NORTH SOUTH UNIVERSITY

CSE 440.3

Faculty: Professor Dr. Md. Sazzad Hossain(Mdsh)

PROJECT TOPIC: "DIABETES PREDICTION THROUGH MACHINE LEARNING."

Group Member:

Name	ID
Shikder Motasim Billah	2011342042
Nazia Kamal	2011075042
Moniruzzaman Shawon	1931485042

Diabetes Prediction Using Machine Learning

Abstract:

The medical industry has quickly developed a strong interest in the concept of machine learning. Medical data sets utilized in research predictions and analyses help with optimal treatment and sickness prevention. The many algorithms that can help with machine learning predictions and judgments. We also discuss many machine learning applications in healthcare, with a focus on machine learning-based diabetes prediction. Diabetes is one of the most rapidly spreading illnesses in the world and must be monitored regularly. Diabetes mellitus (DM) is a metabolic condition characterized by elevated blood sugar levels. The two most common clinical forms of diabetes are type 1 and type 2. Currently, a majority among adolescents. Type-1 diabetes patients have considerably grown. We analyze various machine-learning techniques that will aid in the early detection of this disease to confirm this. This work explores several facets of machine learning and the different kinds of algorithms that can aid in prediction and making decisions.

Introduction:

Diabetes, a serious disease that causes global mortality, is caused by obesity and excessive blood sugar levels. It impacts insulin production, leading to insufficient insulin production. If untreated, diabetes can cause heart, kidney, blood pressure, and eye damage. Early diagnosis can help treat the condition. The World Health Organization reports that diabetes affects millions, especially in low- and middle-income countries, and 463 million globally in 2019. According to the survey, diabetes, an extremely dangerous and chronic ailment, causes 1.5 million fatalities per year. This figure might reach $490 billion by 2030. Bangladesh was ranked eighth by the IDF in 2021, with 13.1 million cases of diabetes among adults aged 20 to 79, and is expected to be ranked seventh by 2045. Diabetes, however, is prominent in several nations, including India and China. Early detection of an illness such as diabetes can help to control it and save a person's life. To do this, the study examines how to predict diabetes by using a range of diabetes-related factors. After collecting the relevant data from Kaggle, we apply a range of machine learning ensemble and classification approaches to predict diabetes. Techniques for explicitly guiding computers and other machinery. Various Machine Learning Techniques generate successful outcomes for acquiring information by constructing multiple classifications and ensemble models from the collected data. The existence of diabetes can be predicted using machine learning algorithms. Several machine learning approaches, such as logistic regression, KNN, SVM, and random forest classifiers, can be used to achieve this goal. Choosing the best approach for a certain dataset is difficult. As a consequence, you will experiment with several algorithms to see which one is the most successful. This project is very important for various reasons:

Healthcare Improvement: Early diabetes detection can improve patient care while lowering healthcare expenses. Clinical Decision Support: This model can help healthcare practitioners make educated judgments. Public Health Impact: Public education activities can help to reduce the prevalence of diabetes. Public Health Impact: Public education activities can help to reduce the prevalence of diabetes.

Literature Review:

Sneha et al. [1] analyzed diabetes mellitus for early prediction using optimal feature selection. The focus of the author's work had been on choosing the characteristics that will aid in the early detection of Diabetes Miletus utilizing predictive analysis and machine learning approaches. The UCI machine repository serves as the source of the data. 15 attributes have been used for classification. The classifiers employed are Support Vector Machine, Random Forest, and Naive Bayes, with accuracy rates of 77.73%, 75.39%, and 73.48% respectively.

Rao et [2] tried to utilize it as a perspective to reduce overfitting in the assessment setting, where adaptive boosting is more consistently applied to problems of medium dimensionality. A help block of tests is removed from the responsiveness block, the categorizers on the models used for fixing are executed differently depending on how the guaranteeing tests are executed, and preparation is complete if performance on the help test significantly decreases even as the execution on the arrangement set keeps getting better.

