Estratégias de Aprendizado de Máquina para Classificação de Diabetes

1st Luiz Carlos Cosmi Filho
Departamento de Engenharia Elétrica
Universidade Federal do Espírito Santo
Vitória, Brasil
luiz.cosmi@edu.ufes.br

2nd Mateus Sobrinho Menines
Departamento de Engenharia Elétrica
Universidade Federal do Espírito Santo
Vitória, Brasil
mateus.menines@edu.ufes.br

***Resumo*—Diabetes is a metabolic disease characterized by elevated blood glucose levels. In the past three decades, the prevalence of type 2 diabetes has dramatically increased affecting approximately 422 million people globally. This study uses data from the Behavioral Risk Factor Surveillance System (BRFSS), collected by the Centers for Disease Control and Prevention (CDC), to classify individuals with diabetes. Using machine learning techniques, we aim to identify patterns and predict the risk of developing diabetes. Using machine learning for diabetes classification shows potential to detect disease risk before any clinical examination, presenting to individuals if their lifestyle can be dangerouos. By using BRFSS database, we compare different classification algorithms and evaluate their performance metrics. Additionally, we addresses the issue of data imbalance and explores strategies to manage it effectively.**

***Index Terms*—diabetes, machine learning, computational intelligence, data imbalance, disease risk prediction.**

I. Introdução

De acordo com a *World Health Organization* [1], a diabetes é uma doença metabólica caracterizada por níveis elevados de glicose no sangue, que ao longo do tempo, causa danos graves ao coração, rins e nervos. Nas últimas três décadas, a prevalência da diabetes tipo 2 aumentou dramaticamente em países, afetando cerca de 422 milhões de pessoas em todo o mundo.

A diabetes, sendo uma doença de grande relevância para a saúde, foi tema de uma pesquisa realizada pelo *Centers for Disease Control and Prevention* (CDC) dos Estados Unidos por meio de um questionário chamado de *Behavioral Risk Factor Surveillance System* (BRFSS). Esta pesquisa coletou dados da população, criando um banco de dados com atributos baseados nas respostas dos participantes, como hábitos de exercícios físicos, hábitos de tabagismo, diagnóstico de diabetes e outros.

A análise dos dados coletados por meio do BRFSS permite a aplicação de algoritmos de inteligência computacional para a classificação de indivíduos diabetes, utilizando técnicas de aprendizado de máquina como redes neurais e regressões para identificar padrões nos dados. Com isso, é possível treinar um modelo que auxilia na identificação de indivíduos com risco de desenvolver diabetes.

Uma das vantagens de utilizar aprendizado de máquina para a classificação de indivíduos com diabetes é a possibilidade de detectar o risco da doença antes de qualquer exame clínico. Isso pode alertar o indivíduo a mudar hábitos que possam causar a doença, servindo como um indicador inicial de risco.

Neste contexto, este artigo propõe um estudo de caso com base no banco de dados da BRFSS, visando realizar a classificação de indivíduos com diabetes utilizando os atributos coletados. O estudo compara diferentes algoritmos de classificação utilizados em inteligência computacional, com o objetivo de avaliar as métricas obtidas por cada modelo e analisar os fatores que podem influenciar a ocorrência de diabetes em humanos. Além disso, também é explorado o desbalanceamento na base de dados e estratégias para lidar com esse problema.

Assim, para um melhor entendimento, as próximas seções irão abordar, respectivamente: trabalhos relacionados (Seção II), a metodologia da proposta desenvolvida nesse trabalho (Seção III), os experimentos e resultados obtidos (Seção IV) e, por fim, as conclusões e trabalhos futuros (Seção V).

II. Trabalhos relacionados

Um exemplo de trabalho que relaciona a diabetes com um modelo preditivo de detecção é o de [2]. Este artigo aborda a construção de modelos de previsão de risco para diabetes tipo 2 utilizando técnicas de aprendizado de máquina, como *Support Vector Machine* (SVM), *Decision Tree*, *Logistic Regression*, *Random Forest*, *Gaussian Naive Bayes* e redes neurais. O estudo também identifica outros fatores de risco para diabetes tipo 2 por meio de métodos estatísticos. Além disso, são discutidas questões relacionadas à cobertura de saúde, atividade física, saúde mental e outras variáveis associadas à diabetes tipo 2. Para lidar com o desbalanceamento, os autores apresentam técnicas para geração de dados sintéticos a fim de gerar amostras da classe minoritária.

