CS 108

Data Science Ethics

Fall 2023

Project Report

Diabetes Predictor: An In-depth Analysis and Predictive Modeling

Aditya Gambhir, Kush Momya, Shaan Palaka, Harjyot Sidhu

March 17, 2024

Introduction

Objective

The primary goal of this project is to develop a machine-learning model capable of accurately predicting whether a patient is at risk for diabetes based on a multitude of health-related attributes [1]. In addition, we are committed to identifying and addressing any potential biases or unethical consequences that may arise from the application of our model, ensuring its fairness and integrity [2, 3].

Dataset

Initial Description

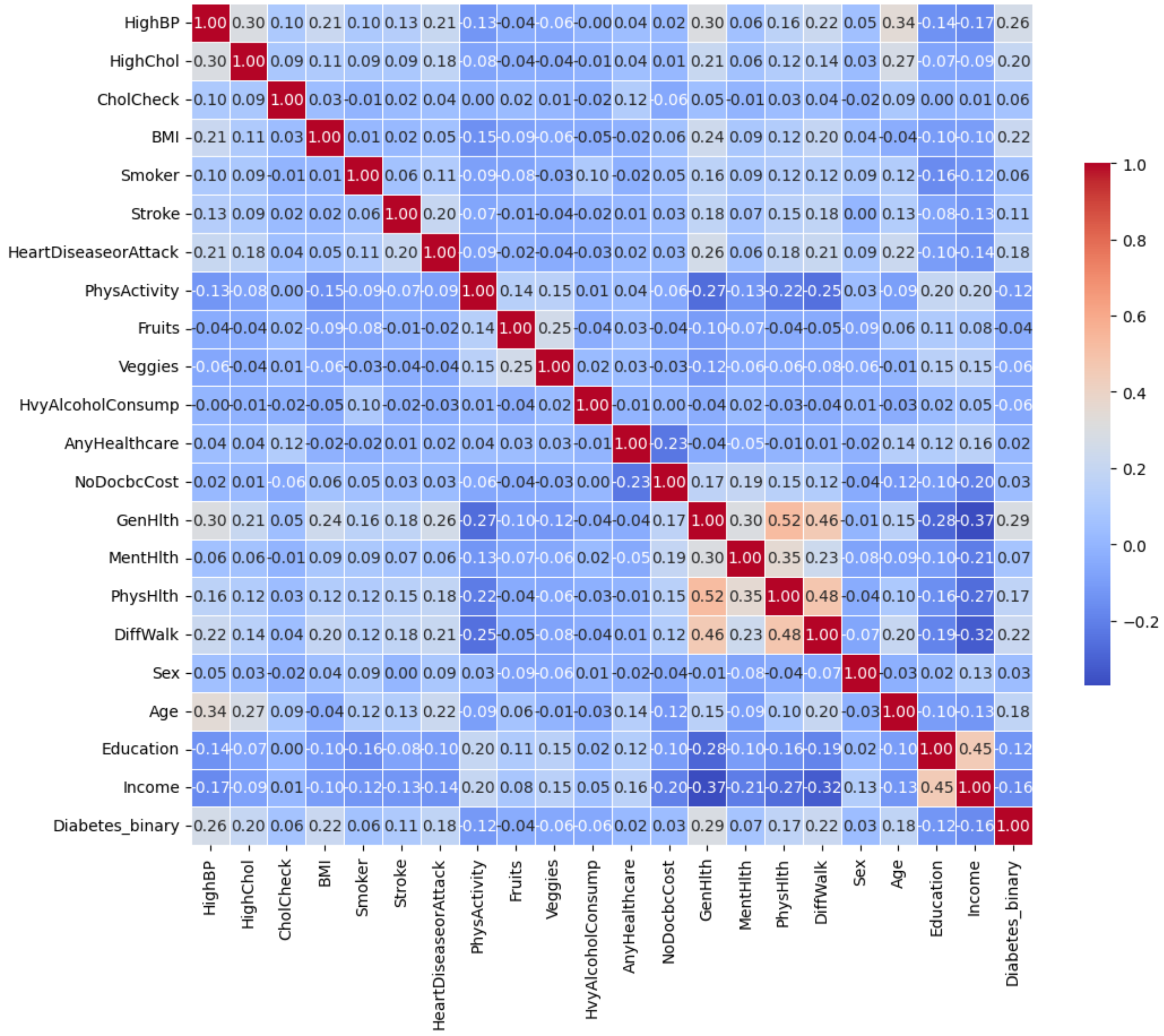
Our research project utilizes an extensive dataset derived from the Centers for Disease Control and Prevention (CDC), encompassing a wide array of health indicators across a substantial cohort. This dataset is comprised of 253,680 entries, each representing individual responses across 22 distinct health indicators, such as blood pressure, cholesterol levels, dietary habits, physical activity, and the presence of various health conditions[4].

A notable aspect of this dataset is its meticulous organization and completeness, with no missing values across all entries, which facilitates a straightforward analysis process[5]. This integrity allows us to focus on the heart of our analysis—feature evaluation and model development without the need for extensive preprocessing to handle missing data.

The dataset encompasses both numerical and categorical variables, ranging from basic demographic information to detailed health-related metrics. For example, indicators like 'HighBP' (High Blood Pressure), 'HighChol' (High Cholesterol), and 'BMI' (Body Mass Index) are quantified alongside lifestyle factors such as smoking status and fruit and vegetable consumption. A critical variable within this dataset is 'Diabetes_binary', a binary indicator denoting the presence or absence of diabetes, which serves as the target variable for our predictive modeling.

Statistical analysis of the dataset reveals a diverse range of values across the variables. For instance, the mean BMI of the cohort is 28.38, with a standard deviation of 6.61, indicating a wide variance in body weight. Similarly, the binary variables, such as 'Smoker' and 'Stroke', show a distribution reflective of the population's health characteristics. The 'Age' variable, treated categorically, spans from 1 to 13, representing various age groups, which could provide insightful correlations with the diabetes outcome.

This dataset's breadth and depth, characterized by its wide range of health indicators and the large sample size, offer a robust foundation for building a machine-learning model[6]. Our objective is to leverage this dataset to predict diabetes risk accurately while ensuring our model's fairness and ethical integrity by addressing and mitigating any potential biases inherent in the data.



