

Gender-Specific Analysis of Diabetes Prediction Using Machine Learning

Shaik Sohail Abrar

*Department of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
sosh23@student.bth.se*

Jeswanth Naidu Padi

*Department of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
jepd23@student.bth.se*

Index Terms—Diabetes, SVM, Adaboost, Random Forest, Accuracy, Precision, Recall, F1-Score.

I. INTRODUCTION

Diabetes mellitus, sometimes known as diabetes, is a set of common endocrine illnesses marked by persistently elevated blood sugar levels. Diabetes is caused by either the pancreas not making enough insulin or the cells in the body becoming resistant to the hormone's effects. Classic symptoms include polydipsia (excessive thirst), polyuria (excessive urine), weight loss, and unclear eyesight [1]. If not addressed, the condition can cause various health consequences, including cardiovascular, eye, renal, and nerve disorders.

Diabetes is responsible for around 4.2 million fatalities per year, with an estimated 1.5 million caused by either untreated or poorly controlled diabetes. The major types of diabetes are type 1 and type 2. The most common treatment for type 1 is insulin replacement therapy (insulin injections), while anti-diabetic medications (such as metformin and semaglutide) and lifestyle modifications can be used to manage type 2. Gestational diabetes, a form that arises during pregnancy in some women, normally resolves shortly after delivery [1].

In 2021, more than 38 million Americans had diabetes, both diagnosed and undiagnosed, according to Ashley Andrews, a population navigator at the St. Joseph's/Candler Center for Diabetes Management, who is also board-certified in Advanced Diabetes Management. She explains that type 2 diabetes is more common in men than women, with 17.7 million more men than women affected worldwide as of 2023. However, women tend to be diagnosed earlier, but they are at a higher risk of experiencing more severe complications from the disease. [2].

The project 'gender-specific diabetes prediction' is a health-related concept that is also relevant to society as it deals with personalized treatment and understanding gender-based risk factors. It addresses a crucial global health concern in which early detection may significantly enhance patient outcomes and reduce healthcare costs. The choice of algorithms for this

project were Random Forest, AdaBoost, and Support Vector Machine.

A. Classification Algorithms:

1) **Random Forest:** Random forests [4] is an ensemble learning that uses the bagging technique method for classification that works by creating a multitude of decision trees during training. For classification tasks, the output of the random forest is the class selected by most trees. Random forests correct for decision trees' habit of overfitting to their training set.

2) **Adaboost:** AdaBoost is an ensemble learning that uses a boosting technique for classification. The main idea behind AdaBoost is to iteratively train the weak classifier on the training dataset with each successive classifier giving more weightage to the misclassified data points. The final AdaBoost model is decided by combining all the weak classifiers that have been used for training with the weightage given to the models according to their accuracies. The weak model which has the highest accuracy is given the highest weightage while the model which has the lowest accuracy is given a lower weightage [5].

3) **Support Vector Machine (SVM):** A Support Vector Machine (SVM) is a versatile machine-learning technique that can perform linear and nonlinear classification and outlier detection. SVM operates by determining the maximum separation hyperplane across classes, making it suitable for both binary and multiclass issues. This outline will look at the SVM method, its applications, and its ability to perform various classification and regression problems [6].

The motivation behind choosing these algorithms is because of their proven effectiveness in handling structured data, their interpretability, and their ability to manage the complexities of predicting diabetes. Random Forest was chosen because of its capacity to handle complicated, non-linear connections in data and offer solid predictions via ensemble learning, making it ideal for medical datasets containing both categorical and numerical characteristics. AdaBoost was chosen for its capacity to improve weak learners by iteratively focusing on misclassified samples, making it useful for datasets

containing subtle patterns, such as the effect of lifestyle factors on diabetes. SVM were chosen because of their capacity to provide exact decision boundaries, which is critical for separating diabetic and non-diabetic patients in complex medical data.

For Random Forest, the number of decision trees used in the model is determined based on the hyperparameter `n_estimators` explored. This directly impacts how stable and accurate the predictions are. The right balance between performance and computational cost was found by testing with different values. The specific range of [50, 100, 150, 250] was chosen to see how increasing the number of trees improves accuracy and at what point the improvements level off.

In the case of AdaBoost, the `n_estimators` parameter was examined since it defines how many weak learners are combined during the boosting process. This parameter is crucial because it influences the model's ability to iteratively correct errors and make better predictions. We tested values in the range of [50, 100, 150, 250] to understand how adding more iterations impacts the model's accuracy and whether it becomes more effective as boosting depth increases.