Sunthur et al. [3] used 3 different types of algorithms in this project which are KNN, Logistic Regression, and Random Forest. KNN gave the highest prediction rate (88.5%) than the other two algorithms. Even though the logistic regression algorithm took less time overall than K-NN and the Random Forest algorithm. K-NN gave more credit for this higher prediction rate because it took roughly the same amount of time as the logistic regression algorithm.

Sindhuma2 et [4] employed to predict diabetes using Machine Learning. The author used three different supervised learning algorithms (SVM, Logistic Regression, and ANN). With the use of a database of diabetic patients, the author suggested an intelligent system for predicting diabetes disease using data mining. They suggested applying techniques like Bayesian and KNN to the database of diabetic patients in this system. and evaluating them using several diabetes-related characteristics to forecast the development of diabetic disease.

Anusha et al [5] tried to approach Machine Learning to predict diabetes. The logistic Regression algorithm basic classifier for classification proposed in this study after discussing several classifiers. To eliminate learning from information, machine learning calculations have been incorporated into information mining pipelines that can combine them with tried-and-true quantifiable methods.

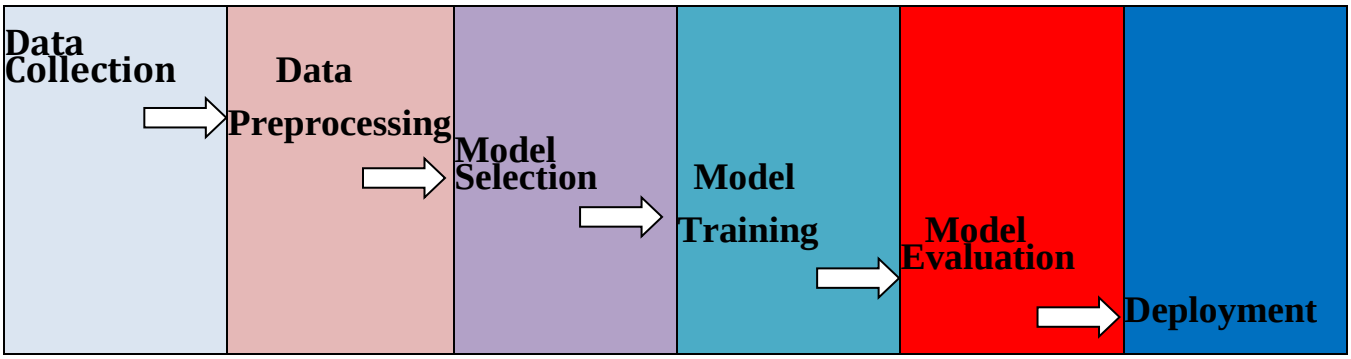
Ramya1 et al [6] The author developed a system that can predict diabetes earlier in patients with a better degree of accuracy. They used the Random Forest algorithm in machine learning. They got an accuracy of 79.44%. The results indicated that the prediction system can forecast diabetes disease effectively, efficiently, and significantly. The suggested model yields the best results for diabetic prediction.

Aiswarya et al. [7] found a way to diagnose diabetes.
They used the PIMA dataset and cross-validation approach. By exploring and examining the patterns that emerge in the data via classification analysis utilizing Decision Tree and Nave Bayes algorithms, The goal of the study is to suggest a quicker and more effective approach to diagnosing the illness that will aid in.

METHODOLOGIES:

System Design:

- **Block Diagram:**



- **Description:**
  - ✓ Data Collection: Collect applicable medical data.
  - ✓ Data Preprocessing: Delete columns or rows if required, clean the data, and make for model training.
  - ✓ Model Selection: Select the best ML algorithms for the given task.
  - ✓ Model Training: Train and optimize the selected models.
  - ✓ Model Evaluation: Evaluate model performance using suitable metrics.
  - ✓ Deployment: Create a user-friendly interface to make predictions.

Required Software:

- **Jupyter Notebook:** Use Jupyter Notebook for data exploration and model development.
- **Libraries:** You'll need libraries like NumPy, Pandas, Matplotlib, Scikit-Learn, TensorFlow, or PyTorch for machine learning.
- **Google collab with Python libraries**: The project will be implemented in Python.

