2417367-1506

2417367_c_16780.docx

University

Document Details

Submission ID
trn:oid:::7727:233590764

Submission Date

Sep 26, 2025, 8:26 AM GMT+3

Download Date

Sep 26, 2025, 8:28 AM GMT+3

File Name

2417367_c_16780.docx

File Size

1.6 MB

20 Pages

2,615 Words

15,020 Characters

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Crossref database
- Crossref posted content database

Match Groups

28 Not Cited or Quoted 11%
Matches with neither in-text citation nor quotation marks

3 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5%	Internet sources
2%	Publications
10%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **28** Not Cited or Quoted 11%
  Matches with neither in-text citation nor quotation marks
- **3** Missing Quotations 1%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

5%	Internet sources
2%	Publications
10%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	Ottawa University (Blackboard LTI 1.3 Prod) on 2025-05-25	1%
2	Submitted works	University of Southampton on 2018-09-08	1%
3	Internet	www.mdpi.com	<1%
4	Publication	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Re...	<1%
5	Submitted works	University of New South Wales on 2025-04-20	<1%
6	Submitted works	PES University on 2024-10-04	<1%
7	Submitted works	University of Greenwich on 2025-08-08	<1%
8	Submitted works	University of Nottingham on 2025-05-02	<1%
9	Submitted works	Whitireia Community Polytechnic on 2025-04-30	<1%
10	Submitted works	University of Ulster on 2024-05-09	<1%

11	Internet	rstudio-pubs-static.s3.amazonaws.com	<1%
12	Internet	jjcit.org	<1%
13	Submitted works	Curtin University of Technology on 2024-10-03	<1%
14	Submitted works	University of Surrey on 2024-01-02	<1%
15	Submitted works	Taylor’s Education Group on 2025-06-11	<1%
16	Submitted works	University of Ulster on 2025-05-02	<1%
17	Submitted works	Vermont State Colleges on 2023-12-13	<1%
18	Publication	Abdulrahim, Walid Abdulla. "Prediction of Diabetes Using Machine Learning.", Ro...	<1%
19	Submitted works	ESC Rennes on 2023-04-16	<1%
20	Internet	carpentries-incubator.github.io	<1%
21	Internet	www.coursehero.com	<1%
22	Submitted works	Georgia Institute of Technology Main Campus on 2021-02-22	<1%

Decision Tree Classification Analysis for Diabetes Prediction in Pima Indian Women

Student Name

Course Code and Name

Instructor

Institution

Date

Abstract

This report presents a decision tree classification analysis to predict Type 2 diabetes in Pima Indian women using diagnostic measurements. The study utilized a dataset of 768 patients with 8 clinical features to develop an interpretable machine learning model. The final decision tree achieved 69.5% accuracy with 92.0% specificity, identifying glucose level 69.9% importance and BMI 25.9% importance as the primary predictive factors. While the model demonstrated good performance in ruling out diabetes, it showed limitations in detecting positive cases 27.8% recall. The results provide clinically meaningful insights for diabetes screening in high-risk populations.

Keywords: Decision Trees, Diabetes Prediction, Machine Learning, Pima Indians, Medical Classification

Word count: 2615

1. Introduction

1.1 Objective

The main goal of this study is to develop a decision tree classification model that can predict Type 2 diabetes in Pima Indian women using diagnostic measurements. We want to answer the question: "Can we use basic medical tests to accurately identify women who have diabetes?" The specific analysis we are doing is binary classification, which means we are trying to put each patient into one of two groups either having diabetes or not having diabetes. Our success will be measured by how accurately the model can make these predictions and how easy it is for doctors to understand and use the decision rules that the model creates.

1.2 Problem Statement

Type 2 diabetes is a serious health problem that affects millions of people around the world. It happens when the body cannot use insulin properly, leading to high blood sugar levels. If not treated early, diabetes can cause many complications like heart disease, kidney problems, and vision loss. The Pima Indians are a Native American group living in Arizona and Mexico who have one of the highest rates of diabetes in the world. Studies show that more than 50% of adult Pima Indians develop Type 2 diabetes, which is much higher than the general population (Nicoa, 2022). This makes them an important group to study because understanding diabetes in this population can help doctors find better ways to predict and prevent the disease. Early detection of diabetes is very important because it allows people to make lifestyle changes and start treatment before serious complications occur. According to health statistics, diabetes costs the US healthcare system over $327 billion every year, so finding ways to predict it early can save both lives and money.

1.3 Method Rationale

We chose decision trees for this analysis because they are easy to understand and interpret. Unlike complex machine learning methods that work like "black boxes," decision trees create simple rules that doctors can easily follow. For example, a decision tree might say "if glucose level is above 150 and BMI is above 30, then the patient probably has diabetes." This interpretability is very important in medical applications because doctors need to understand why a model makes certain predictions (Drousiotis et al., 2023). Decision trees also handle different types of data well and don't require complex data preparation like some other methods. They can find the most important features automatically and show clear decision paths that make sense in a medical context.

2. Analysis

2.1 Data Description

The dataset we used comes from the Kaggle Machine Learning Repository and was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases. It contains information about 768 Pima Indian women who were at least 21 years old when the data was collected (Amini, 2023). The dataset has 8 different measurements that doctors commonly use to assess diabetes risk: number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function which shows family history of diabetes, and age. The target variable tells us whether each woman has diabetes 1 or not 0.

Looking at the class distribution shown in our target variable analysis (see Appendix A, Figure A1), we found that 500 women 65.1% do not have diabetes and 268 women 34.9% do have diabetes. This imbalance is clearly visible in both the pie chart and bar chart representations.

This shows that diabetes is quite common in this population, but non-diabetic cases are still more frequent. One important issue we discovered is that many measurements have zero values, which doesn't make medical sense. For example, 374 women 48.7% have zero insulin levels and 227 women 29.6% have zero skin thickness measurements. These zeros likely represent missing data rather than actual measurements.

2.2 Exploratory Data Analysis

We performed detailed exploratory analysis to understand the patterns in our data. The feature distribution analysis (see Appendix A, Figure A2) shows the histograms for all 8 features, revealing that most variables are right-skewed with some extreme values. Glucose and BMI show relatively normal distributions, while insulin and skin thickness have many zero values creating unusual distribution shapes. The correlation analysis (see Appendix A, Figure A4) showed some interesting relationships between variables. The strongest correlation with diabetes outcome was glucose level r=0.47, which makes sense because high glucose is a main symptom of diabetes. BMI was the second most important predictor r=0.29, followed by age r=0.24. The correlation heatmap clearly shows that pregnancies and age are strongly correlated r=0.54, which is expected since older women typically have had more pregnancies.

