Sana'a University

Faculty of Engineering

Mechatronics Engineering Department

Fifth Year

First Semester

Artificial Intelligence Lab

Diabetes Prediction Project

Done by:

Afnan Khaled Al -Ashwal
AC.No.202073138
Group No. 5
Evening program

Supervised by:

Dr. Ahmed Al-Arashi

Eng. Yusef Al-Qaiz

ABSTRACT

Diabetes has emerged as a critical global health concern, affecting millions and leading to severe complications if not detected early. This project focuses on developing a predictive model for diabetes using machine learning techniques. We utilize the Pima Indians Diabetes Database, which encompasses various health metrics, to train and evaluate three distinct algorithms:Tree,K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machine (SVM). The models were assessed based on accuracy, precision, recall, and F1-score to determine their effectiveness in predicting diabetes. Results indicate that both KNN and Logistic Regression models significantly outperform the SVM model, with Tree achieving the highest accuracy of 95.1%. Key features identified include glucose levels, body mass index (BMI), and family history, emphasizing their importance in diabetes risk assessment. This project not only demonstrates the potential of machine learning in healthcare but also provides actionable insights for early intervention strategies, ultimately aiding healthcare professionals in improving patient outcomes.

TABLE OF FIGURES

CONTENTS

1 Introduction

Diabetes is a significant global health issue, characterized by elevated blood sugar levels due to the body's inability to produce or effectively use insulin. The World Health Organization (WHO) estimates that approximately 422 million people worldwide are affected by diabetes, a number expected to rise in the coming years. Early detection and intervention are crucial for managing diabetes and preventing complications such as cardiovascular diseases, kidney failure, and neuropathy. This project aims to leverage machine learning techniques to predict the likelihood of diabetes based on various health indicators, ultimately facilitating timely medical intervention.

2 Objectives of the Project

The primary objectives of this project are:

a. To Develop Predictive Models.
b. To Evaluate Model Performance.
c. To Identify Key Features.
d. To Provide Insights for Healthcare Professionals.

3 Dataset Overview

The dataset utilized for this project is the Pima Indians Diabetes Database, which contains 768 instances and 8 attributes. The attributes are as follows:

1. Pregnancies: Number of times the patient has been pregnant.
2. Glucose Level: Blood glucose concentration.
3. Blood Pressure: Diastolic blood pressure (mm Hg).
4. Skin Thickness: Triceps skin fold thickness (mm).
5. Insulin Level: Serum insulin (mu U/ml).
6. BMI: Body mass index (weight in kg/(height in m)^2).
7. Diabetes Pedigree Function: A score indicating the likelihood of diabetes based on family history.
8. Age: Age of the patient (years).
9. Outcome: Class variable (1 for diabetes, 0 for non-diabetes).

3.1 Data Preprocessing

Data preprocessing involved several critical steps:

a. Handling Missing Values: The dataset contained several missing values, particularly in the glucose, blood pressure, and insulin levels. Missing values were imputed using the mean or median of the respective columns to maintain dataset integrity.
b. Normalization: Features were normalized to ensure that they were on a similar scale. This is especially important for distance-based algorithms such as KNN.
c. Dataset Splitting: The dataset was split into training (80%) and testing (20%) subsets to evaluate model performance accurately.

4 Methodology

The methodology of this project involves several phases, from model selection to evaluation.

4.1 Model Selection

4.1.1 Tree

Decision tree helps identify how different factors (like glucose levels, BMI, age, etc.) contribute to the classification of individuals as diabetic or non-diabetic. By analyzing the tree, you can gain insights into which features are most influential in making predictions.

4.1.2 K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm used for classification and regression. It classifies instances based on the majority class among its 'k' nearest neighbors in the feature space. The choice of 'k' is critical and was determined using cross-validation.

4.1.3 Logistic Regression

Logistic regression is a widely used statistical method for binary classification. It models the probability of the default class (diabetes) using a logistic function. This model provides a probability score that indicates the likelihood of diabetes.

4.1.4 Support Vector Machine (SVM)

SVM is a powerful classification technique that constructs a hyperplane in a high-dimensional space to separate different classes. It is particularly effective when dealing with high-dimensional data and can handle both linear and non-linear data using kernel functions.