- Anaconda for running environment.
- **Language:** Python

Dataset Description:

Collect large datasets with more features to get more accuracy.

Sample Size: 100000 patients. There are 9 features. No missing values.

Important Features: HbA1c level, Blood pressure level, Glucose level, BMI, age, etc.

Features:

- Age
- Gender
- Body Mass Index (BMI)
- Heart disease
- HbA1c level
- Glucose level
- Hypertension level
- Smoking and alcohol consumption

Source of dataset: Kaggel link:

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?fbclid=IwAR2sb5MyRfea6AzisYGZ2E_SAqv5slOQVk-kPxUMZt67jB8vVOwnt9zXve4

Preprocessing:

- Preprocessing ensures data dependability and quality. This process involves purifying raw data, dealing with missing data, normalizing data, and encoding categorical variables. Furthermore, feature engineering and selection will be employed to enhance model performance.
- **Import Libraries:** It will be used for feature engineering, data manipulation, and visualization.
- **Load Data:** From a file or database into Python data structures like Pandas Data Frames.
- **Investigate the data:** Analyze data to establish distribution, identify outliers, and understand structure.

- **Clean Data**: To accomplish this, the data must be completed by adding any missing values, correcting any errors, and converting it into a format that the machine learning model can understand.
- **Preprocess The Data:** It may involve feature engineering, such as scaling, encoding, and creating new features from existing ones.
- **Handle missing values:** To handle missing values, consider removing them, imputing them using the feature's mean or median, or taking a more complicated method.
- **Transform numerical data into category data:** To use machine learning models, categorical data (e.g. gender or marital status) must be converted to numerical data. One-hot encoding and label encoding are two strategies. That could be utilized to accomplish this.
- **Scale the Data:** A machine learning model may struggle to train if different properties in the data have various sizes. This can be accomplished using a method such as traditional or min-max scaling.
- **Remove the outliers:** Data points that deviate significantly from the rest of the data are considered outliers. It is critical to remove them before training the machine learning model since they have the potential to influence the results. This can be accomplished using an approach known as interquartile range (IQR) outlier identification.

Model:

A. **Logistic Regression:** A supervised machine learning approach used for classification tasks is logistic regression. It is an effective technology that may be used to diagnose medical conditions, identify fraud, and filter spam, among other issues. The logistic regression model is a linear regression model that converts the output of the linear regression function into a probability value using a sigmoid function. A non-linear function called the sigmoid function converts every real integer into a value between 0 and 1. Because of this, it may be used to estimate probabilities and divide data into two groups according to a threshold. The following equation may be used to express the logistic regression model: P (y = 1 | x) = sigmoid (w^T x + b), where:
    - ✓ P (y = 1 | x) is the probability of the target variable y being equal to 1 given the input features x.

- ✓ Sigmoid(x) is the sigmoid function.
- ✓ w^T x is the linear combination of the input features x.

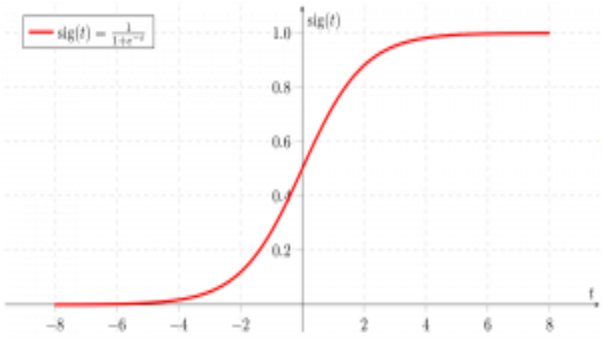
Here are a few instances of machine learning applications for logistic regression:

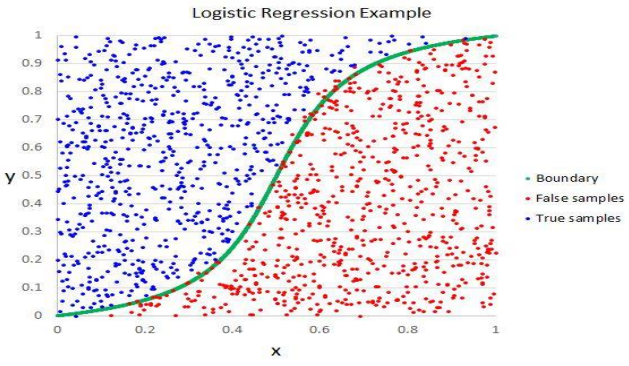
- ✓ **Spam filtering:** Classifying whether an email is spam or not spam by using LR
- ✓ **Fraud detection:** Classifying a fraud detection by using LR

To train a logistic regression model, the steps below are commonly taken:

- ✓ Prepare the data. This includes scaling the data, converting it from a category to a numerical format, and cleaning it.
- ✓ Organize the data into sets for testing and training.
- ✓ Select the loss function.
- ✓ Choose an enhancer.
- ✓ Reset the model's parameters to 0.
- ✓ Reduce the loss function to train the model.
- ✓ Evaluate the model using the testing set.
- ✓ Prediction Applying Logistic Regression

The logistic regression model, a versatile machine learning technique, can predict the probability of a new input variable (x) resulting in a target variable (y) equal to 1, offering a range of classification solutions and being easily understood and applied in most machine learning frameworks.





Logistic regression uses an S-shaped logistic function to fit data, predicting binary outcomes by mapping any real number to a value between 0 and 1.

Here are some of the advantages of using logistic regression:

- It is simple to use and comprehend.
- Because of its processing efficiency, it can handle big datasets with ease.
- We can comprehend how the input features influence the output.
- Prediction since it is interpretable.
- It is resistant to data noise.

Disadvantages:

- It works well on jobs involving binary categorization.
- It may be susceptible to feature selection.
- On datasets where there are non-linear correlations between the features and the output variable, it may perform badly.

B. **Support Vector Machine (SVM):** SVMs are machine learning algorithms used to classify and predict data. They excel at handling binary classification issues, which involve categorizing data into two groups. SVMs function by locating a hyperplane, or dividing line in high-dimensional space, that separates the data points with the greatest possible margin. SVMs are effective at dealing with high-dimensional data and difficult issues, but they can be slow to train on large datasets.

Here are some major points from the paraphrase:

- ✓ SVMs are one form of machine learning algorithm.
- ✓ SVMs can be utilized for classification and prediction applications.
- ✓ SVMs excel at addressing binary classification issues.

- ✓ SVMs use hyperplanes to discover the most significant separation between data points.
- ✓ SVMs are capable of handling difficult challenges and high-dimensional data sets.
- ✓ SVMs may be slow to train on huge datasets.

The advantages of SVM include:

- ✓ Capable of dealing with multi-dimensional data. - Effective with limited training samples. When there are few training samples, they work well.
- ✓ Use kernel functions for non-linear categorization. - Resilient to data noise and outliers.

Drawbacks of SVM include:

- ✓ SVM training on large datasets can be slow.
- ✓ Tuning SVMs can be tricky. The performance of an SVM model can be improved by modifying a range of hyperparameters, but establishing the optimal values for these parameters can be difficult.

SVM applications:

- ✓ Image classification: SVMs are useful for classifying photos, such as dogs, cats, and vehicles.
- ✓ Text classification: Support Vector Machines (SVMs) may be employed to classify text documents into distinct groups, such as spam or not spam, or sentiment (positive or negative).
- ✓ Medical diagnosis: SVMs can be used to identify medical conditions or estimate a patient's chance of contracting one.
- ✓ Fraud detection: SVMs are useful in identifying fraudulent transactions.

In summary,

SVMs are an effective and flexible machine-learning technique that may be used for a wide range of issues. SVMs may be used for regression tasks as well, although they excel at addressing binary classification issues.

C. **The k-nearest neighbors (KNN):** The k-nearest neighbors (KNN) algorithm is a supervised machine learning technique used for regression and classification, predicting a new data point's class or value by finding its k closest neighbors, the most similar data points in the training dataset.

