➔ *Regular Research Paper – NS*

Interpretable Diabetes Prediction using XAI in Healthcare Application

Ilhan Uysal
Burdur Mehmet Akif Ersoy University, Turkey
iuysal@mehmetakif.edu.tr / ilhanuysal@gmail.com

Abstract

Diabetes is a chronic metabolic disorder affecting millions of people worldwide. This study investigates the application of Explainable Artificial Intelligence (XAI) techniques in the prediction and classification of diabetes using a diabetes disease dataset. The effectiveness of various machine learning algorithms such as KNN, Naive Bayes, SVM, Decision Tree, Random Forest Logistic Regression and all models in the Lazy Classifier package are investigated with XAI methods to develop accurate and interpretable models for diabetes prediction. The diabetes dataset used in the study has 768 rows and 9 columns consisting of various medical variables (independent variables) and an outcome variable (dependent variable). Hyperparameters and grid search were utilized for model optimization. The success performances of the models were evaluated with metrics such as F1 Score, accuracy, balanced accuracy, precision, recall, ROC AUC and time taken. SVM and Random Forest stand out as the most successful models. With the most successful models, the impact of different features on diabetes prediction was evaluated with different SHAP plots representing the contribution of each feature to the final prediction compared to the average prediction. Glucose, Age and BMI were found to have a significant and positive effect on the model output. The study aims to uncover important characteristics and patterns that contribute to diabetes risk using XAI and to assist healthcare professionals in providing timely intervention and personalised treatment plans.

Keywords: *Diabetes prediction, Explainable Artificial Intelligence (XAI), Interpretability, Machine learning algorithms, Healthcare applications.*

1. INTRODUCTION

Diabetes is a chronic metabolic disorder affecting millions of people worldwide. Early detection and accurate prediction of diabetes can significantly improve patient outcomes by providing timely intervention and personalised treatment plans. With advances in machine learning and artificial intelligence, predictive models show promise in assisting healthcare professionals in the diagnosis and management of diabetes. However, the black-box nature of traditional machine learning algorithms often makes it difficult to understand the underlying causes of predictions, limiting their interpretability and reliability in critical healthcare applications.

People with diabetes have a high risk of developing diseases such as heart disease, kidney disease, stroke, eye problems and nerve damage. The current practice in the hospital is to collect the necessary information to diagnose diabetes through various tests and provide appropriate treatment based on the diagnosis. Big Data Analytics plays an important role in healthcare sectors. Health sectors have large volumes of databases. Using big data analytics, large data sets can be analysed and hidden information and hidden patterns can be found to discover information from the data and predict outcomes accordingly [1].

Diabetes mellitus is a chronic disease characterised by hyperglycaemia. It can cause many complications. According to the increasing morbidity in recent years, the number of diabetics in the world will reach 642 million in 2040, which means that one in ten adults will have diabetes in the future. This alarming figure requires great attention [2].

Explainable Artificial Intelligence (XAI) techniques have emerged as a powerful tool to bridge the gap between predictive accuracy and interpretability in machine learning models. XAI methods provide insights into the decision-making process of complex models, allowing healthcare professionals to understand and validate the factors driving predictions. In the context of diabetes prediction, XAI can offer valuable insights into key characteristics and risk factors that contribute to predicted outcomes, improving understanding of the disease and facilitating more informed clinical decisions.

This paper aims to investigate the application of XAI techniques in the prediction and classification of diabetes using a diabetes disease dataset. The effectiveness of various machine learning algorithms combined with XAI methods to develop accurate and interpretable models for diabetes prediction is investigated. This study aims to reveal important features and patterns that contribute to the risk of diabetes using XAI and is considered to be a study that can shed light on the underlying mechanisms and help medical practitioners make informed decisions.

2. RELATED WORKS

Mujumdar and Vaidehi have published a study that includes a literature review of various studies on diabetes prediction using health datasets and machine learning algorithms. The study also discusses the application of various prediction models using data mining techniques, machine learning algorithms or a combination of these techniques. For better classification of diabetes, they proposed a diabetes prediction model that includes normal factors such as Glucose, BMI, Age, Insulin etc. as well as several external factors responsible for diabetes. The classification accuracy is improved with the new dataset compared to the existing dataset. Also, a pipeline model for diabetes prediction has been implemented to improve the classification accuracy [1].

Zou et al. used decision trees, random forest and neural network to predict diabetes. The dataset is hospital physical examination data from Luzhou, China. It contains 14 attributes. In this study, five-fold cross-validation was used to examine the models. To verify the universal applicability of the methods, some methods with better performance were selected to conduct independent test experiments. We randomly selected 68994 healthy human and diabetic patient data as the training set, respectively. Due to data imbalance, data were randomly extracted 5 times. The result is the average of these five experiments. In this study, principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) were used to reduce dimensionality. They stated that the results show that prediction with random forest can achieve the highest accuracy (ACC = 0.8084) when all attributes are used [2].

In their study, Aelgani et al. presented a community local explainable agnostic model for predicting diabetes. The study stated that the community voting classifier produced 81% accuracy compared to other traditional prediction models on the Pima Indian diabetes dataset. Then, they applied the explainable artificial intelligence (XAI) technique, which helps medical professionals understand the predictions made by the mode [3].

Soni and Varna conducted a study aiming to predict diabetes in patients using machine learning techniques. They used various classification and ensemble learning methods such as SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers. They collected data from the Pima Indian Diabetes Dataset and achieved a classification accuracy of

77%. The results show that Random Forest achieves the highest accuracy and can be an effective tool for predicting diabetes. The study concluded that early prediction and decision-making can help healthcare providers treat diabetes and save human lives [4].

Lai et al. presented a study focused on developing predictive models to identify patients at risk of Diabetes Mellitus (DM) using machine learning techniques. The study uses patient demographic data and laboratory results to build the models. The study compares the performance of different machine learning techniques such as Logistic Regression, Gradient Boosting Machine, Decision Tree and Random Forest. The study reveals that the Gradient Boosting Machine model outperforms the other models in terms of sensitivity and AROC. The AROC for the proposed GBM model is 84.7% with a sensitivity of 71.6% and the sensitivity of the AROC Logistic Regression model for the proposed GBM model is 84.0% with 73.4%. GBM and Logistic Regression models perform better than Random Forest and Decision Tree models. Fasting blood glucose, body mass index, high-density lipoprotein and triglycerides are the most important determinants in these models [5].

Ghosh et al. compared the effectiveness of various machine-learning techniques in the detection of diabetes. They found that some techniques such as decision trees and random forests are more accurate and efficient than others. They also stated that the use of feature selection techniques can improve the accuracy of the models. Four different machine learning algorithms, namely Gradient Boosting (GB), Support Vector Machine (SVM) AdaBoost (AB) and Random Forest (RF), were evaluated using the Pima Indians diabetes dataset, first against all features and then against the features selected with the Minimal Redundancy Maximal Relevance (MRMR) Feature Selection (FS) approach. Seven different performance evaluation metrics were calculated using a 10-fold cross-validation (CV) approach. Computational complexity was also evaluated. The best results were obtained with the Random Forest approach and an accuracy of 99.35% was achieved [6].

