Yerevan State University
Faculty of Mathematics and Mechanics
Applied Statistics and Data Science

Predicting Diabetes in Patients

Hasmik Margaryan
Hranush Sahradyan

Abstract

The objective of this analysis is to develop a machine learning model to predict the presence or absence of diabetes in patients. A dataset containing various health measurements of individuals was used for this task. After preprocessing the data and exploring its characteristics, several machine learning algorithms, including Logistic Regression, Support Vector Machines, Random Forest, and AdaBoost, were trained and evaluated. The best-performing model was selected based on its performance metrics. The results demonstrate the effectiveness of the chosen model in accurately predicting diabetes. The findings of this study have important implications for early diagnosis and management of diabetes.

1 Introduction

Diabetes is a prevalent chronic disease that affects millions of individuals worldwide. Early detection and accurate prediction of diabetes are crucial for effective disease management and prevention of complications. In this report, we aim to develop a machine learning model that can predict the presence or absence of diabetes based on various health measurements. By leveraging a dataset containing information on pregnancies, glucose levels, blood pressure, BMI, and other factors, we can provide valuable insights into the predictive capabilities of machine learning algorithms in the context of diabetes diagnosis. Accurate prediction models can aid healthcare professionals in identifying individuals at risk and implementing appropriate preventive measures.

2 Data Description and Preprocessing

The dataset used in this analysis is the Pima Indians Diabetes Database, which consists of 768 instances and 9 features.

Pregnancies: This feature represents the number of times an individual has been pregnant. It's important as several studies have indicated that the risk of developing diabetes increases with the number of pregnancies.

Glucose: This represents the plasma glucose concentration a 2 hours in an oral glucose tolerance test. Elevated levels of glucose in the blood, or hyperglycemia, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

BloodPressure: This feature signifies diastolic blood pressure (in mm Hg). Persistent high blood pressure, also known as hypertension, can lead to various health problems including heart disease, kidney disease, stroke, and can also be a risk factor for the development of diabetes.

SkinThickness: This refers to the triceps skin fold thickness (in mm). It's a measure of body fat, and higher values may indicate overweight or obesity, which are known risk factors for diabetes.

Insulin: This is the 2-Hour serum insulin (in mu U/ml). Insulin is a hormone that regulates blood sugar, and problems with insulin production or function can lead to the development of diabetes.

BMI: This feature is the Body Mass Index (weight in $kg/(heightinm)^2$). Like skin thickness, it's a measure of body fat, and high BMI values (overweight or obesity) are associated with an increased risk of diabetes.

DiabetesPedigreeFunction: This is a function that scores likelihood of diabetes based on family history. It's based on the premise that the genetic predisposition to the disease can be quantified and that a family history of the disease increases the risk.

Age: This represents age in years. Aging is associated with changes in body composition, insulin secretion and action, and glucose metabolism, all of which can increase the risk of developing diabetes.

Outcome: This is the class variable (0 or 1). In this dataset, 268 of 768 instances are 1 (representing diabetes), and the rest are 0 (no diabetes). This is our target variable which we aim to predict based on the other features.

The dataset typically contains records of around 768 female Pima Indians. It has no missing values or duplicated values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 1: Data overview

During the data check, we found incorrect values for important features and since they are not small amount compared to the volume of our data, we decide to replace the incorrect values with nans and fill them with appropriate method after examining the distributions of these variables.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2: Statistical measures of numerical features

Pregnancies: The mean number of pregnancies is approximately 3.85, with a standard deviation of 3.37. The minimum value is 0, indicating women who haven't been pregnant, and the maximum is 17, representing the highest number of pregnancies in the dataset.

Glucose: The mean glucose level is around 120.89 mg/dL, with a standard deviation of 31.97. The minimum value of 0 seems incorrect since glucose levels cannot be zero. This might be missing data or an error in data collection.

BloodPressure: The mean blood pressure is approximately 69.11 mm Hg, with a standard deviation of 19.36. The minimum value of 0 seems incorrect as blood pressure cannot be zero. Similar to the glucose column, this could be missing data or measurement errors.

SkinThickness: The mean skin thickness is around 20.54 mm, with a standard deviation of 15.95. The minimum value of 0 also seems incorrect, as it is unlikely for someone to have zero skin thickness. This could be missing data or measurement errors.

