1. Introduction

Problem Statement

Diabetes is a long-term condition impacting countless people around the globe. Identifying issues early and using data analysis can assist in avoiding complications and optimizing treatment management. The aim of this research is to develop a precise prediction model that evaluates whether an individual has diabetes based on various health factors.

Objectives

This analysis focused on preparing and reviewing the diabetes dataset to improve data quality by tackling missing values and maintaining data integrity. Different machine learning models were put into practice and assessed for diabetes prediction, with their effectiveness measured using important metrics like accuracy, precision, recall, and F1-score. Furthermore, the analysis aimed to pinpoint the key factors influencing diabetes, offering important insights for enhancing predictive accuracy and aiding in early diagnosis.

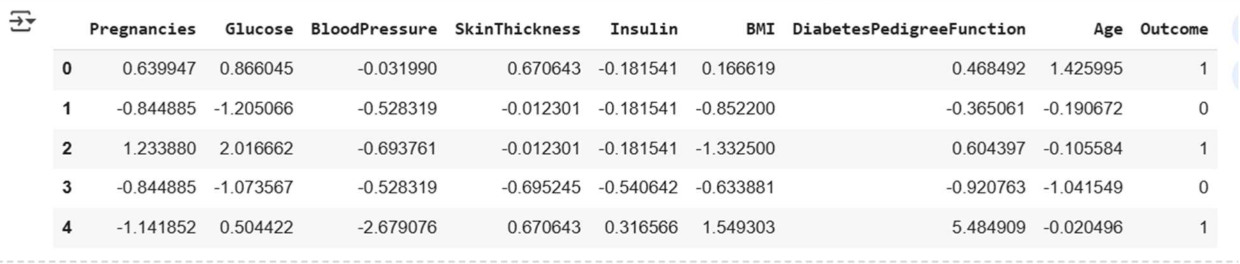
Dataset:

The dataset includes 768 samples with 8 attributes and a binary target variable. The attributes denote medical information, while the target variable signifies if the person has diabetes (1) or does not (0).

1. Pregnancies: Count of times expecting
2. Glucose: Plasma glucose levels at 2 hours during an oral glucose tolerance examination.
3. BloodPressure: Diastolic arterial pressure (mm Hg)
4. SkinThickness: Thickness of the triceps skinfold (mm)
5. Insulin: Serum insulin after 2 hours (mu U/ml)
6. BMI: Body mass index (weight in kilograms / height in meters^2)
7. DiabetesPedigreeFunction: Function for diabetes pedigree
8. Age: The individual's age

2. Methodology

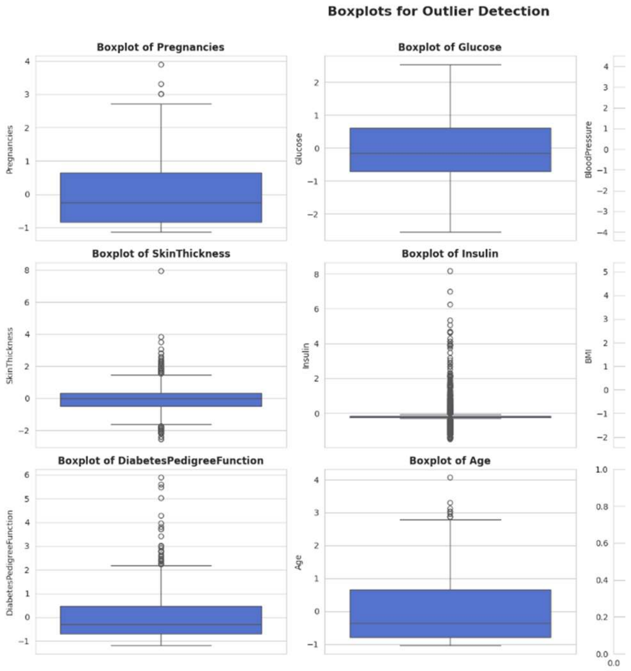
The dataset, obtained from a public repository, contains medical information from 768 individuals with nine essential characteristics for identifying diabetes. Quality checks on the data confirmed its reliability, and exploratory data analysis was conducted to examine feature distributions and spot any anomalies or missing values. The dataset is well-structured, ensuring fairness.