Hasan et al. developed a robust framework for diabetes prediction using a combination of different machine-learning classifiers. The dataset used in the study is the Pima Indian Diabetes (PID) dataset, which contains 768 samples with 8 attributes. The proposed framework consists of several steps such as preprocessing, feature selection and classification. They used different preprocessing techniques, feature selection methods and machine learning classifiers to find the best-performing combination. They compared the performance of different preprocessing techniques, feature selection methods and machine learning classifiers. They also proposed an ensemble classifier by combining the best-performing machine learning models. They preferred the soft-weighted voting method to combine machine learning models. They concluded that their proposed framework achieves high accuracy in diabetes prediction and outperforms other state-of-the-art methods. They stated that their proposed framework can be used for early prediction of diabetes, which can help in the prevention and management of the disease. The proposed fusion classifier was the best-performing classifier with sensitivity, specificity, false miss rate, diagnostic odds ratio, and AUC of 0.789, 0.934, 0.092, 66.234, and 0.950 respectively [7].

Kavakiotis et al. conducted a systematic review of the application of machine learning and data mining techniques and tools in the field of diabetes research concerning Prediction and Diagnosis, Diabetic Complications, Genetic Background, Environment and Health Care and Management and found that the first category was the most popular. They concluded that a wide variety of machine learning algorithms are used and that, in general, 85% of those used are characterised by supervised learning approaches and 15% by unsupervised approaches and, more specifically, association rules. Support vector machines (SVM) emerged as the most successful and widely used algorithm. Regarding the data type, predominantly clinical datasets were used [8].

Lu and Uddin proposed a stack-based model for predicting 30-day hospitalisations of diabetic patients. They also used explainable artificial intelligence (XAI) techniques to identify the most important features for prediction and increase transparency in the model's decision-making process. In the study, they compare the performance of their model with other similar studies in the existing literature using the same data set and conclude that their model outperforms previous forecasting results. According to the permutation feature importance, the strong predictors were the number of inpatients, primary diagnosis, discharge to home with home service and number of emergencies. The interpretable model-agnostic explanations method was also used to demonstrate explainability at the local individual level [9].

Vishwarupe et al. presented a case study using the ELI5 XAI toolkit in combination with the LIME (Local Interpretable Modelagnostic Explanations) and SHAP (Shapley Additive exPlanations) algorithmic frameworks in Python to determine whether a patient is diabetic based on a randomised clinical trial dataset. They also identified trends and the most vital factors that can help clinicians and researchers in analysing patient data in combination with machine learning and artificial intelligence outputs [10]

Nagaraj et al. proposed a diabetes prediction model for better diabetes classification that includes several external factors that cause diabetes as well as common factors such as Glucose, BMI, Age, and Insulin. Various machine learning algorithms such as XGBoost, support vector machine (SVM), random forest classifier and decision tree were used for prediction. In addition, the LIME descriptor was used to tabulate the accuracy of diabetic-positive patients and diabetes-negative patients. In this study, the random forest classifier gave high accuracy with a value of 77%. After estimating the accuracy of diabetes, they proposed a recommendation system to get rid of diabetes, including a diet plan, organic foods that help to get rid of diabetes, exercise and insulin [11].