- ✓ Within the training dataset, identify the k most comparable data points.
- ✓ If classifying data is the task at hand, forecast the new data point's class using the majority class of its k closest neighbors.
- ✓ Predict the value of the new data point based on the average value of the k nearest neighbors if the job requires regression.

Advantages of KNN:

- ✓ It is incredibly adaptable due to its non-parametric nature and ease of understanding and implementation.
- ✓ It applies to jobs involving both regression and classification.
- ✓ It can withstand certain noise and anomalies in the data.

Disadvantages:

- ✓ Training on big datasets can be computationally costly.
- ✓ The selection of the k value may have an impact on it.
- ✓ The model might be challenging to comprehend.

KNN Applications:

- ✓ Classification of images.
- ✓ Text categorization.
- ✓ Systems of recommendations.
- ✓ Fraud identification.
- ✓ Medical evaluation.

Here are a few KNN model examples:

- ✓ **Simple KNN:** This is the most basic kind of KNN model; it gauges data point similarity using the Euclidean distance.
- ✓ **Weighted KNN:** In this kind of KNN model, the k closest neighbors are given varying weights based on how far away they are from the new data point.
- ✓ **Kernel KNN:** This kind of KNN model gauges the similarity between data points using a kernel function.

D. Random Forest Classifier:

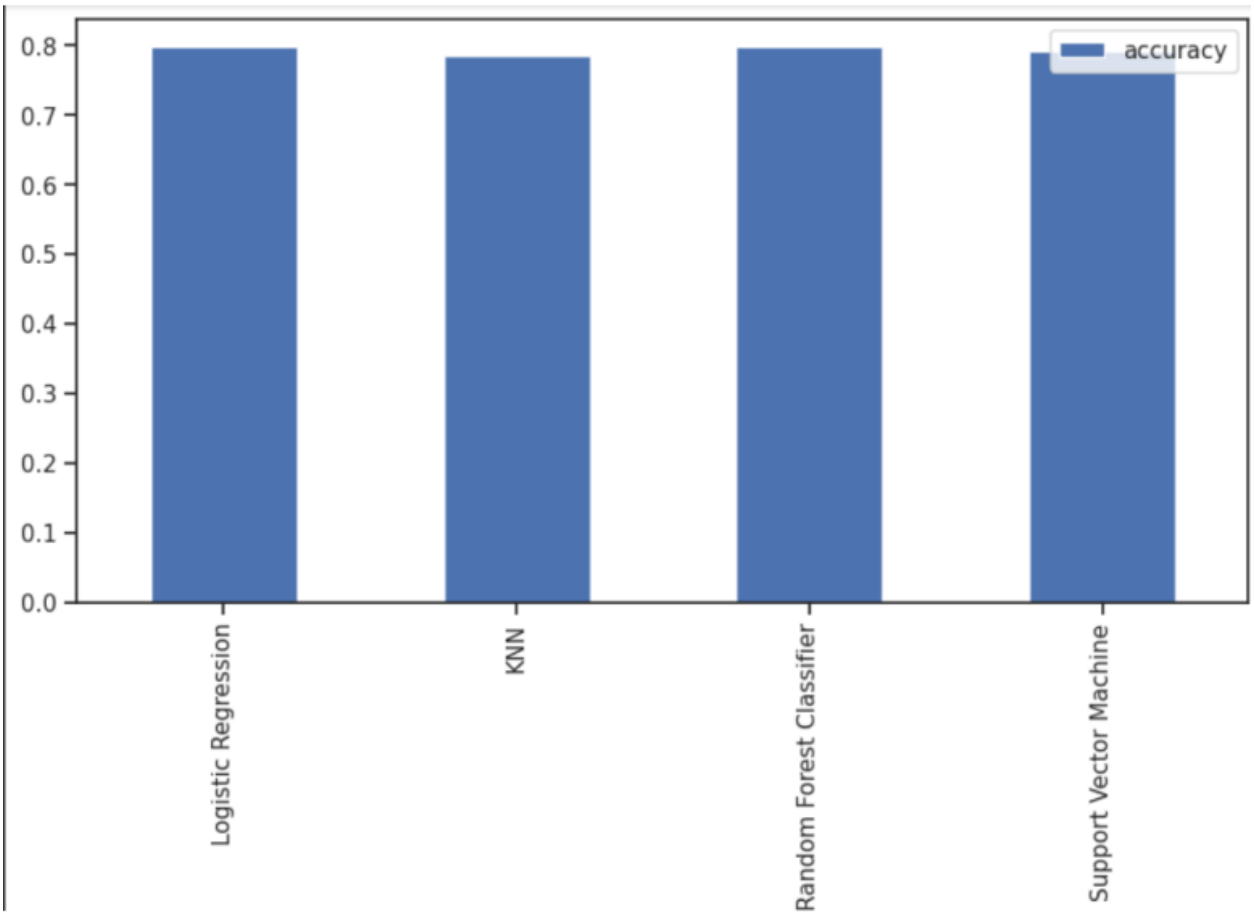
- ✓ A machine-learning model that is a part of the ensemble technique family is the random forest classifier. This indicates that it integrates the forecasts from several models to produce a more precise outcome.
- ✓ Decision trees are used as the individual models in this instance. Using various subsets of the input, the random forest trains a large number of decision trees while adding some unpredictability to their construction. Next, to make a forecast, a new data point is assigned to a class based on a vote by all the trees. The vote with the most number of votes wins!

