

Interpretability & Explainability in AI (DSAI 305)

Project – Phase

Team Name: ExplainAI

Team Members:

Sama Mohamed - 202201867

Bosy Ayman - 202202076

Zeyad Sherif 202201220

GitHub Repository: [Link](#)

Google Drive: [Link](#)

Phase 1

Research Problem:

The problem we are addressing is the early and accurate prediction of diabetes using machine learning. Diabetes is a serious and chronic disease that leads to high blood sugar levels, and if left undiagnosed or untreated, it can result in severe complications. Traditional diagnosis methods require patients to visit medical centres, which can be time-consuming and sometimes inaccessible. By leveraging machine learning, we aim to develop a predictive model that can assess a patient's likelihood of having diabetes with high accuracy. This approach can assist in early detection, enabling timely medical intervention and reducing the risk of severe health issues.

Student 1 paragraphs (Sama Mohamed):

1. Logistic Regression:

Logistic Regression is a classification model which is commonly used in medical diagnosis which includes diabetes prediction. The research employed Logistic Regression in the study on diabetes detection, utilizing the Pima Indians dataset which is a widely used dataset for diabetes classification [1]. An 80% accuracy was achieved by the Logistic Regression model achieved in predicting diabetic and non-diabetic patients. The model was trained on medical features, including glucose levels, BMI, insulin levels, and age to predict the probability of diabetes [1]. The dataset classes were nearly linear separable which helps the Logistic Regression perform well and achieve 80% accuracy [1]. There is a strong positive relation between glucose levels and the probability of diabetes occurrence.

2. Neural Network:

Neural Networks is a classification model which is commonly used in medical diagnosis due to its ability to predict complex, nonlinear relationships between the training features and the probability of diabetes occurrence [2]. Women were selected with a number of 15,000 that were aged between 20 to 80 to be involved in the study. The model was trained on medical features including plasma glucose level, BMI, insulin level, and age [2]. The study involved 15,000 women aged 20 to 80. A 90% accuracy, with a precision of 83.8% and a recall of 89% was achieved by the Neural Network model in predicting diabetic and non-diabetic patients [2]. Neural Networks have the ability to capture complex relations between features which is different from the linear models that only capture a simple relationship between features and target. This helps the model to perform well and get higher accuracy [2].

3. XGBOOST:

Extreme Gradient Boosting (XGBoost) is a commonly used machine learning model in medical diagnosis. XGBOOST has performed well with diabetes prediction due to its ability to predict complex datasets with high dimensions efficiently [3]. The model was trained on the medical features of the Pima Indians diabetes dataset to predict the probability of diabetes occurrence. The XGBOOST model achieved an Area Under the Curve (AUC) score of 94% in predicting diabetic and non-diabetic patients [3]. XGBOOST uses an extreme gradient boosting mechanism which helps in minimizing error by sequentially refining weak learners in the same time of reducing overfitting by using regularization techniques, leading to performing better than other models [3].

Student 2 paragraphs (Bosy Ayman):

1. KNN:

K-Nearest Neighbour (KNN) is a supervised learning algorithm that classifies diabetes based on patient health attributes. It determines the class of a new data point by measuring its distance from existing data and assigning it to the majority class among its k nearest neighbours [4]. Using the Pima Indians Diabetes Dataset, studies highlight the importance of glucose levels, BMI, and blood pressure in prediction [4]. Data preprocessing, including scaling and normalization, improves accuracy [4]. While effective, KNN's performance depends on k selection, dataset distribution, and computational cost [4].

2. SVM:

Support Vector Machine (SVM) is a machine learning algorithm used for diabetes prediction. A study using NHANES 1999-2004 data applied SVM to detect diabetes and pre-diabetes cases, achieving AUC scores of 83.5% and 73.2%. Key features included family history, age, BMI, waist circumference, and hypertension. The Radial Basis Function (RBF) kernel yielded the best results. Researchers also developed a web-based Diabetes Classifier tool, demonstrating SVM's potential for early screening using non laboratory clinical data [5].

3. Naïve Bayes:

Naïve Bayes is a simple and fast way to classify data using probabilities. The method assumes that different features don't affect each other, but it still works well in medical research. A study with the Pima Indians Diabetes Dataset showed 76.30% accuracy in identifying diabetic and non-diabetic patients. The model handles missing data and uneven datasets well, making it useful for early diagnosis. Researchers tested accuracy with ROC curves and found the approach reliable for predicting diabetes. With ease of use, speed, and clarity, Naïve Bayes remains a valuable tool in medical diagnosis [6].

Student 3 paragraphs (Zevad Sherif):

1. Decision Tree:

Decision Tree is a supervised machine learning algorithm which is commonly used for classification of medical diagnosis as in our case diabetic, or non-diabetic [7]. A decision tree model was implemented by researchers to predict the occurrence of diabetes after training on Pima Indians diabetes dataset. The features with higher correlation with the target value and affects the classification process most were glucose levels, BMI, and age. An 73.82% accuracy, with a precision of 87.31% and a F-Score of 81.87% was achieved by the Decision tree model in predicting diabetic and non-diabetic patients [7]. Decision trees are effective in splitting the data into meaningful subsets based on feature importance. This helps the model to perform well and get higher accuracy [7].

2. Random Forest:

Random Forest (RF) is a machine learning algorithm which is commonly used for classification of medical diagnosis as it enhances predictive accuracy by aggregating multiple decision trees [8]. A Random Forest model was implemented by researchers to predict the occurrence of diabetes after training on hospital physical examination data from Luzhou, China, and the Pima Indians Diabetes dataset [8]. An 73.82% accuracy was achieved by the Random Forest model in predicting diabetic and non-diabetic patients [8]. Random Forest has the ability to reduce variance and prevent overfitting through up-sampling and feature randomness. This helps the model to perform well, get higher accuracy, be more generalized with different datasets [8].

3. LightGBM:

Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework which is mainly optimized for speed and efficiency [9]. This makes LightGBM commonly used for medical classifications as diabetes prediction. A LightGBM model was implemented by researchers to early predict the occurrence of diabetes after training on a labelled dataset from Bangladesh [9]. The research used techniques as feature selection, missing value imputation, and hyperparameter tuning via grid search to enhance the performance of the model's prediction. An 83.30% area under the curve (AUC) was achieved by the LightGBM model in predicting diabetic and non-diabetic patients [9].

Research Gap:

Explainable AI has significantly improved in healthcare filed but there is still a gap in the combination of interpretability and performance of machine learning models. Previous studies have used multiple machine learning models with different architecture, structure, and parameters to be trained on the patient's data and be able to give prediction based on that. These studies lack comparing the explainability and effectiveness of these models, as they focus on the specific performance metrics. Our research aims to fill this gap by using 9 different models, applying them on one dataset to be able to see the difference of performance on using each model, and evaluating the models by using interpretability and explainability frameworks. It aims to provide insights into the optimal the combination between model accuracy, other performance metrics and interpretability, explainability which improves the adoption of Explainable AI in the real-world decision making in the medical healthcare filed.

Phase 2

1- Preprocessing

1. Importing needed libraires:

- NumPy, pandas: Used for data manipulation. Essential for structuring and understanding tabular data.
- Matplotlib, PyPlot, Seaborn: Used for visualization. These help us see patterns in the data a core part of explainability.
- scipy, scipy.stats: Offers advanced statistics tools like hypothesis testing. They are also used for correlation and chi-square analysis which helps in interpreting relationships between variables.
- sklearn.model, sklearn libraries: They are tools used for data splitting, scaling, and feature selection and they are necessary for preparing the model and understanding which features matter most.
- statsmodels.api: It is used for deeply statistical modeling which helps with regression interpretability.
- warnings.filterwarnings("ignore"): It prevents clutter from warning messages and it doesn't affect explainability but keeps the output readable.
- %matplotlib inline: It is a Colab command to make sure that the plots display in the notebook.

2. Loading the dataset:

- It downloads the dataset from Kaggle and loads it into a pandas data frame.
- The data contains real patient metrics like BMI, glucose levels, insulin levels, skin thickness, blood pressure, pregnancies, Diabetes Pedigree Function, Age, and the target feature is Outcome.

3. Initial Exploration:

- data.head(): Shows first 5 rows.
- data.describe(): Summarizes numerical features as mean, standard deviation, minimum, maximum which is important for detecting outliers, and identifying skewness.
- data.info(): Checks non null values and data types.
- data.shape(): Gives dataset dimensions (768, 9).
- data.value_counts(): Helps identify class imbalance.
- data.dtypes(): Confirms types of columns of the dataset.
- data.columns(): Confirms columns of the dataset.

4. Univariate analysis:

A. Checking and Handling Missing Values:

- The data at first seems to have no nan values in it but it suffers from problem of zeros in column of insulin, glucose, skin thickness, blood pressure, and age.

B. Handling problem of Zeros:

- Replaces zeros with NAN to be filled as it is unrealistic that person can have zero insulin, glucose, skin thickness, blood pressure, or age.
- Fill the missing using the median as median is more robust to outliers than mean, which Keeps the distribution more realistic without making skewness from extreme values.

C. Checking for duplicates:

- It is important to check for duplicates, as duplicates can affect certain patterns which misleads the model, making it interprets the repeated rows as stronger evidence.

D. Checking outliers using Box plots:

- Boxplots helps in identifying outliers, which is shown as points far from the median or the interquartile range, and the results are:
 - Pregnancies: Right-skewed. Outliers indicate rare high values.
 - Glucose: Higher values more common in diabetics. Model should learn these patterns.
 - Blood Pressure: Needs validation for extreme low, and high values.
 - Age: Right-skewed.
 - Insulin: skewed, as it might require normalization.
 - BMI: concentrated between 25-45, and there are outliers at extremes.
 - Diabetes Pedigree Function: Most values are low; and a few extreme highs which are important for risk-based predictions.