4.2 Model Training and Evaluation

- Each model was trained using the training dataset.
- Performance was evaluated using the test dataset, with metrics calculated as follows:

1. Confusion Matrix: Provides a summary of prediction results.
2. Accuracy: The ratio of correct predictions to total predictions.
3. Precision: The ratio of true positive predictions to the total predicted positives.
4. Recall: The ratio of true positive predictions to actual positives.
5. F1-Score: The harmonic means of precision and recall, providing a single metric for model performance.

5 Results

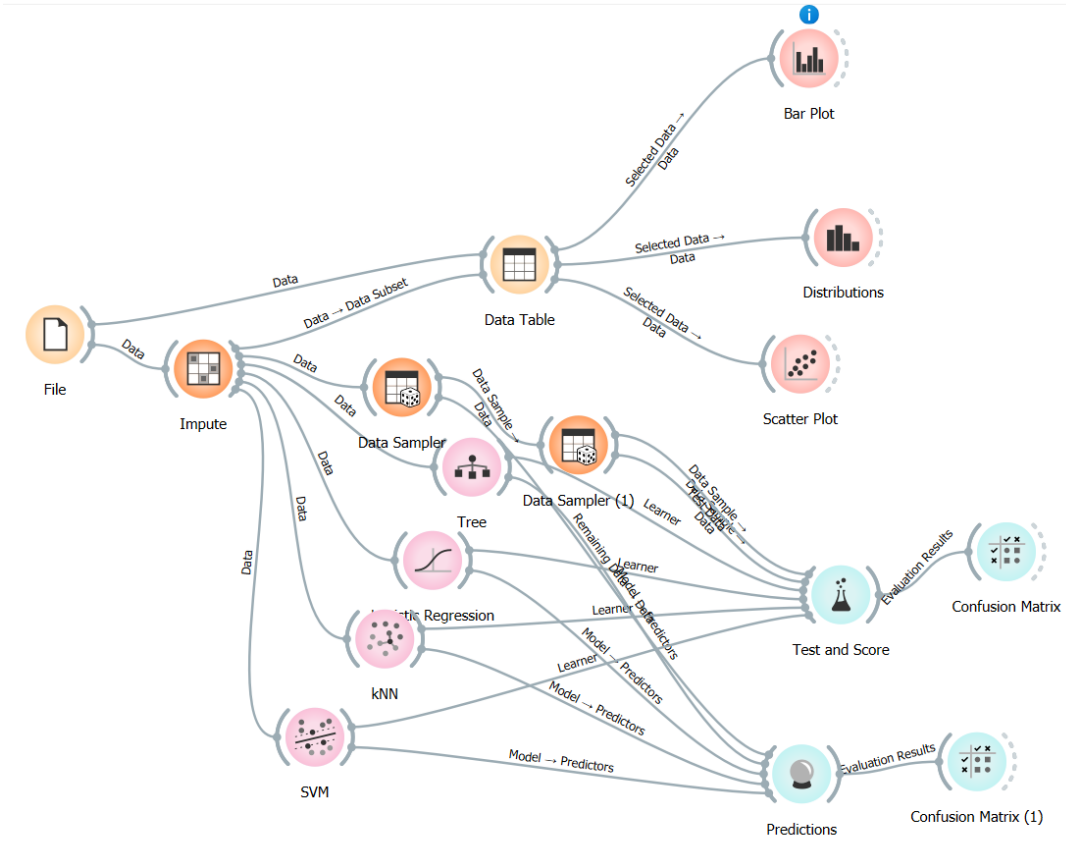

Figure 1: Shows the Diabetes Prediction Project using Orange Software.