Insulin: The mean insulin level is approximately 79.80 mu U/ml, with a standard deviation of 115.24. The minimum value of 0 seems incorrect since insulin levels cannot be zero. This could be missing data or measurement errors.

BMI: The mean BMI (Body Mass Index) is approximately 31.99 kg/m^2, with a standard deviation of 7.88. There are no obvious incorrect values in this column.

DiabetesPedigreeFunction: The mean diabetes pedigree function value is 0.47, with a standard deviation of 0.33. This column represents a numerical value related to family history of diabetes, and there are no obvious incorrect values.

Age: The mean age is around 33.24 years, with a standard deviation of 11.76. The minimum age is 21, and the maximum age is 81, indicating the age range of the patients in the dataset.

Outcome: This column represents the target variable, where 0 indicates no diabetes and 1 indicates diabetes. The mean of 0.35 suggests that the dataset is slightly imbalanced, with slightly more non-diabetic cases.

	Zero Count
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374

Figure 3: Incorrect data counts

3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to gain a deeper understanding of the dataset. Visualizations such as histograms, box plots, and scatter plots were utilized to explore the relationships between the features and the target variable. One interesting finding was that the glucose levels showed a noticeable difference between diabetic and non-diabetic patients, indicating its potential importance as a predictor. Additionally, age and BMI demonstrated varying distributions among the two groups. These observations suggest that these features may play a significant role in determining the presence of diabetes and warrant further investigation in the modeling phase.

There is only 0 and 1 values in target variable so there is no incorrectness, and we also can see that Outcome is imbalanced, there are more observations about healthy patients.

There is high correlation coefficient between Age and Pregnancies then between Insulin and Glucose features. And let's see correlation between target variable and features.[fig.5]

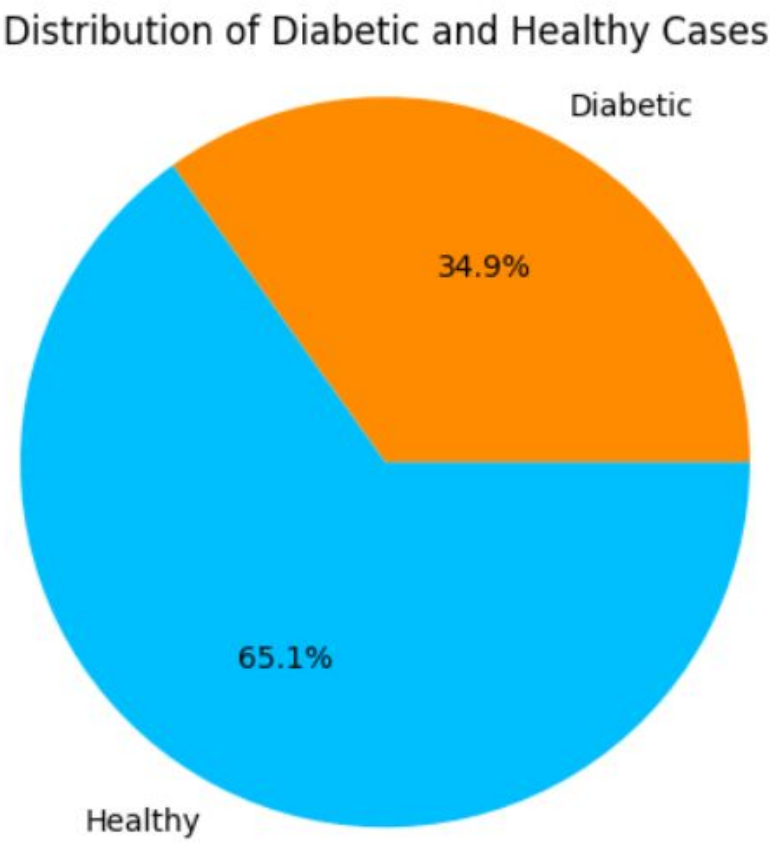

Figure 4: Distribution of target variable

We see that target is more correlated with Glucose and BMI.[fig. 6]

Here we can see for patienc who has diabet the features values medean and interquartile range(50 percent of the data) differs, such as in variables Glucose, Insulin, Skinthickness but not for Diabets Predigree Function values.[fig.7]

Here what says National Library of Medicine about Glucose levels:[fig.8]

From plot above we understand that glucose levels alone may not be a definitive indicator of diabetes. While high glucose levels are commonly associated with diabetes, they can also occur in non-diabetic individuals due to various factors such as diet, stress, and other medical conditions.[fig 9]