3. MATERIAL METHOD

The diabetes dataset used in the study has 768 rows and 9 columns [12]. The data set consists of several medical variables (independent variables) and an outcome variable (dependent variable). The independent variables in this dataset are: 'Pregnancies', 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', 'BMI', 'Diabetes Pedigree Function', and 'Age'. The value of the outcome variable is either 1 or 0, indicating whether a person has diabetes or not. A summary of the dataset is given in Figure 1.

```
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. Diabetes dataset

The statistical information in Figure 2 is provided to help understand how the data is spread across the table. Some columns have a minimum value of 0, which is not medically possible, so in the data cleaning process we replaced them with the median/mean value depending on the distribution. Also in the maximum column, there are outliers, i.e. insulin levels as high as 846. Therefore, data preprocessing steps such as dropping duplicate values, checking NULL values, checking for 0 values and replacing them were performed.

	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 information about the dataset

The Pearson Correlation Coefficient helps to find the relationship between two quantities. It gives a measure of the strength of the relationship between two variables. The value of Pearson Correlation Coefficient can be between -1 and +1. 1 means that they are highly correlated and 0 means that there is no correlation. A heat map is a two-dimensional representation of information with the help of colours. Heat maps allow the user to visualise simple or complex information. Figure 3 shows the heat map of the dataset. According to the heatmap, Glucose, BMI and Age have the highest correlation with Outcome. Blood Pressure, Insulin, and Diabetes Pedigree Function have the least correlation, so they do not contribute much to the model, so they were dropped.

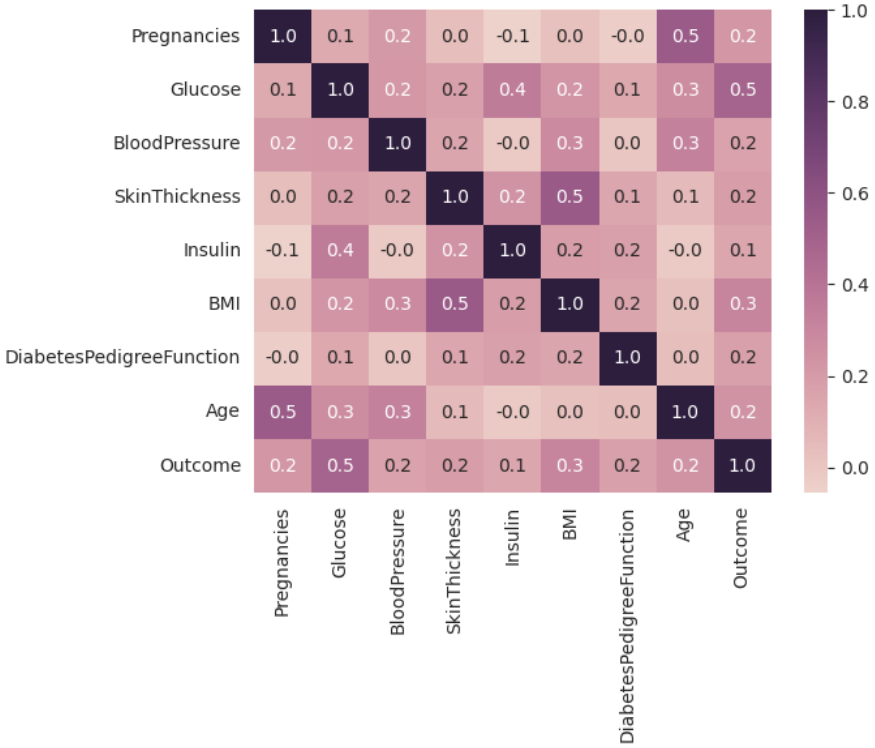

Figure 3. Correlation Matrix

To test the performance of the models by applying machine learning algorithms, eighty per cent of the dataset that has passed the preprocessing steps is divided into training and twenty per cent as test data. Therefore, 614 rows of data allocated for training and 154 rows of data allocated for testing were obtained.

Hyperparameters are variables that are usually specified when building a machine-learning model. Therefore, hyperparameters are specified before specifying the parameters or hyperparameters are used to evaluate the optimal parameters of the model. The best part of the hyperparameters is that their values are decided by the user creating the model. For example, max_depth in Random Forest Algorithms, k in KNN Classifier. Hyperparameter tuning is the process of adjusting the parameters found as tuples when building machine learning models.

Grid Search uses a different combination of all specified hyperparameters and their values and calculates the performance for each combination. Therefore, it selects the best value for hyperparameters.

Metrics such as Confusion Matrix, F1 Score, Precision Score, Recall Score, Accuracy, and Balanced Accuracy were used as performance evaluation metrics.

- Confusion Matrix: It is a table used to evaluate the performance of a classification model and shows the relationship between true classes and predicted classes (True Positive, False Positive, True Negative, False Negative).
    - True Positive (TP): Number of positive instances correctly classified as positive by the model.
    - False Positive (FP): Number of negative instances incorrectly classified as positive by the model.
    - True Negative (TN): Number of negative instances correctly classified as negative by the model.
    - False Negative (FN): Number of positive instances incorrectly classified as negative by the model.

- F1 Score: A performance measure used to balance precision and recall; it balances the effect of false positives and false negatives.

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (1)$$

- Accuracy: The ratio of correct predictions of the classification model to the total number of samples; measures overall success.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (2)$$

- Balanced Accuracy: In multi-class data sets, if there is a sampling imbalance between classes, it is the measure used instead of accuracy; it averages the accuracy of all classes.

$$\text{Balanced Accuracy} = (\text{Recall\_Positive} + \text{Recall\_Negative}) / 2 \quad (3)$$

