Equitable Diabetes Diagnosis: Tackling Ethnic and Gender Disparities

Lívia Ruback[1], **Luisa Felix** [1], **and Ariel Soares Teles** [2]

[1]Faculdade de Tecnologia, Universidade Estadual de Campinas (UNICAMP)
Limeira, SP, Brazil

[2]Federal Institute of Maranhão, Araioses, Brazil

liviacr@unicamp.br, luisa.felix.oliveira@gmail.com, ariel.teles@ifma.edu.br

Abstract. *Machine Learning (ML) has advanced disease diagnosis in healthcare, but raises fairness concerns, as model biases can perpetuate social inequalities. This study aims to evaluate and mitigate bias in diabetes diagnosis prediction models. We conducted experiments considering ethnicity and gender as protected attributes, evaluating bias using the fairness metrics Statistical Parity Difference, Equal Opportunity Difference, and Average Odds Difference. We applied the bias mitigation techniques Reweighing and Prejudice Remover, which showed improvements in fairness metrics, with a reduction in disparities between groups, while maintaining model accuracy. These findings reinforce the need to integrate fairness considerations into ML models for healthcare applications.*

1. Introduction

ML models have been widely applied across various domains in healthcare, particularly in disease diagnosis and prognosis, enabling early detection, risk assessment, and personalized treatment strategies [Khanam and Foo 2021]. One prominent application is diabetes prediction, a disease that affects millions globally [GBD 2021 Diabetes Collaborators 2023]. Diabetes is typically classified into two main types: type 1, an autoimmune condition usually diagnosed in childhood or adolescence, and type 2, which is more common and often associated with lifestyle and metabolic factors.

However, recent studies have shown that ethnic minorities are disproportionately affected by certain diabetes-related complications, which may contribute to reduced access to care [Huang et al. 2022]. Similarly, other studies have identified potential disparities in the performance of diabetes prediction models across different genders [Talebi Moghaddam et al. 2024].

Recognizing the importance of fairness in predictive modeling, recent guidelines recommend integrating bias mitigation into the standard ML pipeline in healthcare [Klement and El Emam 2023]. In this context, fairness refers to the absence of discrimination against individuals or groups based on sensitive attributes such as age, race/ethnicity, gender, or socioeconomic status [Caton and Haas 2024]. These attributes are referred to as protected attributes, as they represent individual characteristics that require ethical and legal safeguards and are central to identifying and addressing bias in ML [Raza 2023].

This study aims to promote fairness in ML models for the diagnosis of diabetes by evaluating and mitigating bias that leads to discrimination against individuals or groups. The main contributions of this work are as follows:

- We conduct an experiment using a diabetes diagnostic model, evaluating fairness using the metrics Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD). We also interpret the results for each fairness metric.
- We apply the bias mitigation techniques Reweighing (a pre-processing method) and Prejudice Remover (an in-processing method), and evaluate their effects on both fairness metrics and model accuracy.

The remainder of this paper is organized as follows. Section 2 reviews related work on fairness and bias mitigation in ML models for the diagnosis of diabetes. Section 3 describes the materials and methods used in this study. Section 3.7 presents the results. Section 4 discusses the findings and their implications. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Related work

In the health domain, several reviews have focused on biases, fairness metrics, mitigation methods, and tools for evaluating and mitigating bias [Huang et al. 2024]. Some of these works particularly address racial bias in ML applications [Huang et al. 2022], highlighting disparities in access to care, diagnosis accuracy, and treatment outcomes among minority populations. These reviews also point to a growing consensus on the need to evaluate ML models not only in terms of predictive performance but also fairness across protected attributes. dataset.

In the diabetes domain, recent studies have assessed disparities in prognostic models. Raza [Raza 2022] developed a predictive model for forecasting hospital readmissions and mitigating disparities found related to race, socioeconomic status, and other demographic factors. Cronjé et al. [Cronjé et al. 2023] evaluated type 2 diabetes risk models using NHANES data, investigating how they perform for historically marginalized racial groups, finding systematic underestimation of diabetes risk for non-Hispanic Black individuals.

Pias et al. [Pias et al. 2025] mitigated bias in the context of type 2 diabetes prediction, using the BRFSS data. They propose resampling strategies as a solution for reducing disparities across subgroups by applying techniques including SMOTE (Synthetic Minority Over-sampling Technique), ADASYN, and undersampling, to balance the dataset before training.