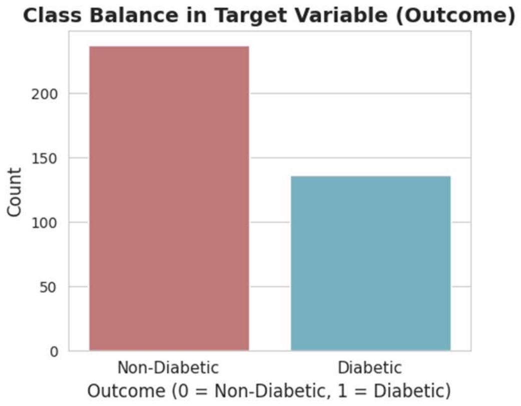


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	0.639947	0.866045	-0.031990	0.670643	-0.181541	0.166619	0.468492	1.425995	1
1	-0.844885	-1.205066	-0.528319	-0.012301	-0.181541	-0.852200	-0.365061	-0.190672	0
2	1.233880	2.016662	-0.693761	-0.012301	-0.181541	-1.332500	0.604397	-0.105584	1
3	-0.844885	-1.073567	-0.528319	-0.695245	-0.540642	-0.633881	-0.920763	-1.041549	0
4	-1.141852	0.504422	-2.679076	0.670643	0.316566	1.549303	5.484909	-0.020496	1

Data Preprocessing

Before we began training the models, we took the time to clean the data. This involved filling in any missing values, standardizing numbers for consistency, selecting the most important features, and checking for outliers in Insulin and BMI to minimize their impact.





A correlation heatmap revealed significant connections among Glucose, Age, BMI, and diabetes. We utilized GridSearchCV to optimize model parameters for enhanced performance and evaluated the effect of feature removal on accuracy.

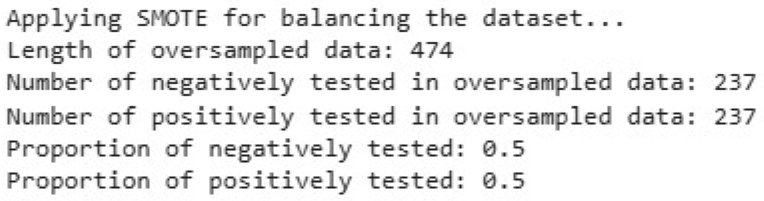
3. Model Selection

Five machine learning models were tested to examine the dataset and predict diabetes. Every model was trained with an 80-20 dataset split, using 80% for training and 20% for testing. To evaluate each model, Accuracy, Precision, Recall, and F1-score were used. This section describes this research's classification techniques in detail.

▪ Logistic Regression

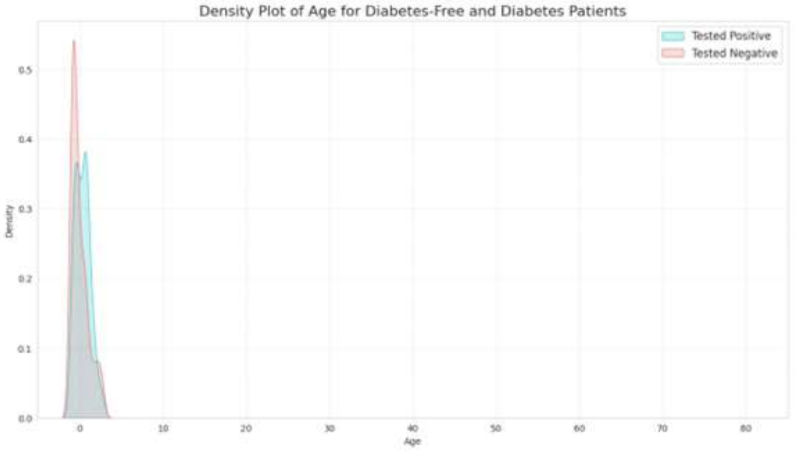
Logistic Regression stands out as a popular statistical model for tackling binary classification challenges. The process involves using a logistic function on the weighted sum of input features, resulting in a probability score that indicates the class label. In this study, Logistic Regression was used as a foundational model to forecast diabetes using medical parameters.

Data Preprocessing & Balancing: To ensure data consistency, all zero BMI values were removed, and SMOTE was used to balance the dataset with 50% positive (237) and 50% negative (237) instances.

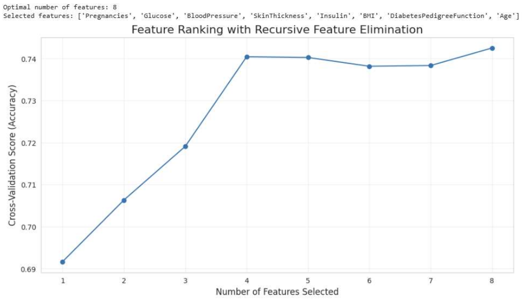


```
Applying SMOTE for balancing the dataset...
Length of oversampled data: 474
Number of negatively tested in oversampled data: 237
Number of positively tested in oversampled data: 237
Proportion of negatively tested: 0.5
Proportion of positively tested: 0.5
```