- ✓ This method aids in mitigating the overfitting issue that arises in machine learning when a model becomes overly dependent on the training set and underperforms when exposed to new data.

Decision and Result Analysis:

From the four models:

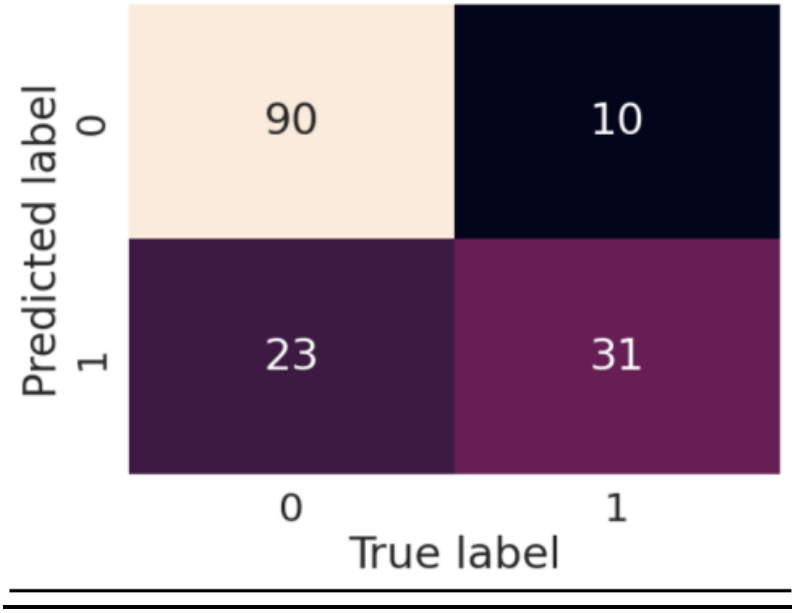
	Logistic Regression	KNN	Random forest	SVM
Accuracy	80%	78.5%	79.9%	79.5%



So, we get the highest accuracy from the Logistic Regression Model. Then find the Random Search CV and Grid Search CV scores:

	Random Search	Grid Search
Score	78.57%	78.57%

Confusion matrix:



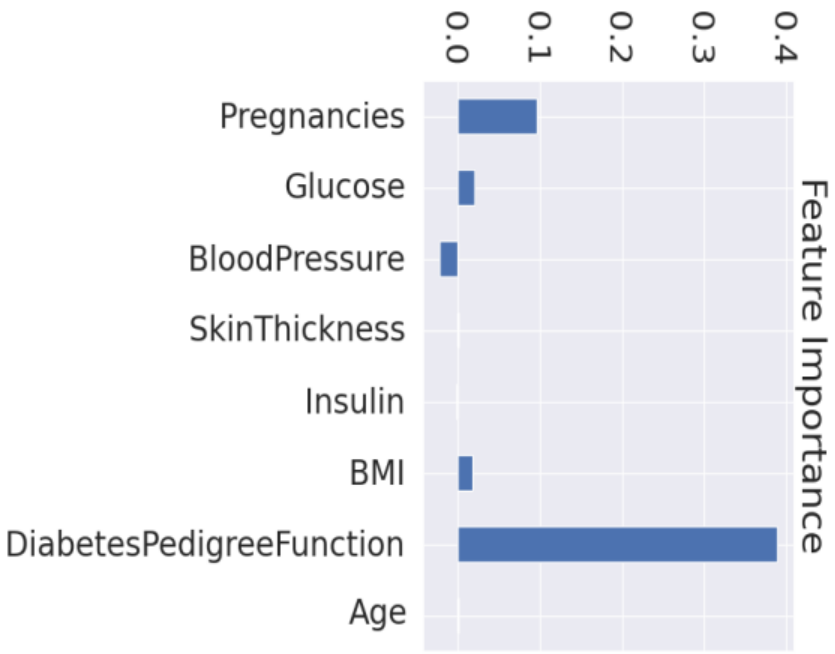
From the matrix the precision, recall, and f-1 Score values:

	precision	recall	f1-score	support
0	0.80	0.90	0.85	100
1	0.76	0.57	0.65	54
accuracy			0.79	154
macro avg	0.78	0.74	0.75	154
weighted avg	0.78	0.79	0.78	154

Cross-validation Score is:

Accuracy	Precision	Recall	f-1 Score
75%	71.5%	48.1%	57.1%

After that, we also found the more important feature from the dataset:



Diabetes pedigree function is the most important feature here. Then, we give some random value to the model and the model gives us the prediction result for Diabetics:

Feature	Values
Pregnancies	3
Blood pressure	64
Skin Thickness	32
Insulin	153
BMI	38
Diabetes Pedigree Function	1.599
Age	48
Glucose	233

We got the result 1. So, this man is a Diabetes patient.

Conclusion:

➢ This project is based on a database related with the help of a machine learning model. We used here four types of models: Logistic Regression, KNN, SVM, and RF Classifier. Finally, we get the maximum result through logistic regression.

➢ We know the dataset is tabular so there is a high chance of error for some values, but we tried our best to get the errorless result by using outlier checking. This model doesn't give results for the image data.

➢ Lastly, for future development, we used another ML model here to check, whether the score is maximum or not. Also, used Image processing by using deep learning here. So that, we'll hopefully get the best result on that time

Reference:

- ✓ Rani, K. J. (2020). Diabetes prediction using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 294–305
- ✓ Deep Learning based System Design for Diabetes Prediction. (2021, October 29). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9645906/
- ✓ Diabetes Prediction using Machine Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/diabetes-prediction-using-machine-learning
- ✓ N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, no. 1, Feb. 2019, doi: 10.1186/s40537-019-0175-6.
- ✓ D. Vidya Sagar Rao, and S. N. Chandra Shekhar, "DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS," Volume: 52, Issue 5, No. 3, May: 2023, ISSN: 0970-2555.
- ✓ Mr SENTHUR R, Mr PAVITHRAN M M, Mr RAGUL P, "Diabetes Prediction using Machine Learning," May 2022
- ✓ Chilupuri Anusha, "A Machine Learning Approach for Prediction of Diabetes Mellitus," Volume 11. No.6, June 2023, ISSN 2347 – 3983.
- ✓ S. Ramya and D. Kalaivani, "Machine Learning Approach for diabetes prediction," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 7, pp. 4444–4448, Jul. 2022, doi: 10.22214/ijraset.2022.46012.
- ✓ D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, Jan. 2018, doi: 10.1016/j.procs.2018.05.122.