To the best of our knowledge, our study is the first to systematically apply and compare both pre-processing (*Reweighting*) and in-processing (*Prejudice Remover*) bias mitigation techniques on the BRFSS dataset in the context of type 2 diabetes prediction. We evaluated these methods separately for ethnicity and gender, allowing for a more nuanced analysis of fairness. By evaluating fairness separately across ethnicity and gender, and applying mitigation strategies from different stages of the machine learning pipeline, our approach enables a more interpretable analysis of bias and mitigation effects, particularly relevant for public health decision-making.

3. Materials and Methods

This section presents the materials and methods employed in this study. We begin by describing the ML pipeline we followed for integrating fairness into the traditional ML tasks. Next, we detail the dataset used, including the selection of relevant attributes and the preprocessing procedures. Finally, we outline the model development process, the fairness evaluation metrics adopted, and the bias mitigation strategies implemented.

3.1. ML Pipeline Ensuring Fairness

We describe an ML pipeline designed to assess and mitigate bias in ML models for diabetes diagnosis, as illustrated in Figure 1, inspired by the pipeline proposed in [Raza 2023]. We followed this pipeline to evaluate and mitigate bias in a diabetes diagnosis ML model, described as follows.

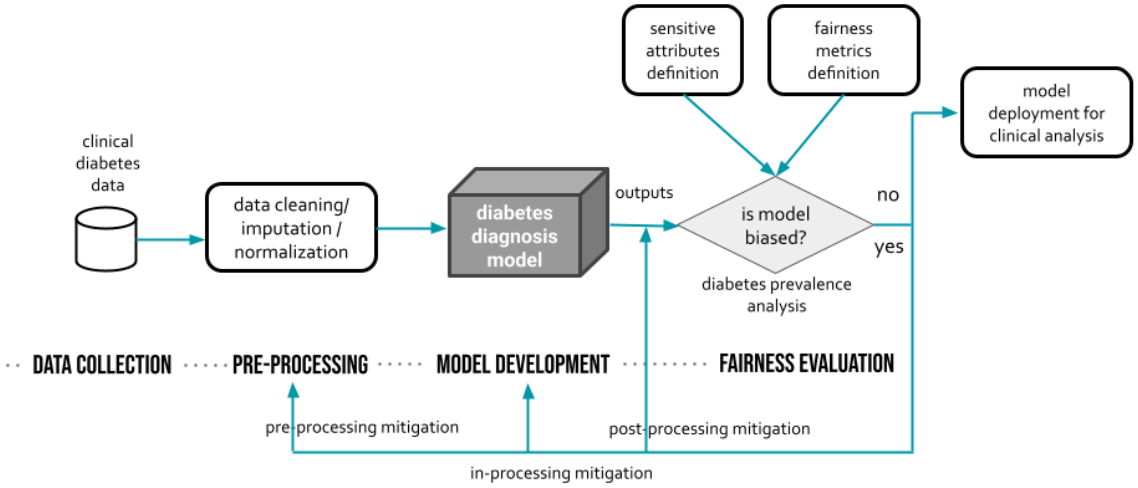

Figure 1. ML Pipeline for Bias Mitigation in Diabetes Diagnosis.

The ML pipeline begins with collecting clinical data related to the diabetes diagnosis ML task. There are many publicly available datasets for this task, such as the Pima Indians Diabetes Dataset, the NHANES, the BRFSS, and the MIMIC-IV.

The pre-processing step consists of preparing the data for learning by cleaning, normalizing, and transforming it. The model development step involves selecting the ML algorithm and training the model using the pre-processed data, which is evaluated using performance classification metrics, such as accuracy, recall, precision, F1-score, and the Area Under the Receiver Operating Characteristic curve (AUC-ROC) [Wang et al. 2024].

The fairness evaluation step includes defining the sensitive attributes (e.g, race/ethnicity, gender, age, income) and selecting appropriate fairness metrics for classification ML tasks. The literature offers multiple definitions and categorizations of bias [Mehrabi et al. 2021, Verma and Rubin 2018, Ruback et al. 2022], along with a range of fairness metrics for its assessment [Varshney 2022, Hardt et al. 2016, Dwork et al. 2012]. In this context, fairness refers to the ability of predictive models to avoid discrimination against individuals or groups based on sensitive attributes such as age, race/ethnicity, sex/gender, or socioeconomic status. Nonetheless, there is no universal

standard for measuring fairness, nor consensus on which metrics are most appropriate [Caton and Haas 2024].