E. Distribution of Numerical Features:

- The distribution of the numerical features is visualized and resulted in:
 - Pregnancies: Right-skewed, and most women had 0–6 pregnancies. Rare cases up to 17.
 - Glucose: Approximately normally distributed, higher glucose levels dominate among diabetics, and there is a noisy tail at the end.
 - Blood Pressure: It has a long-left tail, which indicates that some entries are suspiciously low.
 - Skin Thickness: Right skewed.
 - Insulin: Right skewed, many values are near zero, with a long tail after 600.
 - BMI: Approximately normally distributed, with a peak around 30–35, and some patients have BMI over 50 which can be outliers.
 - Diabetes Pedigree Function: Right-skewed, as the majority of the values is less than 0.5, with rare very high values more than 1.0.
 - Age: Right-skewed, as most patients are aged less than 40, and Patients over 60 are less common.
 - Outcome:
 - ❖ 0 (No Diabetes): More frequent, as the dataset is slightly imbalanced.
 - ❖ 1 (Diabetes): Less frequent, but it is still substantial representative, as no major class imbalance.

5. Multivariate analysis:

A. Correlation Matrix, and Heatmap:

- It filters only numerical features.
- It computes Pearson correlation matrix.
- It visualizes the correlation matrix with a heatmap (`sns.heatmap`) with color gradients.
- It is used in detecting multicollinearity, as highly correlated features can be dropped, and features highly correlated with outcome (target) are strong and important features.

B. Raw Correlation Matrix: `data.corr()`:

- It outputs a raw table of the correlation matrix in text format

C. Histograms for All Features:

- Histogram plots for all features are visualized, they check distributions side-by-side, and compare feature spread & skewness visually.
- It shows results of:
 - Insulin, Diabetes Pedigree Function, and Skin Thickness are heavily skewed.
 - Age, and Pregnancies are right-skewed.
 - Glucose, and BMI are more symmetrical.

D. Pair plot for Mean Features:

- It Plots scatterplots, and histograms for all feature combinations.
- The diagonal shows the distribution of each feature.
- The off diagonal shows the pair relationships.

E. Pair plot with Outcome Hue:

- It adds the target of outcome as a color hue, which helps in separating diabetic and non-diabetic patterns.
- It shows that diabetic cluster have higher Glucose, BMI, Age values.
- It shows that non diabetic cluster have lower values.

6. Splitting training and testing data:

- The dataset will be split into: training set (80%), and testing set (20%) with a random state=42 to ensures reproducibility.
- We train the model on x_{train} and test how well it generalizes to x_{test} , as the split ensures that the model will be tested on unseen data to make the evaluation of the model generalized.

7. Feature Selection:

A. Feature importance for all features in your dataset using (Fisher's Score):

- It is also called ANOVA F-test and it used to select features that have the strongest relationship with the target.

B. Feature importance for all features in your dataset using (Correlation Coefficient):

- It calculates Pearson correlation between scaled features and target to measure the linear dependence between each feature and the target feature.

C. Feature importance for all features in your dataset using (Variance Threshold):

- It uses the Variance Threshold to remove features with low variance, which means that this feature has low information content.

D. Features that are dependent on each other:

- It uses pairwise correlation to detect and remove multicollinear features.

E. The most correlated 3 features with it using Chi-Square Test Scores:

- It uses the chi-Square contingency to assess independence between the categorical features and target.

F. Backward Elimination Feature Selection:

- It uses the OLS regression with iterative removal algorithm based on:
 - High p-values which means non-significant features.
 - High VIF which means that there is multicollinearity.
- It Refines feature set for linear models while ensuring statistical relevance and no multicollinearity.

2- Sama Mohamed / 202201867.

Model 1: logistic Regression:

Logistic Regression model is implemented which is a simple linear classification model that can be explained and interpreted using its six assumptions:

Assumption 1 - Appropriate outcome type:

- Logistic regression requires the target features to be categorical and to have binary (2) values only. The assumption is satisfied as the code `(y.nunique()) = 2`.

Assumption 2 - Linearity of independent variables and log odds

- Logistic regression assumes that log-odds are linearly related to input variables. Scatter plots of all features vs. log-odds should look linear after being visualized.
- The results show that:
 - Pregnancies: Not linear.
 - Glucose: Linear.
 - Blood Pressure: It can be approximately linear with some adjustments.
 - Skin Thickness: Not linear.
 - Insulin: Not linear.
 - BMI: Linear.
 - Diabetes Pedigree Function Not linear.
 - Age: It can be approximately linear with some adjustments.

Assumption 3 - No strongly influential outliers:

- It uses Cook's distance and standardized residuals to detect points that influence on the model.
- If the Cook's distance less than the defined threshold, this means that it is potentially influential data points.
- If the standardized residual more than 3, this means that it is an outlier in terms of residuals.

- This helps in deciding whether to remove points or investigate them searching for data errors, and rare events.
- The used threshold in the model is 0.005208333333333333, and the proportion of data points that are highly influential = 6.4%.
- Sort descending the extremes to see the most influential points for manual inspection.

Assumption 4 - Absence of multicollinearity:

- It checks for the correlation between the independent variables.
- It checks for the multicollinearity, as it can decrease the consistency and stability of logistic regression coefficients.
- The colors of Bright red or green near diagonal means that there is strong correlation with values more than 0.8 or less than -0.8.
- This helps in identifying if you should:
 - Remove some features.
 - Use regularization techniques like Lasso or Ridge.
 - Perform dimensionality reduction.
- If the VIF value is more than 10, this means that there is a severe redundancy.
- Features with high VIF should be handled by removing it, or concatenating it with others.
- The results show high VIF values in Glucose, Blood Pressure, Skin Thickness, BMI, Age, so the assumption is not satisfied.

Assumption 5 - Independence of observations:

- The visualized plot is used to check of the residual behavior, as it should appear randomly scattered around 0.
- The Patterns or trends suggest model misspecification which does not satisfy the assumption.
- The results shows that there is no trend or pattern observed so the assumption is satisfied.

Assumption 6 - Sufficiently large sample size:

- Since the threshold of the number of sufficient sample size is 500, and the total number of observations is 768 in the dataset, then the assumption is satisfied.

Testing and evaluating the logistic regression model:

Mapping y_{pred} as values more than 0.5 will be labelled as 1, otherwise it will be 0.

The model scored precision of 0.65, recall of 0.65, f1score of 0.65, Mean Squared Error of 0.35064935064935066, R2 score of -0.6536090674090276, and accuracy of 64.93%

3- Zeyad Sherif / 202201220.

Model 1: Random Forest:

Random Forest (RF) which is a machine learning algorithm which is commonly used for classification of medical diagnosis as it enhances predictive accuracy by aggregating multiple decision trees is implemented and can be explained and interpreted by using Local and Global model Agnostic methods:

A. Global Model Agnostic Methods:

1- Partial Dependence Plot (PDP):

- PDP shows the average predicted outcome as a function of one or two features, marginalizing over all other features.
- It helps in identifying whether increasing a feature will increase or decrease predictions.
- The results show that:
 - Glucose: As glucose increases, partial dependence rises, which identifies strong positive correlation with diabetes outcome risk.
 - BMI: Moderate increase in risk with higher BMI.
 - Age: Gradual rise, which indicates that the age is a cumulative risk factor.
 - Diabetes Pedigree Function: Non-linear, showing stable line risk.
 - Insulin: It shows flat regions that is followed by rise, indicating that the model is sensitive to very high insulin levels.
 - Blood Pressure: Mostly flat, indicating low influence overall on the model.
 - Pregnancies: Risk increases with number of pregnancies increases.
 - Skin Thickness: positive trend, which shows minor contributor as it is slightly increasing.

2- Individual Conditional Expectation (ICE) Plot:

- ICE plots are similar to PDPs, but they show one line per instance in the dataset, showing how the model's prediction changes as a single feature varies for that specific value, as it gives the individual behavior, not the average.
- The results show that:
 - ICE plots show the variation for individual samples.
 - Features like Glucose and Age show consistent positive effects.
 - The more spread there means the more consistency in how the model treats that feature.
 - Insulin and Glucose show greater variance.

3- Accumulated Local Effects (ALE) Plot:

- ALE plots show the average local effect of a feature over its range, which is computed in small intervals.
- ALE doesn't assume feature independence, so it provides a more reliable interpretation when features are correlated.
- It calculates how much a prediction changes as you move from one bin of feature values to the next.

- The results show that:
 - Glucose, BMI, Age show consistent monotonic increases.
 - ALE captures interaction free marginal effects, better than PDP when features are correlated.
 - Flat or noisy ALE lines for Blood Pressure, Skin Thickness indicates a minor contribution.

4 -Permutation Feature Importance:

- It quantifies how important a feature is by randomly shuffling its values and seeing how much the model's performance degrades.
- The larger the performance drop means the more important the feature is.
- The results show that:
 - Measures drop in performance when a feature is shuffled.
 - Glucose, BMI, Age scores were on the top which indicates the consistency across methods.

5 -LOFO (Leave One Feature Out) Importance:

- It trains the model multiple times, each time excluding one feature, and compares the performance.
- The difference in performance tells you how much that feature contributed.
- It directly measures how much ROC, and AUC drops when a feature is removed.
- The results show that:
 - Glucose, Age, BMI have the largest drop which means that they are the most important for model predictions.

6 -Global Surrogate Model:

- Fits an interpretable model like linear model to mimic the predictions of a black-box model in this case the random forest model.
- It approximates the behavior of the original model across the input space.
- The results show that:
 - High surrogate accuracy of 1.00 which means that the complex model can be approximated linearly.

B. Local Model Agnostic Methods:

1- LIME:

- LIME for Tabular Data is used as the dataset is tabular.
- LIME selects an instance, create small variations of the instance, get predictions for these variations using the ML model. train a simple interpretable model (like linear regression) on the created variance data, and explain the prediction using the simple model's weights.
- The results show that:
 - It Shows top 5 influential features per instance.
 - It indicates the consistent features of Glucose, BMI, Age.
 - It Visualizes both impact and actual feature values.

- SHAP is used to explain the output of Machine Learning models.
- It is based on Shapley values, which use game theory to assign credit for a model's prediction to each feature or feature value.
- SHAP decomposes the output of a model by the sums of the impact of each feature, calculates a value that represents the contribution of each feature to the model outcome, uses these values to understand the importance of each feature and to explain the result of the model to a human.
- The results show that:
 - SHAP summary bar plot confirms that Glucose, BMI, Age as most influential.
 - The force plots show how features push predictions toward positive class which is the diabetic or negative class which is the non-diabetic.
 - SHAP aligns well with LIME and PDP.

