não são afetadas por essa alteração.

De forma geral, ao comparar as três métricas: F1-Score, acurácia balanceada e ROC-AUC, observa-se que diferentes modelos se destacam em cada uma delas, evidenciando comportamentos complementares. Essas diferenças reforçam a importância de avaliar múltiplas métricas em bases desbalanceadas, pois cada uma oferece uma perspectiva distinta sobre o desempenho do modelo.

Tabela 3. Os 10 melhores modelos na métrica área sob a curva ROC.

Modelo	Base	Ensamble	Imblearn	Limiar	Acurácia	Recall	Precisão	F1-Score	**ROC-AUC**	BalancedAcc
RF	Todas	8	RandomOverSampler	Nao	0.875251	0.596341	0.259092	0.360874	**0.858571**	0.744512
RF	Todas	8	RandomOverSampler	Sim	0.943472	0.393902	0.541919	0.449397	**0.858571**	0.685861
RF	Todas	4	Default	Nao	0.948565	0.324390	0.624684	0.424745	**0.858248**	0.655983
RF	Todas	4	Default	Sim	0.946341	0.392683	0.581896	0.463453	**0.858248**	0.686814
RF	Todas	8	Default	Nao	0.905811	0.530488	0.321212	0.399457	**0.857844**	0.729878
RF	Todas	8	Default	Sim	0.944620	0.382927	0.586397	0.449630	**0.857844**	0.681326
RF	Todas	4	RandomOverSampler	Nao	0.933716	0.425610	0.435398	0.429646	**0.857617**	0.695541
RF	Todas	4	RandomOverSampler	Sim	0.940459	0.425610	0.506303	0.456790	**0.857617**	0.699123
XGB	Todas	16	Default	Sim	0.937948	0.412195	0.502346	0.442583	**0.854673**	0.691502
XGB	Todas	16	Default	Nao	0.767432	0.769512	0.171506	0.280443	**0.854673**	0.768407

Foi realizada a otimização de hiperparâmetros dos três melhores modelos das métricas F1-Score, Acurácia balanceada e AUC-ROC, apresentada pela Tabela 4. Observa-se uma pequena melhoria das métricas: F1-Score de 0,4678 para 0,4728, acurácia balanceada de 0,7700 para 0,7704 e ROC-AUC de 0,8585 para 0,8589.

Tabela 4. Os 3 melhores modelos após otimização de hiperparâmetros a fim de maximizar suas melhores métricas, respectivamente.

n_estimators	max_depth	min_samples_split	min_samples_leaf	max_features	Accuracy	Recall	Precision	F1-Score	ROC-AUC	BalancedAcc
150	30	3	1	'sqrt'	0.945624	0.415854	0.560618	0.472853	0.850530	0.697294
150	None	2	1	'log2'	0.774534	0.765854	0.176206	0.286377	0.855184	0.770465
150	None	2	1	'sqrt'	0.874319	0.593902	0.257075	0.358407	0.858998	0.742873

A análise de importância por permutação indica a contribuição relativa de cada atributo para o desempenho preditivo dos modelos, não implicando relação causal com o desfecho observado. De acordo com a Figura 3, os três modelos apresentaram uma ordem diferente. Observa-se que o modelo de melhor F1-Score formado pela ausência de *ensemble* e técnicas de balanceamento de amostra tem como principal atributo a conduta de afastamento do trabalho, seguido da idade, se foi encaminhado a um serviço especializado em transtornos mentais (CAPS), tempo de exposição e a semana do ano da notificação.

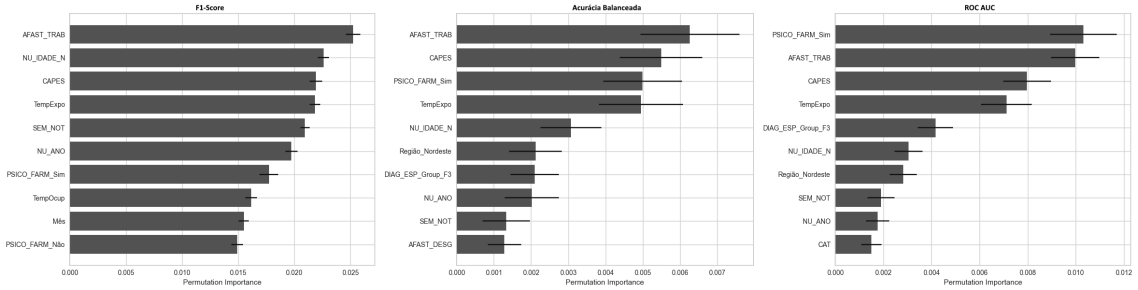

Figura 3. Os dez atributos com maior importância por permutação dos três melhores modelos das métricas F1-Score, Acurácia balanceada e ROC-AUC, respectivamente.

Muito similar ao modelo de melhor F1-Score, o modelo de melhor acurácia balanceada tem como diferença a presença de uso de psicofármacos. Já o modelo de melhor área sob a curva ROC tem como diferença a presença do diagnóstico específico F3, que representa episódio maníaco, transtorno afetivo bipolar, episódios depressivos e transtornos do humor.