When fairness evaluations reveal discrimination, various mitigation strategies can be applied to reduce bias. These strategies are commonly grouped according to where they intervene in the ML pipeline: pre-processing (before model training), in-processing (during model training), and post-processing (after model predictions). Regardless of the chosen approach, their effectiveness is evaluated by measuring changes in fairness metrics after application. These metrics are used iteratively to ensure that fairness goals are achieved and bias is adequately addressed.

We used the AI Fairness 360 (AIF360) toolkit [Kush Varshney 2018] in this experiment. AIF360 is an open-source Python library developed to assess and mitigate bias in ML applications. It provides a range of quantitative metrics for fairness evaluation, along with several algorithms for bias mitigation.

3.2. Dataset

We use the dataset collected from the Behavioral Risk Factor Surveillance System (BRFSS), made available by the Centers for Disease Control and Prevention (CDC) [Xie et al. 2019], and has been used by other studies to develop ML models for diabetes diagnosis [Chang et al. 2022]. The data was collected in the 2015 BRFSS Survey Data and includes information from U.S. adults concerning their health behaviors and how these may impact long-term medical conditions. It covers data from the 50 U.S. states, the District of Columbia, Guam, and Puerto Rico. The raw dataset consists of 441,456 instances (i.e., individual respondents), each described by 330 attributes.

Table 1 lists the 24 attributes selected for use in this study. These include clinical variables commonly associated with diabetes, such as high blood pressure and Body Mass Index (BMI) [Pias et al. 2025], as well as social determinants of health, such as gender, ethnicity, income, and access to health insurance, that may contribute to disparities in diagnosis and care. Among these, gender and ethnicity were considered protected attributes for fairness evaluation and bias mitigation. This study specifically focuses on the diagnosis of type 2 diabetes, as it is the target attribute in the dataset.

The distribution of protected attributes to the target variable reveals important demographic patterns. Among individuals diagnosed with type 2 diabetes, 52.66% are women. Regarding ethnicity, the breakdown is as follows: 73.13% are White, 11.29% are Black, 8.64% are Hispanic, 4.74% are Other, and 2.21% are Multiracial. Considering a binary classification into White (1) and Non-White (2–5), Non-White individuals represent 26.87% of the positive diabetes cases.

3.3. Pre-processing

We removed rows with responses such as “Blank”, “Don’t know/Not sure”, or “Refused” in any of the 24 selected attributes (see Table 1). Rather than imputing missing values, we chose to work only with complete entries to preserve the integrity of the original responses and avoid introducing artificial patterns that could bias the model.

Recent studies have shown that imputation methods can influence algorithmic fairness. For example, [Bhatti et al. 2025] empirically demonstrates that different missing

Attribute	Description
Diabetes	Indicates whether the respondent has diabetes (0 = No, 1 = Yes).
HighBloodPressure	Presence of high blood pressure (0 = No, 1 = Yes).
HighCholesterol	Presence of high cholesterol (0 = No, 1 = Yes).
CholesterolCheck	Had cholesterol checked in the past 5 years (0 = No, 1 = Yes).
BMI	Body Mass Index of the respondent.
Smoker	Smoked at least 100 cigarettes in life (0 = No, 1 = Yes).
Stroke	Has had a stroke (0 = No, 1 = Yes).
HeartDisease	Coronary heart disease or myocardial infarction (0 = No, 1 = Yes).
PhysicalActivity	Any physical activity in the past 30 days (0 = No, 1 = Yes).
FruitsDaily	Eats one or more fruits daily (0 = No, 1 = Yes).
VegetablesDaily	Eats one or more vegetables daily (0 = No, 1 = Yes).
AlcoholConsumption	Heavy drinking: > 14 drinks/week (men) or > 7 drinks/week (women) (0 = No, 1 = Yes).
HealthCoverage	Has any kind of health insurance (0 = No, 1 = Yes).
NoDoctorDueToCost	Needed to see a doctor but couldn't due to cost (0 = No, 1 = Yes).
GeneralHealth	Self-assessed health (1 = Excellent to 5 = Poor).
MentalHealthDays	Days with poor mental health in the past 30 days (0–30 scale).
PhysicalHealthDays	Days with poor physical health in the past 30 days (0–30 scale).
DifficultyWalking	Serious difficulty walking or climbing stairs (0 = No, 1 = Yes).
Sex	0 = Female, 1 = Male.
Age	Age group (e.g., 1 = 18–24, ..., 13 = 80+).
EducationLevel	1 = Never attended school to 6 = College graduate.
Income	1 = < \$10,000 to 8 = $\$\geq 75{,}000$.
Ethnicity	1 = White, 2 = Black, 3 = Hispanic, 4 = Other, 5 = Multiracial.
State	U.S. state of residence.