Outro trabalho que se relaciona a diabetes é o de [3] que aborda a aplicação de técnicas de aprendizado de máquina para o diagnóstico de diabetes em um conjunto de dados desbalanceado, utilizando o conjunto de dados BRFSS. Ele destaca a importância da identificação precoce de indivíduos em risco de desenvolver diabetes e como as abordagens tradicionais de teste podem ser custosas e demoradas. O estudo explora a utilização de técnicas de amostragem, como *oversampling*, *undersampling* e técnicas híbridas, para lidar com o desequilíbrio de classes no conjunto de dados.

Diferentemente dos artigos citados, nesse trabalho pretende-se lidar com o problema de desbalanceamento entre as classes através de técnicas disponíveis à nível de algoritmo. Bem como, explorar o desempenho desses algoritmos adaptados para lidar com o desbalanceamento sobre a base de dados.

III. Metodologia

A base de dados utilizada nesse trabalho corresponde a (BRFSS)[1] de 2015 [4]. Abrange os comportamentos de saúde, as condições e os aspectos socioeconômicos dos residentes dos EUA. Além disso, possui 229.474 amostras e os seus atributos são:

1) *DiabetesBinary*: você tem diabetes? Sim (1) ou não (0);
2) *HighBP*: você tem pressão alta? Sim (1) ou não (0);
3) *HighChol*: você tem colesterol alto? Sim (1) ou não (0);
4) *CholCheck*: você realizou exame de colesterol nos últimos cinco anos? Sim (1) ou não (0);
5) *BMI*: qual seu índice de massa muscular? Variável inteira;
6) *Smoker*: você é fumante? Sim (1) ou não (0);
7) *Stroke*: você já teve um AVC? Sim (1) ou não (0);
8) *HeartDiseaseorAttack*: você tem doença coronariana (DCC) ou infarto do miocárdio? Sim (1) ou não (0);
9) *PhysActivity*: você praticou atividade física nos últimos 30 dias? Sim (1) ou não (0);
10) *Fruits*: você consume uma ou mais frutas por dia? Sim (1) ou não (0);
11) *Veggies*: você consume uma ou mais verduras por dia? Sim (1) ou não (0);
12) *HvyAlcoholConsump*: você consume grandes quantidades de álcool (homens adultos que bebem mais de 14 drinques por semana e mulheres adultas que bebem mais de 7 drinques por semana)? Sim (1) ou não (0);
13) *AnyHealthcare*: você tem algum plano de saúde? Sim (1) ou não (0);
14) *NoDocbcCost*: houve algum momento nos últimos 12 meses em que você precisou consultar um médico, mas não pôde por causa do custo? Sim (1) ou não (0);
15) *GenHlth*: você diria que, em geral, o quão boa é a sua saúde? Escala de 1 à 5;
16) *MentHlth*: sobre sua saúde mental, que inclui estresse, depressão e problemas emocionais, por quantos dias durante os últimos 30 dias sua saúde mental não foi boa? Escala de 0 à 30;
17) *PhysHlth*: sobre sua saúde física, que inclui doenças e lesões físicas, por quantos dias durante os últimos 30 dias sua saúde física não foi boa? Escala de 0 à 30;
18) *DiffWalk*: você tem muita dificuldade para andar ou subir escadas? Sim (1) ou não (0);
19) *Sex*: qual o seu sexo? Feminino (0) ou masculino (1);
20) *Age*: qual a sua idade? Escala de 0 à 14;
21) *Education*: qual é a série ou ano de escolaridade mais alto que você concluiu? Escala de 1 à 6;
22) *Income*: qual a sua renda familiar anual? Escala de 1 à 8;

[1]Disponível para baixar no repositório de banco de dados do UCI: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

Ao iniciar os trabalhos com a base de dados, foi possível observar um forte desbalanceamento entre as classes (diabéticos e não diabéticos). Na Figura 1, observa-se que os diabéticos correspondem a apenas $15,3\%$ do total de pessoas. Em números absolutos, isso significa que 35.097 indivíduos são diabéticos, de um total de 229.474 pessoas na base de dados.