Observa-se que variáveis relacionadas à conduta assistencial, como afastamento do trabalho (AFAST_TRAB) e uso de psicofármacos (PSICO_FARM_Sim), estão entre as mais relevantes nos três modelos avaliados. Esses atributos podem refletir a gravidade percebida do caso no momento da notificação. Variáveis sociodemográficas, como idade (NU_IDADE_N) e região geográfica, podem refletir diferenças estruturais no perfil populacional, no acesso aos serviços ou na dinâmica de registro do sistema. Da mesma forma, atributos relacionados ao tempo de exposição ou tempo na ocupação podem estar associados à cronificação do quadro. Esses resultados sugerem que os modelos foram capazes de capturar tanto características individuais quanto aspectos estruturais das notificações.

5. Discussão e Conclusão

Este estudo explora as dificuldades de se trabalhar com bases desbalanceadas em modelos preditivos, utilizando dados de pacientes diagnosticados com transtornos mentais relacionados ao trabalho, disponibilizados pelo SUS. Foram testadas e categorizadas diversas técnicas, avaliadas sob diferentes métricas voltadas para bases desbalanceadas. Por fim, foram comparados os resultados a fim de identificar os melhores modelos para cada classe proposta. A principal contribuição foi a análise do impacto da escolha da métrica de otimização no comportamento dos modelos com dados reais.

Dessa forma, foram identificados três modelos distintos, que apresentaram valores de 0,47 de F1-Score, 0,77 de acurácia balanceada e 0,86 de ROC-AUC. Os resultados evidenciam que a definição prévia da métrica de otimização é determinante para o comportamento do modelo final. Enquanto a maximização do F1-Score favoreceu modelos com maior equilíbrio entre precisão e *recall*, a maximização da acurácia balanceada levou a classificadores com alto *recall*, porém baixa precisão, alterando significativamente o perfil de erros. Esse achado reforça que, em aplicações com bases públicas de saúde, a escolha da métrica deve preceder a escolha do modelo, e não o contrário.

A Tabela 5 evidencia que a maioria dos trabalhos em aprendizado de máquina com ênfase na saúde trata problemas com dados desbalanceados, apresentando proporções da classe minoritária frequentemente inferiores a 5%, chegando a valores significativamente baixos, como 0,81% e 1,63%. Em relação aos algoritmos, nota-se a predominância de métodos baseados em árvores, especialmente *Random Forest* e XGBoost, além do LightGBM. Esses modelos apresentam desempenho competitivo mesmo em bases severamente desbalanceadas. Contudo, há grande variação nas métricas reportadas, especialmente no F1-Score e no ROC-AUC, sugerindo que o grau de desbalanceamento impacta diretamente a capacidade do modelo em capturar adequadamente a classe minoritária.

O desempenho obtido por modelos de aprendizado de máquina em dados públicos de saúde mostrou-se fortemente dependente do contexto específico do problema, da proporção entre as classes e das estratégias metodológicas adotadas. Esses resultados reforçam que não existe uma solução única, sendo necessária uma combinação de decisões para que o modelo selecionado esteja alinhado à aplicação.

As limitações deste trabalho incluem o uso de dados com possível subnotificação (no período em análise, a notificação não era mandatória e existe um atraso em seu lançamento) e ausência de variáveis clínicas mais detalhadas. Além disso, mudanças temporais no padrão de registro ou nas políticas públicas podem influenciar o comportamento observado nos modelos.

Tabela 5. Comparação dos resultados das métricas em função do grau de desbalanceamento das bases em outros trabalhos

Referência	Problema	Algoritmo	Imb.	% Minoritária	Acurácia	F1-Score	BalancedAcc	ROC-AUC
-	**Predição de Desfechos em Dados de Saúde Ocupacional**	**RF**	**Sim**	**6,25%**	**94,56%**	**47,28%**	**77,04%**	**85,89%**
[Moreira et al. 2021]	Predição de diabetes tipo 1 na gestação	RF	Sim	1,63%	N/A	97,56%	N/A	N/A
[Rodrigues et al. 2024]	Predição da mortalidade por tuberculose	lightgbm	Sim	33,33%	N/A	N/A	N/A	71,50%
[Morsoleto et al. 2025a]	Predição da mortalidade infantil	XGBoost	Sim	0,81%	>90%	44,88%	N/A	N/A
[Jesus et al. 2020]	Previsão da mortalidade em gêmeos recém-nascidos	XGBoost	Sim	3,91%	97,40%	57,30%	N/A	95,00%
[Barros et al. 2025]	Classificação tratamento de tuberculose	RF	Sim	28,08%	85,70%	85,70%	N/A	N/A

Diante das limitações inerentes a dados secundários e do perfil de desbalanceamento da base, é esperado que os modelos não consigam predizer com perfeição ambas as classes. Portanto, em uma aplicação real, é imprescindível o alinhamento dos modelos às prioridades clínicas, preferencialmente em colaboração com especialistas da área da saúde, considerando o custo e o impacto de cada tipo de erro de predição. Trabalhos futuros podem aprofundar o pré-processamento dos dados, aplicando técnicas de seleção de atributos e de engenharia de variáveis.

6. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG (APQ-06570-24) e UFOP.

Referências

Barros, M. H. L. F. d. S., Silva, J. M. N. d., Vilhena, V., Melo, J. R. F., França, L. S., Freitas, L. R. S. d., Maia, L. T. d. S., Endo, P. T., and Ramalho, W. M. (2025). Machine learning classification of favorable vs unfavorable tuberculosis treatment outcomes using clinical and sociodemographic data from brazil's sinan-tb (2001–2023). *Research Square*. Preprint.

Brasil. Ministério da Saúde and Fundação Oswaldo Cruz (2024). Saúde mental dos trabalhadores dos serviços de saúde: diretrizes para formulação de políticas públicas em emergências em saúde pública. `https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/saude-do-trabalhador/saude-mental-dos-trabalhadores-dos-servicos-de-saude`. Acesso em: 25 Jan. 2026.

Campêlo, M. A. (2023). Estresse financeiro: causas, consequências e estratégias de enfrentamento. `https://www.gov.br/investidor/pt-br/penso-logo-investido/estresse-financeiro-causas-consequencias-e-estrategias-de-enfrentamento`. Acesso em: 25 Jan. 2026.

G1 (2025). Crise de saúde mental: Brasil tem maior número de afastamentos por ansiedade e depressão em 10 anos. `https://g1.globo.com/trabalho-e-carreira/noticia/2025/03/10/crise-de-saude-mental-brasil-tem-maior-numero-de-afastamentos-por-ansiedade-e-depressao-em-10-anos.ghtml`. Acesso em: 25 Jan. 2026.

Gomes, P., Almeida, D., Reis, N., Franca, J., Santos, P., Neto, J. S., Alves, E., and Oliveira, M. (2019). Predição de ações judiciais de consumo não registrado: uma abordagem para o problema de classes desbalanceadas. In *Anais da VII Escola Regional de*

Computação do Ceará, Maranhão e Piauí, pages 158–165, Porto Alegre, RS, Brasil. SBC.

ILO (2022). Perspectivas sociais e do emprego no mundo: Tendências 2022. https://www.ilo.org/global/research/global-reports/weso/trends2022/WCMS_834081/lang--en/index.htm. Acesso em: 25 Jan. 2026.

Instituto Nacional do Seguro Social - INSS (2025). Cuidados com a saúde mental podem ser amparados pela previdência. Acesso em: 25 Jan. 2026.

Jesus, E., Calais-Ferreira, L., and Barreto, M. (2020). Matched-pair analysis using machine learning to predict 1-year mortality in newborn twins. In *Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 215–225, Porto Alegre, RS, Brasil. SBC.

Ministério da Saúde (2025). Saúde incorpora câncer e transtornos mentais relacionados ao trabalho à lista de notificação compulsória. https://www.gov.br/saude/pt-br/assuntos/noticias/2024/agosto/saude-incorpora-cancer-e-transtornos-mentais-relacionados-ao-trabalho-a-lista-de-notificacao-compulsoria. Acesso em: 25 Jan. 2026.

Moreira, J., Bernardino, H., Barbosa, H., and Vieira, A. (2021). Modelos de aprendizado de máquina na predição de diabetes tipo 1 na gestação usando dados do sistema Único de saúde. In *Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 392–403, Porto Alegre, RS, Brasil. SBC.

Morsoleto, R., Silva, V., Caliari, J., Miranda, S., and Ferreira, H. (2025a). Combating class imbalance for infant mortality risk modeling: Resampling strategies in brazil's unified health system. In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*, pages 57–64, Porto Alegre, RS, Brasil. SBC.

Morsoleto, R., Silva, V., Caliari, J., Miranda, S., and Ferreira, H. (2025b). Prediction of infant mortality in brazil using machine learning and entity matching on brazilian unified health system's data. In *Anais do XIII Symposium on Knowledge Discovery, Mining and Learning*, pages 113–120, Porto Alegre, RS, Brasil. SBC.

Organização Pan-Americana da Saúde (OPAS) (2022). Pandemia de covid-19 desencadeia aumento de 25% na prevalência de ansiedade e depressão em todo o mundo. https://www.paho.org/pt/noticias/2-3-2022-pandemia-covid-19-desencadeia-aumento-25-na-prevalencia-ansiedade-e-depressao-em. Acesso em: 25 Jan. 2026.

Organização Pan-Americana da Saúde (OPAS) (2025). Transtornos mentais. Disponível em: https://www.paho.org/pt/topicos/transtornos-mentais. Acesso em: 25 Jan. 2026.

Rodrigues, M. M. S., Barreto-Duarte, B., Vinhaes, C. L., et al. (2024). Machine learning algorithms using national registry data to predict loss to follow-up during tuberculosis treatment. *BMC Public Health*, 24:1385.

WHO (2024). Saúde mental no trabalho. https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work. Acesso em: 25 Jan. 2026.