The box plots for outlier detection (see Appendix A, Figure A3) shows that most features contain outliers, particularly in insulin, glucose, BMI, and blood pressure measurements. These outliers might be due to underlying medical conditions or measurement errors, but we kept them since they could represent real clinical cases. The pairplot analysis (see Appendix A, Figure A5) examined relationships between key features ie glucose, BMI, age, pregnancies colored by diabetes outcome. This figure shows clear separation between diabetic and non-diabetic groups, especially for glucose levels where the two classes form distinct clusters.

The violin plots by outcome (see Appendix A, Figure A6) provide another view of how feature distributions differ between diabetic and non-diabetic groups. These plots show both the distribution shape and the differences in central tendencies, with diabetic patients generally showing higher values for most features. Finally, the overlapping histograms by outcome (see Appendix A, Figure A7) give the clearest view of how each feature separates the two classes. Glucose shows the best separation, with diabetic patients having a much higher distribution than non-diabetic patients. BMI also shows good separation, while features like blood pressure show considerable overlap between classes.

2.3 Data Preprocessing

Based on our exploratory analysis, we handled the missing data problem by replacing zero values with median values for variables where zero doesn't make medical sense. We focused on glucose, blood pressure, skin thickness, insulin, and BMI because having zero values for these measurements is medically impossible (Mishra, 2024). For example, we replaced 5 zero glucose values with the median of 117.0 mg/dL, 35 zero blood pressure values with 72.0 mmHg, and 374 zero insulin values with 125.0. This approach helps preserve the overall data patterns while providing realistic values for missing measurements. We didn't need to scale the features because decision trees are not sensitive to different measurement scales. Lastly, we split the data into training and testing sets using an 80/20 split, which gave us 614 samples for training and 154 samples for testing. We used stratified sampling to make sure both sets had similar proportions of diabetic and non-diabetic cases.

2.4 Decision Tree Algorithm

Decision trees work by asking a series of yes/no questions about the data to split patients into groups. The algorithm starts with all patients in one group and finds the best question to ask that separates diabetic from non-diabetic patients as much as possible. The "best" question is determined using a measure called entropy, which calculates how mixed up the groups are. The formula for entropy is: $H(S) = -\sum(pi \times log2(pi))$, where pi is the proportion of each class. When entropy is 0, all patients in a group have the same outcome, which is perfect (Zhou, 2022). When entropy is 1, the groups are completely mixed. Information gain measures how much better the separation becomes after asking a question: $IG = H(parent) - \sum(|Sv|/|S| \times H(Sv))$. The algorithm picks the question with the highest information gain and repeats this process for each new group until it meets stopping conditions like maximum depth or minimum number of samples per group.

2.5 Model Development and Hyperparameter Tuning

We used grid search with 5-fold cross-validation to find the best settings for our decision tree. We tested different combinations of parameters including the splitting criterion gini vs entropy, maximum tree depth 3, 5, 7, 10, or unlimited, minimum samples needed to split a node 2, 5, 10, or 20, and minimum samples required in each final group 1, 2, 5, or 10. After testing 160 different combinations, the best model used entropy as the splitting criterion, had a maximum depth of 3 levels, required at least 2 samples to split a node, and needed at least 10 samples in each final group. This model achieved a cross-validation accuracy of 75.7%, which was the highest among all tested combinations.

The grid search process took about 800 model fits across 5 folds, which ensured that our final model parameters were robust and not just lucky results from one particular data split.

3. Results

3.1 Model Performance

Our final decision tree model achieved several important performance metrics on the test data. An accuracy of 69.5%, meaning the model correctly classified about 7 out of every 10 patients. While this might seem modest, it's actually reasonable for medical screening applications. Looking at more detailed metrics shown in the performance summary (see Appendix A, Figure A11), the precision was 65.2%, which means that when the model predicted diabetes, it was correct about 65% of the time. The recall sensitivity was 27.8%, indicating that the model only caught about 28% of actual diabetes cases. The specificity was much higher at 92.0%, showing that the model was very good at correctly identifying patients without diabetes.

The F1-score, which balances precision and recall, was 38.9%. The ROC-AUC score was 77.1%, which indicates good discriminative ability much better than random guessing which would be 50%. The ROC curve analysis (see Appendix A, Figure A9) shows the model performs well above the random classifier line. The confusion matrix (see Appendix A, Figure A8) shows the detailed breakdown: 92 true negatives, 8 false positives, 39 false negatives, and 15 true positives. This pattern shows that the model is quite conservative, rarely predicting diabetes unless it's very confident.

3.2 Model Interpretation

The decision tree structure (see Appendix A, Figure A12) shows exactly how the model makes decisions. The most important split happens at glucose level 154.5 mg/dL. Patients with glucose above this threshold are more likely to have diabetes, while those below are more likely to be healthy. For patients with lower glucose levels, the model looks at BMI and number of pregnancies to make finer distinctions. The feature importance analysis (see Appendix A, Figure A13) confirms that glucose is by far the most important predictor, accounting for 69.9% of the model's decision-making power. BMI comes second with 25.9% importance, followed by pregnancies at 3.2%. Interestingly, several features like age, insulin, blood pressure, and skin thickness had zero importance in the final model. This suggests that once glucose and BMI are considered, these other factors don't add much predictive value in this particular dataset. According to the American Diabetes Association (2024), the glucose threshold of 154.5 mg/dL that the model found is clinically meaningful, as it falls between normal fasting glucose less than 100 mg/dL and diabetes diagnosis levels 126 mg/dL or higher for fasting glucose.

3.3 Model Validation

Cross-validation results (see Appendix A, Figure A10) showed that the model performance was fairly consistent across different data subsets. The 10-fold cross-validation achieved an average accuracy of 74.3% with a standard deviation of 8.5%. Most folds performed between 70-80% accuracy, with the best fold reaching 83.6% and the worst achieving 68.9%. This consistency suggests that our model is not overfitting to the training data and should generalize reasonably well to new patients. The performance shows some variation across folds, which is normal and expected in cross-validation analysis (Lumumba et al., 2024).

4. Conclusion

4.1 Summary of Findings

2 This study successfully developed an interpretable decision tree model for predicting Type 2 diabetes in Pima Indian women. The main finding is that glucose level is overwhelmingly the most important predictor, accounting for nearly 70% of the model's decision-making process (Barakeh, 2024). This aligns well with medical knowledge since high glucose is a primary symptom of diabetes. The model achieved 69.5% accuracy with very high specificity 92.0%, making it potentially useful as a screening tool to rule out diabetes in patients with low glucose levels. The simple tree structure with only 3 levels means that doctors can easily understand and apply the decision rules in clinical practice.

4.2 Limitations

Several important limitations should be considered when interpreting these results. First, the low recall 27.8% means the model misses about 72% of actual diabetes cases, which could be dangerous in a medical setting where failing to detect disease is worse than false alarms. The dataset limitations include the relatively small sample size of 768 patients and the high percentage of missing values, especially for insulin measurements. The model is also specific to Pima Indian women and may not work well for other populations with different genetic backgrounds and lifestyle factors. From a methodological perspective, we only tested one type of algorithm decision trees and didn't explore more advanced techniques that might capture complex relationships between features that a simple tree might miss.