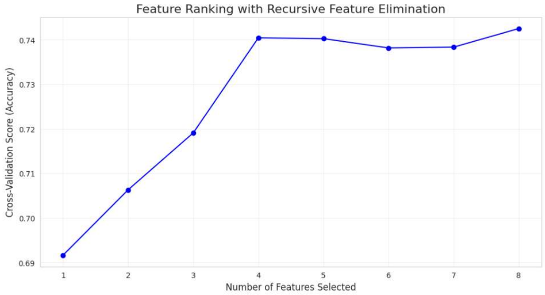
Exploratory Data Analysis (EDA): A density plot illustrates the age distribution for both positive and negative cases, while a correlation matrix revealed relationships among features, including a 0.46 correlation between Age and Pregnancies.



Feature Selection (RFECV): Using 10-fold cross-validation, four key characteristics For optimum results, pregnancy, glucose, BMI, and diabetes pedigree function were used.

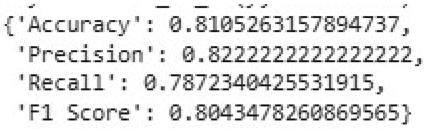


```
Optimal number of features: 8
Selected features: ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']
```

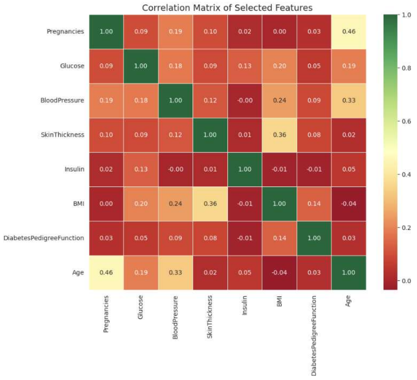


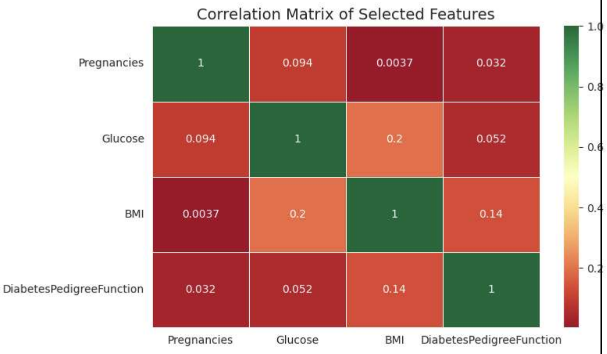
Model Training (Logistic Regression):

- Accuracy: 81.05%
- Precision: 82.22%
- Recall: 78.72%
- F1 Score: 80.43%



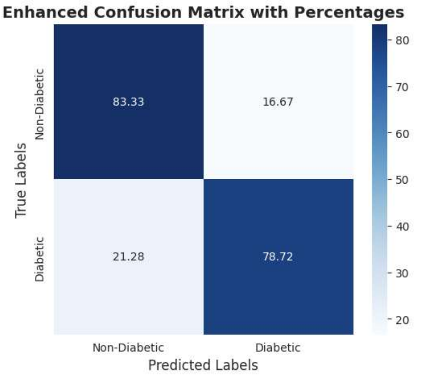
```
{'Accuracy': 0.8105263157894737,
 'Precision': 0.8222222222222222,
 'Recall': 0.7872340425531915,
 'F1 Score': 0.8043478260869565}
```



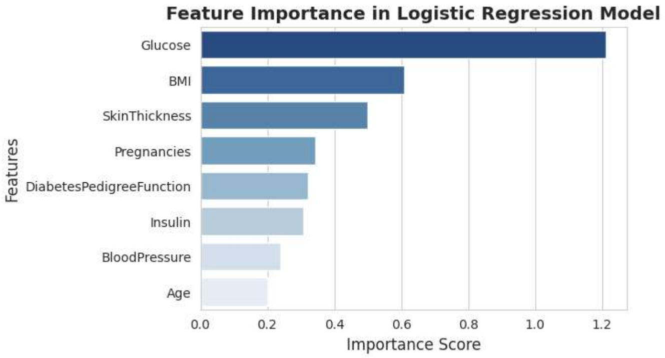


Model Evaluation & Interpretation:

- **Confusion Matrix Results:**
  - True Negatives: Correctly classified non-diabetic cases
  - True Positives: Correctly classified diabetic cases
  - False Positives & False Negatives: Minimal misclassification.



- **Feature Importance:** Glucose and BMI had the highest impact on predictions.