Table 1. Description of attributes used in the dataset.

data handling techniques can significantly affect fairness: while simple techniques or case-wise deletion tend to better preserve fairness, more complex imputations - often optimized for predictive performance — may degrade equity in model outcomes.

By retaining only fully observed data, we aimed to preserve the underlying disparities that are crucial for fairness-aware ML. After this filtering step, the final dataset contained 251,467 instances.

The target attribute was initially divided into four categories: individuals with diabetes (1), those who had diabetes only during pregnancy (2), individuals without diabetes (3), and those with pre-diabetes (4). We transformed this into a binary classification, where 1 indicates a patient has diabetes or prediabetes, and 0 means the patient does not have diabetes or had it only during pregnancy. We consider prediabetes as a positive case due to its clinical relevance. According to the American Diabetes Association(ADA), prediabetes is a high-risk state for developing diabetes and associated complications, and early detection is critical for prevention [Association 2023].

The ethnicity attribute was originally represented with multiple categories. It was later binarized into two groups: white (1, privileged) and non-white (0, non-privileged), with the non-white group including Black, Hispanic, other races, and multiracial individuals. This binarization was applied exclusively for the fairness evaluation and the subsequent bias mitigation steps. On the other hand, the gender attribute was already binary, so no modifications were needed.

3.4. Model development

For the ML classification task on diagnosing type 2 diabetes, we trained a Logistic Regression (LR) algorithm. Although LR may not always perform optimally on imbalanced datasets, it remains a widely used and interpretable model in fairness research for identifying and mitigating bias [Mehrabi et al. 2021]. LR is also particularly compatible with several bias mitigation techniques, as some post-processing methods—such as Prejudice Remover—internally rely on the learning dynamics of LR.

To ensure a fair and consistent comparison between mitigation approaches, we explicitly used the SAGA solver for all LR models, including those trained with reweighting, since the Prejudice Remover implementation in AIF360 is also based on LR with SAGA. This guarantees that any differences in results are attributable to the mitigation technique itself, not to discrepancies in the underlying optimization algorithm or hyperparameters.

We used the default configuration of the AI Fairness 360 (AIF360) toolkit for LR, based on the scikit-learn implementation, which uses the following default hyperparameters: penalty = l2 (L2 regularization), C = 1.0 (regularization strength), fit intercept = True (including an intercept term), and max_iter = 10 (maximum iterations for optimization). The LR model was evaluated using classification performance metrics, including accuracy, recall, precision, F1-score, and AUC-ROC. To ensure robust evaluation across different data partitions, the model was assessed using a 70-30 hold-out split and 5-fold cross-validation.

3.5. Fairness Evaluation

Fairness metrics can be computed as either absolute differences (where 0 indicates no disparity) or relative ratios (with 1 indicating fairness). In our experiments, we assess fairness across demographic groups based on two protected attributes: ethnicity and gender. For ethnicity, white individuals are considered the privileged group, and non-white individuals the unprivileged group; for gender, males are treated as the privileged group and females as the unprivileged group.

In this study, we employed three fairness metrics to assess potential disparities in the prediction outcomes of the developed models: Statistical Parity Difference (SPD) [Hardt et al. 2016], Equal Opportunity Difference (EOD) [Hardt et al. 2016], and Average Odds Difference (AOD) [Hardt et al. 2016]. All three metrics are difference-based, meaning they calculate the absolute difference between two groups for specific prediction outcomes. An ideal value of 0 for each metric indicates perfect fairness, signifying no disparity between the groups. These metrics are described as follows.