4.3 Future Improvements

Several improvements could enhance this analysis. First, using ensemble methods like random forests or gradient boosting could improve accuracy while maintaining some interpretability. Implementing cost-sensitive learning could help address the low recall by penalizing false negatives more heavily than false positives. Additional clinical features like HbA1c levels, detailed family history, or lifestyle factors could improve prediction accuracy. Collecting data from larger and more diverse populations would help create models that work better across different groups of people. Finally, developing a two-stage screening process where high-risk patients identified by this model receive additional testing could combine the benefits of simple screening with more accurate diagnosis for borderline cases.

Reference

American Diabetes Association. (2024). Standards of Care in Diabetes—2024. Diabetes Care, 47(Supplement_1), S20–S42. https://doi.org/10.2337/dc24-S002

Amini, Z. (2023). Pima Indians Diabetes. Kaggle. Retrieved from https://www.kaggle.com/datasets/aminizahra/pima-indians-diabetes

Barakeh, R. (2024). Leveraging machine learning for precise prediction of Type 2 diabetes. Diabetes, 73(Supplement_1), 59-PUB. https://doi.org/10.2337/db24-59-PUB

Drousiotis, E., Joyce, D., Varsi, A., Spirakis, P., & Maskell, S. (2023). Intrinsically Interpretable Decision Trees for Healthcare Applications. University of Liverpool. Retrieved from

https://assets-eu.researchsquare.com/files/rs-4608203/v1_covered_feab1db3-fcd2-48ac-bd05-13216124023d.pdf?c=1720492046

Mishra, S. R. (2024). Predictive Analysis On Diabetes Detection Using Pima Indian Diabetes Dataset. IJRAR. Retrieved from https://www.academia.edu/123715945/Predictive_Analysis_On_Diabetes_Detection_Using_Pima_Indian_Diabetes_Da

Nicoa, D. (2022). Type II Diabetes in Pima Indians of Arizona. ArcGIS StoryMaps. Retrieved from https://storymaps.arcgis.com/stories/b8b857c36cbb4476a744fcb81b3f232d

Zhou, V. (2022). A Simple Explanation of Information Gain and Entropy. victorzhou.com. Retrieved from https://victorzhou.com/blog/information-gain/