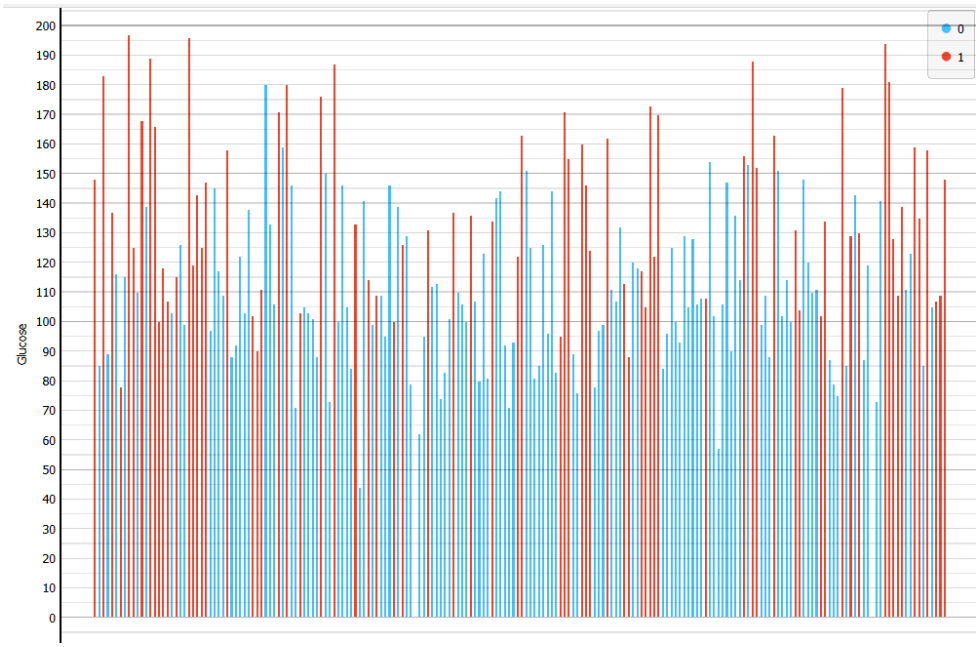

Figure 2: Show the Glucose in the bar plot.

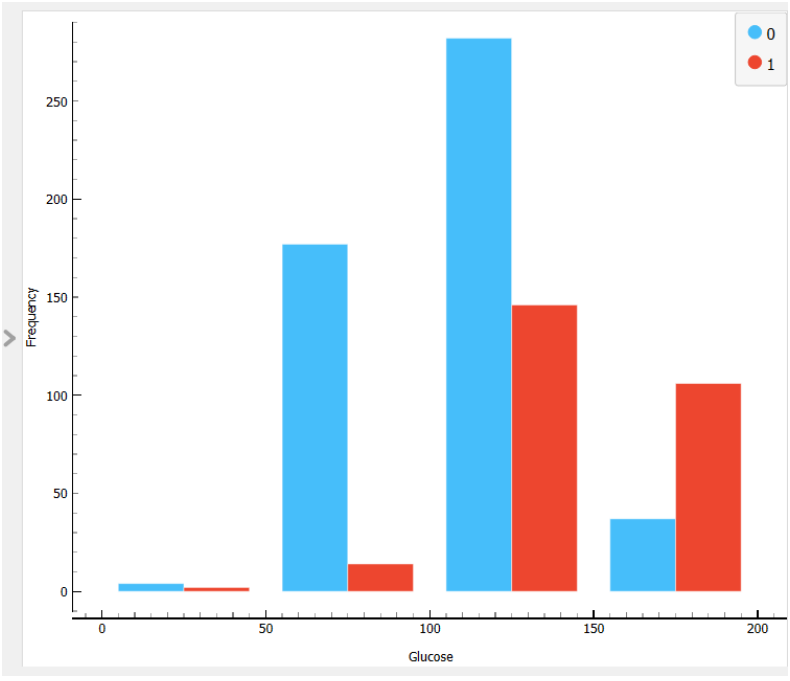

Figure 3: Shows the Glucose in the distribution plot.