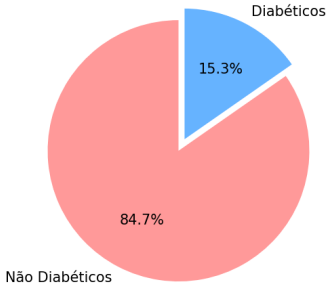


Figura 1: Proporção de diabéticos e não diabéticos na base de dados.

A. Problema de desbalanceamento entre classes

O desbalanceamento que contém a base de dados é intrínseco ao problema. Em tarefas de classificação, se houver um desbalanceamento significativo entre as classes, a classe majoritária pode ser enfatizada. Isso significa que um modelo de classificação pode ter um desempenho superior para a classe majoritária, resultando em classificações errôneas para a classe minoritária. Em aplicações na área da saúde, a classificação incorreta da classe minoritária tem um custo maior do que o contrário, pois pode significar deixar de iniciar o tratamento em uma pessoa com a doença.

Sendo assim, é possível abordar o problema de classificação em base de dados desbalanceadas por meio duas abordagens: a nível de dados e a nível de algoritmos.

Ao tentar resolver o problema a nível de dados, pode-se tentar coletar mais exemplos das classes minoritárias. No entanto, nem sempre é possível realizar isso. Assim, pode-se tentar remostrar a base de dados, removendo exemplos da classe majoritária [5] ou adicionando exemplos sintéticos às classes minoritárias [6], [7].

À nível de algoritmos, há a possibilidade de adaptar os algoritmos de aprendizado de máquina para serem capazes de lidar com essa problemática. Por exemplo, realizando a escolha de um viés indutivo apropriado (aprendizado sensível ao custo), onde associa-se custos diferentes ao erro de classificação de exemplos em classes diferentes [8]. Assim, os métodos de aprendizagem sensíveis aos custos permitem que os algoritmos deem prioridade à classe minoritária e obtenham um melhor desempenho em problemas críticos de classificação.

B. Algoritmos paar treinamento do modelo

Nesse trabalho, serão abordados técnicas à nível de algoritmo a fim de otimizar os modelos de aprendizado de

máquina. Num primeiro momento, verificou-se o desempenho de alguns algoritmos na base de dados sem nenhuma adaptação para o problema de desbalanceamento, sendo eles: Regressão Logística (RL), SVM Linear (SVM) e *Random Forest* (RF). Em um segundo momento, abordou-se os seguintes algoritmos com adaptações para o problema de desbalanceamento:

- Regressão logística: pode ser adaptado para que atribua custos diferentes ao erro de classificação para classes diferentes. Nesse trabalho, utilizou-se o algoritmo de gradiente descendente estocástico associado com a função de perda logarítmica (do inglês, *log loss*). Assim, ajustou-se automaticamente os custos inversamente proporcionais às frequências de cada classe;
- SVM Linear: também pode ser adaptado para que atribua custos diferentes ao erro de classificação para classes diferentes. Nesse trabalho, utilizou-se o algoritmo de gradiente descendente estocástico associado com a função de perda de articulação (do inglês, *hinge loss*). Assim, utilizou-se a mesma metodologia que à aplicada a regressão logística, ajustando automaticamente os custos inversamente proporcionais às frequências da classe;
- *Random Forest*: é um classificador que ajusta uma série de árvores de decisão em vários sub-conjuntos do conjunto de dados original e utiliza o cálculo da média para melhorar a precisão da classificação e controlar o sobreajuste. Dessa forma, utilizou-se a metodologia onde, para cada sub-conjunto gerado durante o treinamento, os pesos de cada classe são ajustados inversamente proporcionais às frequências da classe;
- *Extreme Gradient Boosting*: possui hiperparâmetros que permite ajustar o comportamento do algoritmo para problemas de classificação em base de dados desbalanceadas. Assim, utilizou-se o hiperparâmetro que permite escalar o gradiente para a classe positiva.