- Recall: A performance metric that divides true positives by the total number of positive samples; important for minimising false negatives.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (4)$$

- Precision: A performance metric that divides true positives by the total number of predicted positive samples; important for minimising false positives.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (5)$$

KNN, Naive Bayes, SVM, Decision Tree, Random Forest, and Logistic Regression models from machine learning techniques were used in the study. In addition, the lazy classifier from Python packages was also used to compare many models.

3.1. K-Nearest Neighbourhood (KNN)

K-Nearest Neighbour is a simple classification algorithm that relies on the labels of its nearest neighbours in the training data to determine the class of a new sample. For example, to predict the class of a new data point, it looks at the class labels of the K nearest neighbours of that point and predicts the class of the majority. The classification report and the confusion matrix obtained with the KNN algorithm are given in Figure 4. The value accuracy of the KNN classifier is 0.81, the F1 score is 0.67, the precision score is 0.70 and the recall score is 0.64. Additionally, the true positive value is 30, the true negative value is 94, the false positive value is 13 and the false negative value is 17.

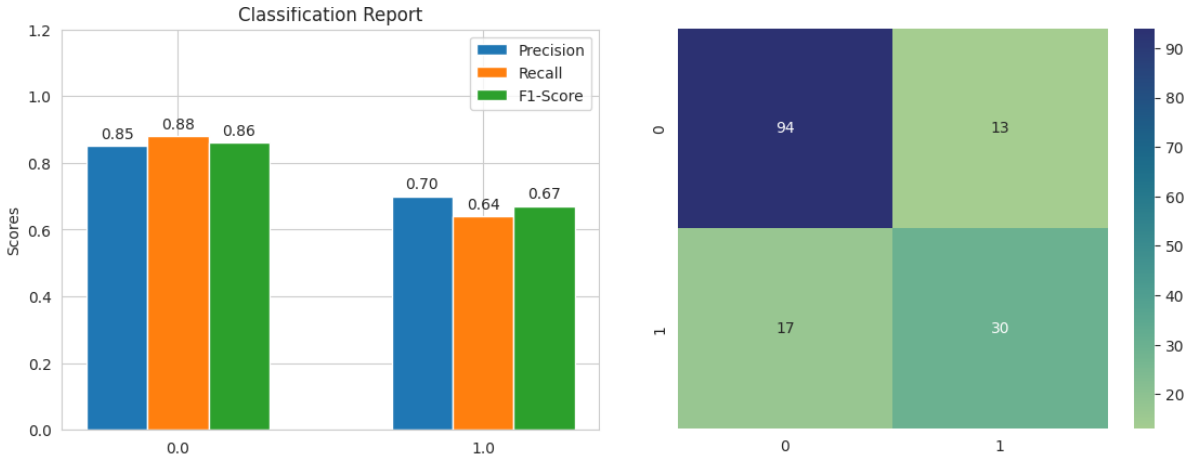

Figure 4. Classification Report and Confusion Matrix for KNN

3.2. Naive Bayes

Naive Bayes is an algorithm for classification based on basic probability theory. Based on the assumption of independence, it assumes that the features in the data are independent of each other. Therefore, it is called "Naive" (pure). It calculates class probabilities using training data and classifies new data with these probabilities. The classification report and the confusion matrix obtained with the Naïve Bayes algorithm are given in Figure 5. The value accuracy of the Naïve Bayes classifier is 0.77, the F1 score is 0.58, the precision score is 0.64 and the recall score is 0.53. Additionally, the true positive value is 25, the true negative value is 93, the false positive value is 14 and the false negative value is 22.

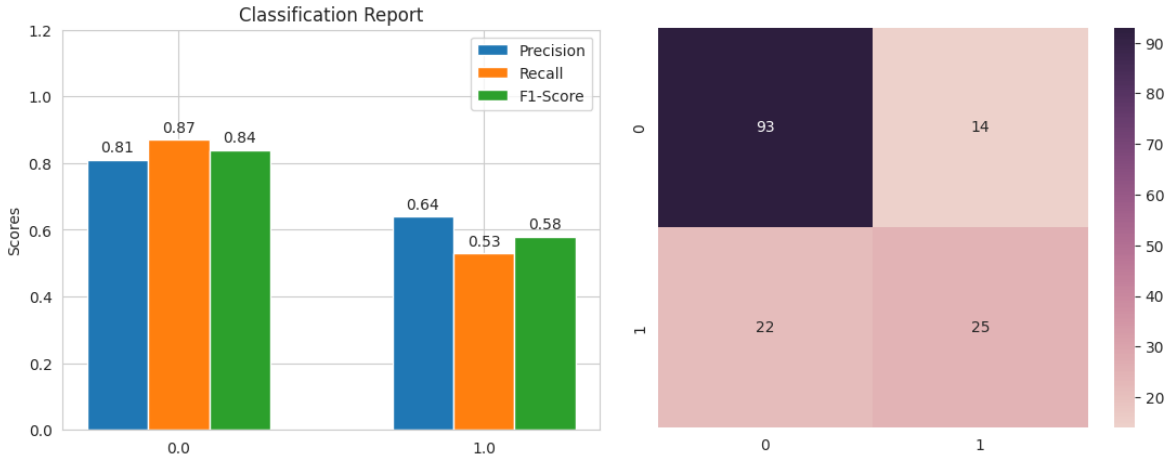

Figure 5. Classification Report and Confusion Matrix for Naïve Bayes

3.3. Support Vector Machines (SVM)

SVM is a powerful algorithm used for classification and regression problems. It creates an optimal hyperplane to partition the data into classes. The goal is to find a hyperplane that maximises the margin between classes. Kernel methods can be used to solve non-linear problems. The classification report and the confusion matrix obtained with the SVM algorithm are given in Figure 6. The value accuracy of the SVM classifier is 0.82, the F1 score is 0.67, the precision score is 0.70 and the recall score is 0.64. Additionally, the true positive value is 32, the true negative value is 95, the false positive value is 12 and the false negative value is 15.

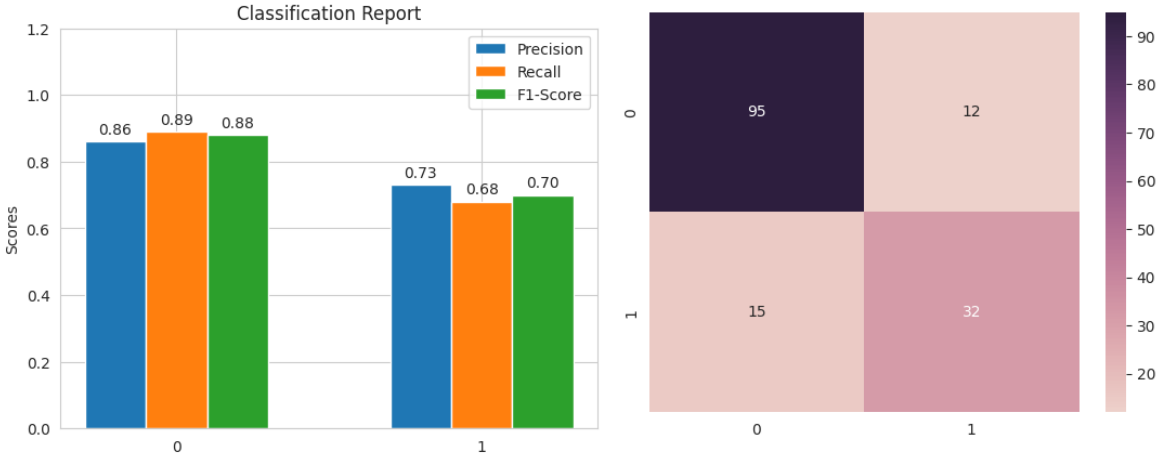

Figure 6. Classification Report and Confusion Matrix for SVM

3.4. Decision Trees

Decision trees are classification algorithms that create a tree structure to best predict the target variable by sequentially evaluating features in the dataset. Each internal node represents a feature value and child nodes according to the branches. Leaf nodes are used to assign data to a class. The classification report and the confusion matrix obtained with the Decision Trees algorithm are given in Figure 7. The value accuracy of the Decision Trees classifier is 0.79, the F1 score is 0.57, the precision score is 0.78 and the recall score is 0.45. Additionally, the true

positive value is 21, the true negative value is 101, the false positive value is 6 and the false negative value is 26.

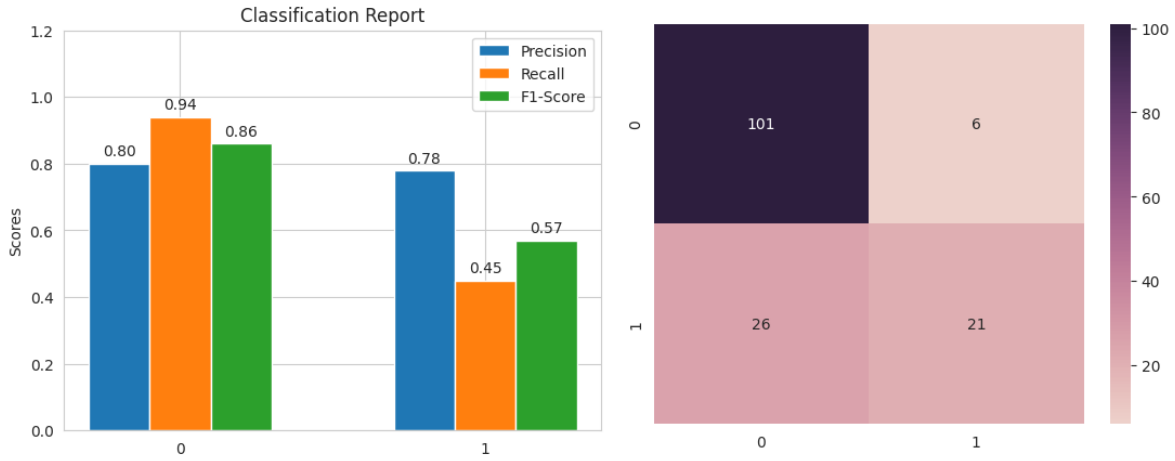

Figure 7. Classification Report and Confusion Matrix for Decision Trees

3.5. Random Forest

Random Forests is a classification method consisting of many decision trees. It trains multiple decision trees by randomly sampling and selecting random features. As a result, a more accurate and stable classification is obtained by combining the predictions of all trees. The classification report and the confusion matrix obtained with the Random Forest algorithm are given in Figure 8. The value accuracy of the Random Forest classifier is 0.79, the F1 score is 0.67, the precision score is 0.70 and the recall score is 0.64. Additionally, the true positive value is 31, the true negative value is 90, the false positive value is 17 and the false negative value is 16.