4 Model Training and Evaluation

For this classification task, we employed several machine learning models, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and AdaBoost. Logistic Regression is a popular choice for binary classification tasks, while SVM is known for its effectiveness in separating classes with complex decision boundaries. Random Forest, GradientBoosting and AdaBoost are ensemble methods that combine multiple weak classifiers to improve predictive performance. The dataset was split into training and testing sets, and cross-validation techniques, such as k-fold

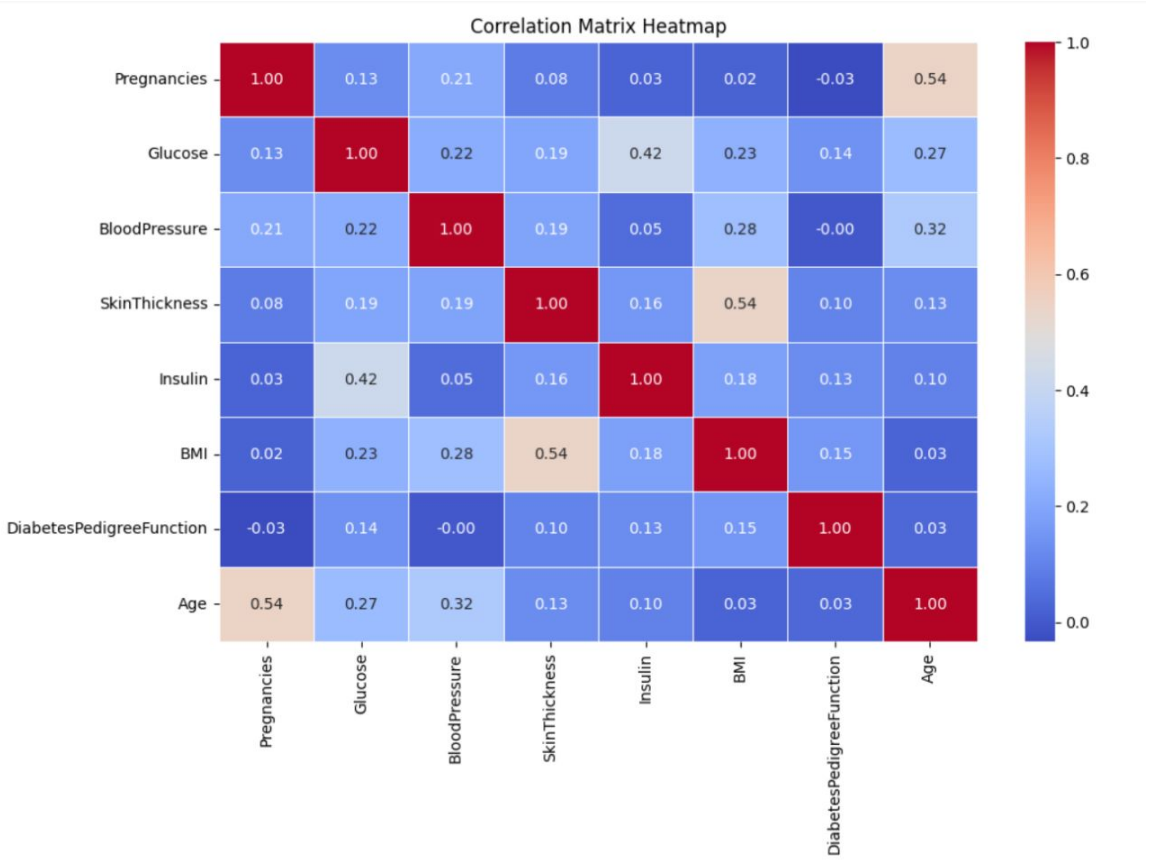

Figure 5: Correlation matrix

cross-validation, were employed to ensure robustness of the models. Evaluation metrics, including accuracy, precision, recall, and F1-score, were utilized to assess the models' performance and generalize their predictive capabilities to unseen data.