Two hyperparameters were explored i.e., `C` and `kernel` for SVM. The `C` parameter was chosen because it controls the trade-off between allowing some misclassifications and creating a larger margin for classification. This balance is essential to avoid underfitting or overfitting. The `kernel` was studied because it determines how data is transformed into higher-dimensional spaces to handle nonlinear relationships. The values for `C`—[0.1, 1, 10]—were tested to see how different levels of regularization impact the model's performance. Additionally, the `linear` and `rbf` kernels were selected because they are effective for handling both simple and complex data patterns.

The main purpose of the project is to evaluate the performance of the proposed machine-learning algorithms with the help of evaluation metrics in predicting diabetes disease separately for male and female subsets of the dataset. The proposed solution includes extensive data preprocessing, and managing class imbalances with techniques such as SMOTE, and hyperparameter tuning for the proposed algorithms. The problem is solvable due to the dataset's structure, and the application of machine learning models will produce encouraging results with better performance across important metrics and high accuracy.

II. METHODS

A. Data Collection:

The study's dataset [3] was obtained from Kaggle, a diabetes prediction dataset that collects patient demographic and medical information as well as the patient's diabetes status (positive or negative). Age, gender, body mass index (BMI),

blood pressure, heart disease, smoking history, HbA1c level, and blood glucose level are among the characteristics included in the data. Using this dataset, machine learning models that predict diabetes in patients based on their demographics and medical history can be developed. Healthcare providers may find this helpful in determining whether individuals are at risk of acquiring diabetes and in creating individualized treatment programs.

B. Data Preprocessing:

The pre-processing techniques were carefully developed to ensure that the data was balanced, clean, and suitable for machine learning modeling. In the first phase, known as "**data cleaning**," missing or abnormal values—which are frequently seen in medical datasets—were found and addressed. Missing values were marked as "No Info" for the **Smoking History** feature to maintain the dataset's integrity and avoid needless data loss. To make it compatible with machine learning techniques, the Smoking History variable was also converted into a binary format, with smokers being represented as 1 and non-smokers as 0. The dataset was then divided into two gender-based subsets: one for men and one for women. To minimize any biases in the analysis, this split was carefully carried out to maintain balanced class distributions within each subset.

To address the problem of class imbalance, the Synthetic Minority Oversampling Technique (**SMOTE**) was applied separately to the male and female datasets to preserve the integrity of the gender-based analysis. SMOTE improved the model's ability to predict cases of diabetes by creating synthetic samples for the minority class. Additionally, **scaling** was used to standardize numerical parameters including blood glucose level, HbA1c level, and BMI. By providing consistency among characteristics, this standardization improved overall prediction accuracy by preventing any one property from unnecessarily affecting the model's performance.

C. Model Training:

The machine learning models were applied separately to the two groups of the dataset as it helps and allows for an independent evaluation of model performance in each group.

The initialization of the **Random Forest** model required setting the number of estimators (`n_estimators`) to 50, 100, 150, and 250. In the same way, the **AdaBoost** model was started with various numbers of estimators (`n_estimators`), which were set to 50, 100, 150, and 250.

Support Vector Machine (SVM), the regularization parameter (`C`) was varied across three values—0.1, 1, and 10. The other hyperparameter was `kernel` types, radial basis function ('rbf') and linear, were evaluated .

A Grid Search method was used to determine which hyperparameter combination was optimal for each model. The best parameters were found by systematically testing every

possible combination of initial hyperparameters. To separate the data into three subsets, a 3-fold cross-validation technique was employed. Every subset was used once for validation, while the other two subsets were utilized for training. By using this method, bias was reduced variability was lessened, and a solid and reliable model selection was maintained.

D. Evaluation Metrics:

After we train our machine learning models, it's important to understand how well our model has performed. Evaluation metrics [7] are used for this same purpose.