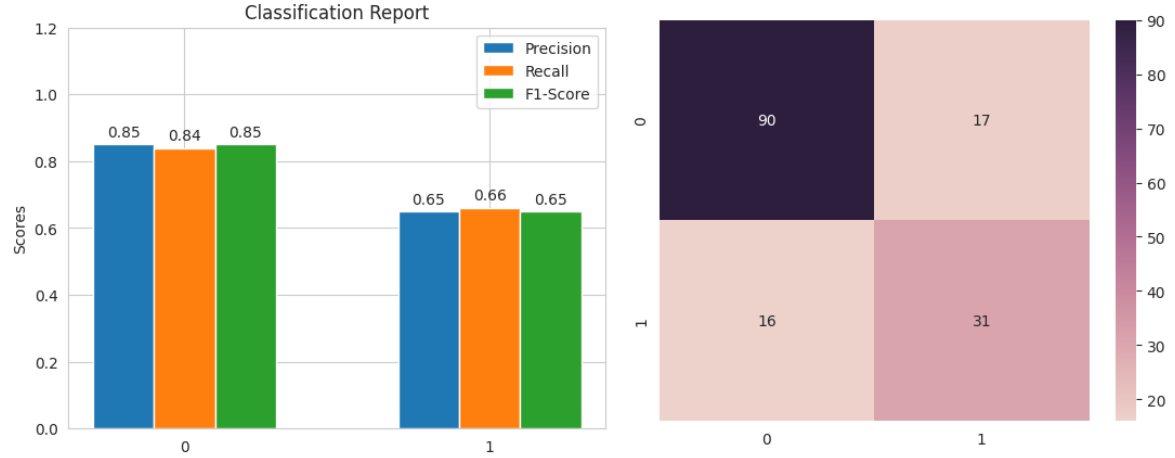

Figure 8. Classification Report and Confusion Matrix for Random Forest

3.6. Logistic Regression

Logistic Regression is a basic linear model used for classification problems. It interprets the output as a probability value and is often used to separate two classes. It creates a decision boundary between classes using input features and weights and estimates the output with a

logit function. The classification report and the confusion matrix obtained with the Logistic Regression algorithm are given in Figure 9. The value accuracy of the Logistic Regression classifier is 0.79, the F1 score is 0.63, the precision score is 0.69 and the recall score is 0.57. Additionally, the true positive value is 27, the true negative value is 95, the false positive value is 12 and the false negative value is 20.

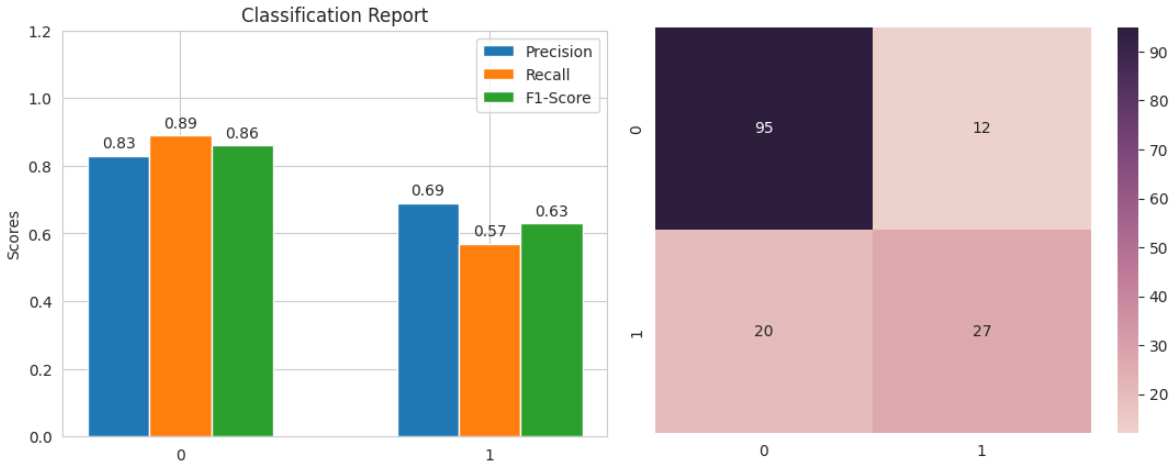

Figure 9. Classification Report and Confusion Matrix for Logistic Regression

3.7. Lazy Classifier

Lazy Classifier is a Python library that enables rapid evaluation of machine learning models. By automatically applying various classification algorithms, this library allows us to obtain key performance metrics for data. The table comparing the model performances obtained with the LazzyClassifer package for the test data is given visually in Figure 10. Accordingly, the most successful model was Random Forest.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
RandomForestClassifier	0.81	0.78	0.78	0.81	0.23
NearestCentroid	0.75	0.76	0.76	0.76	0.02
XGBClassifier	0.78	0.75	0.75	0.78	0.11
LGBMClassifier	0.79	0.74	0.74	0.79	0.06
CalibratedClassifierCV	0.81	0.74	0.74	0.80	0.07
RidgeClassifierCV	0.79	0.74	0.74	0.79	0.02
RidgeClassifier	0.79	0.74	0.74	0.79	0.02
GaussianNB	0.77	0.74	0.74	0.77	0.01
NuSVC	0.78	0.73	0.73	0.78	0.04
KNeighborsClassifier	0.78	0.73	0.73	0.78	0.02
LinearDiscriminantAnalysis	0.79	0.73	0.73	0.78	0.02
BaggingClassifier	0.79	0.73	0.73	0.79	0.06
LinearSVC	0.78	0.73	0.73	0.78	0.03
ExtraTreesClassifier	0.78	0.73	0.73	0.78	0.17
AdaBoostClassifier	0.77	0.72	0.72	0.77	0.12
LogisticRegression	0.78	0.72	0.72	0.77	0.02
SVC	0.77	0.71	0.71	0.77	0.03
LabelSpreading	0.73	0.71	0.71	0.73	0.05
SGDClassifier	0.73	0.71	0.71	0.73	0.02
QuadraticDiscriminantAnalysis	0.77	0.71	0.71	0.76	0.01
DecisionTreeClassifier	0.71	0.69	0.69	0.72	0.02
BernoulliNB	0.73	0.68	0.68	0.73	0.02
Perceptron	0.70	0.68	0.68	0.71	0.01
LabelPropagation	0.71	0.68	0.68	0.71	0.04
ExtraTreeClassifier	0.69	0.65	0.65	0.69	0.01
PassiveAggressiveClassifier	0.65	0.60	0.60	0.65	0.02
DummyClassifier	0.69	0.50	0.50	0.57	0.01

Figure 10. Performance metrics of models with Lazy Classifier for test data

The table comparing the model performances obtained with the LazzyClassifer package for the training data is given visually in Figure 11. Accordingly, one of the most successful models was again Random Forest.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LabelPropagation	1.00	1.00	1.00	1.00	0.05
XGBClassifier	1.00	1.00	1.00	1.00	0.13
RandomForestClassifier	1.00	1.00	1.00	1.00	0.24
DecisionTreeClassifier	1.00	1.00	1.00	1.00	0.04
ExtraTreeClassifier	1.00	1.00	1.00	1.00	0.02
ExtraTreesClassifier	1.00	1.00	1.00	1.00	0.29
LabelSpreading	1.00	1.00	1.00	1.00	0.07
LGBMClassifier	0.99	0.99	0.99	0.99	0.14
BaggingClassifier	0.98	0.98	0.98	0.98	0.26
AdaBoostClassifier	0.82	0.79	0.79	0.82	0.42
KNeighborsClassifier	0.81	0.79	0.79	0.81	0.03
NuSVC	0.79	0.77	0.77	0.79	0.04
SVC	0.78	0.75	0.75	0.78	0.04
GaussianNB	0.76	0.75	0.75	0.76	0.01
QuadraticDiscriminantAnalysis	0.76	0.74	0.74	0.76	0.04
NearestCentroid	0.73	0.74	0.74	0.74	0.02
LinearSVC	0.77	0.74	0.74	0.76	0.03
LogisticRegression	0.77	0.74	0.74	0.76	0.02
RidgeClassifierCV	0.77	0.73	0.73	0.76	0.02
CalibratedClassifierCV	0.77	0.73	0.73	0.76	0.30
RidgeClassifier	0.76	0.73	0.73	0.76	0.02
LinearDiscriminantAnalysis	0.76	0.73	0.73	0.76	0.05
SGDClassifier	0.72	0.71	0.71	0.73	0.02
Perceptron	0.71	0.71	0.71	0.71	0.02
BernoulliNB	0.70	0.68	0.68	0.70	0.03
PassiveAggressiveClassifier	0.65	0.63	0.63	0.65	0.02
DummyClassifier	0.64	0.50	0.50	0.50	0.02

Figure 11. Performance metrics of models with Lazy Classifier for training data

3.8. Explainable Artificial Intelligence (XAI)