Appendix A: Visualizations

Exploratory Data Analysis

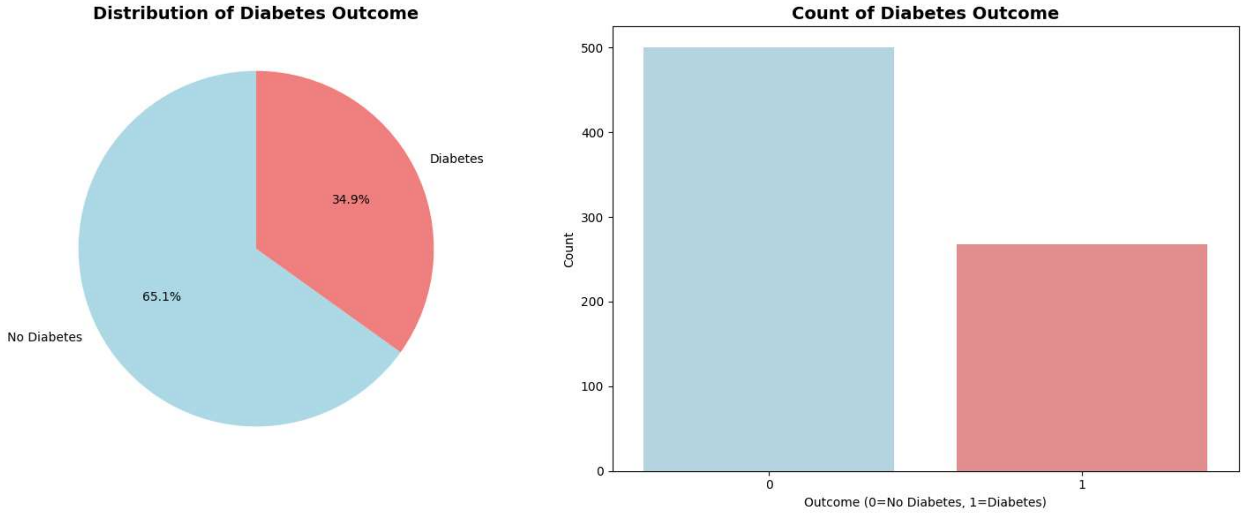

Figure A1: Figure 1: Target Variable Distribution

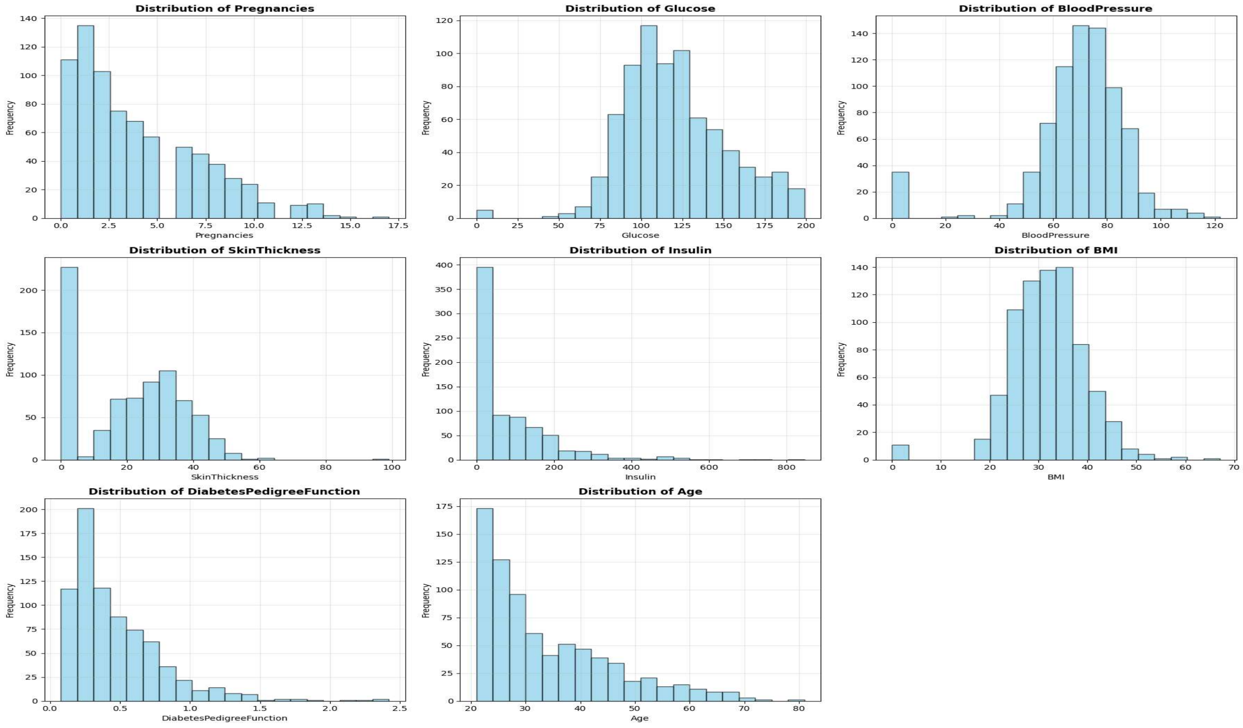

Figure A2: Figure 2: Feature Histograms

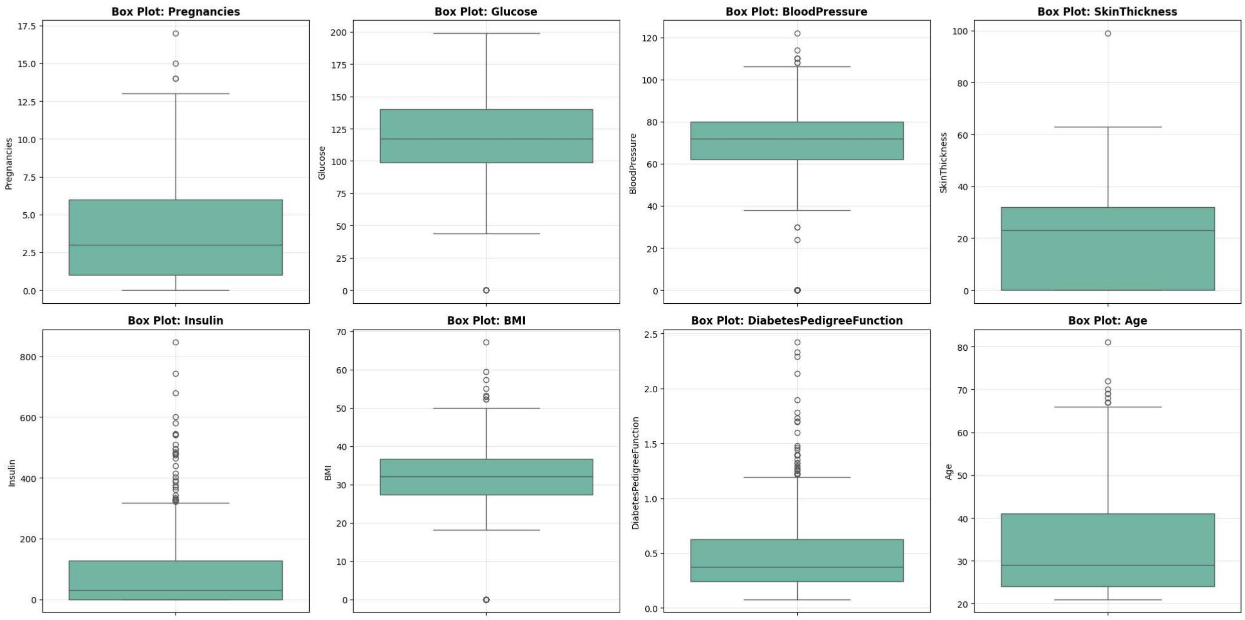