Model Selection and Hyperparameter Tuning

Hyperparameter tuning was conducted to optimize the performance of the machine learning models. Grid search, a widely used hyperparameter optimization technique, was applied to explore different combinations of hyperparameters for each model. The models were evaluated using cross-validation, and the best-performing model was selected based on the highest average validation score.(fig.10) For example, in the case of Random Forest, hyperparameters such as the number of estimators, maximum depth, and minimum sample split were tuned. The performance of the tuned models significantly improved compared to their default configurations. After careful consideration of the performance metrics, the Random Forest model was chosen as the best-performing model for this task

5 Explainable AI

To gain insights into the decision-making process of the Random Forest model, explainable AI techniques were employed. Global explanations, such as feature importance scores(fig 11), were utilized

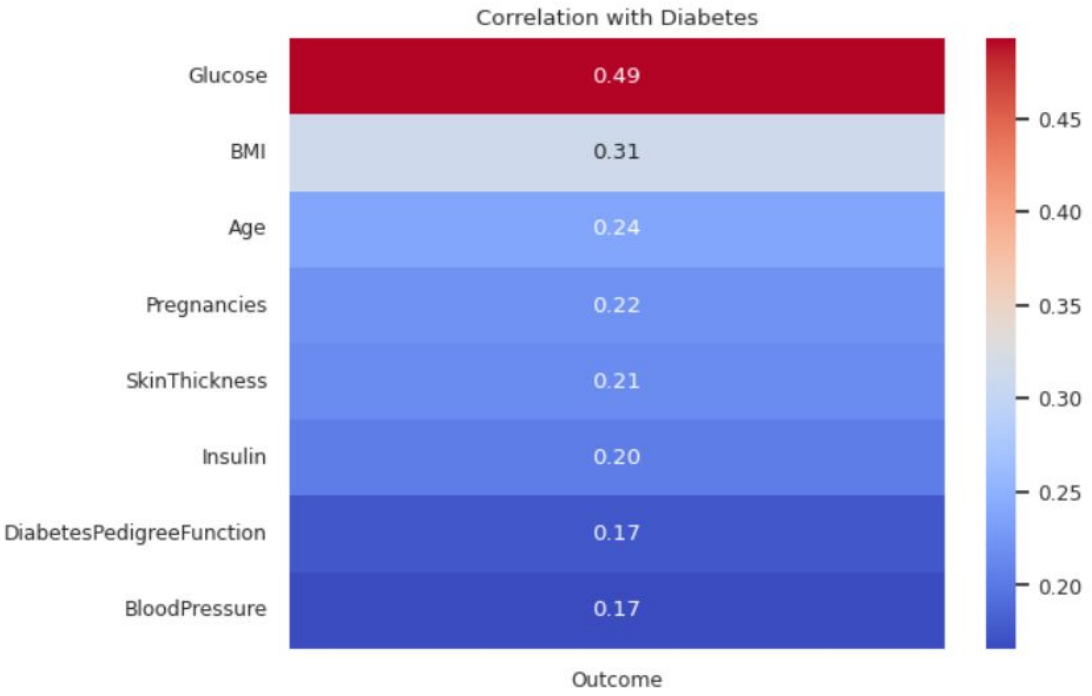

Figure 6: Correlation coefficients for target variable

to identify the most influential features for the overall model performance. These explanations provide valuable insights into the factors considered by the model when predicting the presence or absence of diabetes, enhancing transparency and interpretability.

Results and Discussion

The Random Forest model was selected as the best model based on the highest validation score obtained during the grid search process. The model demonstrated strong performance on the test dataset, achieving an accuracy of 0.73. The precision, recall, and F1-score for each class are as follows:(fig. 12)

The model demonstrated higher precision for class 0, indicating a higher ability to correctly identify non-diabetic individuals. However, the recall and F1-score for class 1 were relatively lower, suggesting that the model had more difficulty in correctly identifying diabetic cases.(fig 13)

6 Conclusion

Based on the comprehensive evaluation of various machine learning models, the Random Forest model was selected as the best-performing model for the task of predicting diabetes. It achieved the highest validation score during the grid search process and demonstrated satisfactory performance on the test dataset. The model exhibited promising precision for non-diabetic cases and reasonable overall accuracy. However, further improvements are required to enhance the model's ability to correctly identify individuals with diabetes, as reflected in the relatively lower recall and F1-score for class 1.