C. Testing and evaluating the Random Forest model:

The model was trained on x_{train} features and y_{train} labels, then it was tested on the x_{test} to get y_{pred} and y_{pred_proba} which will be used in understanding the model confidence. The model scored precision of 0.80, recall of 0.81, f1score of 0.80, Mean Squared Error of 0.19480, R2 score of 0.08132, and accuracy of 80.519%

4- Bosy Ayman / 202202076.

Model 1: SVM:

This project implements a Support Vector Machine (SVM) model based on the paper:

“Application of SVM modeling for prediction of common diseases: the case of diabetes and pre-diabetes.”

Key Elements:

- **10-fold Cross Validation:** To ensure robustness and generalization.
- **Kernels Used:**
 - Linear
 - Radial Basis Function (RBF)
 - Polynomial
 - Sigmoid

Explainability Techniques

Precision: Proportion of positive identifications that were actually correct.

Recall (Sensitivity): Proportion of actual positives correctly identified.

F1 Score: Harmonic mean of Precision and Recall.

Accuracy: Overall proportion of correct predictions.

AUC-ROC: Area Under the Curve of Receiver Operating Characteristic – shows model's ability to distinguish between classes.

Confusion Matrix: Gives detailed insight into TP, FN, FP, and TN.

Results when $C = 10$

[1] Linear kernel : Best performer

Metric	Class 0 (Non diabetic)	Class 1 (Diabetic)
Precision	0.80	0.72
Recall	0.86	0.62
F1 Score	0.83	0.67

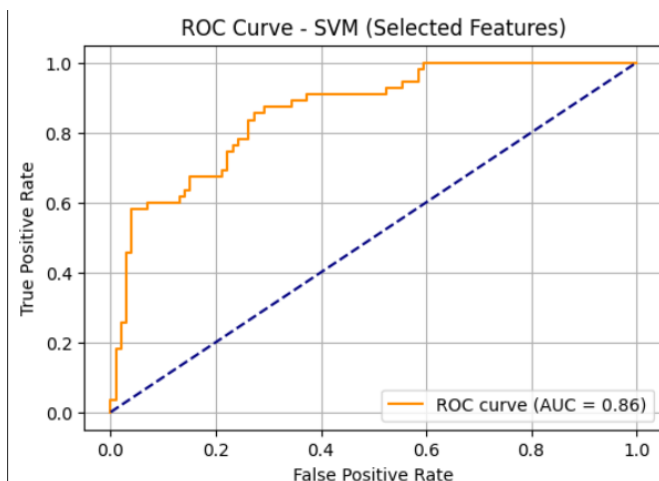
The model performs better predicting the non diabetic based on the high Recall (0.86) and F1 score (0.62) that indicate a balanced performance between (Precision & Recall) for class 0.

Which means that the model is more confident and accurate when predicting **non-diabetic** patients. It is **less effective** at identifying **diabetic** patients.

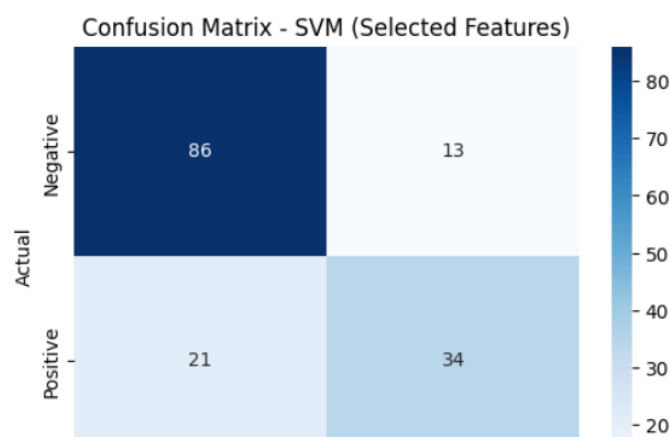
Accuracy: 78%

AUC-ROC:

The insights show a result of 0,86 which indicates that the model performs well.



Confusion Matrix:



TP: 34 diabetic cases were **correctly** predicted as diabetic.

FN: 86 This means the model **misses a lot of diabetic patients**, which is **critical in healthcare**.

FP: 13 non diabetic cases were wrongly predicted as diabetic.

TN: 21 non diabetic cases were correctly predicted as non-diabetic.

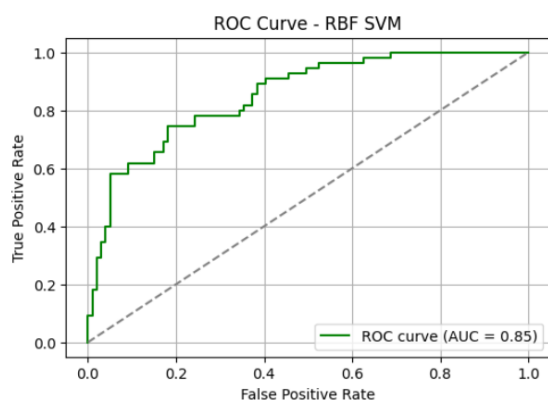
[2] RBF Kernel

Metric	Class 0 (Non diabetic)	Class 1 (Diabetic)
Precision	0.81	0.72
Recall	0.87	0.62
F1 Score	0.83	0.67

The model with the RBF kernel shows strong precision, meaning it is good at confirming diabetic cases when it makes a positive prediction. On the contrary, the recall is moderate in prediction of the diabetic individuals.

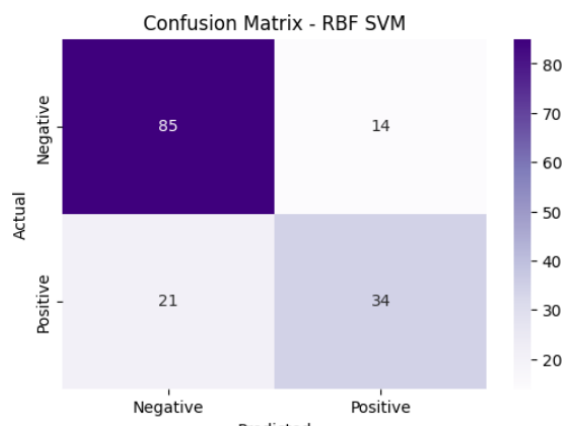
Accuracy: 73%

AUC-ROC:



The insights show a result of 0.85 which indicates that the model performs well.

Confusion Matrix:



TP: 34 diabetic cases were **correctly** predicted as diabetic.

FN: 85 This means the model **misses a lot of diabetic patients**, which is **critical in healthcare**.

FP: 13 non diabetic cases were wrongly predicted as diabetic.

TN: 14 non diabetic cases were correctly predicted as non-diabetic.

[3]Poly Kernel

Metric	Class 0 (Non diabetic)	Class 1 (Diabetic)
Precision	0.76	0.76
Recall	0.92	0.47
F1 Score	0.83	0.58

The model with the polynomial kernel shows strong precision for both classes, meaning that its positive predictions are fairly reliable. However, the recall for diabetic individuals is low, which indicates that many diabetic patients are missed by the model. On the other hand, the model is very effective at identifying non-diabetic individuals.

Accuracy : 75.97%

[4]Sigmoid Kernel

Metric	Class 0 (Non diabetic)	Class 1 (Diabetic)
Precision	0.80	0.67
Recall	0.83	0.62
F1 Score	0.81	0.64

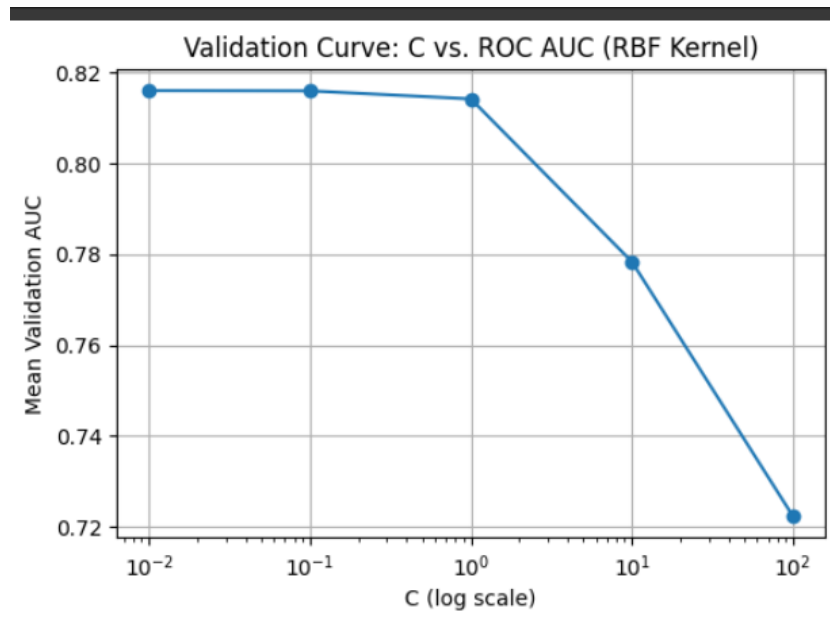
The model with the sigmoid kernel performs reasonably well in identifying both classes. Precision is slightly lower for diabetic predictions, indicating a few false positives.

Recall is better balanced between both classes compared to the poly kernel.so the model is moderately successful in detecting diabetic patients while maintaining good detection of non diabetic individuals.

Accuracy : 75.32%

Assumptions:

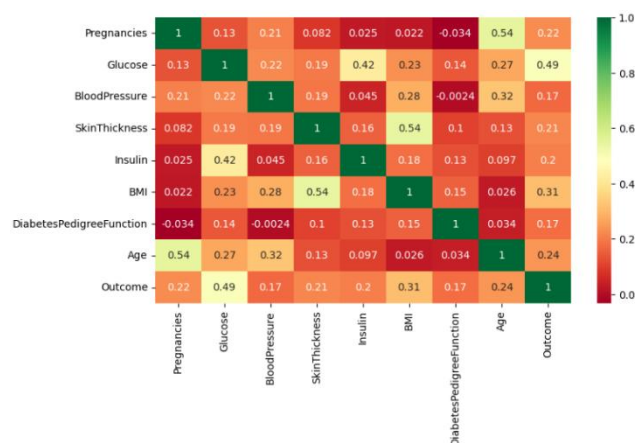
Assumption 1:



-AUC curve analysis shows best performance when C is between 1 and 10.