	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Diabetes_binary
HighBP	1.00	0.30	0.10	0.21	0.10	0.13	0.21	-0.13	-0.04	-0.06	-0.00	0.04	0.02	0.30	0.06	0.16	0.22	0.05	0.34	-0.14	-0.17	0.26
HighChol	0.30	1.00	0.09	0.11	0.09	0.09	0.18	-0.08	-0.04	-0.04	-0.01	0.04	0.01	0.21	0.06	0.12	0.14	0.03	0.27	-0.07	-0.09	0.20
CholCheck	0.10	0.09	1.00	0.03	-0.01	0.02	0.04	0.00	0.02	0.01	-0.02	0.12	-0.06	0.05	-0.01	0.03	0.04	-0.02	0.09	0.00	0.01	0.06
BMI	0.21	0.11	0.03	1.00	0.01	0.02	0.05	-0.15	-0.09	-0.06	-0.05	-0.02	0.06	0.24	0.09	0.12	0.20	0.04	-0.04	-0.10	-0.10	0.22
Smoker	0.10	0.09	-0.01	0.01	1.00	0.06	0.11	-0.09	-0.08	-0.03	0.10	-0.02	0.05	0.16	0.09	0.12	0.12	0.09	0.12	-0.16	-0.12	0.06
Stroke	0.13	0.09	0.02	0.02	0.06	1.00	0.20	-0.07	-0.01	-0.04	-0.02	0.01	0.03	0.18	0.07	0.15	0.18	0.00	0.13	-0.08	-0.13	0.11
HeartDiseaseorAttack	0.21	0.18	0.04	0.05	0.11	0.20	1.00	-0.09	-0.02	-0.04	-0.03	0.02	0.03	0.26	0.06	0.18	0.21	0.09	0.22	-0.10	-0.14	0.18
PhysActivity	-0.13	-0.08	0.00	-0.15	-0.09	-0.07	-0.09	1.00	0.14	0.15	0.01	0.04	-0.06	-0.27	-0.13	-0.22	-0.25	0.03	-0.09	0.20	0.20	-0.12
Fruits	-0.04	-0.04	0.02	-0.09	-0.08	-0.01	-0.02	0.14	1.00	0.25	-0.04	0.03	-0.04	-0.10	-0.07	-0.04	-0.05	-0.09	0.06	0.11	0.08	-0.04
Veggies	-0.06	-0.04	0.01	-0.06	-0.03	-0.04	-0.04	0.15	0.25	1.00	0.02	0.03	-0.03	-0.12	-0.06	-0.06	-0.08	-0.06	-0.01	0.15	0.15	-0.06
HvyAlcoholConsump	-0.00	-0.01	-0.02	-0.05	0.10	-0.02	-0.03	0.01	-0.04	0.02	1.00	-0.01	0.00	-0.04	0.02	-0.03	-0.04	0.01	-0.03	0.02	0.05	-0.06
AnyHealthcare	0.04	0.04	0.12	-0.02	-0.02	0.01	0.02	0.04	0.03	0.03	-0.01	1.00	-0.23	-0.04	-0.05	-0.01	0.01	-0.02	0.14	0.12	0.16	0.02
NoDocbcCost	0.02	0.01	-0.06	0.06	0.05	0.03	0.03	-0.06	-0.04	-0.03	0.00	-0.23	1.00	0.17	0.19	0.15	0.12	-0.04	-0.12	-0.10	-0.20	0.03
GenHlth	0.30	0.21	0.05	0.24	0.16	0.18	0.26	-0.27	-0.10	-0.12	-0.04	-0.04	0.17	1.00	0.30	0.52	0.46	-0.01	0.15	-0.28	-0.37	0.29
MentHlth	0.06	0.06	-0.01	0.09	0.09	0.07	0.06	-0.13	-0.07	-0.06	0.02	-0.05	0.19	0.30	1.00	0.35	0.23	-0.08	-0.09	-0.10	-0.21	0.07
PhysHlth	0.16	0.12	0.03	0.12	0.12	0.15	0.18	-0.22	-0.04	-0.06	-0.03	-0.01	0.15	0.52	0.35	1.00	0.48	-0.04	0.10	-0.16	-0.27	0.17
DiffWalk	0.22	0.14	0.04	0.20	0.12	0.18	0.21	-0.25	-0.05	-0.08	-0.04	0.01	0.12	0.46	0.23	0.48	1.00	-0.07	0.20	-0.19	-0.32	0.22
Sex	0.05	0.03	-0.02	0.04	0.09	0.00	0.09	0.03	-0.09	-0.06	0.01	-0.02	-0.04	-0.01	-0.08	-0.04	-0.07	1.00	-0.03	0.02	0.13	0.03
Age	0.34	0.27	0.09	-0.04	0.12	0.13	0.22	-0.09	0.06	-0.01	-0.03	0.14	-0.12	0.15	-0.09	0.10	0.20	-0.03	1.00	-0.10	-0.13	0.18
Education	-0.14	-0.07	0.00	-0.10	-0.16	-0.08	-0.10	0.20	0.11	0.15	0.02	0.12	-0.10	-0.28	-0.10	-0.16	-0.19	0.02	-0.10	1.00	0.45	-0.12
Income	-0.17	-0.09	0.01	-0.10	-0.12	-0.13	-0.14	0.20	0.08	0.15	0.05	0.16	-0.20	-0.37	-0.21	-0.27	-0.32	0.13	-0.13	0.45	1.00	-0.16
Diabetes_binary	0.26	0.20	0.06	0.22	0.06	0.11	0.18	-0.12	-0.04	-0.06	-0.06	0.02	0.03	0.29	0.07	0.17	0.22	0.03	0.18	-0.12	-0.16	1.00

Figure 1: Correlation Heatmap for Entire Dataset

Feature Analysis

The analysis of features entailed a detailed examination of all 22 health indicators, aiming to comprehend their distribution patterns, identify any outliers, and investigate their correlation with the diabetes outcome. Through meticulous correlation analysis, we identified several features that exhibited significant associations with the outcome of diabetes, guided by recent findings in healthcare predictive analytics[7]. These findings guided our feature selection strategy, which was carefully designed to minimize multicollinearity and maximize the interpretability of our model.

Data Preprocessing

Data preprocessing emerged as a crucial step in refining our dataset for the predictive modeling process. We embarked on this phase by performing a meticulous examination for missing values across the dataset and subsequently eliminated any such instances to guarantee a smooth and uninterrupted training phase. This proactive measure ensured the integrity and completeness of our data, setting a solid foundation for the subsequent stages of model preparation.