- **SPD**: Measures fairness as the absolute difference in Positive Prediction Rates (PPR) between groups, ensuring that each group has the same likelihood of being classified with a positive outcome [Caton and Haas 2024]. For example, in diagnosing diabetes, SPD assesses whether the probability of receiving a positive prediction is independent of race or ethnicity.
- **EOD**: Evaluates whether the privileged and unprivileged groups have equal True Positive Rates (TPR), quantifying the absolute difference in TPR between them [Hardt et al. 2016]. In the diabetes context, EOD checks if individuals with diabetes from both groups are equally likely to be correctly identified as having the condition.

- **AOD**: Computes the average absolute difference in both the TPR and the False Positive Rate (FPR) between groups. This metric captures disparities in both error types and reflects the overall difference in model behavior across groups. For example, in diagnosing diabetes, AOD assesses whether the disparity in both the rate of correct diagnoses (TPR) and incorrect diagnoses (FPR) differs between demographic groups.

3.6. Bias mitigation strategies

The bias mitigation strategies, commonly grouped into pre-processing (before model training), in-processing (during model training), and post-processing, are used to ensure that fairness goals are achieved. Each approach has its strengths and limitations, and the choice of which strategy to apply depends on the specific context, the nature of the data, and the desired outcomes for fairness.

Pre-processing techniques aim to reduce underlying biases by modifying the data before model training [Mehrabi et al. 2021, Caton and Haas 2024], and are especially suitable when direct adjustments to the training data are feasible. One well-known approach is the Reweighting technique [Blow et al. 2024], which adjusts the weights of individual samples to promote fairer representation. These weights are then incorporated during training to influence the model's learning process. This method is particularly effective in addressing imbalances when certain demographic groups are underrepresented in the dataset.

In-processing techniques, on the other hand, adjust the learning process itself to minimize bias. These methods require direct access to the model, making them suitable for scenarios where the model is not treated as a black box [Caton and Haas 2024]. An example is the Prejudice Remover [Kamishima et al. 2012], which introduces a regularization term into the objective function of a classifier (specifically LR) to penalize dependence between the predicted outcome and a sensitive attribute (e.g., ethnicity or gender). This approach discourages the model from learning biased associations, thereby reducing discriminatory behavior during training.

Lastly, post-processing techniques focus on adjusting the predictions after model training to reduce unfair outcomes [Caton and Haas 2024]. Because these methods do not modify the model itself, they are often employed when altering the learning process or when data is not available.

3.7. Results

This section presents the evaluation results of the LR model for diabetes diagnosis using the BRFSS dataset. We first assess the model's performance in terms of accuracy, recall, precision, F1-score, and AUC-ROC. Next, we evaluate fairness by analyzing the model's behavior concerning ethnicity and gender, the two protected attributes. We then present the results of applying the bias mitigation techniques, Reweighting and Prejudice Remover, with detailed outcomes for each approach. Finally, we discuss the trade-off between fairness and accuracy. All the code used in this study is publicly available on GitHub [1], including data preprocessing, graphic visualization, and bias mitigation. Addi-

[1] `https://github.com/luisafelixx/Fairness_Diabetes`

tional graphics that could not be included in this paper are also available there, providing further insight into the results and data distributions.

3.7.1. Performance evaluation

The diabetes diagnostic model using Logistic Regression (LR) achieved high accuracy (0.84728) and ROC-AUC (0.8181), but relatively low precision (0.54950), recall (0.18622), and F1-score (0.2781). These performance results largely reflect the imbalanced distribution of positive diabetes cases, which represent only 15% of the dataset.

While class balancing techniques, such as subgroup-based resampling, can effectively improve performance metrics like precision [Chang et al. 2022, Pias et al. 2025], these techniques fall outside the scope of this study. Our primary objective is not to enhance predictive performance but to address fairness concerns. Thus, we focus on assessing and mitigating algorithmic bias directly within the inherently imbalanced BRFSS dataset, emphasizing fairness metrics and the application of bias mitigation techniques.

3.7.2. Assessing Fairness

Table 2 reports fairness metrics before bias mitigation, focusing on ethnicity and gender. The privileged groups are defined as white (for ethnicity) and male (for gender). In contrast, non-white and female individuals are considered unprivileged.