XAI is an area of research and practice that focuses on the ability of artificial intelligence systems to explain their work and decisions understandably. Today, complex and powerful artificial intelligence models such as deep learning are increasingly used. However, the content and decision mechanisms of these models can often be perceived as a magic box, meaning that it is not fully understood how they produce results. The main objective of XAI is to explain the working processes and results of these complex models transparently and clearly. Clear explanations should be presented in a way that people can trust and understand. This increases the reliability of AI systems while ensuring that users have the chance to challenge or correct the decisions made by the algorithm. XAI provides benefits in several areas such as understanding ethically important cause-effect relationships, identifying biases in the training data of the model and improving the accuracy of the model. Especially in critical application areas such as medical diagnosis, financial decision making and the automotive industry, creating understandable and reliable artificial intelligence systems for humans increases the importance of XAI.

SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are two popular techniques used for explainable artificial intelligence (XAI) in machine learning models. They serve similar purposes but have some differences in their approach and capabilities. SHAP is based on cooperative game theory and assigns to each feature a unique value (SHAP value) that represents the contribution of the feature to the model's prediction for a given instance. It provides global and local explanations, i.e. it can explain the overall behaviour of the model (global) as well as individual predictions (local). SHAP values guarantee consistency, i.e. the sum of SHAP values for all features is equal to the difference between the model output for a given sample and the average model output. SHAP is computationally more intensive, especially for complex models and large datasets.

LIME is used to locally approximate the behaviour of a complex model by training an interpretable surrogate model (e.g. linear regression) around the forecast sample and to explain this particular forecast. While focusing on local explanations, it also makes it easier to understand the model behaviour of specific instances by explaining individual predictions separately. It is model-independent, meaning it can be used with any machine learning model, including complex black box models such as deep neural networks. It is computationally less expensive than SHAP because it only needs to train a simple surrogate model on a small subset of the data.

SHAP is more appropriate if global descriptions are needed to understand the overall behaviour of the model and the importance of features across the entire dataset. If there is concern about ensuring consistency in feature contributions and having strong theoretical underpinnings, the game-theoretic approach of SHAP offers these guarantees.

If one is interested in understanding how a particular prediction is made and needs localised explanations for individual instances, LIME is a better choice. If working with computationally expensive models or large datasets, the computational efficiency of LIME may make it a more practical choice.

4. RESULTS AND DISCUSSIONS

In this study, Explainable Artificial Intelligence (XAI) techniques are applied in the prediction and classification of diabetes using a diabetes disease dataset. After comparing the model performances, SHAP values were calculated with the most successful model, random forest. SHAP values represent the contribution of each characteristic to the final prediction compared to the mean prediction. Positive SHAP values indicate an increase in the probability of having diabetes, while negative values mean a decrease in the probability. In the context of diabetes prediction, SHAP values help to understand how much each input attribute contributes to the final prediction of whether a person has diabetes. The visualisation prepared with a bar graph is given in Figure 12. Accordingly, glucose level has the highest positive effect on diabetes prediction. A higher glucose level is associated with an increased risk of diabetes. This is in line with medical understanding that high blood glucose levels are an important risk factor for diabetes. Body Mass Index has a moderate positive effect on diabetes prediction. As obesity is a known risk factor for diabetes, higher BMI values are generally associated with an increased risk of diabetes. Age also has a moderate positive effect on diabetes prediction. As individuals age, the risk of developing diabetes tends to increase. This may be due to lifestyle factors, metabolic changes and other age-related health problems. Number of pregnancies in women has a slight positive effect on diabetes prediction. A higher number of pregnancies may indicate a higher risk for gestational diabetes or an association with other factors that increase the risk of diabetes. Skin thickness has a small positive effect on diabetes prediction. However, it should be noted that the link between skin thickness and diabetes risk is not as well-founded as the other characteristics mentioned.