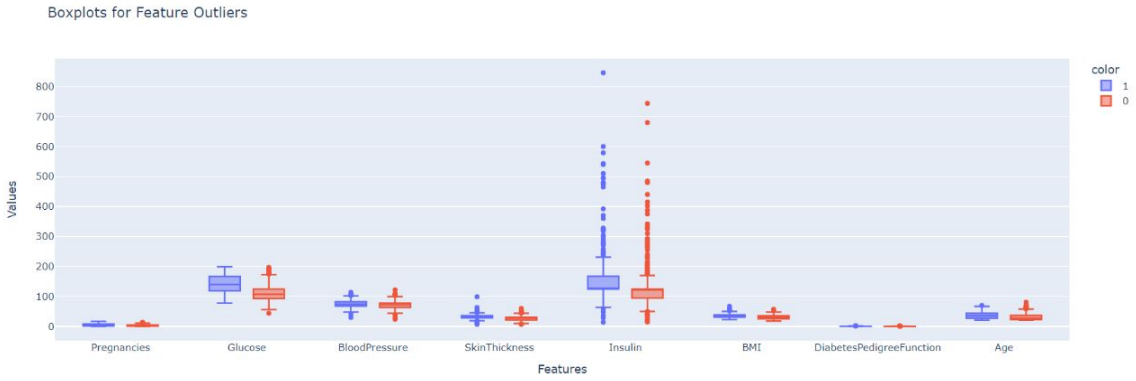

Figure 7: Box Plots

The results of the OGTT as a screening test for type 2 diabetes can be interpreted as follows:

- The 2-hour plasma glucose level <140 mg/dL is considered normal
- The 2-hour plasma glucose level of 140-199 mg/dL indicates impaired glucose tolerance
- The 2-hour plasma glucose level ≥200 mg/dL indicates diabetes

Figure 8: National Library of Medicine about Glucose levels

References

[1] https://labpedia.net/insulin-level-insulin-assay.

[2] https://www.ncbi.nlm.nih.gov/books/NBK532915

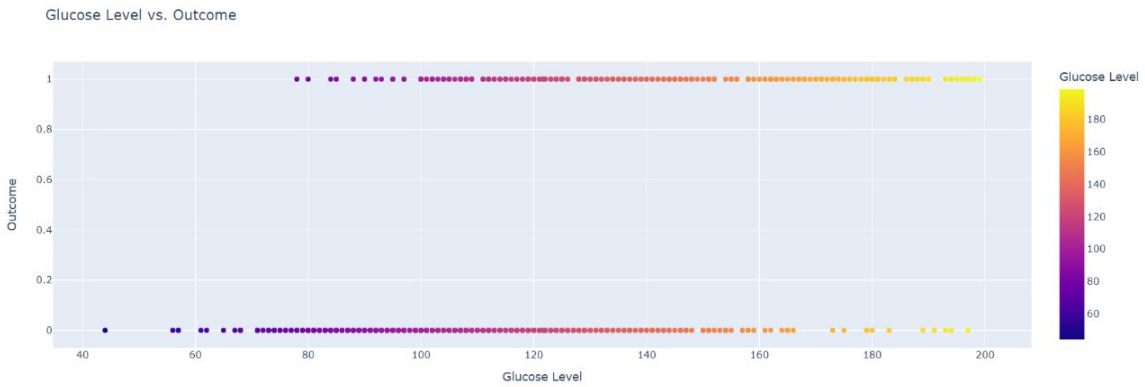

Figure 9: Glucose vs Outcome

	Model	Validation Score
0	Logistic	77.2
1	SVM	76.7
2	AdaBoost	75.7
3	RandomForest	78.3
4	GradientBoosting	77.5