Following the initial cleanup, we directed our focus toward the normalization of a selected subset of features, specifically targeting variables that significantly influence the model's learning capacity and predictive accuracy[8]. Utilizing the MinMaxScaler, we normalized the values of key variables including 'BMI', 'GenHlth' (General Health), 'MentHlth' (Mental Health), 'PhysHlth' (Physical Health), 'Age', 'Education', and 'Income'. This scaling process transformed the selected features to a uniform range between 0 and 1, thereby facilitating a more effective learning process by equalizing the scale of numerical inputs.

In the final step of our preprocessing journey, we addressed the categorical variables within our dataset. Through the application of appropriate encoding techniques, we transformed these categorical variables into a machine-readable format, enabling their utilization in predictive modeling[9]. This transformation is crucial for incorporating the rich, categorical information contained within our dataset into the predictive modeling process, further enhancing the model's ability to discern and learn from complex patterns and relationships in the data.

Collectively, these preprocessing steps—ranging from the elimination of missing values and normalization of numerical features to the encoding of categorical variables—have meticulously prepared our dataset for the intricate demands of machine learning model development. This comprehensive approach to data preparation not only facilitates a more robust and effective modeling process but also underscores our commitment to leveraging data in a manner that maximizes predictive accuracy and model reliability[10].

Exploratory Data Analysis

General EDA

In the exploratory phase of our analysis, we embarked on a detailed exploration of the dataset to identify underlying patterns or significant predictors for the ‘diabetes_binary‘ outcome. Initial visualizations were focused on variables anticipated to be crucial in predicting diabetes risk.

A notable observation was the presence of disparities, particularly related to the ‘Sex‘ attribute, which we marked for further evaluation regarding its impact on model bias and fairness. Additionally, we encountered a pronounced scarcity of positive instances within the dataset, a crucial factor that warranted subsequent adjustment strategies. To mitigate the challenge posed by the dataset's high dimensionality and facilitate a more efficient model fitting, Principal Component Analysis (PCA) was implemented as a preliminary step before model initialization[11].

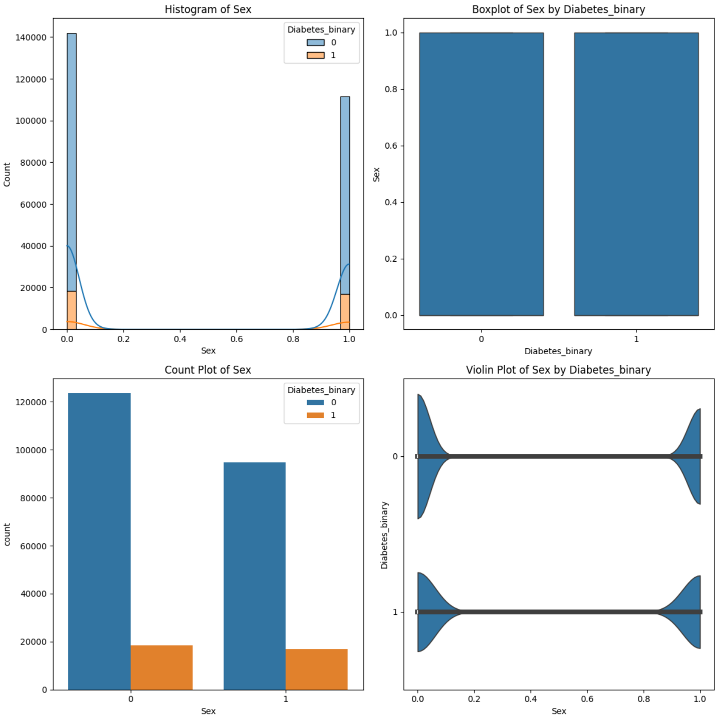


Figure 2: Sample of EDA

Protected Attributes

Our investigation into protected attributes highlighted significant disparities that could potentially influence the model's fairness and bias. The analysis was conducted across multiple attributes, namely 'Sex', 'Age', 'Education', and 'Income', with each showing varying levels of impact on the model's predictions.

'Sex' Attribute Analysis: The Disparate Impact Analysis yielded a ratio of 0.8554 (Ideal: 1), indicating a notable disparity. Specifically, the positive outcomes ratios for Groups 0 and 1 were 0.9307 and 1.0881, respectively, suggesting an imbalance that merits attention for fairness corrections. Demographic Parity Differences and Equality of Opportunity Differences were also computed, showing deviations from the ideal values, thereby confirming the presence of bias in this attribute[12].

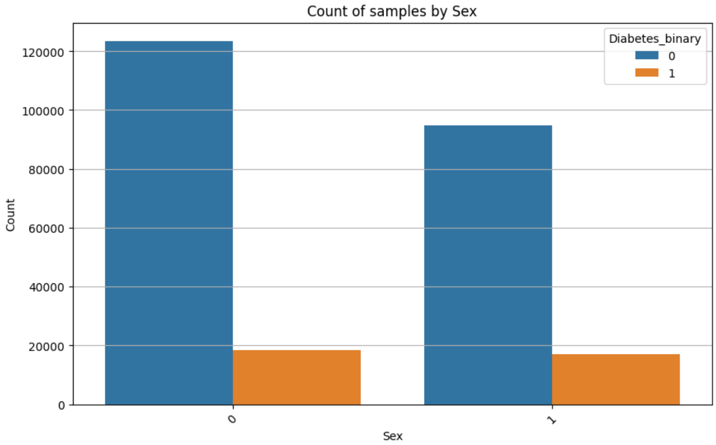


Figure 3: Count plot for Sex Attribute

'Age' Attribute Analysis: The analysis of the 'Age' attribute revealed a Disparate Impact Analysis ratio of 0.0626, significantly deviating from the ideal. This analysis detailed the progression of positive outcomes ratios across age groups, illustrating a clear trend that could affect the model's fairness. The Demographic Parity Differences and Equality of Opportunity Differences across all age groups consistently highlighted disparities[13], emphasizing the need for careful consideration of age as a protected attribute in our model.