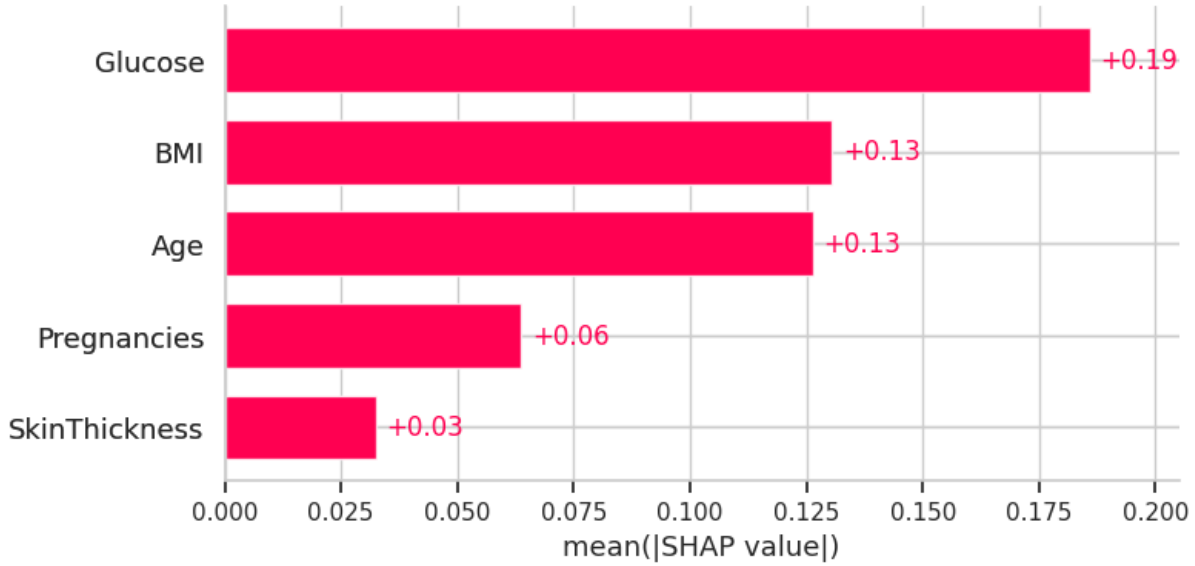

Figure 12. Bar graphic

The waterfall plot is a visual representation used in the context of model descriptions, especially in combination with SHAP values, to show how the contributions of different attributes lead to a deviation from the baseline prediction for a single sample. It contributes to the understanding of the cumulative effect of each attribute on the final prediction. The waterfall plot with SHAP values for diabetes prediction is given in Figure 13. The bottom part of the waterfall plot starts with the expected output value of the model, E[f(X)]. Each row shows how the positive (red) or negative (blue) contribution of each feature moves from the expected output value to the output value of the model. According to the results, the expected model output value is 0.39 while the current model output value is 1. The grey numbers in front of the feature names represent the value of each feature in this example. While the expected output value of the model was 0.39, the output value of the model was 1 due to the positive contribution of the attributes to the output of the model. For example, the age variable moved the output of the model from 0.7 to 1 with a Shap value of 0.926 and a positive contribution of +0.3.

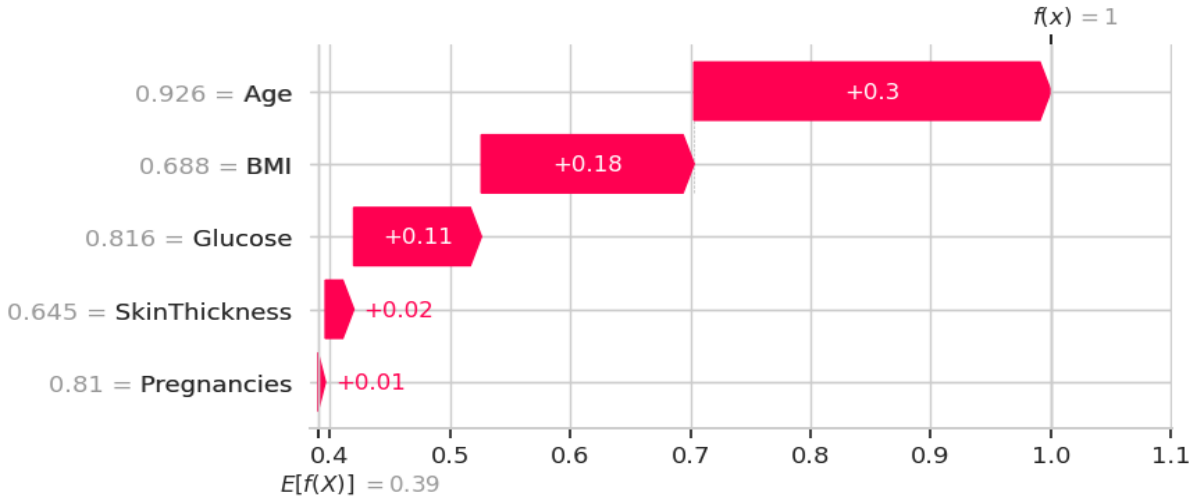

Figure 13. Waterfall graphic

The force graph is a visualisation method used in combination with the waterfall graph. In the force graph, red arrows represent positive effects while blue arrows show negative effects. The size of the arrows represents the magnitude of the effect of the feature. The grey-coloured "base value" indicates the average prediction value of the model on the training set. The "output value" is the predicted result of the model. The force plot provides an effective summary for prediction, making it easier to understand the effects of features. The force plot for diabetes prediction is given in Figure 14. Accordingly, glucose, BMI and age variables contributed positively to the model. While the base value of the model was 0.39, the output value was 1.

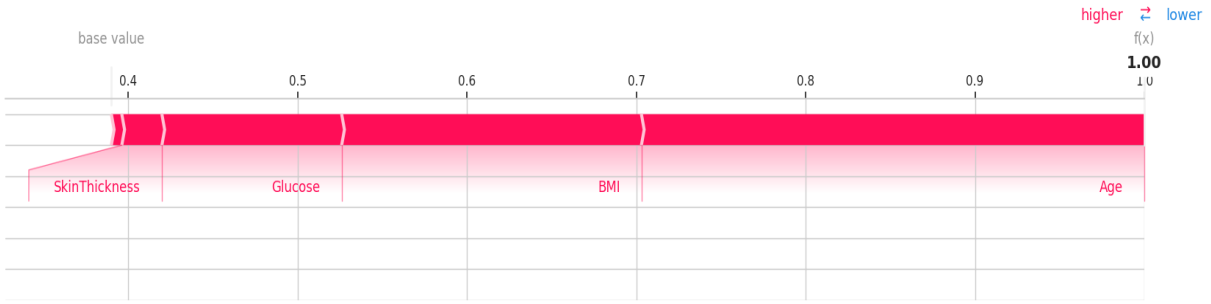

Figure 14. Force graphic

In contrast to the force plot, the decision plot provides insight into the model's baseline value through a straight vertical line. The coloured lines in the plot represent the predicted values associated with the features. Visualising the SHAP values (feature effects), the decision plot shows the path from the base value to the final score of the model. It provides a comprehensive understanding of how each feature contributes to the overall prediction. Decision and force plots support each other and express similarities. The decision plot is given in Figure 15. Looking at the graph from bottom to top, while the expected output value of the model was 0.39, the output value of the model reached 1 thanks to the high positive contribution of BMI, glucose and age variables.

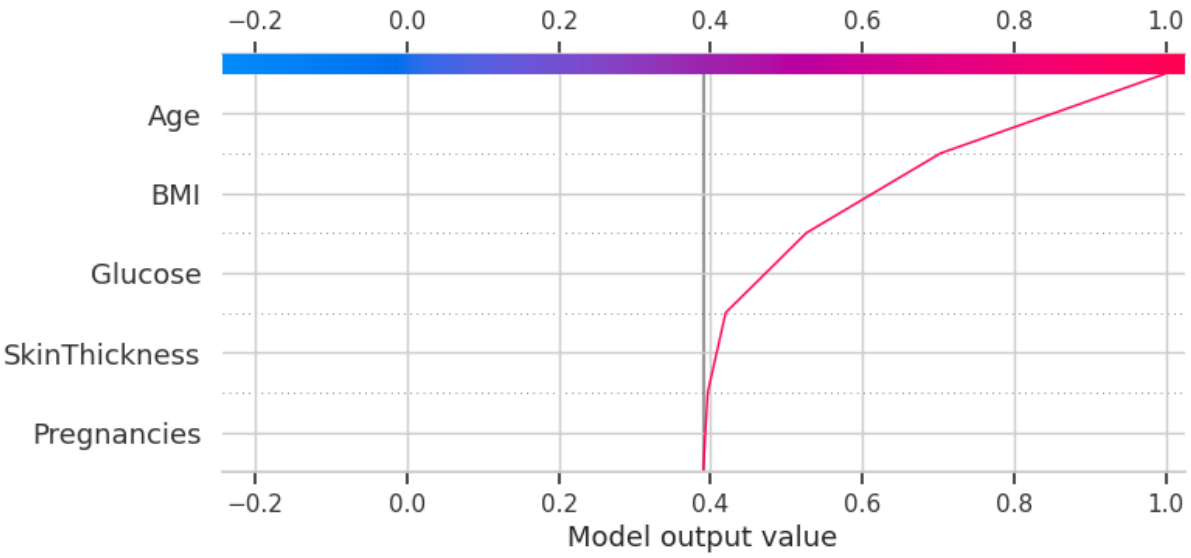

Figure 15. Decision graphic – 1

A beeswarm plot is a useful tool for understanding the impact of features on output and visualising the relationship between different values. By combining density, distribution and original values, it provides comprehensive information to understand the important features of

a dataset. The horizontal position of the points is determined by the SHAP value of the feature. That is, it shows the influence of a feature on the output. The density of the dots increases with each feature row, indicating the strength of the feature's influence on the output. The original value of each dot is also shown using colour. The colour is used to visualise the original value of the feature. This helps to see the different feature values in the original data set. The beeswarm plot prepared with SHAP values in diabetes prediction is given in Figure 16. When the feature value and the Shap values affecting the model are considered, it can be seen that BMI, glucose and age variables have a high positive effect on the output of the model with the aggregation of the points. In pregnancies and skin thickness, it is understood that the points are more scattered and have a lower effect on the output of the model.

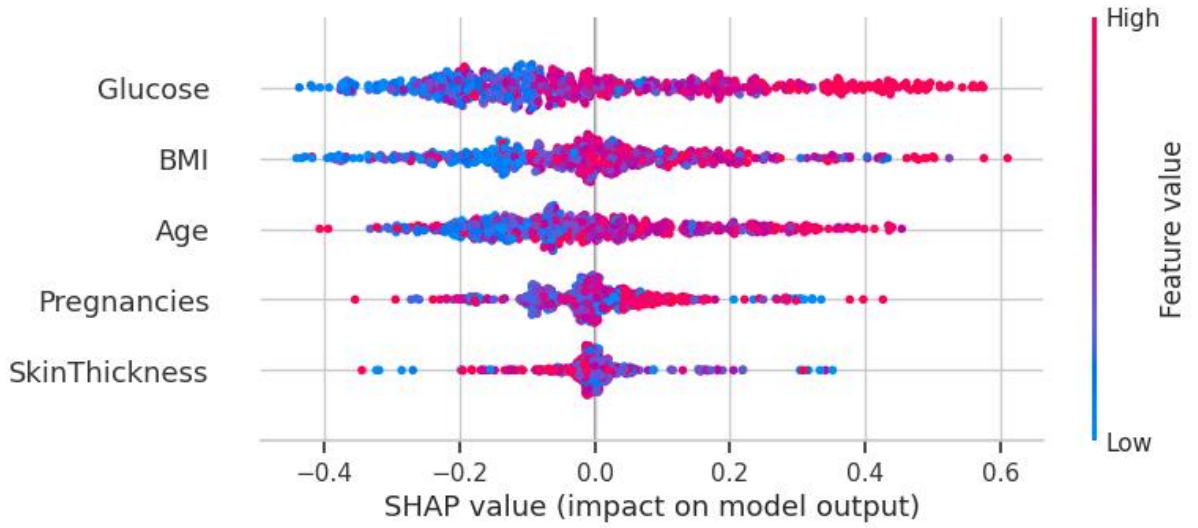

Figure 16. Beeswarm graphic