Fairness Metric	Ethnicity	Gender
SPD	0.075	-0.005
EOD	0.162	0.017
AOD	0.101	0.006

Table 2. Fairness metrics without bias mitigation for ethnicity and gender.

The SPD score for ethnicity (0.075) indicates that the unprivileged group (non-white) receives fewer favorable outcomes compared to the privileged group (white). This value suggests a moderate disparity, favoring white individuals in overall prediction rates. In contrast, the SPD score for gender (-0.005) indicates negligible bias in the overall rate of diabetes predictions, suggesting that the model treats male and female individuals similarly in terms of outcome distribution.

The EOD score for ethnicity (0.162) indicates a substantial disparity in TPR between the groups. Specifically, non-white individuals with diabetes are less likely to be correctly identified compared to white individuals. In contrast, the EOD score for gender (0.017) suggests a very small disparity, with females being slightly less likely to be accurately diagnosed than males.

The AOD score for ethnicity (0.101) suggests that non-white individuals (unprivileged group) have both a lower true positive rate (TPR) and potentially a higher false positive rate (FPR) compared to white individuals (privileged group). In other words, the model is less accurate in correctly identifying diabetes in non-white individuals and is more likely to wrongly classify them as having diabetes. In contrast, the AOD score for

gender (0.006) indicates an insignificant disparity, suggesting that the model's accuracy in true and false positive predictions is nearly equal between males and females.

3.7.3. Bias mitigation results

We applied the Reweighting pre-processing and the Prejudice Remover in-processing mitigation strategies for the protected attribute ethnicity, while we did not apply any bias mitigation strategy for gender because the fairness metrics did not indicate significant bias against this protected attribute. Table 3 shows the effectiveness of each bias mitigation strategy in reducing disparities across various fairness metrics (SPD, EOD, and AOD), with ethnicity as a protected attribute, where White is considered the privileged group and Non-White is the unprivileged group.

	No mitigation	Reweighting	Prejudice Remover
SPD	0.075048	0.014945	**0.014301**
EOD	0.162955	**0.002936**	0.006501
AOD	0.101394	**0.003856**	0.004988

Table 3. Bias Mitigation Results for Ethnicity.

Both the Reweighting bias mitigation strategy and the Prejudice Remover performed very well overall, improving all three fairness metrics and effectively mitigating the bias previously identified, as shown in Table 3. Figure 2 compares fairness scores across the two mitigation strategies and no mitigation for ethnicity.

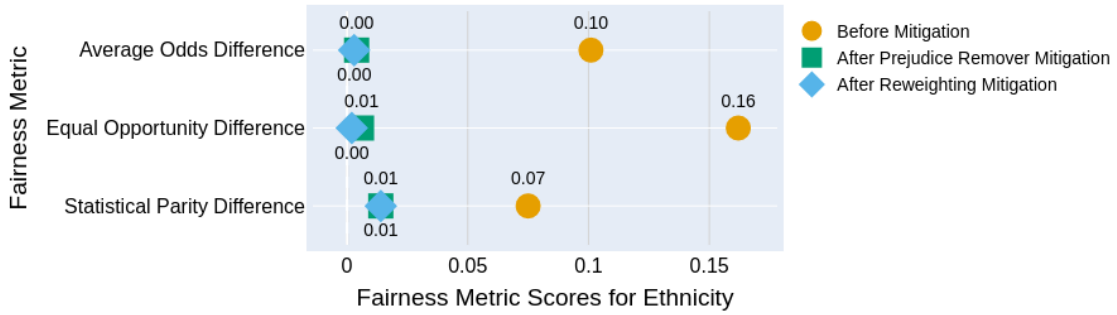

Figure 2. Fairness improvements for Ethnicity.

3.8. The Fairness-Accuracy trade-off

The trade-off between fairness and accuracy refers to the balance between improving fairness and maintaining model accuracy. A minimal change in accuracy suggests that the bias mitigation strategies had a limited impact on the model's predictive performance while still addressing fairness concerns. In our experiment, the Reweighting bias mitigation strategy showed no effect on the model's accuracy, remaining at 0.847284. Similarly, the Prejudice Remover strategy demonstrated minimal impact on accuracy, changing slightly from 0.847284 to 0.849405.