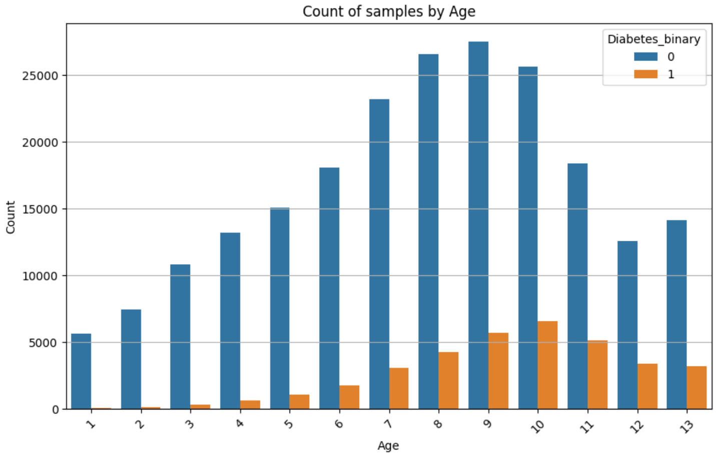


Figure 4: Count plot for Age Attribute

'Education' Attribute Analysis: For the 'Education' attribute, a Disparate Impact Analysis ratio of 0.3312 was observed, indicating potential bias[14]. The positive outcomes ratios across educational groups varied widely, with higher education levels generally correlating with better outcomes. This trend underscores the influence of education on model predictions and highlights the importance of adjusting for such biases.

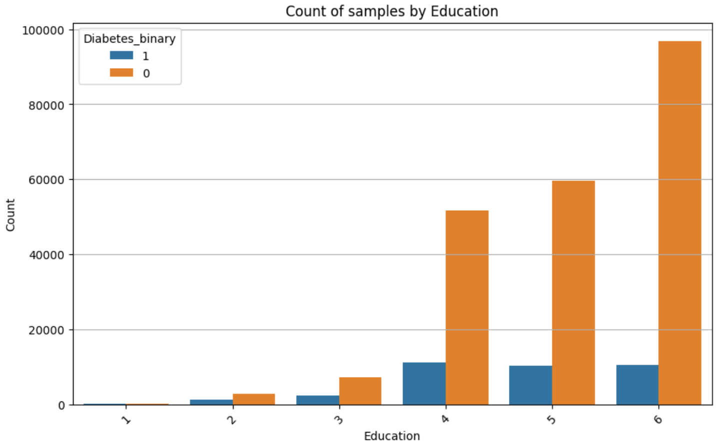


Figure 5: Count plot for Education Attribute

'Income' Attribute Analysis: The 'Income' attribute's Disparate Impact Analysis ratio stood at 0.3039, suggesting significant disparities based on income levels. The analysis showed a gradient in positive outcomes ratios, with higher income groups generally experiencing better predictions. These findings necessitate adjustments in the model to mitigate income-based disparities and ensure equitable predictions across all income groups.

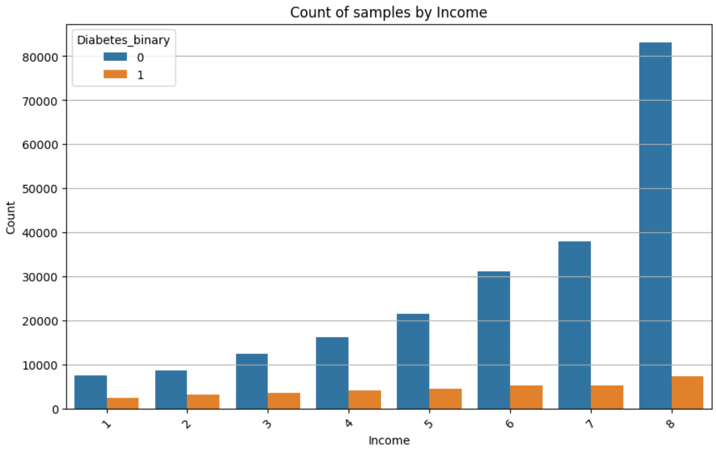


Figure 6: Count plot for Income Attribute

Stacking Ensemble

Model Preparation

Principal Component Analysis(PCA)

The implementation of PCA in our project was pivotal for dimensionality reduction and feature extraction. To retain significant variance while transforming the feature space, PCA reduced the dataset to two principal components[15].

The cumulative explained variance by these components was 25%[15], as evidenced by the following plot:

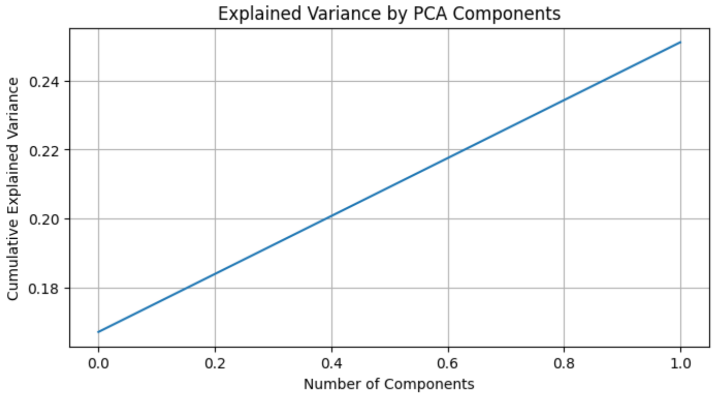


Figure 7: Explained Variance by PCA Components

Moreover, we visualized the dataset in the reduced two-dimensional space, enhancing the aesthetics to aid interpretability. The following scatter plot illustrates the distribution of the dataset across the first two principal components, color-coded by the diabetes label:

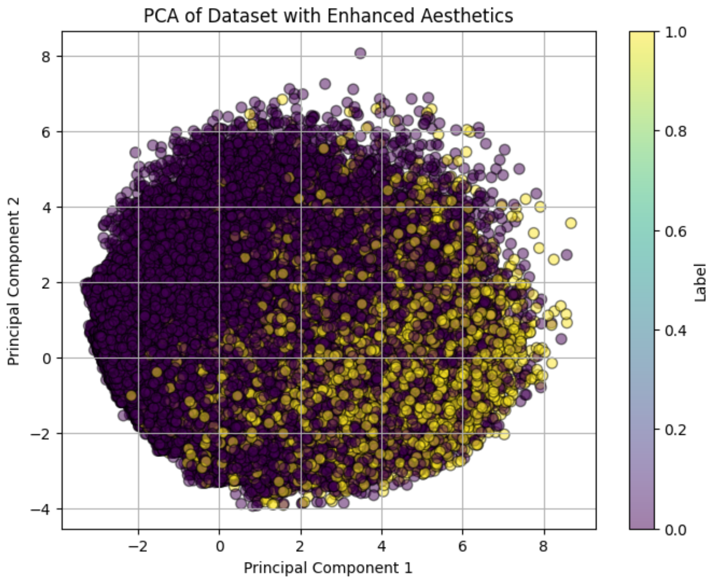


Figure 8: PCA of Dataset with Enhanced Aesthetics

To interpret the principal components, we examined the PCA loadings[15], which reflect the correlation between the original features and the components. The following bar chart showcases the loadings for the first principal component, highlighting the influence of each feature on this component:

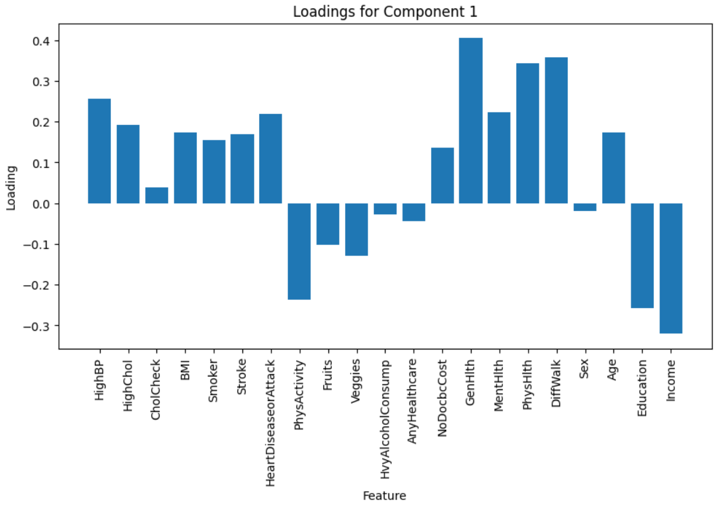


Figure 9: Loadings for Component 1

Stacking Classifiers

In this segment, we elucidate the construction and training of a stacking ensemble model, designed to harness the collective prowess of diverse machine learning algorithms for our binary classification task. Stacking, or stacked generalization[16], is an ensemble learning paradigm that amalgamates the predictions from multiple base classifiers via a final estimator, thereby amplifying the predictive acumen of the model.

Base Classifiers: We commence with the instantiation of an array of classifiers, each equipped with a meticulously selected suite of hyperparameters. These classifiers, functioning as the foundational learners in our stacking ensemble, include:

- *Logistic Regression (LR)*[17]: A linear classification model employing L1 regularization and the SAGA solver for optimization purposes.
- *Random Forest (RF)*[18]: An ensemble technique founded on decision trees, which utilizes the Gini impurity criterion for node splitting and a square root function for feature selection at each bifurcation.
- *Multi-layer Perceptron (NN)*[19]: A neural network architecture comprising two hidden layers, deploying the ReLU activation function and an adaptive learning rate scheme.
- *XGBoost (XGB)*[20]: A gradient boosting framework underpinned by decision trees, with hyperparameters, fine-tuned to mitigate overfitting and bolster model efficacy.
- *Gaussian Naive Bayes (NB)*[21]: A probabilistic classifier predicated on Bayes' theorem, with an underlying assumption of feature independence.

The selection of these classifiers is predicated on their ability to capture a diverse array of patterns from the data, thereby encompassing linear, ensemble, neural network, and probabilistic approaches.

Final Estimator: The ensuing phase involves the base classifiers' predictions (probability estimations) being utilized as input for the final estimator—another XGBoost model[22]. This final estimator is tasked with the optimization of the base classifiers' combined predictive output.

Stacking Ensemble Classifier: We employ the `StackingClassifier` from the `scikit-learn` ensemble module to synthesize the individual classifiers with the final estimator into a unified model[23]. The `stack_method='predict_proba'` parameter is set to leverage the base classifiers' probabilistic predictions for the training of the final estimator, which is essential for binary classification tasks.

Training the Model: The ensemble classifier is trained on the complete training dataset, with each base classifier learning from the data and the final estimator being trained on their predictions. This training regimen is computationally demanding, entailing the sequential training of multiple models. Consequently, we monitor and report the duration of the training to evaluate performance.

The objective of this stacking ensemble model is to amalgamate the individual strengths of each classifier while counterbalancing their inherent limitations, culminating in a robust and precise model. The convergence of these disparate models is anticipated to deliver superior predictive performance compared to the potential of a solitary classifier.

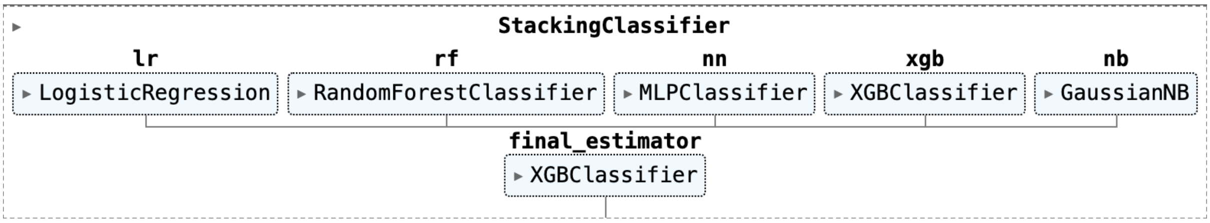


Figure 10: Trained Stacking Classifer

Bias Detection and Mitigation Strategies

Fairness by Unawareness

In pursuit of fairness through unawareness, we adapted our dataset to curtail potential biases, thus fostering a more equitable representation of classes[24]. We explored dual strategies: adjusting the training data by excising samples, and amplifying the minority class through oversampling[25]. The chosen tactics, grounded in the configuration of boolean flags—adjust_train_data and over-sample—yielded two distinct datasets.

Our process initiated with the original dataset, from which we derived a modified copy, designed to undergo bias mitigation. Should adjust_train_data be enacted without oversampling, we deliberately reduced the majority class instances, thereby attenuating class imbalance. Post-adjustment, we scrutinized the dataset, affirming its alignment with fairness principles by excluding sensitive attributes such as Sex, Age, Education, and Income, thus safeguarding against their undue influence in model training.

Balancing Target Class Subgroups

The imbalance within target classes posed an additional bias risk. To counteract this, we employed the Synthetic Minority Over-sampling Technique (SMOTE), which synthetically generated new instances of the minority class, thereby presenting a balanced perspective of the classes to the model[25]. In scenarios where both adjust_train_data and over-sample flags were activated, we further refined the training data, ensuring the minority class outnumbered the majority, compelling the model to accord greater consideration to the former.

This comprehensive approach to data preparation was visually validated; histograms depicting the class distribution before and after adjustments elucidated the alterations made to achieve class balance.

Threshold Calibration

Post-detection of biases within the dataset, we resorted to threshold calibration, a post-processing technique, to mitigate biases across various classes[3]. By adjusting the decision thresholds according to the disparate impacts identified, we aspired to realize a more nuanced balance of model predictions across demographic groups.