The violin plot is a useful tool to compare the effect of features on the output and to see the variation between different feature values. It consists of one or more juxtaposed violin-shaped regions to show the SHAP values of each feature. The width of each region represents the intensity of the SHAP values. That is, a wider region represents a denser distribution of SHAP values. The centre line of the graph shows the mean SHAP value of the feature. The shapes spanning the width on the left and right sides indicate the distribution of SHAP values. A wider shape indicates that the SHAP values have a more diffuse distribution, while a narrower shape represents a more focused distribution. The violin plot is given in Figure 17. The result of the Viloin chart also supports the other Shap charts.

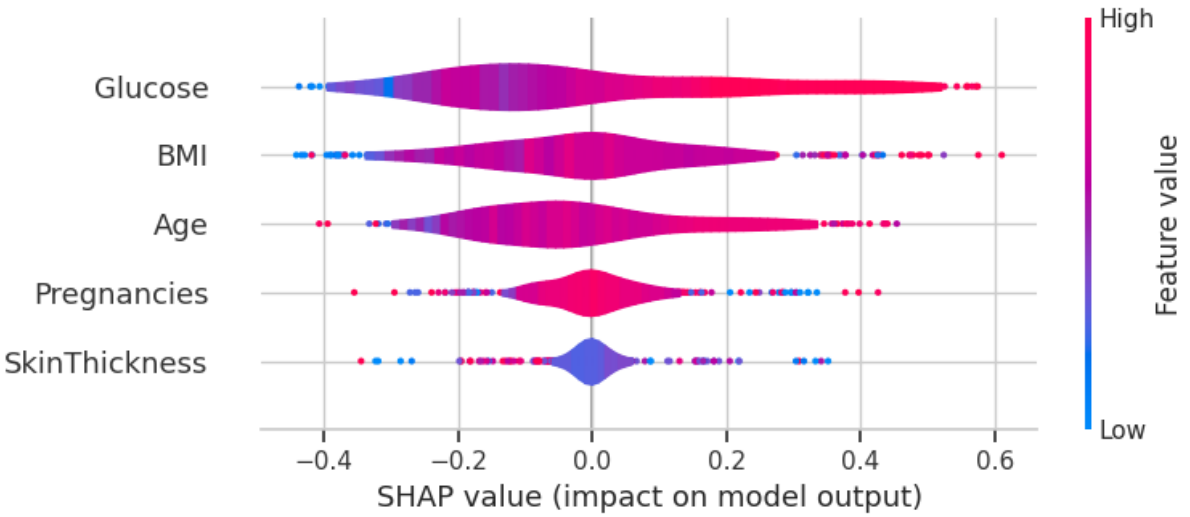

Figure 17. Violin graphic

A heatmap graph is a type of graph that clearly shows the effect of each feature on the target variable. Regions where features have a high impact are shown in darker colour and intensity, while regions with low impact are shown in lighter colour and intensity. This graph can be used to better understand the features of the dataset, to identify the features with significant effects and to improve the interpretability of the model. The heatmap graph prepared with SHAP values in diabetes prediction is given in Figure 18. It is noteworthy that, especially after sample 400, the red colour intensifies in the wet glucose and BMI features and there is a sharp increase in the upper f(x) line.

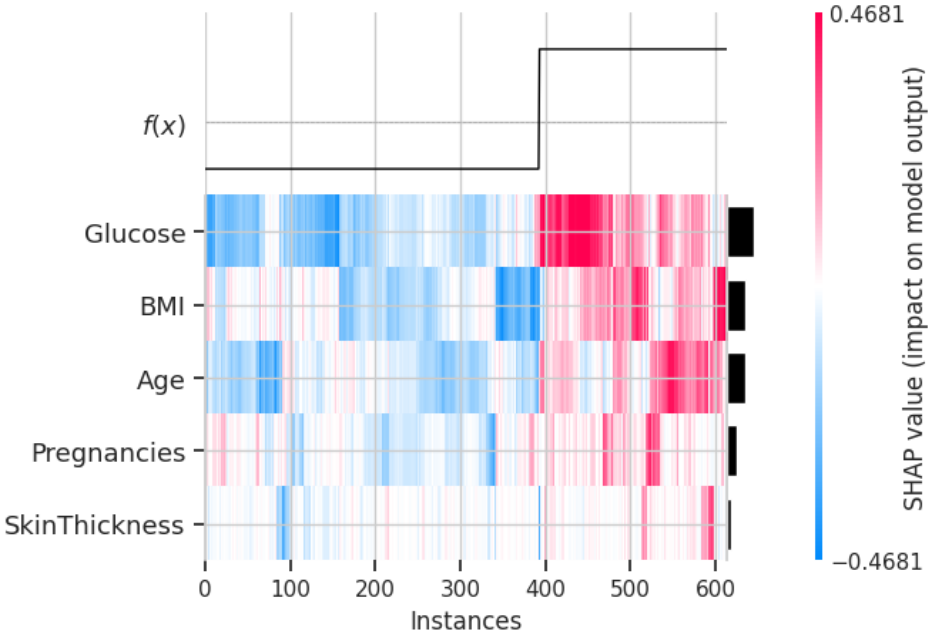

Figure 18. Heatmap graphic

5. CONCLUSION

In this study, XAI techniques were applied to diabetes prediction and classification using a diabetes disease dataset. We investigated the effectiveness of various machine learning algorithms in combination with XAI methods to develop accurate and interpretable models for diabetes prediction. Using XAI, important features and patterns contributing to diabetes risk were uncovered, shedding light on the underlying mechanisms and helping medical practitioners make informed decisions.

The success performances of the models were evaluated with metrics such as F1 Score, accuracy, balanced accuracy, precision, recall, ROC AUC and time taken. SVM and Random Forest stand out as the most successful models. With the most successful models, the impact of different features on diabetes prediction was evaluated with different SHAP plots representing the contribution of each feature to the final prediction compared to the average prediction. Glucose, Age and BMI were found to have a significant and positive effect on the model output.

In future studies, better results can be obtained with larger data sets, different models and different explainable model plots.

REFERENCES

[1] Mujumdar, A., Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.

[2] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.

[3] Aelgani, V., Gupta, S.K., Narayana, V.A. (2023). Local Agnostic Interpretable Model for Diabetes Prediction with Explanations Using XAI. In: Reddy, K.A., Devi, B.R., George, B., Raju, K.S., Sellathurai, M. (eds) Proceedings of Fourth International Conference on Computer and Communication Technologies. Lecture Notes in Networks and Systems, vol 606. Springer, Singapore. https://doi.org/10.1007/978-981-19-8563-8_40.

[4] Soni, M., Varma, S. (2020). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (Ijert) Volume, 9.

[5] Lai, H., Huang, H., Keshavjee, K., Guergachi, A., Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders, 19, 1-9.

[6] Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., Jonkman, M. (2021). A comparative study of different machine learning tools in detecting diabetes. Procedia Computer Science, 192, 467-477.

[7] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[8] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

[9] Lu, H., Uddin, S. (2022). Explainable Stacking-Based Model for Predicting Hospital Readmission for Diabetic Patients. Information, 13(9), 436.

[10] Vishwarupe, V., Joshi, P. M., Mathias, N., Maheshwari, S., Mhaisalkar, S., Pawar, V. (2022). Explainable AI and Interpretable machine learning: A Case Study in Perspective. Procedia Computer Science, 204, 869-876.

[11] Nagaraj, P., Muneeswaran, V., Dharanidharan, A., Balananthanan, K., Arunkumar, M., Rajkumar, C. (2022, April). A Prediction and Recommendation System for Diabetes Mellitus using XAI-based Lime Explainer. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 1472-1478). IEEE.

[12] Kahn M. Diabetes. UCI Machine Learning Repository. https://doi.org/10.24432/C5T59G.