4. Discussion

This section discusses the key findings of our study regarding the ML model applied to diabetes diagnosis. We interpret the results of the fairness metrics in light of demographic imbalances, evaluate the effectiveness of the applied bias mitigation techniques, and reflect on the trade-offs between fairness and accuracy. Additionally, we consider the implications of disease prevalence on fairness evaluation and discuss the limitations of our approach.

4.1. Data imbalance and bias

While the gender distribution in the BRFSS dataset is relatively balanced, the ethnicity distribution is notably skewed, with white individuals comprising the majority. This imbalance may contribute to the observed disparities in the model's predictions, particularly regarding fairness metrics. Notably, our findings revealed more significant bias related to ethnicity than gender, which had no significant bias found. This aligns with prior literature suggesting that models trained on imbalanced data tend to favor the majority class or group.

It is important to highlight that data balancing techniques, such as resampling and reweighting, are closely related to bias mitigation approaches. Reweighting, for instance, not only addresses class imbalance but also corrects for demographic imbalances that may otherwise propagate or amplify bias. However, since our study focuses on fairness mitigation rather than improving classification performance, we limited our balancing efforts to fairness-aware strategies like Reweighting and Prejudice Remover rather than class-balancing techniques.

4.2. Reweighting versus Prejudice Remover

The Reweighting bias mitigation approach acts in the pre-processing ML stage, directly modifies the instance weights before model training, which has a stronger influence on correcting disparities in imbalanced datasets. In contrast, the Prejudice Remover bias mitigation strategy acts during the in-processing stage and may have more limited capacity to fully compensate for strong underlying data imbalances. Although they are very different strategies and complement each other, in our experiment, both strategies completely mitigated bias.

4.3. Fairness Metrics and Disparities

Fairness metrics such as SPD, EOD, and AOD are influenced not only by model behavior but also by the underlying distribution of the target variable (disease prevalence) across demographic groups. For instance, if diabetes prevalence differs significantly between groups, fairness metrics might reflect both model bias and actual disparities in health outcomes. These disparities are also driven by biological factors, in addition to social determinants of health (SDOH) such as income, education, access to healthcare, and environmental conditions, which contribute to unequal disease burdens across populations. Therefore, interpreting fairness metrics requires considering the context of real-world demographic, biological, and social differences. This type of interpretation benefits from the involvement of disease specialists and public health experts, who can contextualize the results within broader epidemiological and social frameworks.

4.4. Limitations

This study presents several limitations. First, while we focused on gender and ethnicity as protected attributes, other relevant factors such as socioeconomic status, geographic region, or age were not analyzed for fairness. Second, we binarized ethnicity into White and Non-White, which simplifies complex social identities and may mask important differences within diverse Non-White groups. While useful for analysis, the results should be interpreted cautiously, acknowledging these limitations. Third, our model used default hyperparameters and did not undergo extensive optimization, which may affect both performance and fairness outcomes. Fourth, we employed only two bias mitigation strategies (Reweighting and Prejudice Remover); other techniques, including post-processing or hybrid methods, might yield different results. Finally, interpreting fairness metrics in health applications is inherently complex and requires interdisciplinary expertise. Our analysis did not involve clinical or public health professionals, which may limit the contextualization of disparities identified by the model.

5. Conclusion

This study investigated fairness in ML models for diabetes diagnosis using the BRFSS dataset. While the logistic regression model achieved high accuracy and AUC-ROC, performance metrics such as precision, recall, and F1-score were limited due to the imbalanced distribution of diabetes cases. More importantly, fairness evaluations revealed disparities across protected attributes, particularly ethnicity, with non-white individuals receiving less favorable outcomes. We applied the Reweighting and Prejudice Remover bias mitigation strategies to address these disparities. Reweighting yielded better overall fairness improvements, especially for the ethnicity attribute, while it had minimal impact on model accuracy.

Our findings highlight the importance of integrating fairness assessments into health-related predictive modeling, especially when working with imbalanced or demographically skewed data. However, interpreting fairness metrics in healthcare contexts must go beyond technical evaluation, requiring input from domain experts in epidemiology and public health. In future work, we will explore additional mitigation techniques, including broader sets of protected attributes, and incorporate domain expertise to strengthen fairness analyses in health prediction models.