- **Confusion Matrix** is a technique for summarizing the performance of a classification algorithm by comparing expected and actual results. It uses several key terms, including True Positives(TP)-correctly predicted positives, True Negatives(TN)-correctly predicted negatives, False Positives(FP)-incorrectly predicted positives, and False Negatives(FN)-incorrectly predicted negatives.
- **Accuracy** can be defined as the percentage of correct predictions made by our classification model.

$$Accuracy = TP + TN / TP + TN + FP + FN$$

- **Precision** indicates out of all positive predictions, how many are positive. It is defined as a ratio of correct positive predictions to overall positive predictions.

$$Precision = TP / TP + FP$$

- **Recall** indicates out of all positive values, how many are predicted positive. It is a ratio of correct positive predictions to the overall number of positive instances in the dataset.

$$Recall = TP / TP + FN$$

- **F1-score** is defined as the harmonic mean of precision and recall. When avoiding both false positives and false negatives is equally important for our problem, we need a trade-off between precision and recall. This is when we use the f1 score as a metric.

$$F1-Score = 2.Precision.Recall / Precision + Recall$$

III. RESULTS AND ANALYSIS:

A. Male Dataset Analysis:

In class distribution, with 37,391 samples in the majority class (0) and only 4,039 in the minority class (1), the dataset was severely unbalanced before the application of SMOTE. The dataset was balanced with 26,174 samples in each class following the use of SMOTE.

The optimal hyperparameters for the male dataset are as follows:

- **Random Forest:** n_estimators = 250
- **AdaBoost:** n_estimators = 250
- **SVM:** C = 10, kernel = rbf

TABLE I
MODEL PERFORMANCE METRICS FOR MALE DATASET

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.95	0.95	0.95	0.95
AdaBoost	0.92	0.94	0.92	0.93
SVM	0.88	0.94	0.88	0.90

The best-performing model was Random Forest, which produced balanced predictions with few false positives or false negatives and few misclassifications. Compared to Random Forest, AdaBoost produced more false negatives despite having a slightly lower recall. SVM produced the most false negatives and the lowest recall and F1-scores when compared to the other models, despite having decent precision. It also had trouble correctly classifying the minority class. Figure 1 shows the performance visualization of male dataset.

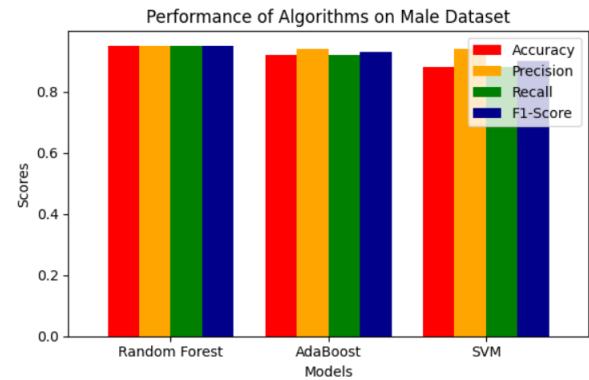


Fig. 1. Performance visualization of Male dataset

B. Female Dataset Analysis:

With 54,091 samples in the majority class (0) and just 4,461 in the minority class (1), the female dataset had a notable class imbalance before the use of SMOTE. The dataset was balanced with 37,863 samples in each class following SMOTE.

TABLE II
MODEL PERFORMANCE METRICS FOR FEMALE DATASET

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.96	0.96	0.96	0.96
AdaBoost	0.94	0.96	0.94	0.95
SVM	0.89	0.95	0.89	0.91

The optimal hyperparameters for the female dataset are as follows:

- **Random Forest:** n_estimators = 150
- **AdaBoost:** n_estimators = 250
- **SVM:** C = 10, kernel = rbf

In terms of model performance comparison, Random Forest outperformed all other models with a low number of false positives and negatives, achieving almost perfect classification. The recall and F1-score of AdaBoost were impacted by a few

more false negatives, which made it slightly less accurate and precise than Random Forest despite its strong performance. Despite having outstanding precision, SVM had trouble recognizing minority class cases, as demonstrated in the male sample. This led to the highest amount of false negatives, which negatively impacted its recall and F1-score. Figure 2 shows the performance visualization of female dataset.

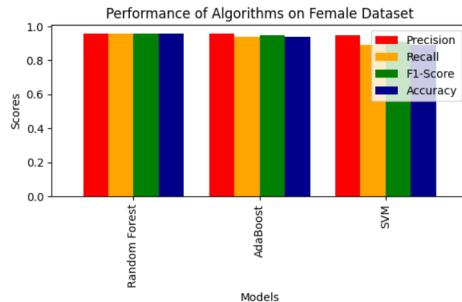


Fig. 2. Performance visualization of Female dataset

C. Comparative Analysis:

In our analysis, Random Forest consistently performed better than the other two models on both datasets. With outstanding precision, recall, and F1-scores for both majority and minority classes, it turned out to be the most dependable option. Although it dropped a bit behind Random Forest in terms of recall and F1-score, AdaBoost also did well, providing good accuracy and precision. Because of this, AdaBoost is a good substitute, especially in situations where interpretability or computational effectiveness are crucial. However, SVM struggled with recall and F1-scores, particularly for minority classes, but demonstrated respectable precision. Its accuracy was further affected by its vulnerability to unusual values, which reduced its usefulness in situations where precise minority class prediction is essential.