Figure A3: Figure 3:Box Plots for Outlier Detection

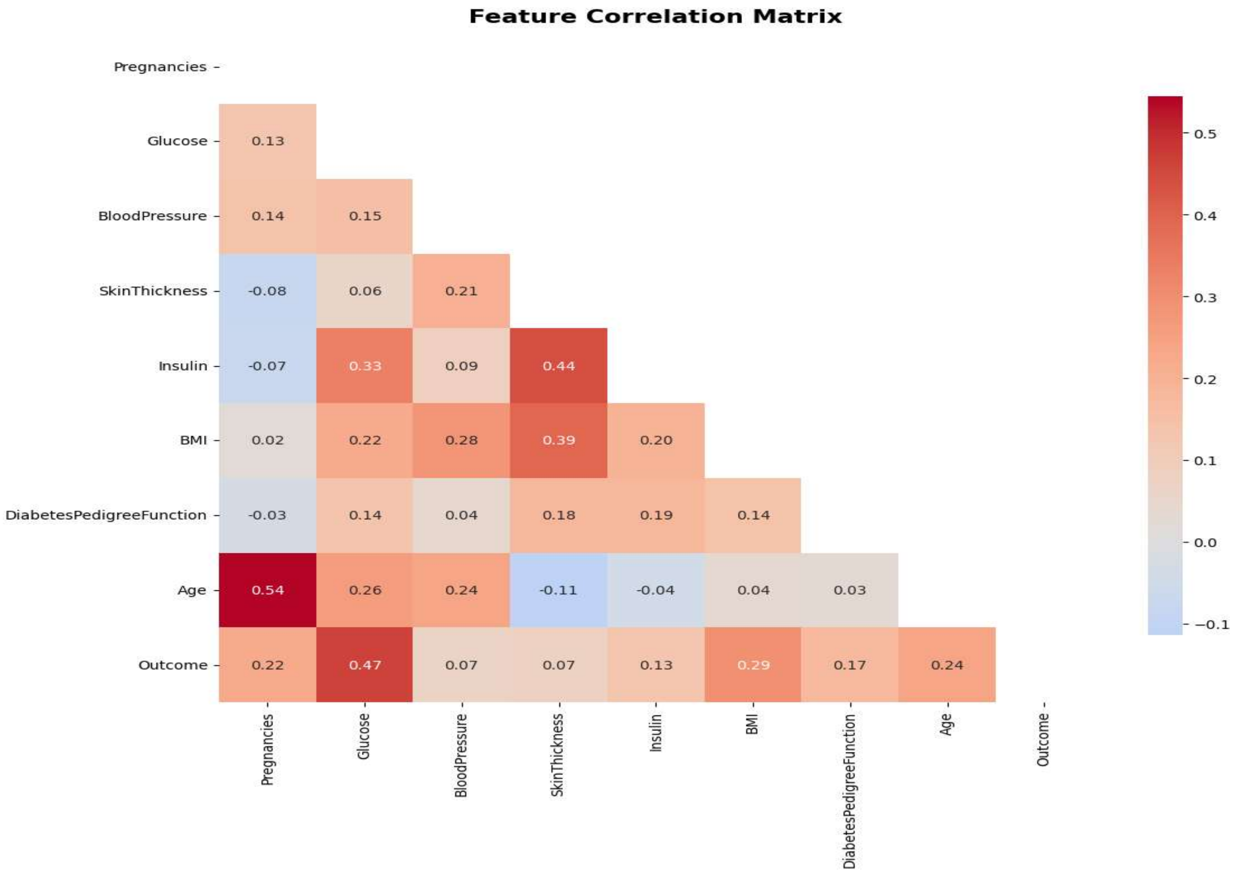

Figure A4: Figure 4: Correlation Heatmap

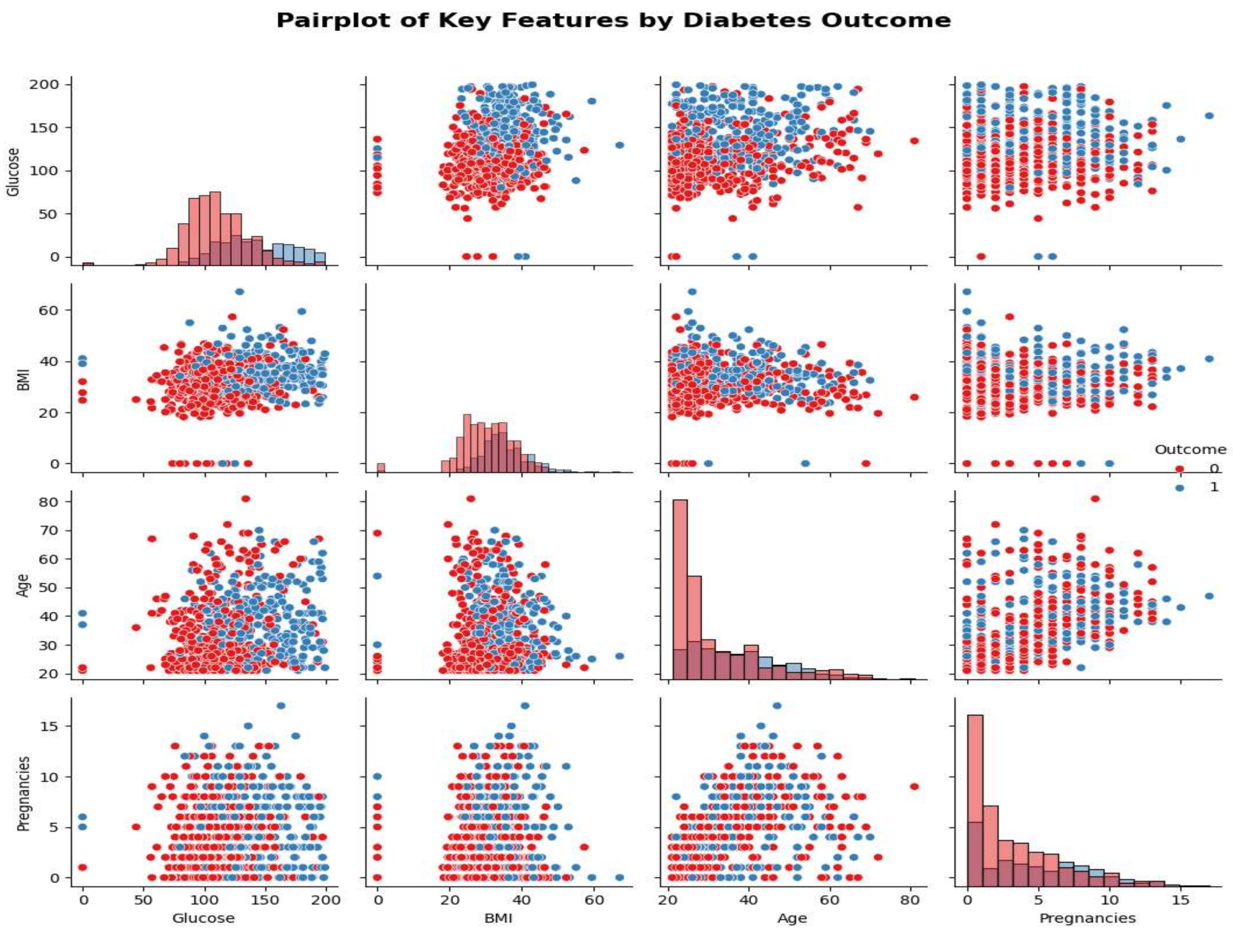

Figure A5: *Pairplot of Key Features by Outcome*