6. Acknowledgments

LR is supported by the PIND (Incentive Program for New Teachers, grant no. 519.287) and the New Faculty Development Program – Unicamp (grant no. 519.287), both funded by FAEPEX (Support Fund for Teaching, Research, and Extension) at UNICAMP. LF is supported by the Scientific Initiation Program funded by UNICAMP. AST receives support from the National Council for Scientific and Technological Development (CNPq), under grants 308059/2022-0 and 441817/2023-8.

References

[Association 2023] Association, A. D. (2023). Standards of medical care in diabetes. *Diabetes Care*, 46(Supplement 1):S1–S291.

[Bhatti et al. 2025] Bhatti, A., Sandrock, T., and Nienkemper-Swanepoel, J. (2025). The influence of missing data mechanisms and simple missing data handling techniques on fairness. arXiv preprint arXiv:2503.07313.

[Blow et al. 2024] Blow, C. H., Qian, L., Gibson, C., Obiomon, P., and Dong, X. (2024). Comprehensive validation on reweighting samples for bias mitigation via aif360. *Applied Sciences*, 14(9):3826.

[Caton and Haas 2024] Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

[Chang et al. 2022] Chang, V., Ganatra, M. A., Hall, K., Golightly, L., and Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics*, 2:100118.

[Cronjé et al. 2023] Cronjé, H., Katsiferis, A., Elsenburg, L., Andersen, T., Rod, N., et al. (2023). Assessing racial bias in type 2 diabetes risk prediction algorithms. *PLOS Global Public Health*, 3(5):e0001556.

[Dwork et al. 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). Fairness through awareness. In Goldwasser, S., editor, *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, pages 214–226. ACM.

[GBD 2021 Diabetes Collaborators 2023] GBD 2021 Diabetes Collaborators (2023). Global, regional, and national burden of diabetes from 1990 to 2021: a systematic analysis for the global burden of disease study 2021. *The Lancet*, 402(10397):203–234.

[Hardt et al. 2016] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

[Huang et al. 2022] Huang, J., Galal, G., Etemadi, M., and Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Med Inform*, 10(5):e36388.

[Huang et al. 2024] Huang, Y., Guo, J., Chen, W.-H., Lin, H.-Y., Tang, H., Wang, F., Xu, H., and Bian, J. (2024). A scoping review of fair machine learning techniques when using real-world data. *medRxiv*.

[Kamishima et al. 2012] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012)*, pages 35–50. Springer.

[Khanam and Foo 2021] Khanam, J. J. and Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4):432–439.

[Klement and El Emam 2023] Klement, W. and El Emam, K. (2023). Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: Development and validation. *J Med Internet Res*, 25:e48763.

[Kush Varshney 2018] Kush Varshney (2018). Introducing ai fairness 360. `https://research.ibm.com/blog/ai-fairness-360`. Accessed: 2025-04-27.

[Mehrabi et al. 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

[Pias et al. 2025] Pias, T. S., Su, Y., Tang, X., Wang, H., Faghani, S., and Yao, D. (2025). Enhancing fairness and accuracy in diagnosing type 2 diabetes in young adult population. *IEEE Journal of Biomedical and Health Informatics*. Online ahead of print.

[Raza 2022] Raza, S. (2022). A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission. *Healthcare Analytics*, 2:100100.

[Raza 2023] Raza, S. (2023). Connecting fairness in machine learning with public health equity. In *Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 704–708. IEEE.

[Ruback et al. 2022] Ruback, L., Carvalho, D., and Avila, S. (2022). Mitigating bias in machine learning: A socio-technical analysis. *iSys - Brazilian Journal of Information Systems*, 15(1):23:1–23:31.

[Talebi Moghaddam et al. 2024] Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z., Dehghan, A., Khaleghi, M., Sharafi, M., and Nikfar, G. (2024). Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm. *BMC Medical Research Methodology*, 24(1):220.

[Varshney 2022] Varshney, K. R. (2022). *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA.

[Verma and Rubin 2018] Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA. Association for Computing Machinery.

[Wang et al. 2024] Wang, S. C. Y., Nickel, G., Venkatesh, K. P., Raza, M. M., and Kvedar, J. C. (2024). Ai-based diabetes care: risk prediction models and implementation concerns. *NPJ Digital Medicine*, 7(1):36.

[Xie et al. 2019] Xie, Z., Nikolayeva, O., Luo, J., and Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16:E130.