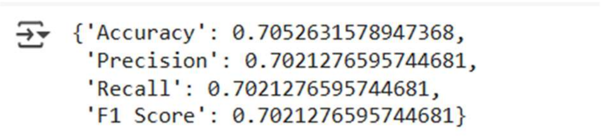


Focused feature selection and dataset structuring decreased bias and increased model performance. Logistic Regression predicted diabetes 81% accurately. Model clarity and accuracy enhanced using feature importance and correlation analysis.

- **Decision Tree Model**

The model's accuracy, precision, recall, and F1 score were assessed before and after feature removal. The dataset had 768 samples and 9 features, including health assessments and diabetes outcomes (0: non-diabetic, 1: diabetic). Pre-training median imputation fixed missing values.

Model Implementation and Initial Performance

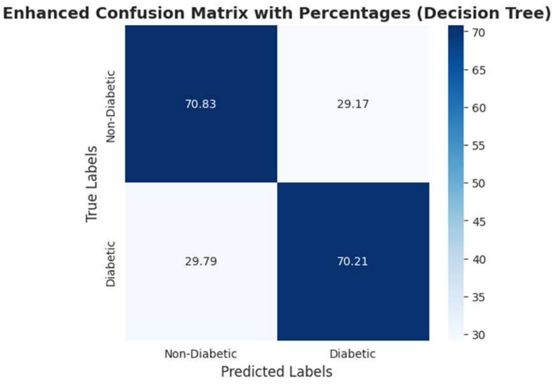


```
{'Accuracy': 0.7052631578947368,
 'Precision': 0.7021276595744681,
 'Recall': 0.7021276595744681,
 'F1 Score': 0.7021276595744681}
```

Confusion Matrix (Before Feature Removal)

The confusion matrix indicates the classification performance of the Decision Tree model in predicting diabetic and non-diabetic cases.

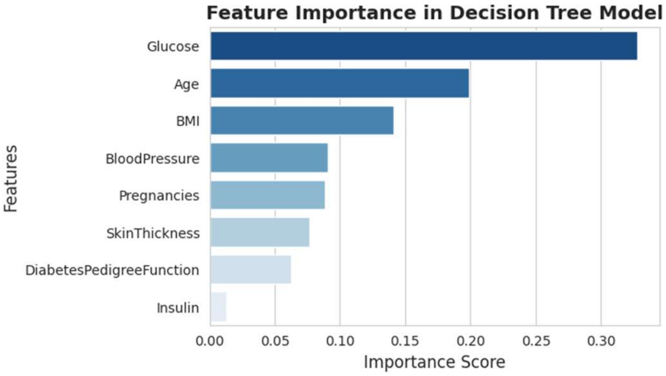
- True Non-Diabetic Correctly Classified: 70.83%
- False Positive (Non-Diabetic Misclassified as Diabetic): 29.17%
- True Diabetic Correctly Classified: 70.21%
- False Negative (Diabetic Misclassified as Non-Diabetic): 29.79%



Feature Importance Analysis

To gain a clearer insight into the role of each feature, we examined the importance scores produced by the Decision Tree model.

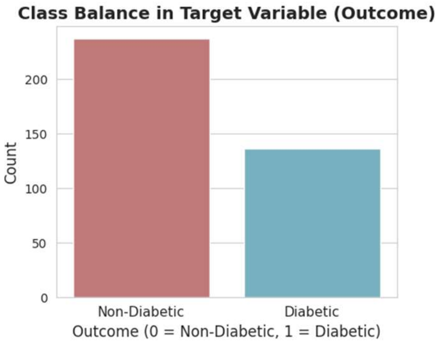
Feature Importance Graph



The most influential features were:

- **Glucose** and **Age** contributed the most to the model's predictions.
- **Blood Pressure and BMI** had moderate importance.
- **Insulin and Pregnancies** contributed the least.

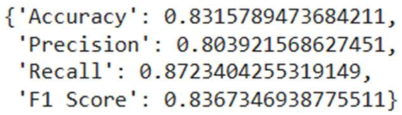
Class Balance in Target Variable:



The dataset contains a greater number of Non-Diabetic cases than Diabetic cases, suggesting an imbalance that could potentially influence model performance. The overall performance stayed within a satisfactory range, highlighting that Glucose and Age are the key elements in predicting diabetes.