The recall and F1 scores for all models, particularly the SVM, were significantly improved by balancing the datasets using SMOTE, demonstrating the significance of addressing class imbalance in binary classification problems. Random Forest was the most dependable model when compared to AdaBoost, with both models doing well on all criteria. However, SVM has trouble predicting the minority class. Similar performance patterns were displayed by both datasets, which represented the male and female distributions, suggesting that the models' behavior stayed constant following the application of SMOTE. These results highlight how important it is to select the appropriate model and apply preprocessing methods such as SMOTE to achieve more balanced and improved predictive performance.

IV. CONCLUSION:

In conclusion, three machine learning algorithms were used in this study to successfully develop and evaluate predictive models for diabetes classification: Random Forest, AdaBoost,

and SVM. With high accuracy, precision, recall, and F1-scores, Random Forest continuously outperformed the other models in both the male and female datasets. It was the most reliable model for this task since it minimized the number of misclassifications. The class imbalance in the original datasets was addressed in large part by the use of SMOTE. Recall and F1-scores were greatly enhanced for all models by balancing the majority and minority classes, especially SVM, which at first had trouble correctly classifying instances of the minority class.

When we analyzed the results by gender, we found consistent performance patterns between male and female datasets. Random Forest remained the top-performing model, while AdaBoost provided competitive results as a simpler alternative. SVM, though precise, struggled with false negatives, especially for the minority class. The models' consistent behavior across both genders suggests they generalized well without favoring either group.

Interestingly, the best hyperparameter settings for Random Forest were different for the male and female datasets. This suggests that gender-specific factors can influence the optimization process, making it essential to consider gender when tuning models for medical predictions. Based on our findings, we recommend using Random Forest with customized hyperparameters for male and female datasets to achieve the most accurate and reliable predictions.

This study underscores the importance of addressing data imbalances and selecting strong machine learning models to ensure dependable predictions. In the future, researchers could explore more features or datasets, test ensemble methods that combine the strengths of Random Forest and AdaBoost, and use interpretability techniques to better understand how individual features contribute to predictions. Additionally, studying the impact of gender-specific data patterns on model optimization could further enhance personalized healthcare solutions.

V. CONTRIBUTION:

Shaik Sohail Abrar was in charge of the project's technical aspects, which included evaluation, training the model, and preparing the data. Sohail worked on the dataset's transformation and cleaning, choosing suitable machine learning models, tuning them, and comparing how they performed for the male and female categories.

Jeswanth Naidu Padi was responsible for the project's documentation, which included gathering appropriate resources, studying previous research, and understanding gender-specific analysis in diabetes prediction. Jeswanth also made sure that the results were presented in an understandable manner and contributed to the development of the report's general structure.

REFERENCES

- [1] Wikipedia contributors, "Diabetes," Wikipedia, 2024. [Online]. Available: <https://en.wikipedia.org/wiki/Diabetes>. [Accessed: 29-Dec-2024].
- [2] St. Joseph's/Candler. (2024, April 25). How diabetes affects women differently than men. Retrieved from <https://www.sjchs.org/living-smart-blog/blog-details/blog/2024/04/25/how-diabetes-affects-women-differently-than-men>.
- [3] M. Taz, "Diabetes Prediction Dataset," Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/tammustafatz/diabetes-prediction-dataset>. [Accessed: 29-Dec-2024].
- [4] Wikipedia contributors, "Random forest," Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest. [Accessed: 29-Dec-2024].
- [5] GeeksforGeeks, "Implementing the AdaBoost Algorithm from Scratch." [Online]. Available: <https://www.geeksforgeeks.org/implementing-the-adaboost-algorithm-from-scratch/>. [Accessed: Dec. 29, 2024].
- [6] GeeksforGeeks, "Support Vector Machine Algorithm." [Online]. Available: <https://www.geeksforgeeks.org/support-vector-machine-algorithm/>. [Accessed: Dec. 29, 2024].
- [7] Analytics Vidhya, "A Tour of Evaluation Metrics for Machine Learning," Nov. 2020. [Online]. Available: <https://www.analyticsvidhya.com/blog/2020/11/a-tour-of-evaluation-metrics-for-machine-learning/>. [Accessed: Dec. 29, 2024].