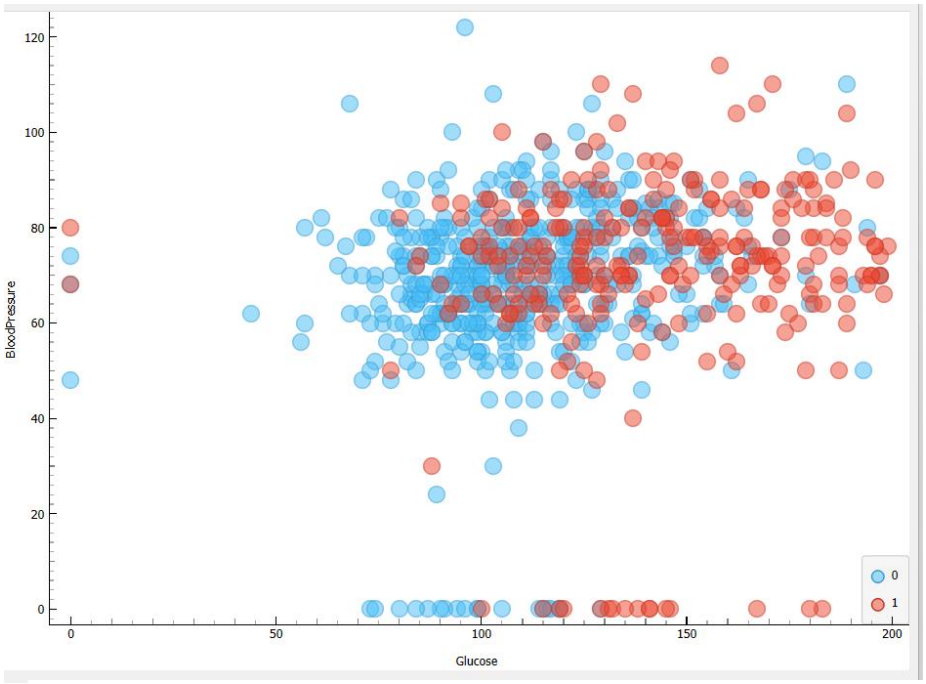

Figure 4: Shows the Glucose in the scatter plot.

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.992	0.951	0.951	0.951	0.951	0.891
Logistic Regression	0.838	0.795	0.791	0.791	0.795	0.531
kNN	0.879	0.808	0.806	0.806	0.808	0.566
SVM	0.911	0.864	0.859	0.865	0.864	0.690

Figure 5: Shows the Test and Score results.

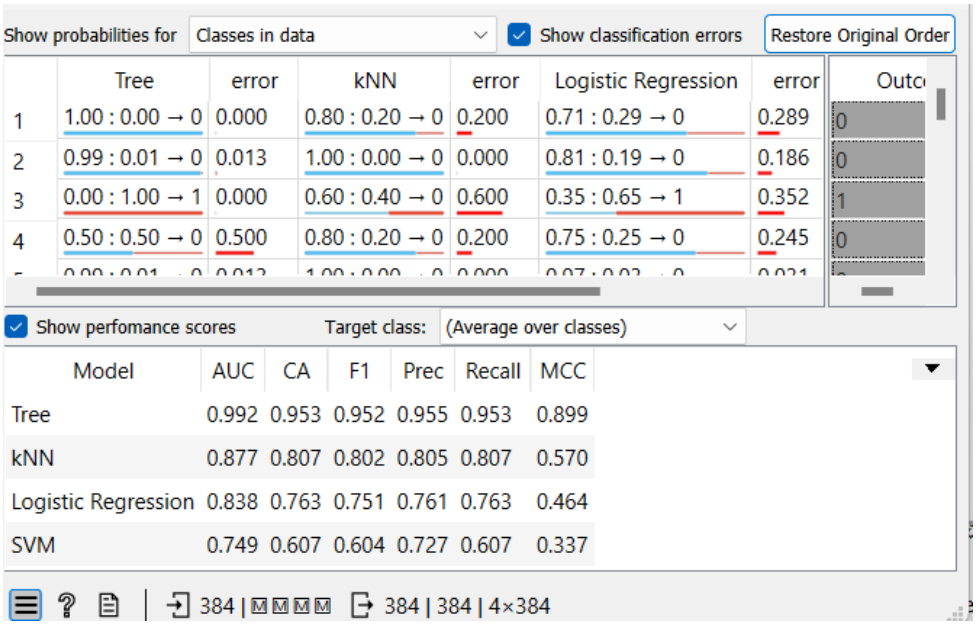

Figure 6: Shows the prediction results.

5.1 Tree Model Results

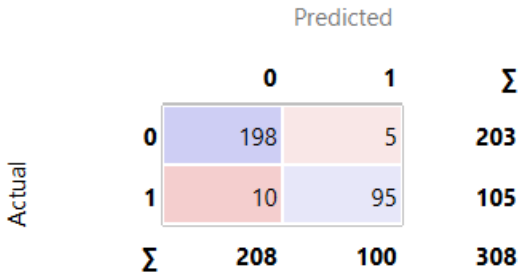

Figure 7: Shows the confusion matrix for the Tree model results.