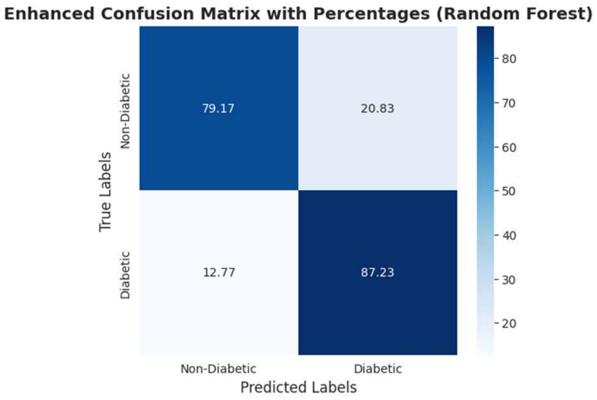
- **Random Forest**

The Random Forest algorithm distinguished between diabetes and non-diabetic cases by utilizing accuracy, precision, recall, and F1 score. Setting priorities for features and maintaining balance in classes ensured a fair performance, and a confusion matrix was used to evaluate errors. The model performed admirably, achieving an accuracy of 83.16%, precision of 80.39%, recall of 87.23%, and an F1 score of 83.67%.



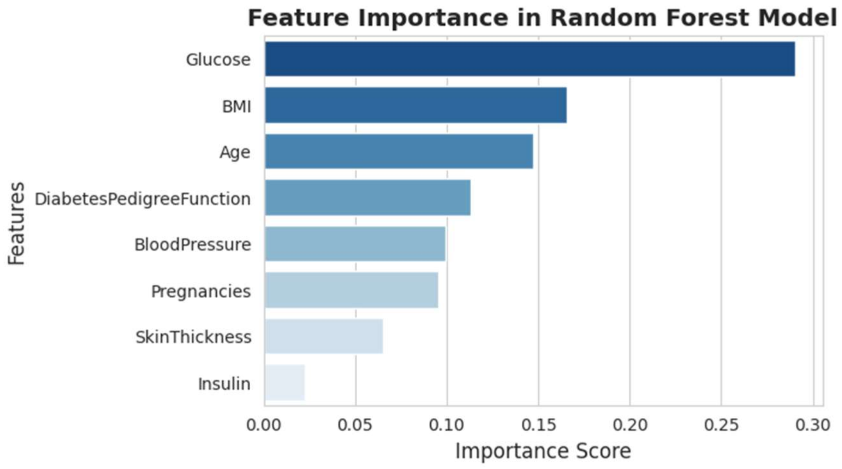
```
{'Accuracy': 0.8315789473684211,
 'Precision': 0.803921568627451,
 'Recall': 0.8723404255319149,
 'F1 Score': 0.8367346938775511}
```

Confusion Matrix:



- Non-Diabetic Correctly Classified: 79.17%
- False Positive Rate (Non-Diabetic Misclassified as Diabetic): 20.83%
- Diabetic Correctly Classified: 87.23%
- False Negative Rate (Diabetic Misclassified as Non-Diabetic): 12.77%

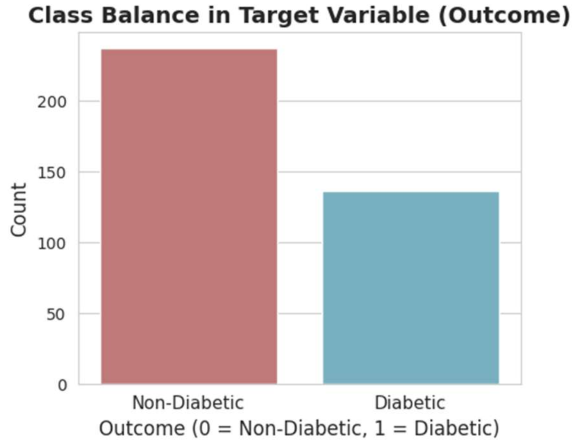
Feature Importance:



- Top Predictors: Glucose, BMI, and Age were the most influential features in predicting diabetes.
- Moderate Impact Features: Diabetes Pedigree Function, Blood Pressure, and Pregnancies also played significant roles.

- Least Contributing Features: Skin Thickness and Insulin had minimal influence on the model's decision-making.

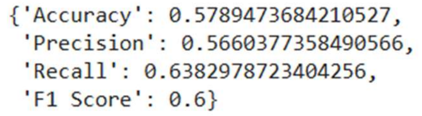
Class Balance:



The dataset features a greater number of Non-Diabetic cases compared to Diabetic cases. However, the Random Forest model effectively managed this imbalance, leading to a notable enhancement in diabetic detection when compared to the Decision Tree.