Model Performance Audit

A rigorous audit of the model's performance, encompassing sensitive attributes, provided insights into the bias landscape[26]. This was accomplished by appending sensitive attributes post-model training, enabling subgroup metrics analysis. The evaluation underscored the essence of threshold calibration, with performance metrics indicating improved fairness in the model's predictive capabilities.

Results and Conclusion

Results Summary

Our multifaceted exploration into diabetes prediction culminated in the construction of a stacking ensemble model[27]. This model judiciously integrates the predictive prowess of Logistic Regression, Random Forest, Multi-layer Perceptron, XGBoost, and Gaussian Naive Bayes classifiers. The ensemble's performance, adjudicated by a secondary XGBoost classifier, yielded commendable results, as delineated by the metrics tables below[28].

Metric	Basic Model	Masked Model
Accuracy	0.8666	0.8662
Precision	0.5686	0.5717
Sensitivity (Recall)	0.1450	0.1296
Specificity	0.9823	0.9844
F1-Score	0.2311	0.2113
Time Taken (s)	2465.1	722.8

Table 1: General Metrics for Basic and Unawareness Models

Metric	Basic Classifier	Masked Classifier
True Positive Rate (TPR)	0.1450	0.1296
False Positive Rate (FPR)	0.0177	0.0156
False Negative Rate (FNR)	0.8550	0.8704
Positive Predictive Value (PPV)	0.5686	0.5717
Negative Predictive Value (NPV)	0.8775	0.8758
Brier Score	0.0964	0.0981

Table 2: Detailed Metrics for Basic and Unawareness Classifiers

Bias assessment unearthed disparities in subgroup metrics, particularly across Age and Income brackets. Measures to counteract this, including the exclusion of sensitive attributes in the training phase, led to significant modifications in model predictions. These were aimed at fostering a balanced representation of diabetes risk, as depicted by the Aquetas plots and calibration curves for both full and masked datasets. [29, 30]