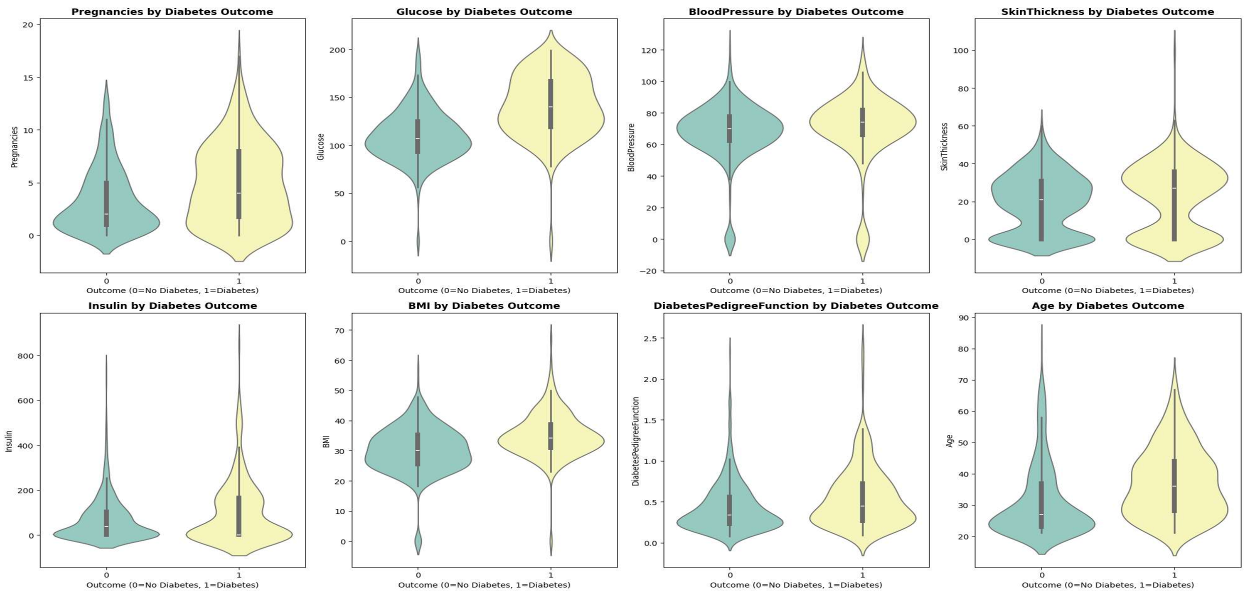

Figure A6: *Figure 5: Violin Plots by Diabetes Outcome*

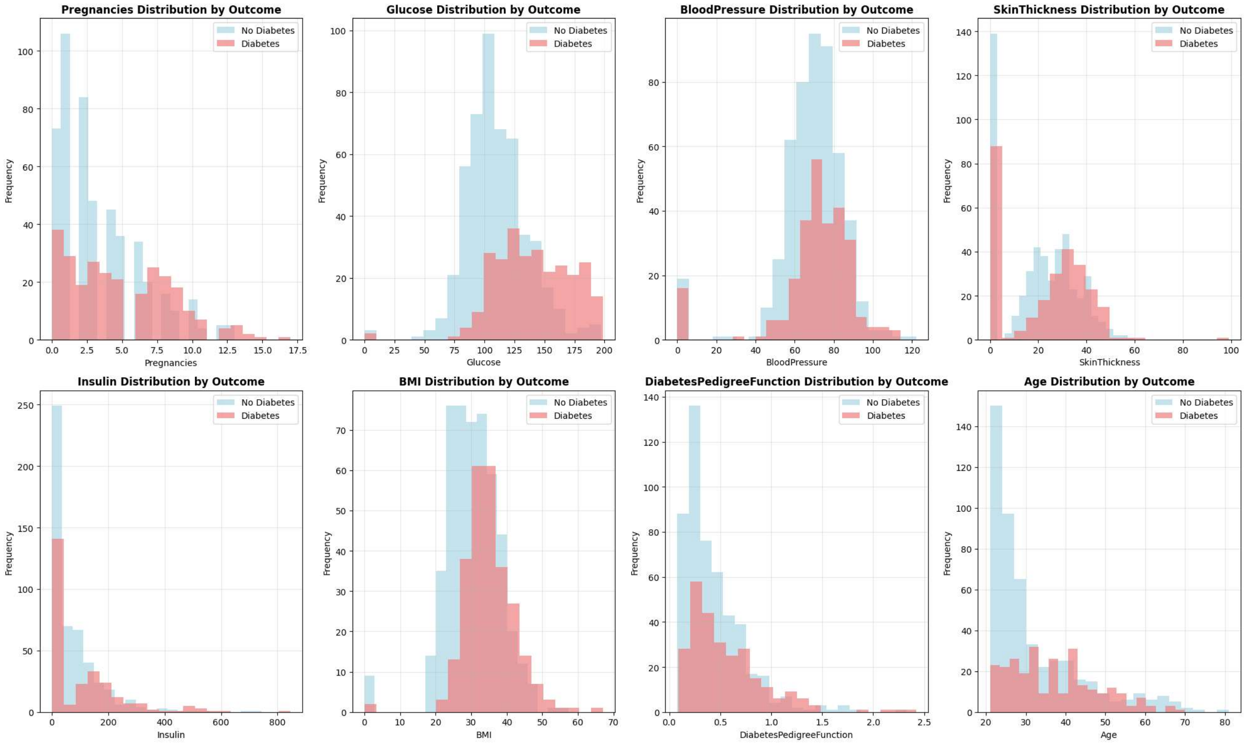

Figure A7: *Figure 6: Overlapping Histograms by Outcome*

Model Results

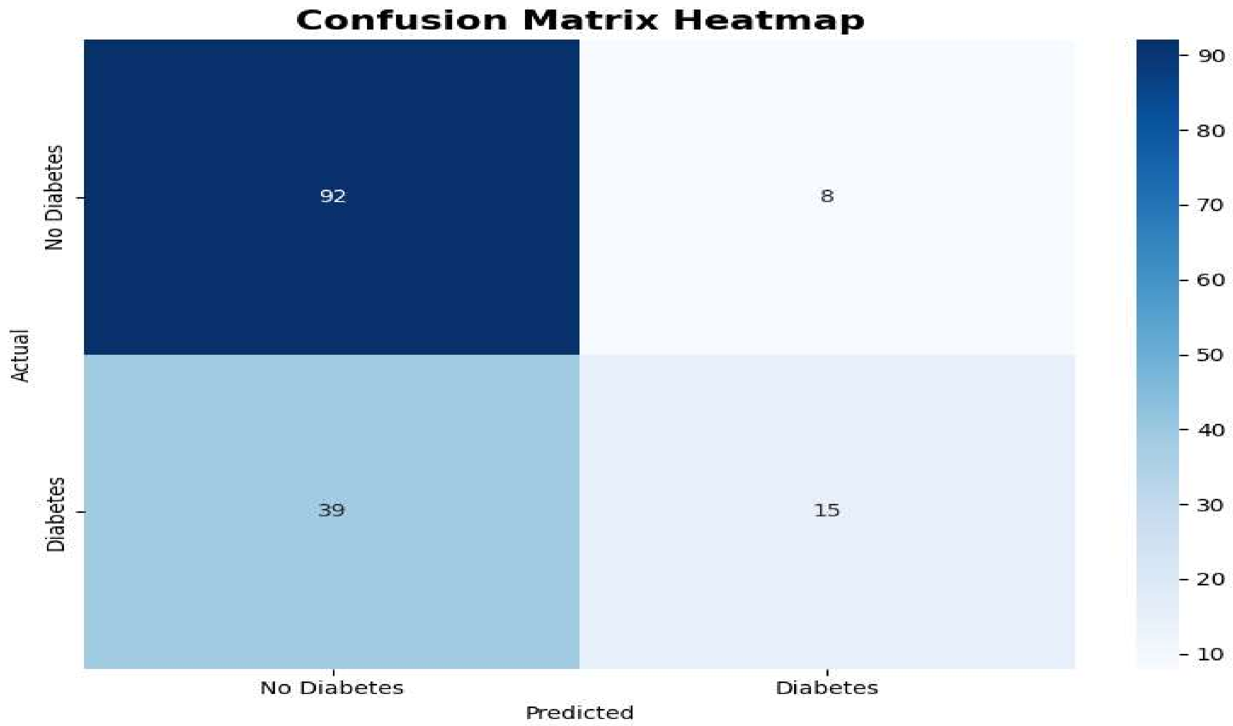