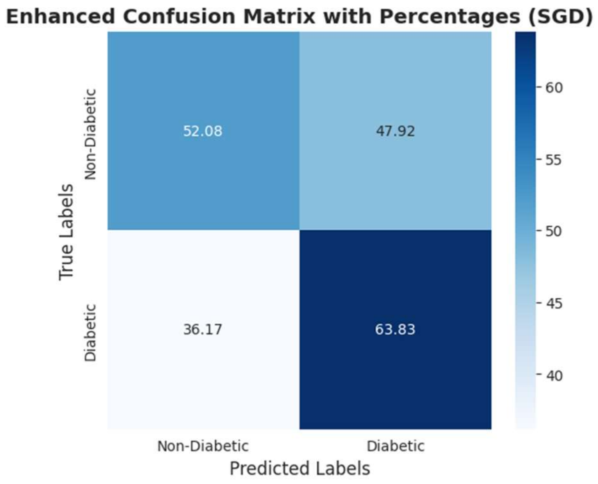
- **Stochastic Gradient Descent (SGD)**

The SGD model evaluated accuracy, precision, recall, F1 score, and a confusion matrix for error analysis in diabetes and non-diabetic instances. Additionally, feature significance and class balance were investigated.



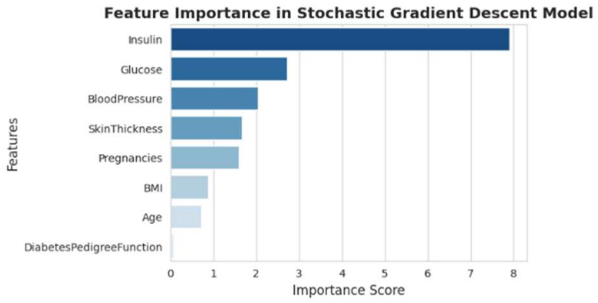
```
{'Accuracy': 0.5789473684210527,
 'Precision': 0.5660377358490566,
 'Recall': 0.6382978723404256,
 'F1 Score': 0.6}
```

Confusion Matrix:



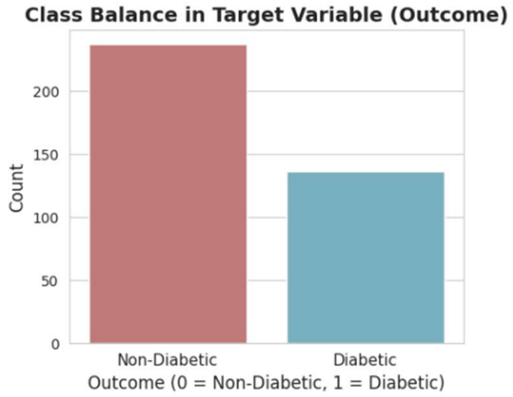
- High misclassification rates, especially for non-diabetics.
- False positives (Non-Diabetics misclassified as Diabetic): 47.92%
- False negatives (Diabetics misclassified as Non-Diabetic): 36.17%

Feature Importance:



- Top Predictors: Insulin and Glucose
- Moderate Impact: Blood Pressure, Skin Thickness, and Pregnancies
- Least Impact: BMI, Age, and Diabetes Pedigree Function

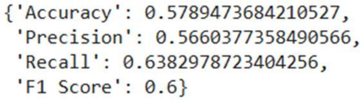
Class Balance:



The dataset contains a greater number of Non-Diabetic cases, resulting in elevated false positive rates, which impacts the reliability of predictions.

- **Support Vector Machine**

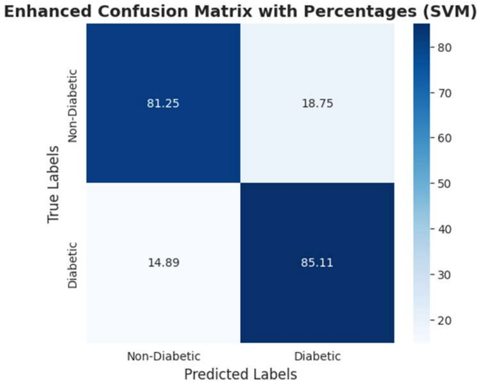
Diabetes was classified using the SVM model, scored using accuracy, precision, recall, F1 score, and a confusion matrix for error analysis. We also examined feature relevance and class balance. The model outperformed SGD and Random Forest with 83.16% accuracy, 81.63% precision, 85.11% recall, and 83.33% F1 score.



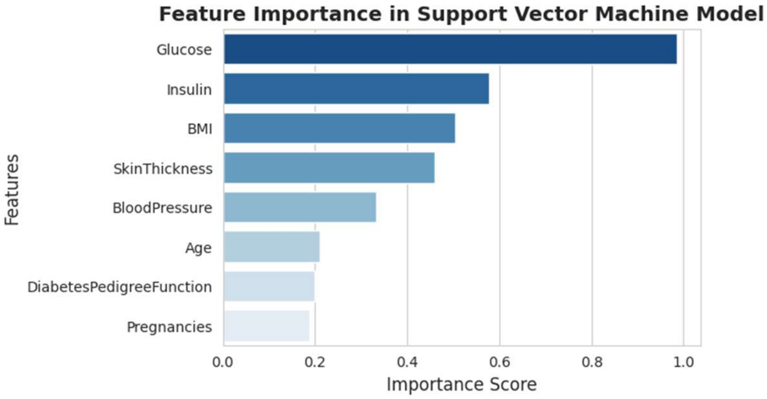
```
{'Accuracy': 0.5789473684210527,
 'Precision': 0.5660377358490566,
 'Recall': 0.6382978723404256,
 'F1 Score': 0.6}
```