C. Métricas para avaliação dos algoritmos

Para avaliar os algoritmos, algumas métricas foram selecionadas. Considere a matriz de confusão para o problema de classificação binária da Tabela I a seguir,

		Predito	
		1	0
Real	1	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	0	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Tabela I: Matriz de confusão para o problema de classificação binária.

A precisão mede a proporção de amostras que foram classificadas como positivas e que realmente são positivas. Ou seja, mede a precisão do algoritmo em informar amostras que são positivas. A Eq. (1) calcula a precisão a partir da matriz de confusão.

$$P = \frac{VP}{VP + FP} \tag{1}$$

Por outro lado, a sensibilidade (*recall*) mede a proporção de amostras positivas que foram classificadas como positivas. Ou seja, mede a capacidade do algoritmo de recuperar todas as amostras positivas. Assim, pode-se calcula-lá por meio da Eq. (2).

$$R = \frac{VP}{VP + FN} \tag{2}$$

No entanto, aumentar a precisão de um algoritmo tende a reduzir o *recall*, e vice-versa. Uma estratégia é combinar as duas métricas em uma só, tal como a média harmônica F_1. Tal métrica é calculada pela Eq. (3).

$$F_1 = \frac{VP}{VP + \frac{1}{2}(FP + FN)} = \frac{2PR}{P + R} \tag{3}$$

A acurácia mede a porcentagem de amostras que foram corretamente classificadas sendo calculada pela Eq. 4.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{4}$$

Por fim, a acurácia balanceada é utilizada para análises em base de dados com classes desbalanceadas, pois a acurácia convencional não é uma métrica adequada nesses casos. Isso ocorre porque os valores de classificação de verdadeiros negativos (VN) podem mascarar a baixa classificação de verdadeiros positivos (VP). A acurácia balanceada, por outro lado, realiza os cálculos considerando a taxa de acerto VP e VN separadamente, proporcionando uma avaliação mais precisa dos acertos do modelo em relação a cada classe. Essa acurácia é calculada pela Eq (5).

$$A_b = \frac{1}{2}\left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP}\right) \tag{5}$$

D. Curvas para avaliação dos algoritmos

A curva ROC consiste em um gráfico que apresenta o desempenho de um modelo de aprendizado de máquina relacionando sua taxa verdadeiros positivos e sua taxa de falsos positivos, conforme se consideram diferentes valores para limite de decisão [9]. Assim, a área sobre a curva ROC é uma medida global da habilidade do classificador em discriminar uma condição específica está ou não presente no indivíduo.

Apesar dos gráficos da curva ROC sejam visualmente interessantes e forneçam uma visão geral do desempenho de um classificador, estudos mostram que no contexto de conjuntos de dados desbalanceados pode ser enganosa a análise da curva ROC no que diz respeito às conclusões sobre a confiabilidade do desempenho da classificação [10]. Nesse sentido, é importante também avaliar a curva PRC. Tal curva, mostra a relação entre precisão e *recall* à medida que o limite de decisão é alterado. À medida que a precisão aumenta, o *recall* tende a diminuir. Dessa forma, mostra-se importante analisar também avaliar a curva PRC para obter uma visão geral do desempenho dos classificadores.

IV. Experimentos e resultados

Os experimentos para a avaliação do aprendizado de máquinas utilizando os algoritmos foram realizados em *Python* com o serviço Google *Colaboratory*. O código desenvolvido está disponível no *Github*.

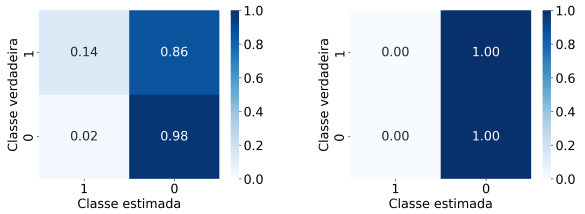
A. *Experimento 1*

No primeiro experimento, não adaptou-se os algoritmos para o problema de desbalanceamento entre as classes. Na Tabela II, é possível observar as médias de cada uma das métricas obtidas por meio do procedimento de validação cruzada. Apesar dos algoritmos apresentarem acurácia maior que 80%, os *recalls* obtidos em todos eles foram muito baixas, indicando a classificação errônea de amostras da classe minoritária. Assim, impactando no cálculo das médias harmônicas F1 e acurácia balanceada.