-A value of C = 10 produced the highest AUC (~0.86), meaning it's a good trade-off between margin maximization and error minimization.

Assumption 2 : Absence of Multicollinearity

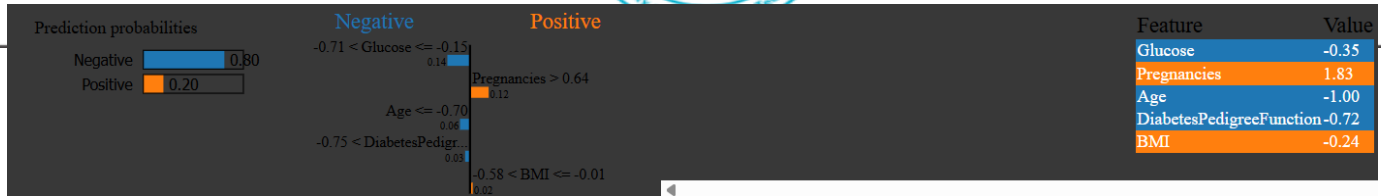


	Feature	VIF
0	Pregnancies	3.273488
1	Glucose	21.723850
2	BloodPressure	31.717477
3	SkinThickness	17.321002
4	Insulin	4.528320
5	BMI	33.678170
6	DiabetesPedigreeFunction	3.153079
7	Age	14.357379

VIF (VIF > 5–10 indicates multicollinearity)

Glucose , BloodPressure, BMI, Age, SkinThickness indicate high collinearity while insulin , Pregnancy and DiabetesPedigreeFunction have acceptable correlation,

LIME

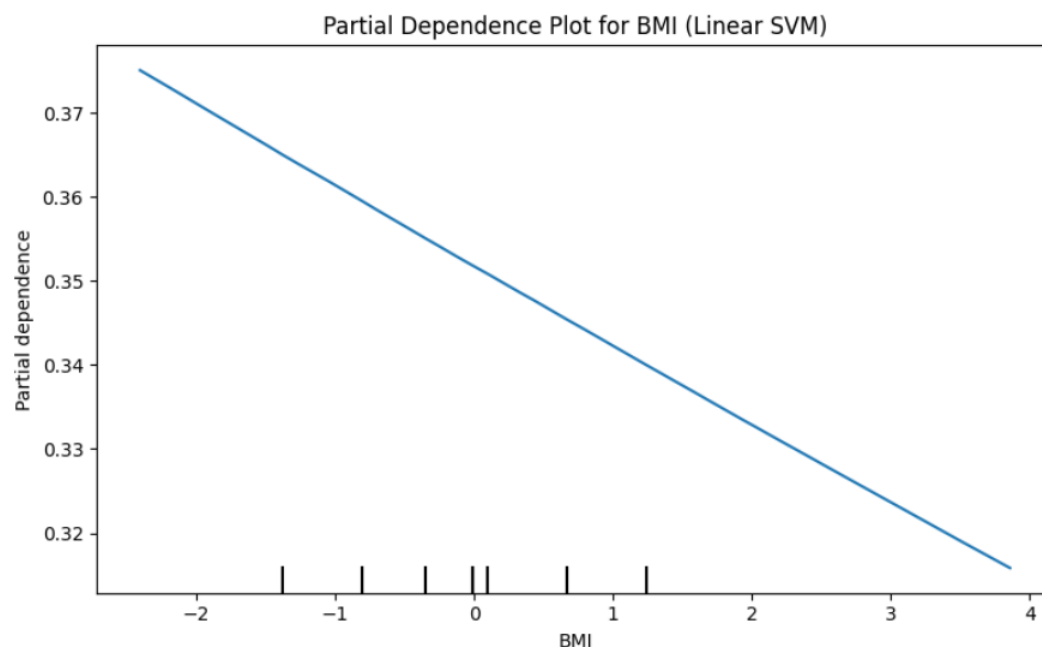


LIME was used to interpret the model's predictions locally by analyzing feature contributions for each prediction.

- **Glucose** had a **strong negative contribution**, indicating high glucose values significantly increase the likelihood of being classified as diabetic.
- **Pregnancies** had a **positive contribution**, especially when the number of pregnancies was higher, contributing more to the prediction of diabetes.

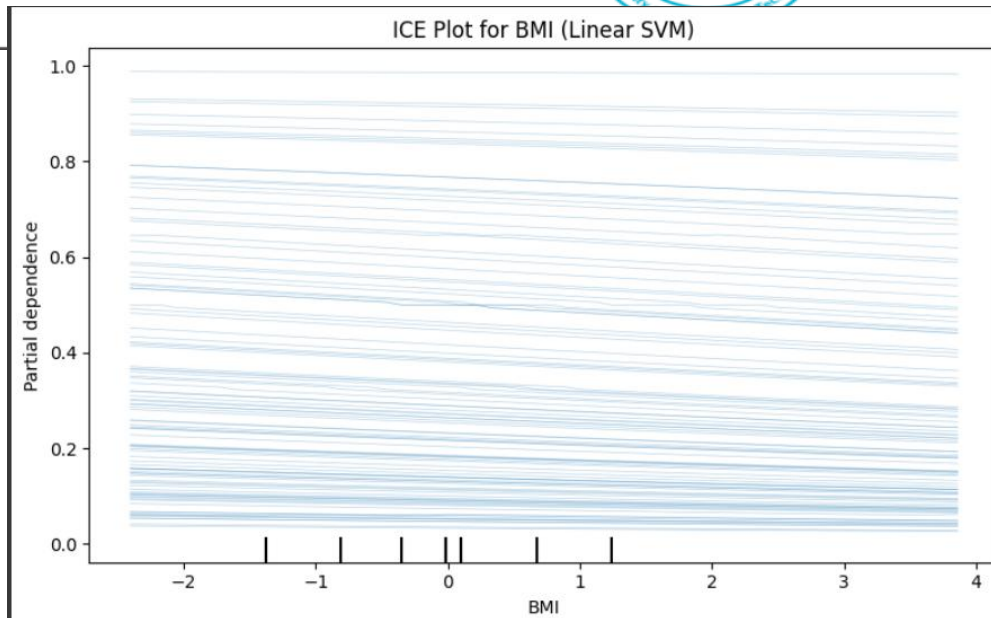
This provides transparency in the black-box SVM model and enhances trust in clinical decision-making.

PDP



The graph indicates that BMI does not contribute well for linear SVM prediction.

ICE



The ICE (Individual Conditional Expectation) lines show how the predicted outcome for each individual data point (or a sampled subset) changes as the feature (BMI) varies while holding all other features constant.

Conclusion

The model demonstrated that SVM model with linear model performs the best for predicting diabetic individuals achieving 78% accuracy and 0.86 AUC ROC, which provide reliable decision making based on its predictions.

Phase 3

1- Sama Mohamed / 202201867.

Model 2: XGB Classifier:

XGB Classifier model is implemented which is a simple classification model that can be explained and interpreted using its global and local agnostic models:

ICE:

- ICE plot shows the model's prediction changes that happen according to each instance of single feature, and the prediction could differ according to the variation of values of the instances in one feature.
- Pregnancies shows less contribution in affecting the prediction.
- Glucose has a consistent positive effect on the predictions.
- Blood pressure, Diabetes degree function has nearly no effect on the prediction, with high variation.
- BMI has variation of instances, and affects the prediction positively
- Insulin, and skin thickness shows a variation of instances.
- Age has variation with positive effect initially, flat area with no effect, and negative effect on predictions.

PFI:

- It measures the importance of the feature by shuffling its instances' values and measure how much degradation happened to the performance of the model, as the more the performance drops, the more importance the feature is.
- Glucose, Insulin, and BMI were the most effecting feature on the predictions of the model.
- Age and Skin Thickness affects the performance with a same value but with different direction as age affects it positively and Skin Thickness affects it negatively.
- Diabetes Degree Function, Blood pressure pregnancies, and skin thickness do not affect the model's prediction at all.

LOFO:

- LOFO works by training the model on several iterations and removing one feature in each iteration, then evaluating the differences in cross validation scores.
- The columns that affect the predictions positively are arranged as Glucose, Age, BMI, Insulin.
- Diabetes Degree Function shows no effect on the model's prediction.
- The columns that affect the predictions negatively are arranged as Pregnancies, Skin Thickness, Blood pressure.

Friedman H-Statistic Method:

- Friedman H-Statistic Method measure how much the interaction between one feature and the rest of them will affect the overall model's prediction, and each iteration one feature is selected.
- All feature has very small effect withing 3 decimal places (0.008).
- The columns that affect the predictions with relatively higher values are arranged as Glucose, Age.
- The columns that affect the predictions with closer values are Insulin, Pregnancies, Skin Thickness, Diabetes Degree Function, BMI.
- Blood pressure shows no effect on the model's prediction.

Lime:

- LIME selects an instance, create small variations of the instance, get predictions for these variations using the model.
- Skin Thickness was the most contributing feature on the model's prediction of class 0.
- Glucose was the most contributing feature by a relatively high value on the model's prediction of class 1.

Testing and evaluating the logistic regression model:

Mapping y_{pred} as values more than 0.5 will be labelled as 1, otherwise it will be 0.

The model scored precision of 0.80, recall of 0.71, f1score of 0.61, Mean Squared Error of 0.2857142857142857, R2 score of -0.3473851660369853, and accuracy of 71.42 %.

Model 3: Neural Network:

Neural Network model is implemented which is a black box classification model that can be explained and interpreted using its global and local agnostic models:

ICE:

- ICE plot shows the model's prediction changes that happen according to each instance of single feature, and the prediction could differ according to the variation of values of the instances in one feature.
- Pregnancies shows less contribution in affecting the prediction.
- Glucose has positive, and negative effect on the predictions.
- Blood pressure has negative effect on the predictions.
- Skin thickness, and Insulin, BMI has positive, and negative effect on the predictions with average of no effect.
- Diabetes degree function has nearly no effect on the prediction.
- Age has variation with positive effect and negative effect on predictions with average of no effect.