Figure A8: *Figure 7: Confusion Matrix Heatmap*

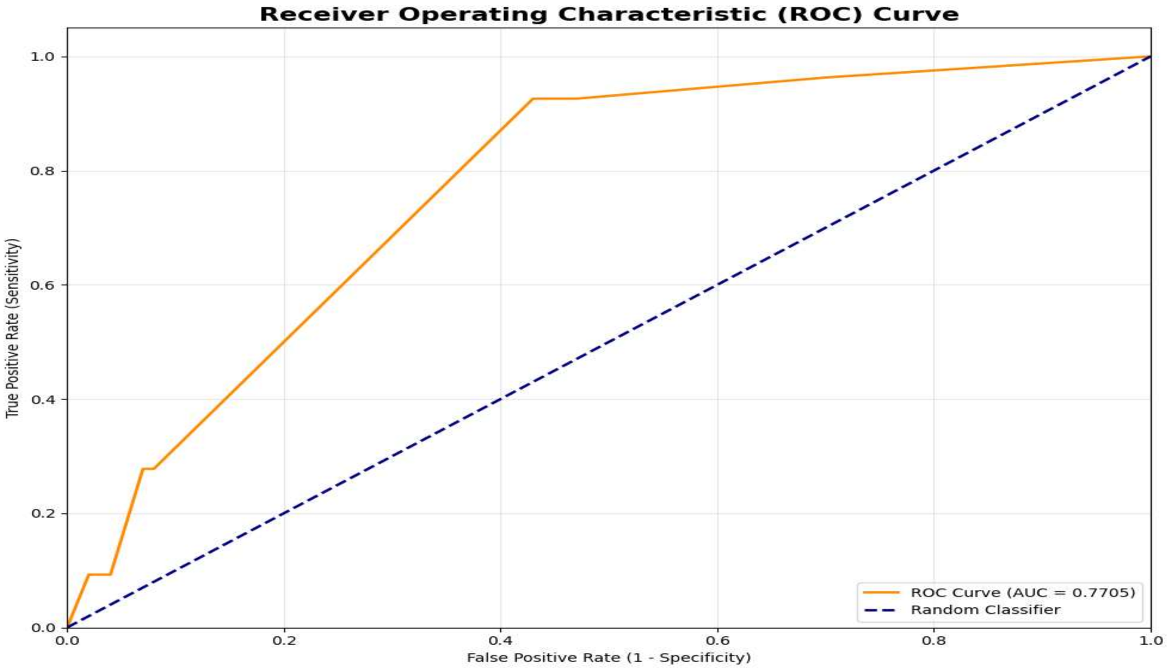

Figure A9: *Figure 8: ROC Curve Analysis*

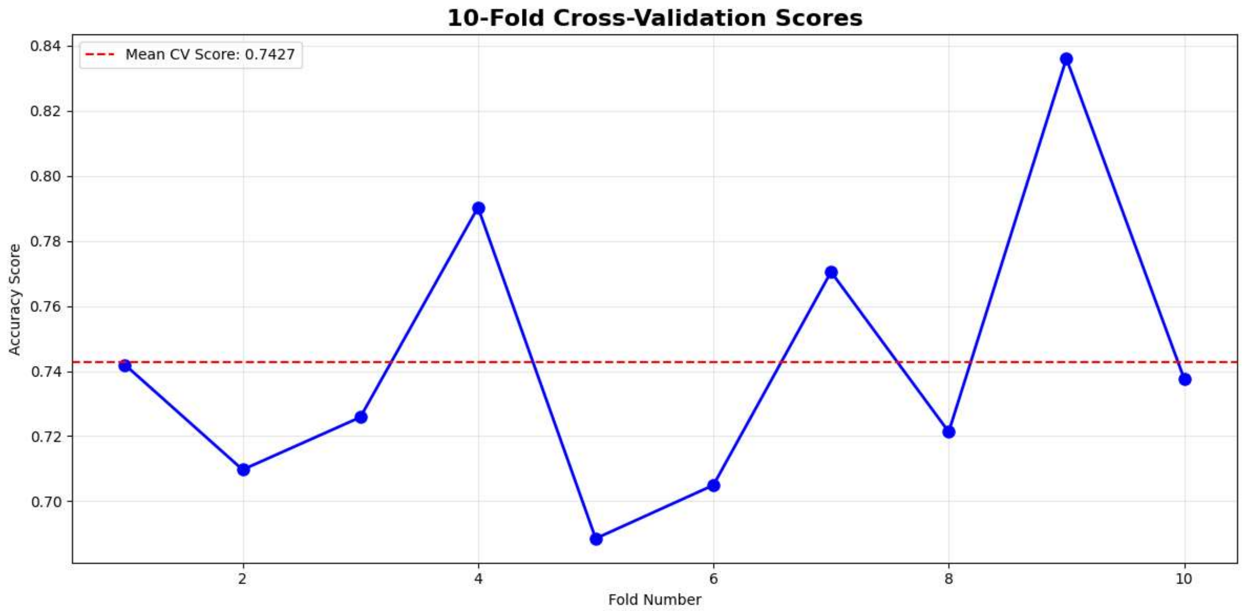

Figure A10: *Figure 9: Cross-Validation Performance Results*

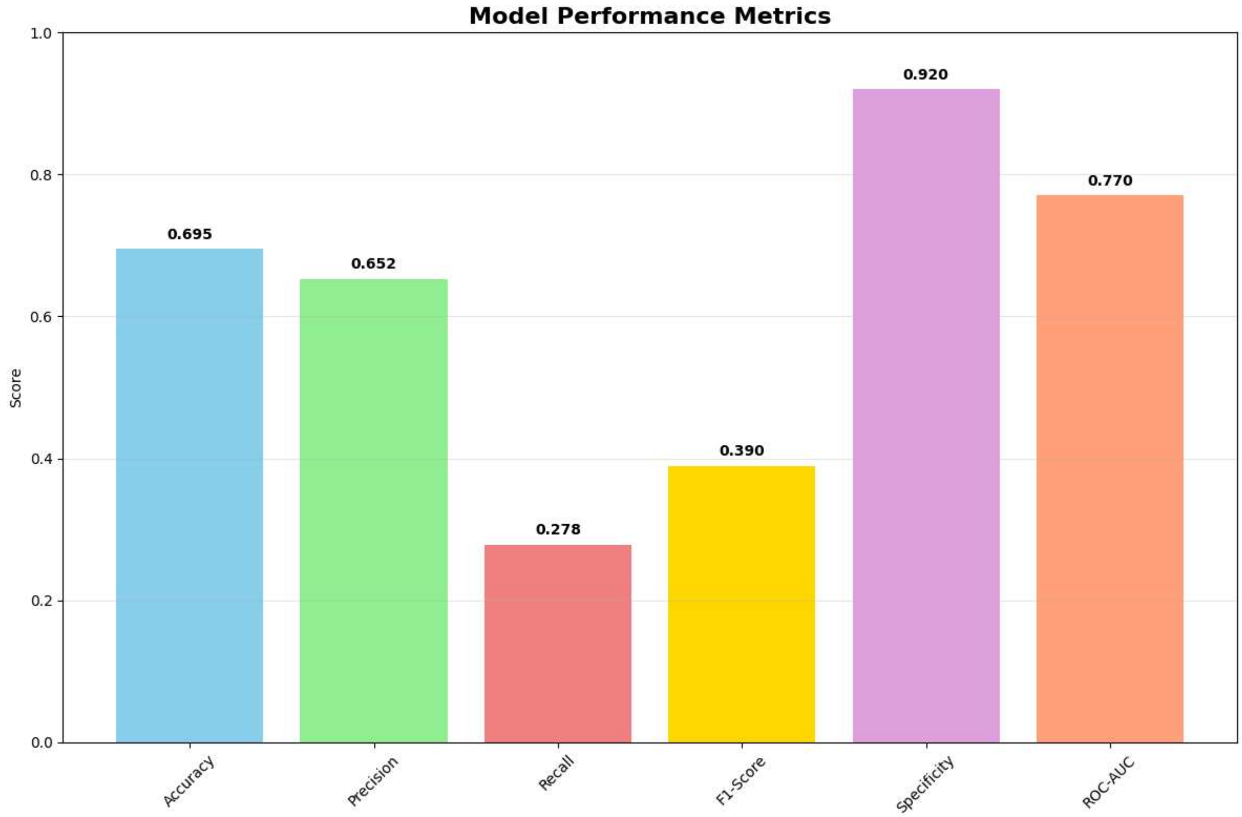