	Acurácia	Acurácia bal.	F1 score	Precisão	Recall
RL	0.849	0.564	0.233	0.529	0.154
SVM	0.847	0.505	0.020	0.104	0.012
RF	0.849	0.512	0.051	0.658	0.027

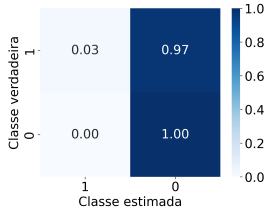
Tabela II: Médias das métricas no processo de validação cruzada desconsiderando o desbalanceamento entre as classes.

Para analisar melhor os resultados, treinou-se os modelos utilizando 70% da base de dados para treino e 30% para teste (estratificando os conjuntos de treino e teste). Na Figura 2, pode-se visualizar as matrizes de confusão de cada modelo sobre o conjunto de teste. É possível observar que os algoritmos convergiram de forma que todos os exemplos são classificados como pertencentes a classe majoritária e possuindo um péssimo desempenho sobre a classe minoritária. As métricas obtidas para esses treinamentos também podem ser analisadas na tabela III.



(a)

(b)



(c)

Figura 2: Matrizes de confusão dos algoritmos treinados com 70% da base de dados para treino e 30% desconsiderando o desbalanceamento entre as classes. (a) Regressão Logística. (b) SVM Linear. (c) *RandomForest*.

	Acurácia	Acurácia bal.	F1 score	Precisão	Recall
RL	0.849	0.558	0.219	0.529	0.138
SVM	0.847	0.500	0.002	0.500	0.001
RF	0.849	0.512	0.052	0.699	0.027

Tabela III: Métricas dos algoritmos com 70% da base de dados para treino para treino e 30% para teste desconsiderando o balanceamento entre as classes.

A métrica acurácia exibida na Tabela III é relativamente alta. A princípio, pode parecer indicar um bom resultado, já que os três algoritmos apresentam uma acurácia próxima de 85%. No entanto, ao observar a Figura 2 com as matrizes de confusão, percebe-se um viés de classificação devido ao desbalanceamento das classes, com a maioria das previsões indicando que todos os indivíduos não têm diabetes. Outra métrica interessante de ser analisada é a acurácia balanceada, que ficou em torno de 50% em todos os algoritmos. Esse valor é relativamente baixo para o aprendizado de máquina, proporcionando uma métrica mais adequada para avaliar o treinamento dos modelos.

Por fim, as métricas *recall* para os modelos estão próximas a zero, mostrando a classificação errada das pessoa com diabetes (classe minoritária) e resultando em uma taxa de verdadeiro positivo próxima a zero. Ou seja, os indivíduos foram erroneamente classificados como não diabéticos.

B. *Experimento 2*

No segundo experimento, adaptou-se os algoritmos para lidar com o desbalanceamento entre as classes. Além disso, foi realizado antes de todos os procedimentos uma busca em grade para se obter os hiperparâmetros que maximizam a métrica F1. Na Tabela IV, pode-se observar as médias de cada uma das métricas obtidas por meio do procedimento de validação cruzada. Apesar de todos os algoritmos apresentarem acurácia menor, em torno de 70%, verifica-se que o *recall* e precisões obtidas aumentaram, indicando uma melhora na classificação das amostras da classe minoritária. Consequentemente, também aumentando a média harmônica F_1 e a acurácia balanceada.

	Acurácia	Acurácia bal.	F1 score	Precisão	Recall
RL	0.711	0.730	0.446	0.316	0.757
SVM	0.702	0.727	0.440	0.310	0.763
RF	0.709	0.732	0.446	0.315	0.766
XGB	0.762	0.731	0.469	0.356	0.687

Tabela IV: Médias de diversas métricas obtidas a partir do processo de validação cruzada considerando o desbalanceamento entre as classes.