PFI:

- It measures the importance of the feature by shuffling its instances' values and measure how much degradation happened to the performance of the model, as the more the performance drops, the more importance the feature is.
- Glucose, and BMI were the most effecting feature on the predictions of the model positively.
- Diabetes Degree Function, and pregnancies affects the model positively but with lower scale.
- Age and Skin Thickness, Insulin, Blood pressure affects the performance negatively.

LOFO:

- LOFO works by training the model on several iterations and removing one feature in each iteration, then evaluating the differences in cross validation scores.
- The columns that affect the models' prediction negatively are Insulin, Diabetes degree function, Skin Thickness, BMI, Blood Pressure.
- The columns that have no effect are Age, Glucose, Pregnancies.

Friedman H-Statistic Method:

- Friedman H-Statistic Method measure how much the interaction between one feature and the rest of them will affect the overall model's prediction, and each iteration one feature is selected.
- The columns that affect the predictions with relatively higher values are arranged as BMI, Glucose, Blood pressure, Insulin, Skin Thickness.
- The columns that affect the predictions with closer values are Age, Pregnancies.
- Diabetes degree function shows no effect on the model's prediction.

Global Surrogate Model:

- Global surrogate Method Simplifies the black box model into decision tree structure to be easier in interpretation and explainability.
- The most affecting feature on the model's predictions are arranged as Glucose Blood pressure, BMI, and Diabetes degree function.

Lime:

- LIME selects an instance, create small variations of the instance, get predictions for these variations using the model.
- Blood Pressure, and Diabetes degree function were the most contributing feature on the model's prediction of class 0.
- Glucose was the most contributing feature by a relatively high value on the model's prediction of class 1.

Testing and evaluating the logistic regression model:

Mapping y_{pred} as values more than 0.5 will be labelled as 1, otherwise it will be 0.

The model scored precision of 0.57, recall of 0.68, f1score of 0.58, Mean Squared Error of 0.3246753246753247, R2 score of -0.5311195068602108, and accuracy of 68.72 %.

2- Zeyad Sherif / 202201220.

Model 2: Decision Tree:

Decision Tree (DT) which is a white-box machine learning algorithm which is commonly used for classification of medical diagnosis as it provides interpretability by representing decisions as a tree. It splits data based on feature thresholds to reach predictions which can be explained and interpreted by using Local and Global model Agnostic methods:

A. Global Model Agnostic Methods:

1- Partial Dependence Plot (PDP):

- PDP shows the average predicted outcome as a function of one or two features, marginalizing over all other features.
- It helps in identifying whether increasing a feature will increase or decrease predictions.
- The results show that:
 - Glucose: Shows an upward trend as predictions increase with glucose levels, indicating a strong positive correlation with diabetes risk.
 - BMI: It shows an upward slope indicating higher BMI increases the model's predicted risk.
 - Age: It shows an upward slope indicating that age is a risk factor.
 - Insulin: Initially flat then rises which indicates sensitivity only at high insulin levels.
 - Pregnancies: It is flat which shows low influence on prediction.
 - Blood Pressure: It is flat which shows low influence on prediction.
 - Skin Thickness: Upward slope which shows that skin thickness is a risk factor.

2- Individual Conditional Expectation (ICE) Plot:

- ICE plots are similar to PDPs, but they show one line per instance in the dataset, showing how the model's prediction changes as a single feature varies for that specific value, as it gives the individual behavior, not the average.
- The results show that:
 - ICE plots show the variation for individual samples.
 - Feature like Glucose shows consistent positive effects.
 - The more spread there means the more consistency in how the model treats that feature.
 - Insulin and BMI show greater variance.
 - ICE helps identify potential interactions in feature effects.

3- Accumulated Local Effects (ALE) Plot:

- ALE plots show the average local effect of a feature over its range, which is computed in small intervals.
- ALE doesn't assume feature independence, so it provides a more reliable interpretation when features are correlated.
- It calculates how much a prediction changes as you move from one bin of feature values to the next.
- The results show that:
 - Glucose shows consistent monotonic increases, which validates the PDP.
 - BMI, Age show an upward slope which validates the contribution to risk.
 - ALE captures interaction free marginal effects, better than PDP when features are correlated.
 - Flat or noisy ALE lines for Blood Pressure, Skin Thickness indicate a minor contribution.

4 -Permutation Feature Importance:

- It quantifies how important a feature is by randomly shuffling its values and seeing how much the model's performance degrades.
- The larger the performance drop means the more important the feature is.
- The results show that:

- Measures drop in performance when a feature is shuffled.
- Glucose, BMI, Age scores were on the top which indicates the consistency across methods.
- Blood Pressure, Skin Thickness scores were the least which indicates the low impact of them.

5 - LOFO (Leave One Feature Out) Importance:

- It trains the model multiple times, each time excluding one feature, and compares the performance.
- The difference in performance tells you how much that feature contributed.
- It directly measures how much ROC, and AUC drops when a feature is removed.
- The results show that:
 - Glucose, Age, BMI have the largest drop which means that they are the most important for model predictions.
 - Blood Pressure, Skin Thickness have the least drop which means that they are the least important for model predictions

B. Local Model Agnostic Methods:

1- LIME:

- LIME for Tabular Data is used as the dataset is tabular.
- LIME selects an instance, create small variations of the instance, get predictions for these variations using the ML model. train a simple interpretable model (like linear regression) on the created variance data, and explain the prediction using the simple model's weights.
- The results show that:
 - It Shows top 5 influential features per instance which are Glucose, Age, BMI, Insulin, Pregnancies.
 - It indicates the consistent features of Glucose, BMI, Age.

C. Testing and evaluating the Random Forest model:

The model was trained on x_{train} features and y_{train} labels, then it was tested on the x_{test} to get y_{pred} and y_{pred_proba} which will be used in understanding the model confidence. The model scored precision of 0.74, recall of 0.75, f1score of 0.74, Mean Squared Error of 0.24675, R2 score of 0.16365, and accuracy of 75.324%.

Model 3: Light GBM:

LightGBM is a gradient boosting framework that uses tree-based learning algorithms and is highly efficient for classification tasks which is commonly used for classification of medical diagnosis, especially on large datasets. It was used to predict diabetes probability based on the input features, and the predictions can be explained and interpreted by using Local and Global model Agnostic methods:

A. Global Model Agnostic Methods:

1- Partial Dependence Plot (PDP):

- PDP shows the average predicted outcome as a function of one or two features, marginalizing over all other features.
- It helps in identifying whether increasing a feature will increase or decrease predictions.
- The results show that:

- Glucose: Shows an upward trend as predictions increase with glucose levels, indicating a strong positive correlation with diabetes risk.
- BMI: It shows an upward slope indicating higher risk with increased body mass.
- Age: It shows an upward slope indicating that age is a risk factor.
- Insulin: Non-monotonic indicating minor influence.
- Pregnancies: Slight increase in risk with number of pregnancies.
- Blood Pressure: It is flat which shows low influence on prediction.
- Skin Thickness: Non-monotonic and flat trends indicating minor influence.

2- Individual Conditional Expectation (ICE) Plot:

- ICE plots are similar to PDPs, but they show one line per instance in the dataset, showing how the model's prediction changes as a single feature varies for that specific value, as it gives the individual behavior, not the average.
- The results show that:
 - ICE plots show the variation for individual samples.
 - Feature like Glucose shows consistent positive effects.
 - The more spread there means the more consistency in how the model treats that feature.
 - Insulin and BMI show greater variance.
 - Pregnancies: Moderate variance between individuals, but consistent positive influence.
 - ICE helps identify potential interactions in feature effects.

3-Accumulated Local Effects (ALE) Plot:

- ALE plots show the average local effect of a feature over its range, which is computed in small intervals.
- ALE doesn't assume feature independence, so it provides a more reliable interpretation when features are correlated.
- It calculates how much a prediction changes as you move from one bin of feature values to the next.
- The results show that:
 - Glucose, Age, BMI: Showed stable and monotonic effects across their range.
 - ALE captures interaction free marginal effects, better than PDP when features are correlated.
 - Flat or noisy ALE lines for Blood Pressure, Skin Thickness indicates a minor contribution.
 - Pregnancies and Insulin: Showed non-linearities but with limited impact.

4 -Permutation Feature Importance:

- It quantifies how important a feature is by randomly shuffling its values and seeing how much the model's performance degrades.
- The larger the performance drop means the more important the feature is.
- The results show that:
 - Measures drop in performance when a feature is shuffled.
 - Glucose, BMI, Age scores were on the top which indicates the consistency across methods.
 - Blood Pressure, Skin Thickness scores were the least which indicates the low impact of them.

5 - LOFO (Leave One Feature Out) Importance:

- It trains the model multiple times, each time excluding one feature, and compares the performance.
- The difference in performance tells you how much that feature contributed.
- It directly measures how much ROC, and AUC drops when a feature is removed.
- The results show that:
 - Glucose, Age, BMI have the largest drop which means that they are the most important for model predictions.
 - Blood Pressure, Skin Thickness have the least drop which means that they are the least important for model predictions

B. Local Model Agnostic Methods:

1- LIME:

- LIME for Tabular Data is used as the dataset is tabular.
- LIME selects an instance, create small variations of the instance, get predictions for these variations using the ML model. train a simple interpretable model (like linear regression) on the created variance data, and explain the prediction using the simple model's weights.
- The results show that:
 - It Shows top 5 influential features per instance which are Glucose, Age, BMI, Insulin, Pregnancies.
 - It indicates the consistent features of Glucose, BMI, Age.
 - Features such as Insulin and Skin Thickness has less impact.

C. Testing and evaluating the Random Forest model:

The model was trained on x_{train} features and y_{train} labels, then it was tested on the x_{test} to get y_{pred} and y_{pred_proba} which will be used in understanding the model confidence. The model scored precision of 0.82 recall of 0.81, f1score of 0.81, and accuracy of 81.168%.