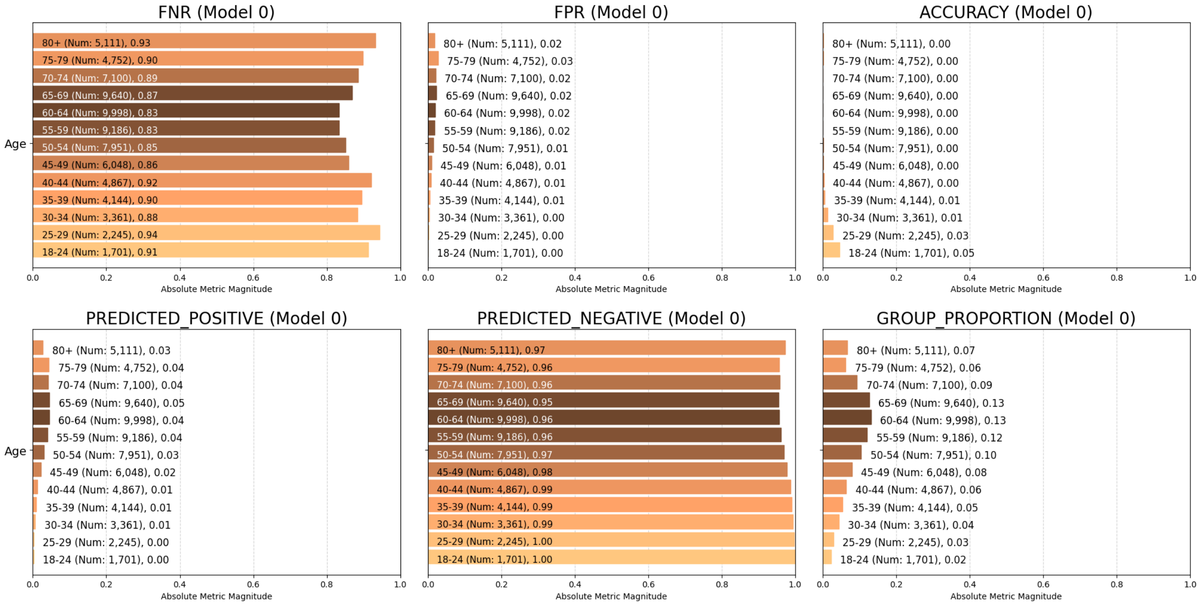


Figure 11: Subgroup Metrics by Age for the Basic Model

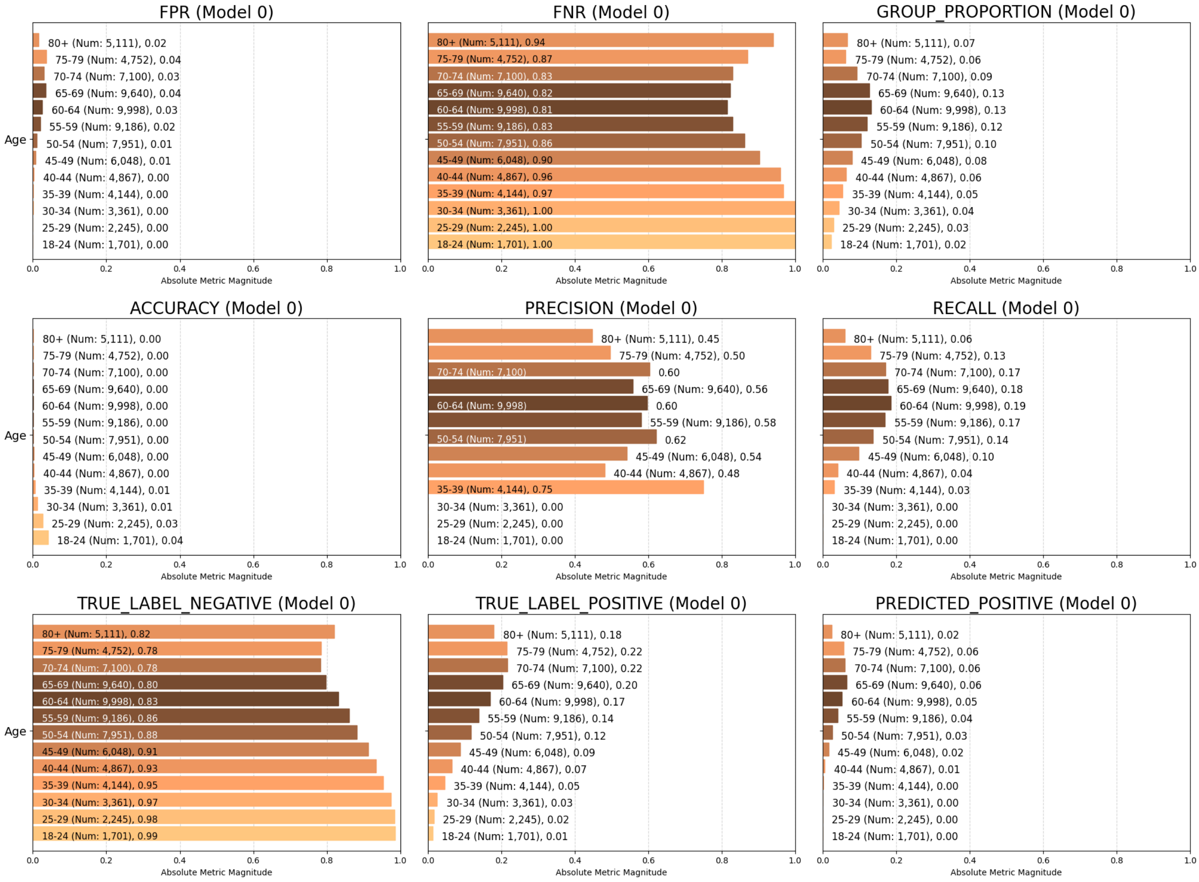


Figure 12: Subgroup Metrics by Age for the Unawareness Model

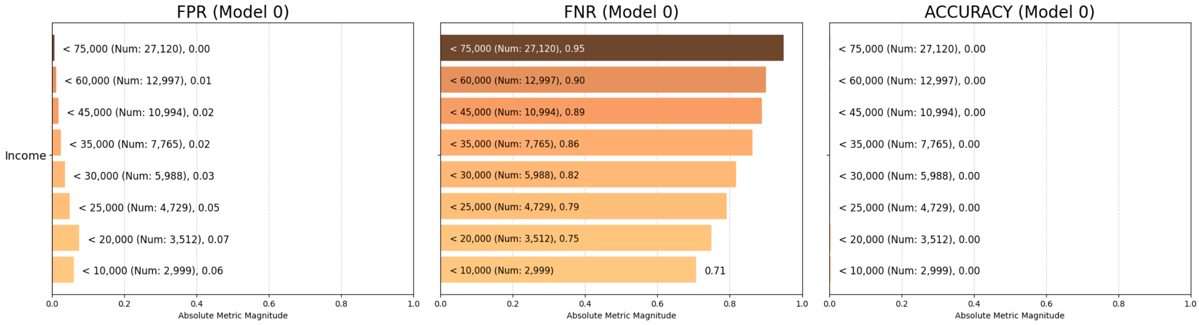


Figure 13: Subgroup Metrics by Income for the Basic Model

Calibration plots for the full dataset and the masked dataset illustrate the model's performance against the ideal of perfect calibration. Both plots demonstrate the models' fidelity in predicting diabetes presence, with the masked dataset model showcasing a slight deviation from perfect calibration. [31]

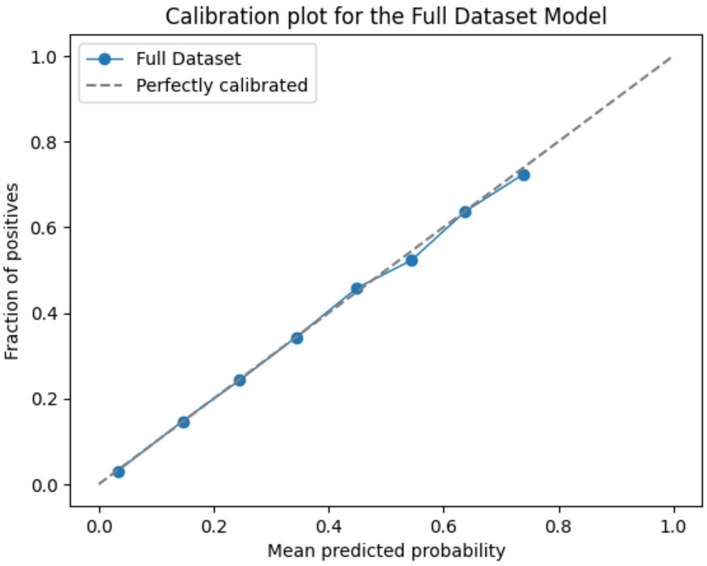


Figure 14: Calibration Plot for the Full Dataset Model

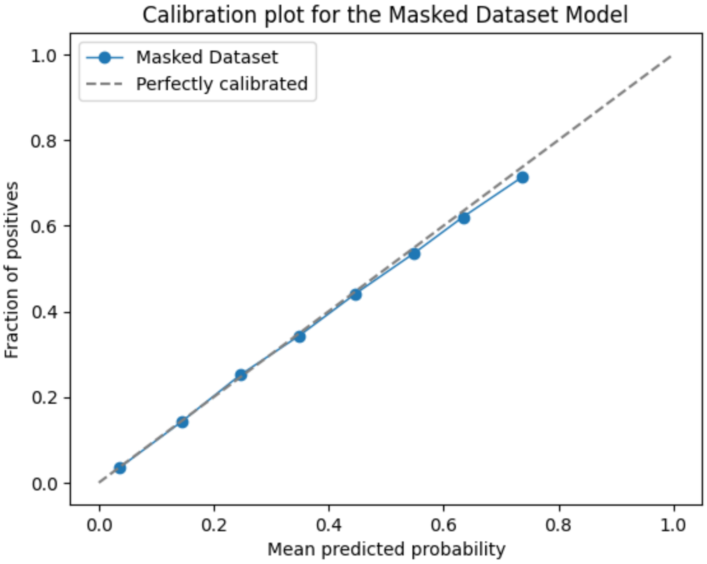
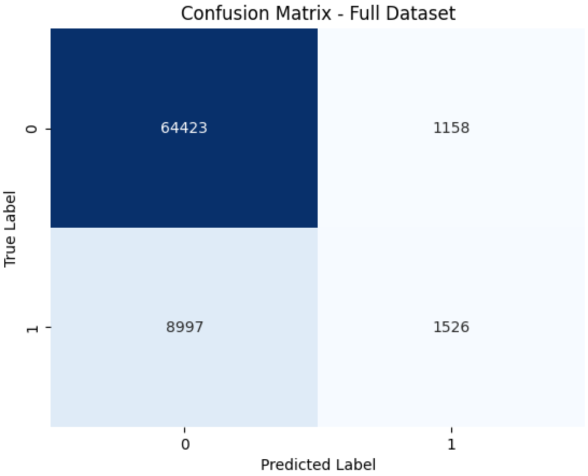