Igualmente ao primeiro experimento, treinou-se os modelos utilizando 70% da base de dados para treino e 30% para teste (estratificando os conjuntos de treino e teste). Então, na

Figura 3, são apresentadas as matrizes de confusão de cada modelo sobre o conjunto de teste. As métricas obtidas para esses treinamentos também podem ser analisadas na tabela V.

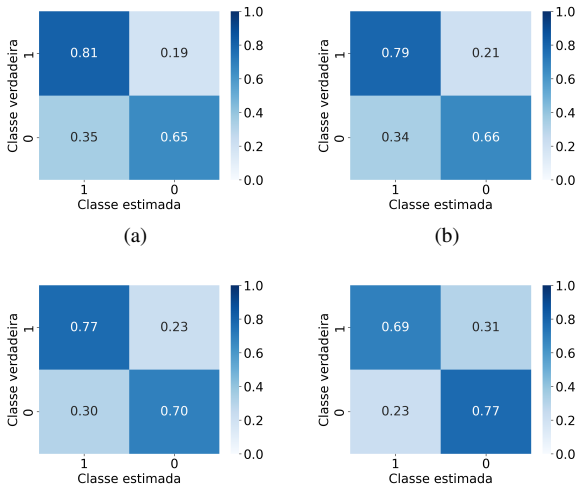


Figura 3: Matrizes de confusão dos algoritmos treinados com 70% da base de dados para treino e 30% considerando o desbalanceamento entre as classes. (a) Regressão Logística. (b) SVM Linear. (c) *RandomForest*. (d) *Extreme Gradient Boosting*.

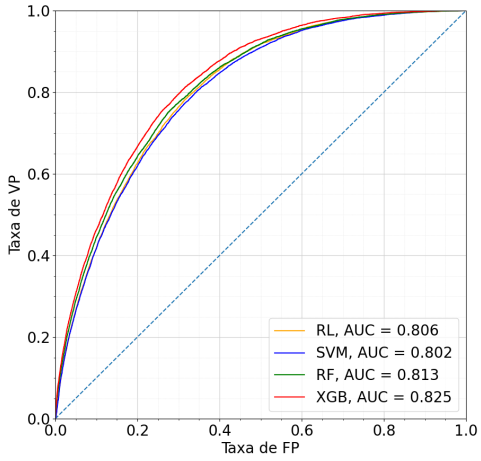
	Acurácia	Acurácia bal.	F1 score	Precisão	Recall
RL	0.674	0.731	0.433	0.295	0.812
SVM	0.679	0.725	0.430	0.295	0.792
RF	0.707	0.734	0.447	0.314	0.773
XGB	0.762	0.734	0.471	0.357	0.693

Tabela V: Diversas métricas dos algoritmos treinados com 70% da base de dados para treino e 30% considerando o desbalanceamento entre as classes.

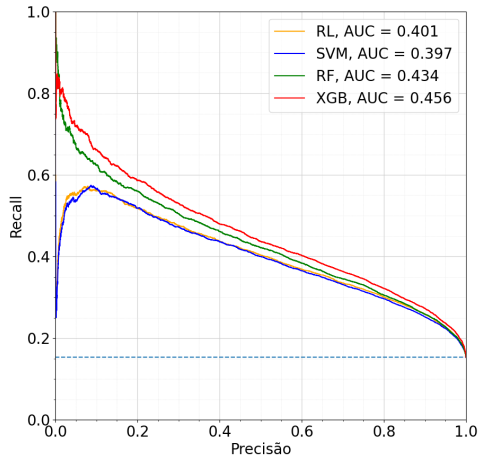
As métricas mostradas na Tabela V apresentaram valores melhores em relação ao primeiro experimento. A acurácia balanceada aumentou em todos os algoritmos, indicando que foi realizado um treinamento capaz de lidar com o desbalanceamento entre as classes. O *recall* também obteve uma melhora significativa, chegando a uma faixa entre 69% a 81%, validando que os modelos obtiveram um acerto maior de verdadeiros positivos e não classificando todos os indivíduos como falsos negativos. Outra análise possível é que os algoritmos conseguiram realizar uma predição melhor entre os indivíduos com diabetes, algo que não ocorria anteriormente devido ao viés nos modelos. Isso pode ser observado nas matrizes de confusão na Figura 3, onde os três algoritmos mostraram semelhanças nos padrões de classificação.