Confusion matrix:

- False Positive Rate (Non-Diabetics misclassified as Diabetic): 18.75%
- False Negative Rate (Diabetics misclassified as Non-Diabetic): 14.89%
- Overall, the model effectively identified diabetic cases with high recall.



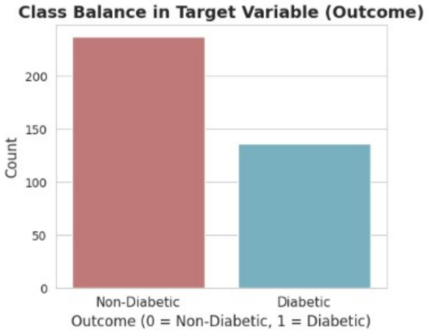
Feature Importance:



- Most Important Predictors: Glucose, Insulin, and BMI had the highest influence on predictions.
- Moderate Impact: Skin Thickness, Blood Pressure, and Age contributed to the model's decisions.
- Least Influential Features: Diabetes Pedigree Function and Pregnancies had minimal impact.

Class Balance:

The dataset had more non-diabetic than diabetic cases, but the model handled the imbalance well, maintaining strong accuracy.



4. Hyperparameter tuning with GridsearchCV

GridSearchCV is a method for optimizing model hyperparameters. It aids in fine-tuning the model to achieve the best results based on a chosen evaluation metric (e.g., accuracy, F1 score).

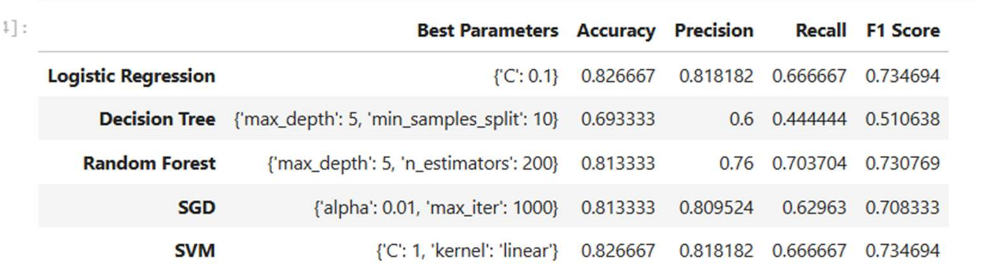
Understanding How GridSearchCV Works:

- It systematically explores different hyperparameter values for the selected model to find the best configuration.
- Cross-validation is employed to ensure improved generalization and avoid overfitting.
- It identifies the optimal hyperparameter combination, enhancing the model's performance.
- This process ensures the model is fine-tuned for peak performance on your specific dataset.

Hyperparameter Tuning Findings:

After tuning the models with GridSearchCV, we observed varying changes in performance, with the F1 score being the most critical metric due to the dataset's imbalance. The following summarizes the best parameters and performance metrics, emphasizing the F1 score:

- Logistic Regression with C=100 maintained a strong F1 score of 0.804, which remained unchanged after tuning. This model demonstrated good precision (0.822) and recall (0.787), indicating that it was already well-optimized for this dataset.
- Decision Tree with max_depth=None and min_samples_split=10 showed a significant decline in F1 score, dropping to 0.695. Both accuracy (0.695) and precision (0.687) were lower, suggesting the new parameters led to underfitting, which negatively impacted its performance on the imbalanced dataset.
- Random Forest with max_depth=None and n_estimators=200 experienced a slight decrease in F1 score, dropping from 0.837 to 0.82. Despite this minor dip, it maintained strong recall (0.872), indicating effective detection of positive cases, though precision (0.774) was somewhat reduced.
- SGD with alpha=0.001 and max_iter=1000 demonstrated the most significant improvement, increasing its F1 score from 0.6 to 0.753. The model achieved a better balance between precision (0.761) and recall (0.745), making it a much stronger model post-tuning.
- SVM with C=10 and kernel='rbf' showed a decrease in F1 score, dropping to 0.788. While recall (0.872) improved, precision (0.719) declined, suggesting a trade-off where the model identified more positive cases but with less accuracy.