Figure A11: *Figure 10: Model Performance Metrics Summary*

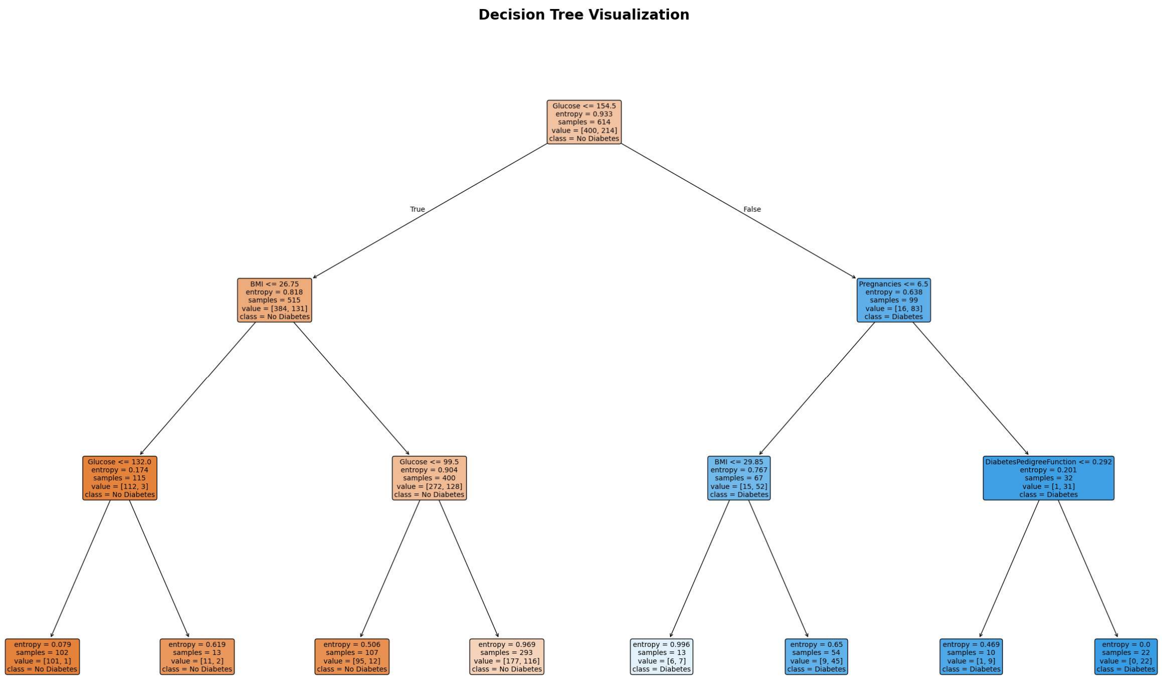

Figure A12: Figure 11: Decision Tree Visualization

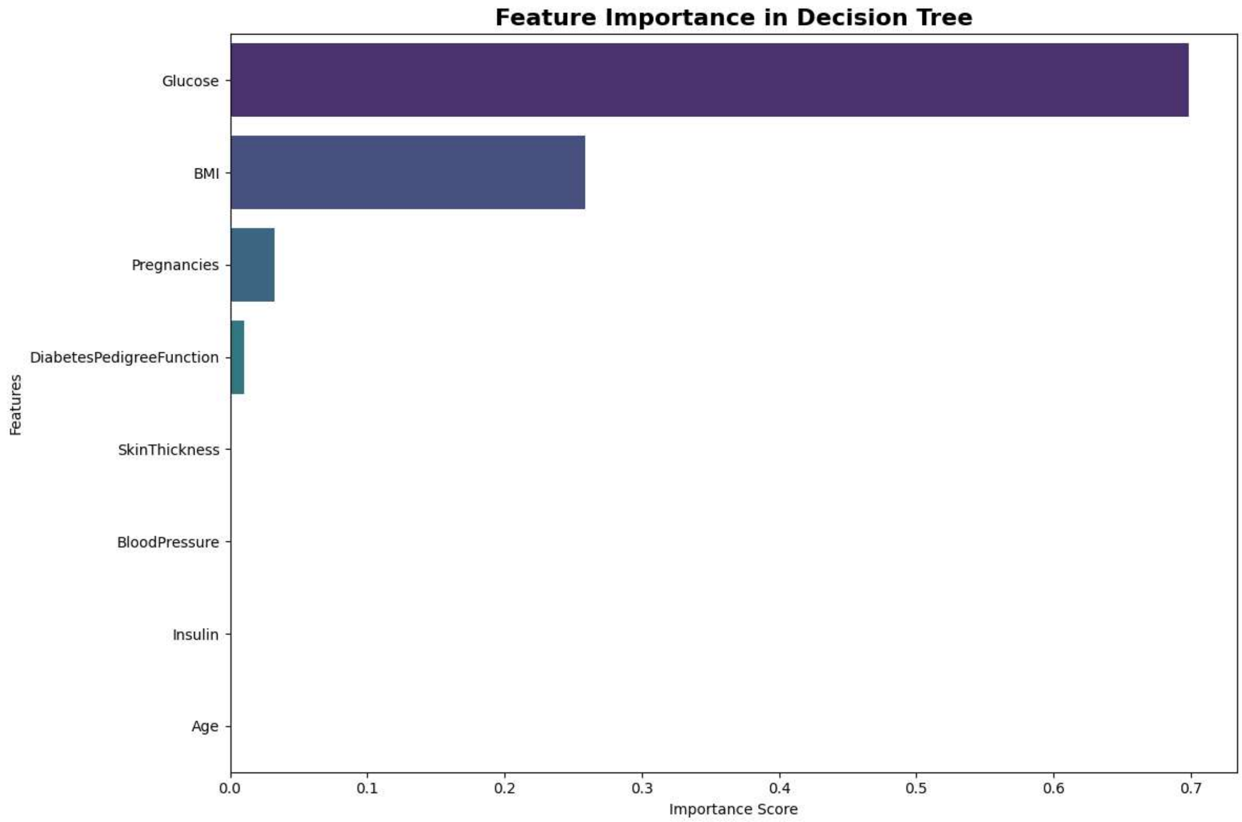

Figure A13: Figure 12: Feature Importance