3- Bosy Ayman / 202202076.

Model 2: Naïve Bayes:

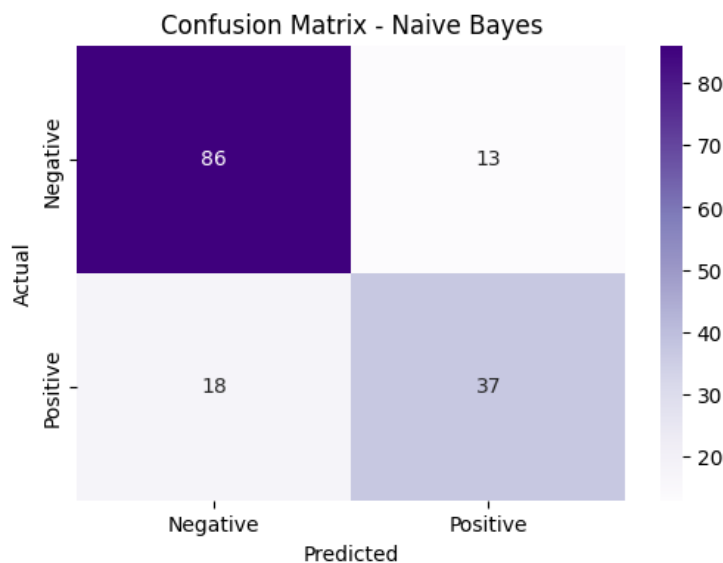
This project implements naive bayes model based on the paper:

“Prediction of Diabetes using Classification Algorithms”

Explainability Techniques

- **Precision:** Proportion of positive identifications that were actually correct.
- **Recall (Sensitivity):** Proportion of actual positives correctly identified.
- **F1 Score:** Harmonic mean of Precision and Recall.
- **Accuracy:** Overall proportion of correct predictions.
- **AUC-ROC:** Area Under the Curve of Receiver Operating Characteristic – shows model's ability to distinguish between classes.
- **Confusion Matrix:** Gives detailed insight into TP, FN, FP, and TN.

Confusion Matrix



TP: 37 diabetic cases were **correctly** predicted as diabetic.

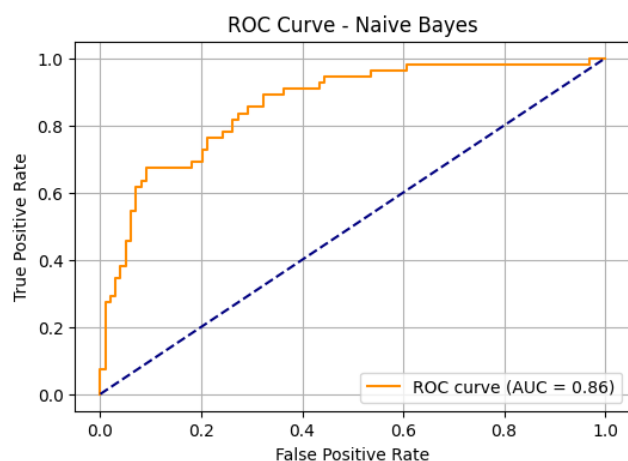
FN: 13 Non diabetic cases were **correctly** predicted as non diabetic.

FP: 18 non diabetic cases were **wrongly** predicted as diabetic.

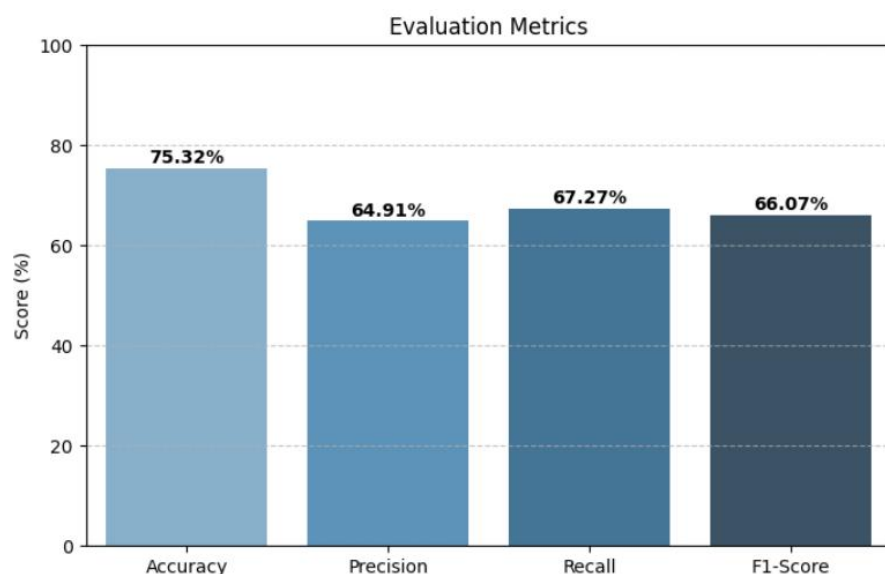
TN: 86 diabetic cases were **correctly** predicted as non-diabetic.

AUC-ROC

The insights show a result of 0.86 which indicates that the model performs well.



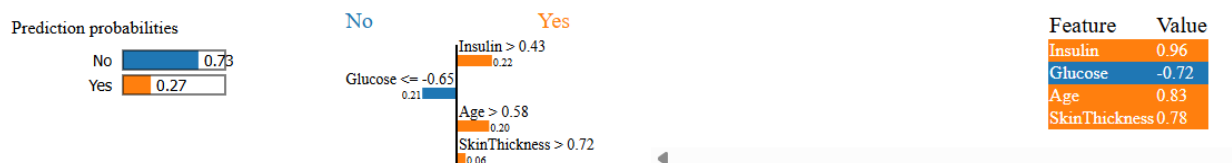
Evaluation Metrics



Metric	Class 0 (Non diabetic)	Class 1 (Diabetic)
Precision	0.81	0.67
Recall	0.82	0.65
F1 Score	0.81	0.66

The model shows strong performance for non-diabetic cases, with balanced and high precision, recall, and F1 score (all around 0.81–0.82). For diabetic cases, performance improves slightly compared to before but remains lower, suggesting the model still needs better sensitivity in detecting diabetes.

1. LIME

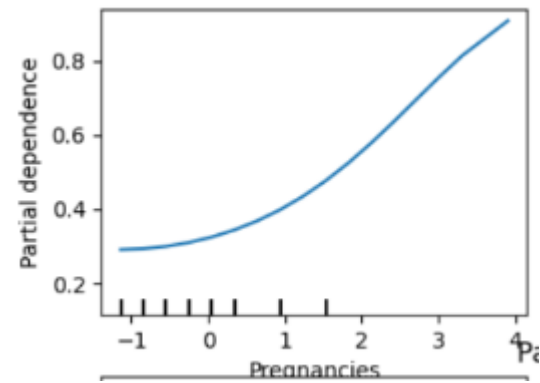
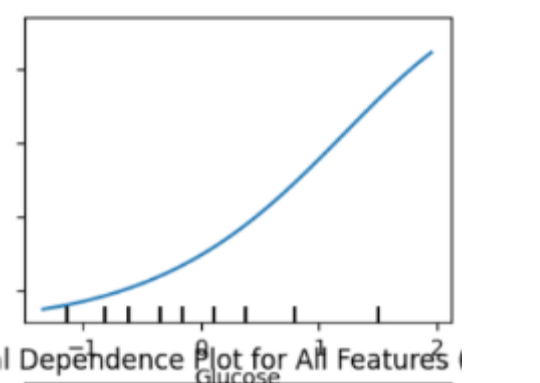
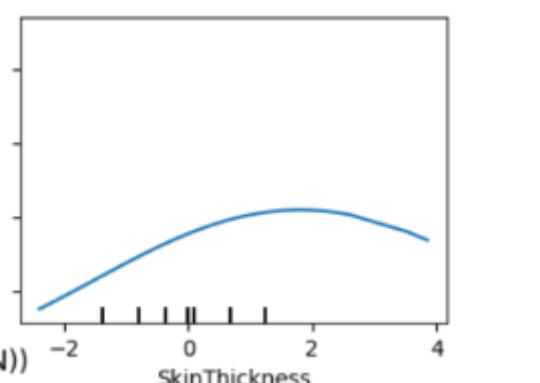
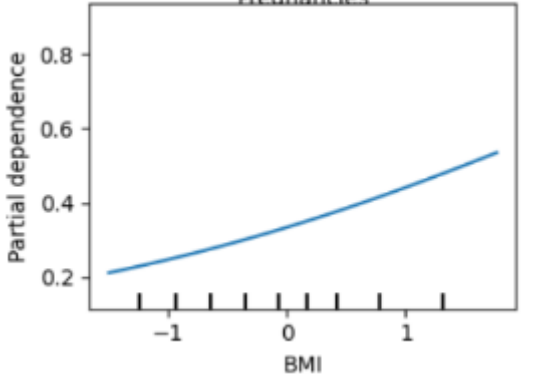


LIME was used to interpret the model's predictions locally by analyzing feature contributions for each prediction.

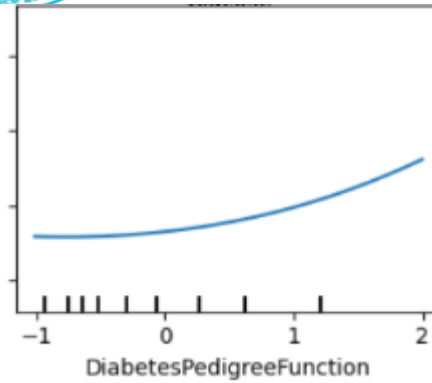
- **Insulin, Age and SkinThickness** Had a **strong negative contribution**, indicating high glucose values significantly increase the likelihood of being classified as diabetic.
- **Glucose** had a **positive contribution**

This provides transparency in the black-box KNN model and enhances trust in clinical decision-making.