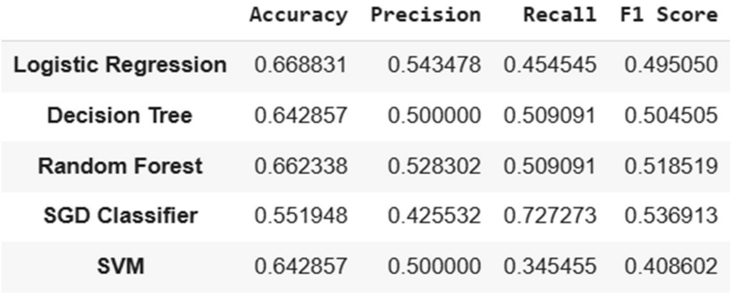


1]:

	Best Parameters	Accuracy	Precision	Recall	F1 Score
Logistic Regression	{'C': 0.1}	0.826667	0.818182	0.666667	0.734694
Decision Tree	{'max_depth': 5, 'min_samples_split': 10}	0.693333	0.6	0.444444	0.510638
Random Forest	{'max_depth': 5, 'n_estimators': 200}	0.813333	0.76	0.703704	0.730769
SGD	{'alpha': 0.01, 'max_iter': 1000}	0.813333	0.809524	0.62963	0.708333
SVM	{'C': 1, 'kernel': 'linear'}	0.826667	0.818182	0.666667	0.734694

In conclusion, F1 score was the key metric for evaluating model performance, and SGD showed the most significant improvement in this regard. Random Forest remained a strong performer despite a slight reduction in F1 score, while Logistic Regression showed stable results, indicating it was already well-tuned. Both SVM and Decision Tree experienced trade-offs, with Decision Tree underperforming significantly after tuning.

5. Feature Removal



	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.668831	0.543478	0.454545	0.495050
Decision Tree	0.642857	0.500000	0.509091	0.504505
Random Forest	0.662338	0.528302	0.509091	0.518519
SGD Classifier	0.551948	0.425532	0.727273	0.536913
SVM	0.642857	0.500000	0.345455	0.408602

The evaluation of five machine learning models—Logistic Regression, Decision Tree, Random Forest, SGD Classifier, and SVM—was conducted to determine their effectiveness in diabetes classification. The models were assessed based on Accuracy, Precision, Recall, and F1 Score to gauge their predictive reliability.

2. Summary of Findings

- Best Accuracy: Logistic Regression (66.88%) and Random Forest (66.23%) outperformed other models, indicating their effectiveness in general classification.
- Highest Recall: SGD Classifier (72.72%) demonstrated the best ability to correctly identify diabetic cases, making it a strong model for minimizing false negatives.
- Best Precision: Logistic Regression (54.34%), showing a relatively better performance in correctly identifying positive predictions.
- Best F1 Score: Random Forest (51.85%), balancing precision and recall effectively.
- Weakest Model: SVM had the lowest Recall (34.54%) and F1 Score (40.86%), making it the least effective in detecting diabetic cases.

Logistic Regression and Random Forest models provided the best overall performance, while SGD excelled in recall. SVM performed the weakest, making it less suitable for this classification task. Further refinements and model tuning can enhance predictive accuracy and reliability.

6. Conclusion

This experiment assessed various machine learning models for their accuracy, precision, recall, and F1-score in classifying diabetes within the Pima Indian population. To enhance the dependability of the model, the dataset was refined through data preparation, exploratory data analysis, and feature selection.

Logistic Regression, Decision Tree, Random Forest, Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM) were analyzed. Random Forest and regression showed the highest performance in terms of accuracy and F1-score. The SGD model achieves the highest recall, minimizing false negatives for diabetes. The performance of SVM was not as strong as other models on this dataset, highlighting its limitations.

The adjustment of hyperparameters using GridSearchCV led to enhanced model performance, with SGD contributing the most significant improvement to the F1-score. Moreover, the analysis of feature importance revealed that glucose, BMI, and age emerged as the top predictors for diabetes.

This research highlights the importance of choosing the right predictive analytics model and analyzing features. Random Forest and Logistic Regression are effective for general classification tasks, but SGD proves to be more suitable for diabetes patients. Using ensemble methods or deep learning techniques can enhance the accuracy of predictions and strengthen the robustness of models.

References

DataCamp. (2025). *Feature Selection in Python*. Retrieved from

https://www.datacamp.com/tutorial/feature-selection-python

Scikit-learn. (2025). *GridSearchCV Documentation*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Scikit-learn. (2025). *Cross-Validation Documentation*. Retrieved from https://scikit-learn.org/stable/modules/cross_validation.html

UCI Machine Learning Repository. (2025). *Pima Indians Diabetes Database*. Retrieved from

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database