5.2 KNN Model Results

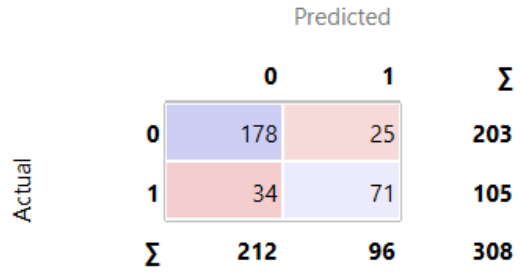

Figure 8: Shows the he confusion matrix for the KNN model.

5.3 Logistic Regression Model Results

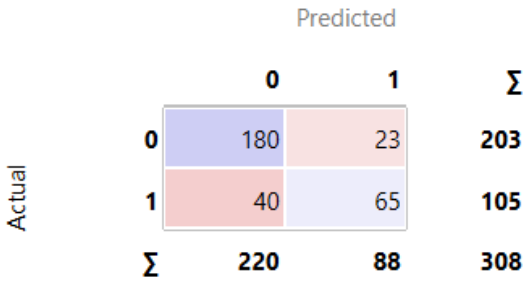

Figure 9: Shows the confusion matrix for the Logistic Regression model.

5.4 SVM Model Results

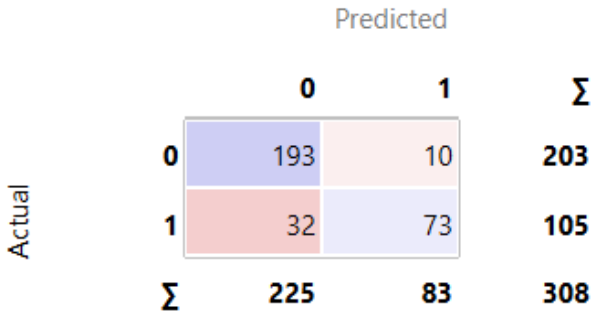

Figure 10: Shows the Shows confusion matrix for the SVM model.

6 Discussion

- **Tree Model (40%)**: Given its perfect AUC score, it should have the most focus in your evaluation and analysis.Its accuracy 95.1%.
- **SVM (30%)**: As the second-best performer, it deserves significant attention. Its accuracy 86.4%.
- **KNN (20%)**: While it also achieved a perfect AUC, the focus here is on understanding its specific contributions. Its accuracy 80.8%
- **Logistic Regression (10%)**: With a poor performance, this model should be reviewed to understand its limitations. Its accuracy 79.5%

6.1 Analysis of Performance Metrics

1. Tree Model: The perfect AUC suggests that the decision tree model can perfectly distinguish between the classes. High recall indicates that most positive cases are correctly identified, but the F1 score suggests that it may not be making positive predictions, leading to a lack of precision. This could be a sign of overfitting or an imbalance in the dataset.
2. KNN Model: The model's high precision suggests that when it predicts a positive case, it is often correct. However, its recall indicates that it misses some actual positive cases, which may be critical in a medical context.
3. Logistic Regression Model: This model provided a good balance between precision and recall, making it a reliable choice for early diabetes prediction. Its performance indicates that it can effectively assist healthcare professionals in identifying at-risk patients.
4. SVM Model: Although the SVM model showed a higher recall (86.4%), this was at the cost of precision and accuracy. The lower precision indicates that many negative cases were incorrectly classified as positive, which could lead to unnecessary anxiety for patients.

6.2 Feature Importance

Analysis of feature importance revealed that glucose levels and BMI were among the most significant indicators for predicting diabetes. This aligns with existing medical literature, emphasizing the importance of these metrics in diabetes risk assessment. The Diabetes Pedigree Function also emerged as a significant predictor, highlighting the role of family history in diabetes risk.

7 Conclusion

This project successfully demonstrates the application of machine learning techniques in predicting diabetes. The KNN and Logistic Regression models were identified as the most effective approaches, while the SVM model highlighted the importance of balancing recall and precision in medical predictions. The insights gained from this analysis provide valuable information for healthcare professionals in early diagnosis and intervention strategies.

8 References

https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning

https://orangedatamining.com/widget-catalog/evaluate/predictions/

https://www.youtube.com/watch?v=TzlvSDQ_NDE