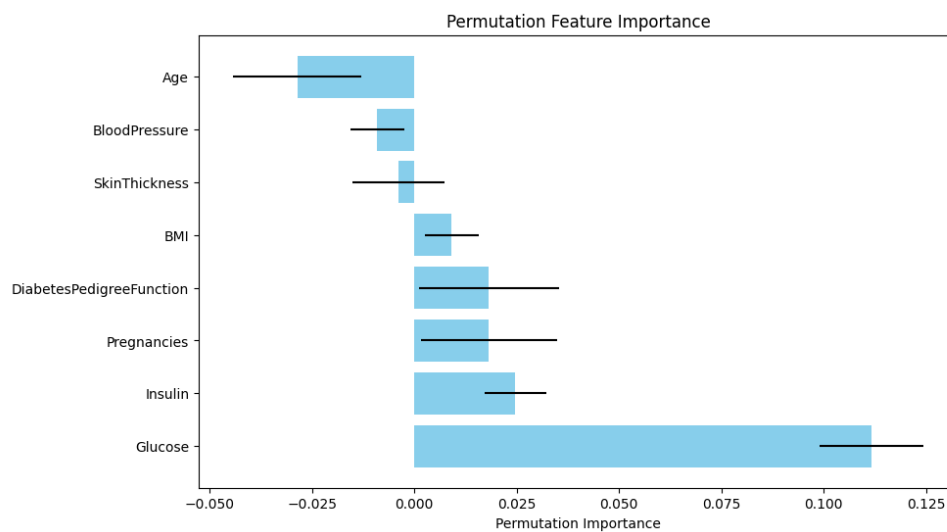
2.PDP

Feature explanation	Graph
The graph indicates that Pregnancies contribute well for linear Naive Bayes prediction. More pregnancies are associated with a higher risk of diabetes according to the Naive Bayes model.	 A line graph showing the partial dependence of the predicted probability of diabetes on the number of pregnancies. The x-axis is labeled 'Pregnancies' and ranges from -1 to 4. The y-axis is labeled 'Partial dependence' and ranges from 0.2 to 0.8. The curve starts at approximately (-1, 0.3) and increases monotonically, reaching approximately (4, 0.9).
The graph indicates that Glucose contributes well for linear Naive Bayes prediction. Higher glucose levels significantly increase the predicted probability of diabetes.	 A line graph showing the partial dependence of the predicted probability of diabetes on glucose levels. The x-axis is labeled 'Glucose' and ranges from -1 to 2. The y-axis is labeled 'Partial dependence' and ranges from 0.2 to 0.8. The curve starts at approximately (-1, 0.2) and increases monotonically, reaching approximately (2, 0.8).
The graph indicates that SkinThickness contributes well for linear Naive Bayes prediction until 2 then slightly decreases. Moderate skin thickness values are associated with a higher diabetes risk, but the effect decreases at very high levels.	 A line graph showing the partial dependence of the predicted probability of diabetes on skin thickness. The x-axis is labeled 'SkinThickness' and ranges from -2 to 4. The y-axis is labeled 'Partial dependence' and ranges from 0.2 to 0.8. The curve starts at approximately (-2, 0.2), rises to a peak of approximately (2, 0.6), and then slightly decreases to approximately (4, 0.5).
The graph indicates that BMI contributes well for linear Naive Bayes prediction. Higher BMI leads to a higher predicted risk of diabetes.	 A line graph showing the partial dependence of the predicted probability of diabetes on BMI. The x-axis is labeled 'BMI' and ranges from -1 to 1. The y-axis is labeled 'Partial dependence' and ranges from 0.2 to 0.8. The curve starts at approximately (-1, 0.2) and increases monotonically, reaching approximately (1, 0.5).

The graph indicates that Glucose **contributes well** for linear Naive Bayes prediction. A higher genetic risk increases the model's predicted probability of diabetes.



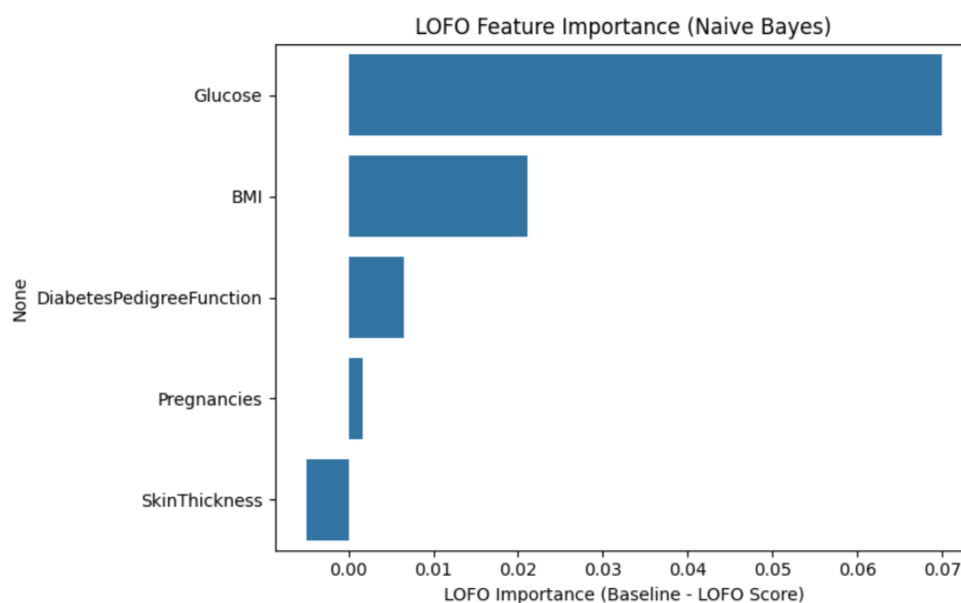
3. Permutation feature importance



It shows how the model degrades when shuffling happens, which indicates that Glucose has **the highest** feature importance while Age has **the lowest** feature importance.

4. LOFO

The goal is to remove features one by one, retrain the model, and evaluate importance. Trains the model multiple times, each time excluding one feature, and compares the performance.



This graph shows that Glucose has **the highest feature contribution** while skin thickness has **the lowest feature contribution**.

Conclusion

The Naive Bayes model performs well for non-diabetic cases, with high precision and recall, but shows weaker performance in identifying diabetic patients, as indicated by many false negatives. Despite this, the AUC-ROC score of 0.86 reflects strong overall discriminative ability. Explainability techniques consistently highlight **Glucose** as the most influential feature, while **SkinThickness** and **Age** contribute less. These insights can guide model improvement and enhance trust in its predictions.

Model 3: KNN:

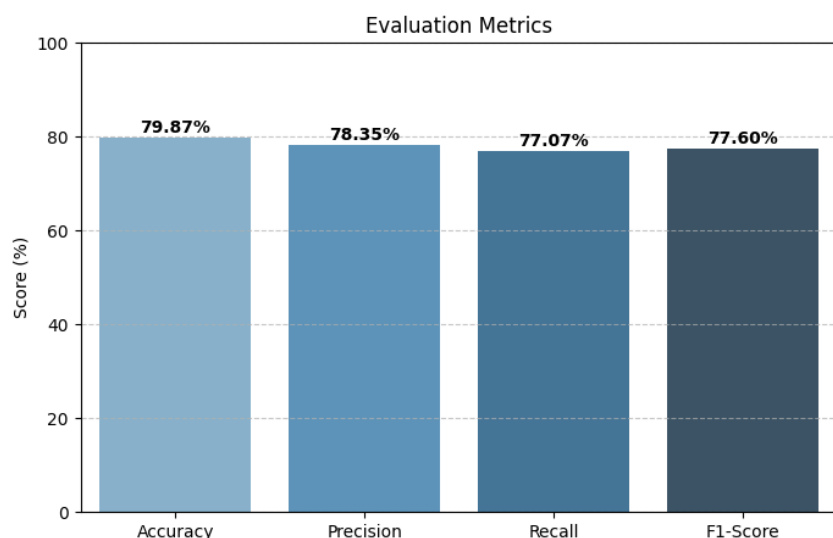
This project implements K nearest neighbor model based on the paper:

“Diabetes prediction using supervised machine learning”

Explainability Techniques

- **Precision:** Proportion of positive identifications that were actually correct.
- **Recall (Sensitivity):** Proportion of actual positives correctly identified.
- **F1 Score:** Harmonic mean of Precision and Recall.
- **Accuracy:** Overall proportion of correct predictions.
- **AUC-ROC:** Area Under the Curve of Receiver Operating Characteristic – shows model’s ability to distinguish between classes.
- **Confusion Matrix:** Gives detailed insight into TP, FN, FP, and TN.

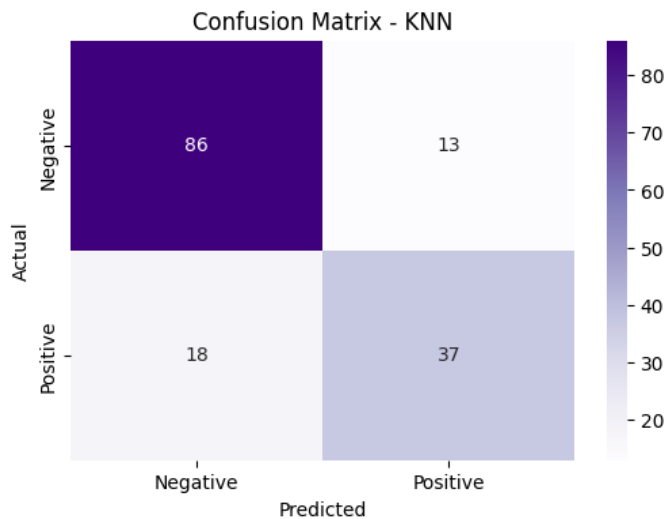
Evaluation Metrics



Metric	Class 0 (Non diabetic)	Class 1 (Diabetic)
Precision	0.79	0.64
Recall	0.81	0.62
F1 Score	0.80	0.63

The model performs better for non-diabetic cases, with higher precision, recall, and F1 score. Its performance on diabetic cases is weaker, indicating a need for improvement in detecting diabetes accurately.