Ao analisar o segundo experimento, é evidente que as métricas são superiores em relação ao primeiro experimento e condicentes ao esperado para tarefa de classificação de indivíduos com diabetes. Além disso, verificou-se na prática que ao utilizar as técnicas adequadas para base de dados desbalanceadas tem-se um melhor resultado, mostrando a importância e relevância de estudos para adaptar algoritmos nesses problemas. Como o segundo experimento obteve um melhor resultado, foram obtidos e analisados os gráficos das curvas ROC e das curvas PRC.

Na Figura 4a, são apresentadas as curvas ROC dos classificadores treinados, bem como a área sobre a curva para cada classificador. Além disso, na Figura 4b exibe as curvas PRC, sendo também apresentada a área sobre a curva para cada classificador.



(a)



(b)

Figura 4: Curva ROC (a) e Curva Precisão/*Recall* (b) dos classificadores Regressão Logística (RL), SVM Linear (SVM), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGB).

Analisando as curvas das duas figuras, é possível observar que o melhor algoritmo de classificação é o *Extreme Gradient Boosting* que possui a maior areá sob a curva, coincidindo nas duas figuras. Apesar de ter obtido um menor *Recall* visto na Tabela V, o algoritmo XGB, obteve a melhor métrica F1 score e a maior precisão para um limiar de decisão de 0.5 entre todos os classificadores.

V. Conclusões

Ao lidar com base de dados desbalanceados, verificou-se a importância de ajustar os algoritmos de aprendizado de máquina para lidar com isso. Nesse trabalho, foi possível explorar algumas técnicas e visualizar a melhora dos resultados.

Além disso, a escolha de métricas adequadas para problemas de classificação em base de dados desbalanceadas também mostrou-se extremamente importante. Através das análises realizadas, mostrou-se que acurácia pode não ser uma métrica boa para representar o bom desempenho dos algoritmos nesses casos. Métricas como a média harmônica F1, que leva em consideração tanto a precisão quanto o *recall*, mostrou-se mais adequada. Verificar também a curva ROC e a curva Precisão/*Recall* também é importante. Bem como, as áreas sobre as curvas de cada algoritmo para ter uma ideia dos desempenhos gerais.

A aplicação de quaisquer um dos algoritmos, levariam a resultados muito semelhantes na prática. No entanto, verificou-se que o algoritmo *Extreme Gradient Boosting* tende a generalizar melhor o problema. Uma vez que, apresenta uma maior área sobre a curva ROC, maior área sobre a curva Precisão/*Recall* e maior média harmônica F1.

Como trabalhos futuros, sugere-se a aplicação de técnicas de seleção de atributos para permitir um melhor desempenho dos algoritmos de aprendizado de máquina. Além disso, técnicas de *undersampling* e *oversampling* podem ser investigadas a fim de analisar se é possível encontrar uma melhora na tarefa de classificação.

Referências

[1] World Health Organization. Diabetes, 2024. Accessed: 2024-06-09.

[2] Zidian Xie, Olga Nikolayeva, Jiebo Luo, and Dongmei Li. Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16, 09 2019.

[3] Mohammad Mihrab Uddin Chowdhury, Ragib Ayon, and Sakhawat Hossain. An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced brfss dataset. *Healthcare Analytics*, 5:100297, 12 2023.

[4] CDC. Behavioral risk factor surveillance system. https://www.cdc.gov/brfss/index.html. Accesso em: 07 de jun. 2024.

[5] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, pages 1–7. ICML, 2003.

[6] N. Chawla, K. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *ArXiv*, abs/1106.1813, 2002.

[7] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.

[8] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2010.

[9] Zhe Hui Hoo, Jane Candlish, and Dawn Teare. What is an roc curve? *Emergency Medicine Journal*, 34(6):357–359, 2017.

[10] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.