Figure 15: Calibration Plot for the Masked Dataset Model

Confusion matrices further delineate the accuracy of predictions for both the full and masked dataset models, emphasizing the modifications enacted post-bias mitigation strategies. [31]


Confusion Matrix - Full Dataset

True Label \ Predicted Label	0	1
0	64423	1158
1	8997	1526

Figure 16: Confusion Matrix for the Full Dataset Model

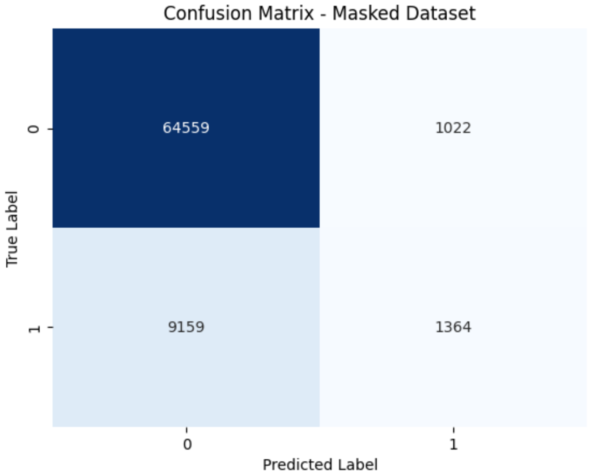


Figure 17: Confusion Matrix for the Masked Dataset Model

Analysis of Findings

Our investigation into diabetes risk prediction has emphasized the crucial balance between achieving model accuracy and ensuring fairness. Adjustments implemented to rectify imbalances and bias incurred a measurable impact on precision—a testament to the ethical conundrum faced in predictive modeling. This necessary trade-off underscores the criticality of equitability in algorithmic decision-making within healthcare, where biases can have significant ramifications. The project's discoveries are a clarion call to meticulously scrutinize and prepare data, advocating for ethical integrity in health-related AI applications.

Ethical Considerations

This endeavor is a reaffirmation of our commitment to embedding ethical tenets within the realm of data science and machine learning. The proactive identification and rectification of bias in our model iterate the broader dialogue on AI's ethical use in healthcare. By championing transparency, equity, and accountability, we endeavor to establish a precedent for the ethical deployment of AI, ensuring its implications are constructive and universally beneficial.

Future Work

The foundation established by our project invites further inquiry into fairness in predictive modeling. Subsequent efforts can build upon our methodologies, harnessing expansive and varied datasets, and broadening the application scope to other medical conditions. This forward-looking approach aims to guarantee that AI's advancements are equitably shared, thereby fostering an inclusive future where AI serves the collective good.

Conclusion

In summation, our venture has highlighted the symbiotic relationship between technical efficacy and moral responsibility in developing healthcare AI. By prioritizing fairness and diligently countering bias, we establish that ethical AI is an attainable reality, achieved through meticulous analysis and deliberate practice. Our contributions advocate for a just and equitable application of technology in healthcare—a paradigm where models are not merely tools but bearers of integrity, aiding every individual impartially.

Bibliography

[1] P. Sonar and K. JayaMalini, "Diabetes prediction using different machine learning approaches," in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019.

[2] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency,* (New York, NY, USA), Association for Computing Machinery, 2019. `https://doi.org/10.1145/3287560.3287598`.

[3] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016. `https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf`.

[4] M. Kahn, "Diabetes." UCI Machine Learning Repository. `https://doi.org/10.24432/C5T59G`.

[5] R. J. Little and D. B. Rubin, "Statistical analysis with missing data," *Wiley Series in Probability and Statistics*, vol. 793, 2019.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, 2017. `https://www.sciencedirect.com/science/article/pii/S2001037016300733`.

[7] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," *Journal of Electrical Systems and Information Technology*, vol. 10, 2023. `https://doi.org/10.1186/s43067-023-00108-y`.

[8] "Preprocessing data — scikit-learn 1.4.1 documentation." `https://scikit-learn.org/stable/modules/preprocessing.html`. Accessed: 2024-03-17.

[9] A. Data, "Data preprocessing: Steps, techniques, and importance in machine learning." `https://blog.arkondata.com/data-preprocessing-steps-techniques-and-importance-in-machine-learning/`. Accessed: 2024-03-17.

[10] J. Brownlee, "Tour of data preparation techniques for machine learning." `https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/`. Accessed: 2024-03-17.

[11] M. Nedyalkova *et al.*, "Diabetes mellitus type 2: Exploratory data analysis based on clinical reading," *Open Chemistry*, vol. 18, pp. 1041–1053, 2020. https://www.researchgate.net/publication/343764625_Diabetes_mellitus_type_2_Exploratory_data_analysis_based_on_clinical_reading.

[12] B. Institution, "Fairness in algorithmic decision-making." https://www.brookings.edu, 2020. Accessed: 2024-03-17.

[13] K. Guruswamy, "Mitigating bias in ai/ml models with disparate impact analysis." https://www.h2o.ai/blog/mitigating-bias-in-ai-ml-models-with-disparate-impact-analysis/, 2019. Accessed: 2024-03-17.

[14] H. Wang, B. Ustun, and F. P. Calmon, "On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning," *arXiv*, 2018. https://arxiv.org/abs/1801.05398.

[15] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," *Nature Methods*, vol. 14, pp. 641–642, 2017. https://doi.org/10.1038/nmeth.4346.

[16] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[17] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1646–1654, 2014.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, 2016.

[21] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, pp. 41–46, 2001.

[22] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine Learning in Python*, 2011.

[24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, ACM, 2012.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[26] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, M. Augustin, G. Simoes, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2869–2871, ACM, 2018.

[27] K. M. Ting, "A comparative study of cost-sensitive boosting algorithms," *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

[28] X. Zhou *et al.*, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 21, no. 1, p. 334, 2020. `https://doi.org/10.1186/s12859-023-05465-z`.

[29] D. Kiyasseh *et al.*, "Bias in ai-based models for medical applications: challenges and mitigation strategies," *npj Digital Medicine*, vol. 4, no. 1, p. 62, 2021. `https://www.nature.com/articles/s41746-023-00858-z`.

[30] D. Char *et al.*, "Considerations for addressing bias in artificial intelligence for health equity," *npj Digital Medicine*, vol. 4, no. 1, p. 129, 2021. `https://www.nature.com/articles/s41746-023-00913-9`.

[31] J. Brownlee, "How and when to use a calibrated classification model with scikit-learn," 2020. `https://machinelearningmastery.com/calibrated-classification-model-in-scikit-learn/`.