Confusion Matrix



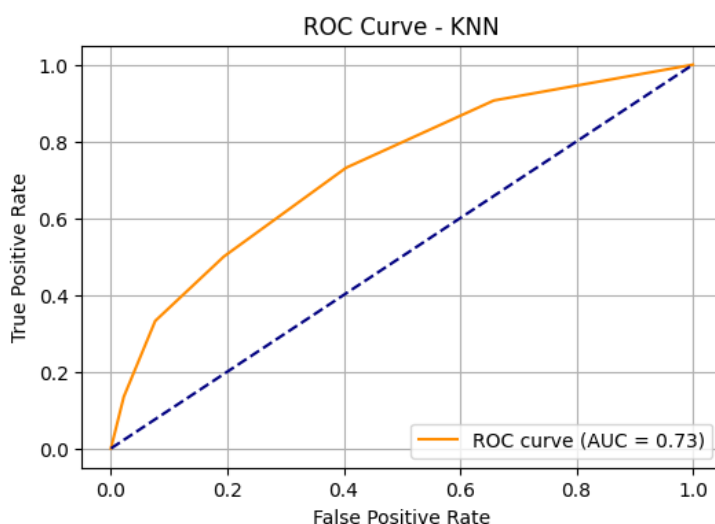
TP: 37 diabetic cases were **correctly** predicted as diabetic.

FN: 86 This means the model **misses a lot of diabetic patients**, which is **critical in healthcare**.

FP: 13 non diabetic cases were wrongly predicted as diabetic.

TN: 18 non diabetic cases were correctly predicted as non-diabetic.

AUC - ROC



The insights show a result of 0.73 which indicates that the model performs well.

1. LIME

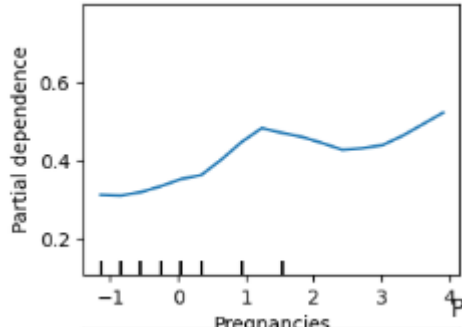
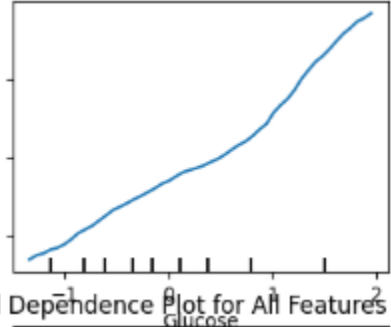
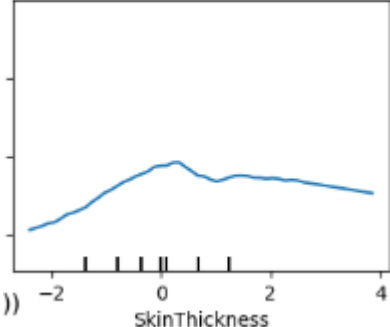


LIME was used to interpret the model's predictions locally by analyzing feature contributions for each prediction.

- **Glucose** had a **strong negative contribution**, indicating high glucose values significantly increase the likelihood of being classified as diabetic.
- **Blood Pressure** had a **positive contribution**

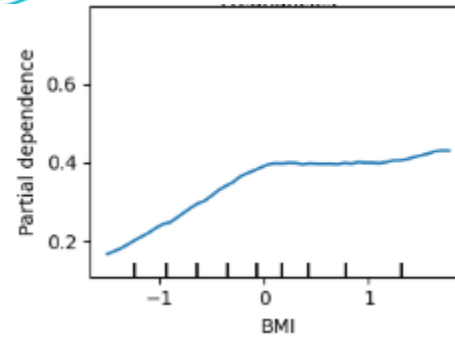
This provides transparency in the black-box KNN model and enhances trust in clinical decision-making.

2.PDP

Feature explanation	Graph
The graph indicates that Pregnancies contributes with Mild positive impact for linear KNN.	 <p>A line graph showing the partial dependence of the model's output on the number of pregnancies. The x-axis is labeled 'Pregnancies' and ranges from -1 to 4. The y-axis is labeled 'Partial dependence' and ranges from 0.2 to 0.6. The curve starts at approximately 0.32 at x=-1, rises to about 0.38 at x=0, peaks at about 0.48 at x=1.2, dips slightly to about 0.42 at x=2.5, and then rises to about 0.52 at x=4.</p>
<p>- The graph indicates that Glucose contributes well for linear KNN.</p> <p>- Most influential feature.</p>	 <p>A line graph showing the partial dependence of the model's output on glucose levels. The x-axis is labeled 'Glucose' and ranges from 0 to 2. The y-axis is labeled 'Partial dependence' and ranges from 0.2 to 0.6. The curve starts at approximately 0.25 at x=0 and rises steadily to about 0.55 at x=2.</p>
<p>The graph indicates that SkinThickness does not contribute well for linear KNN.</p> <p>Moderate skin thickness increases the prediction probability, but at higher levels, the effect stabilizes or drops.</p>	 <p>A line graph showing the partial dependence of the model's output on skin thickness. The x-axis is labeled 'SkinThickness' and ranges from -2 to 4. The y-axis is labeled 'Partial dependence' and ranges from 0.2 to 0.6. The curve starts at approximately 0.25 at x=-2, rises to a peak of about 0.45 at x=0.5, dips slightly to about 0.42 at x=1.5, and then remains relatively flat around 0.42 for higher values of skin thickness.</p>

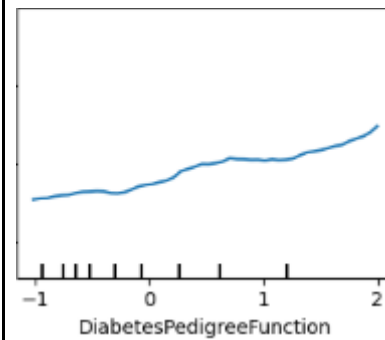
- The graph indicates that BMI **contributes well** for linear KNN.

- Strong positive contributor.

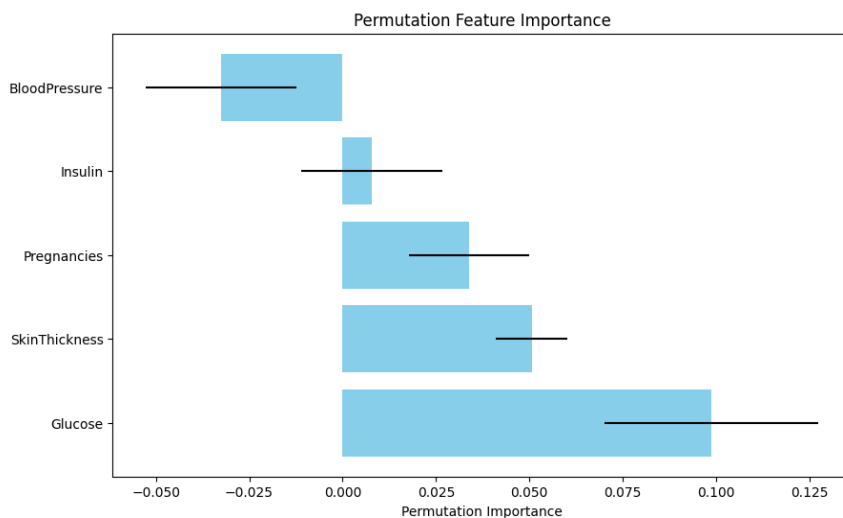


- The graph indicates that DiabetesPedigreeFunction **contributes well** for linear KNN.

- Moderate but consistent positive impact.

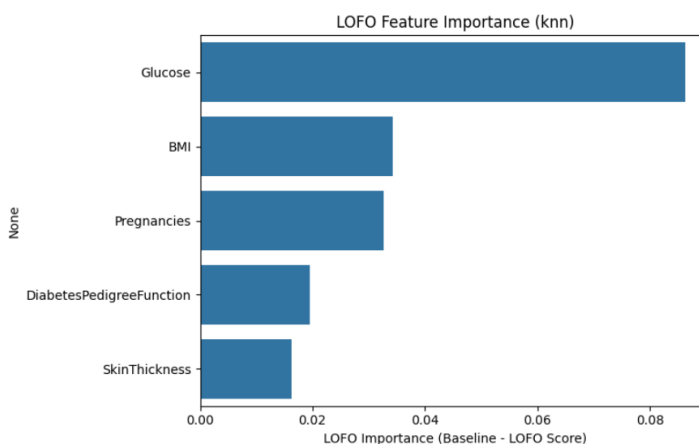


3. Permutation Feature importance



It shows how the model degrades when shuffling happens, which indicates that Glucose has **the highest feature importance** while BloodPressure has **lowest feature importance**.

4. LOFO



This graph shows that Glucose has **the highest feature contribution** while skin thickness has **the lowest feature contribution**.

Conclusion

The KNN model performs well for non-diabetic cases but struggles with detecting diabetic patients, shown by lower metrics and many false negatives. An AUC of 0.73 indicates fair performance. Explainability methods highlight **Glucose** as the most important feature, while **SkinThickness** and **BloodPressure** contribute least.

Three references to previous related work for each team member:

Sama Mohamed:

1. https://www.researchgate.net/profile/Amjed-Almousa/publication/353487060_Diabetes_Detection_Using_Machine_Learning_Classification_Methods/links/62e181e43c0ea8788762247d/Diabetes-Detection-Using-Machine-Learning-Classification-Methods.pdf
2. <https://www.mdpi.com/2075-4426/13/3/406>
3. <https://ieeexplore.ieee.org/abstract/document/9076634>

Bosy Ayman:

4. <http://sciencedirect.com/science/article/pii/S1877050918308548>
5. <https://link.springer.com/article/10.1186/1472-6947-10-16#Sec11>
6. <https://www.sciencedirect.com/science/article/pii/S1877050922021858#cebibl>

Zeyad Sherif:

7. https://www.researchgate.net/profile/Amandeep-Sharma-12/publication/350390088_Prediction_of_Diabetes_Disease_Using_Machine_Learning_Model/links/60b0d396a6fdcc1c66e8e36b/Prediction-of-Diabetes-Disease-Using-Machine-Learning-Model.pdf
8. <https://www.mdpi.com/1660-4601/19/19/12378>
9. <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2018.00515/full>