Figure 10: Models result

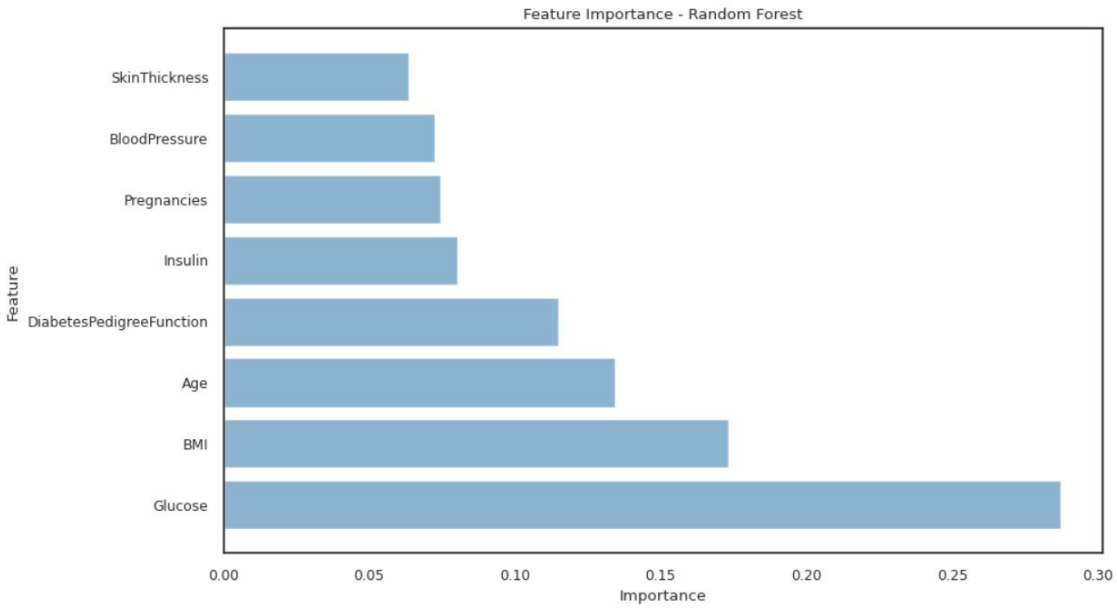

Figure 11: Feature importance

Class	Precision	Recall	F1-Score	Support
0	0.82	0.80	0.81	108
1	0.55	0.59	0.57	46

Figure 12: Evaluation metrics

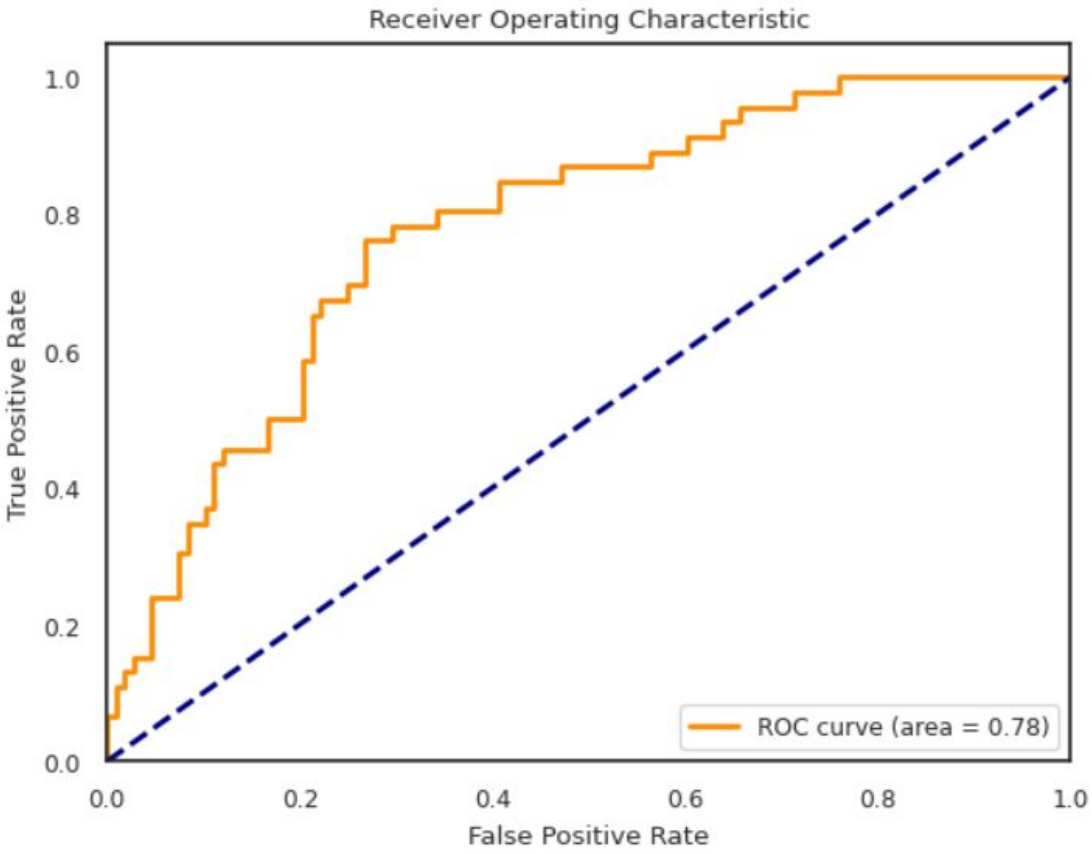

Figure 13: ROC curve