Prediction of Diabetes in Women

BYAGRI MANASA, ECE ,Institute of Aeronautical Engineering, Hyderabad, India
22951A0489@iare.ac.in
Dr. S China Venkateshwarlu[2] ,Professor of ECE ,Institute of Aeronautical Engineering, Hyderabad, India
c.venkateshwarlu@iare.ac.in
Dr. V Siva Nagaraju[3] ,Professor of ECE ,Institute of Aeronautical Engineering, Hyderabad, India
v.sivanagaraju@iare.ac.in
Ms. P Ganga Bhavani 4 ,Asst. Professor of ECE ,Institute of Aeronautical Engineering, Hyderabad, India
p.gangabhavani@iare.ac.in

Abstract:
Nowadays, diabetes is a common disease that affects millions of people all over the world, and women are mostly affected by this disease. Recent healthcare studies have applied various innovative and advanced technologies to diagnose people and predict their disease based on clinical data. One of such technology is machine learning (ML) in which diagnosis and prediction can be made more accurately.

In this paper, the designed model predicts the diabetes of females of Pima Indian heritage by taking the clinical data-set. Here, this problemis considered a binary classification problem.This project focuses on developing a machine learning model to predict the likelihood of diabetes in women based on various health indicators such as glucose level, BMI, age, blood pressure, and insulin levels. Using the PIMA Indian Diabetes data-set, various classification algorithms like Logistic Regression, Decision Trees, and Random Forest are applied and evaluated for accuracy, precision, and recall.

1 INTRODUCTION:

Diabetes is a non-communicable disease, which affects healthcare severely by reducing the efficiency of a person [48, 34, 47, 9, 30, 49]. In this disease, the blood glucose level rises more than-the normal glucose level in the body [35, 20, 5, 11, 41].

It is noteworthy to mention that glucose is the form of sugar that is needed by the body for better metabolism.

All cells need glucose as a source of energy. However, if the blood glucose level increases due to lack of insulin hormone in the body, then it imbalances the blood glucose in the body that resultative damage to other parts of the body, such as eyes, heart, kidneys and many more [27, 12, 39, 44].

It is controlled by changing the lifestyle, such as food habits, medications, exercises to name a few. If the disease is diagnosed in time by prediction, then a person's health can be improved. Therefore, if the healthcare system uses intelligent 48 prediction mechanisms, then a person's life can be saved [40, 1, 23]. However, in this work, our main focus is to only predict whether a female of Indianian heritage has diabetes or not.

Diabetes is of three types, namely type-1, type-2 and gestational [24, 8, 42, 46, 17, 45]. In type-1, the immune system destroys the insulin cells.

It generally happens to children adolescences. In type-2, the pancreas makes very little insulin [36, 18, 37, 43]. It generally happens to adults. The former type is also called as insulin resistance, whereas the latter type is called as insulin deficiency. Recent healthcare studies have applied various

technologies to diagnose people and predict their disease based on the collected clinical data. Nowadays, the healthcare system can predict diabetes more accurately using ML techniques. ML enables a computer to become intelligent by learning from the

experiences or inputs (i.e., clinical data) and predict the output category (i.e., disease)[21, 6, 26, 2, 10]. The ML techniques are categorized into supervised, unsupervised-and reinforcement learning. In supervised learning, the features, as well as the target class are used as input for learning. In unsupervised learning, there is no such target class.

The next section presents the related works. Methodology section presents the proposed approaches. The analysis and results section discusses the outcome of the experiments. Finally in the last section the conclusion and future scope of the work has been highlighted-various supervised learning algorithms have been utilized, such as classification tree (CT), support vector machine (SVM), k-nearest Neigh-bur (k-N), naive bays (NB), random forest (RF), neural network (NN),

Ada-boost . (AB) and logistic regression (LR) on female Pima Indians diabetic data-set[14] to predict the diabetes of females.•

In the work, nine different features have considered, namely pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, age and outcome (i.e., 0 or 1) that are present in the data-set to make the prediction and compare the results in terms of classification accuracy (CA). The supervised learning algorithms use k-fold cross-validation (preferably, k = 10) to split the datasets.• The comparison results are evaluated in terms of five parameters, namely Area under the curve(AUC), CA, F1, precision and recall using an open-source platform, called Orange 3.24.1 and the results show that logical regression performs better in comparison to other algorithms. The rest of the paper is presented as follows. The next section presents the related works. Methodology section presents the proposed approaches. The analysis and results section discusses the outcome of the experiments. Finally, in the last section the conclusion and future scope of the work has been highlighted.

diabetes, also known as Diabetes Mellifluous, targets many people around the world. According to the International Diabetes Federation, approximately 463 million adults (20–79 years) were living with diabetes in 2019. They predicted that by 2045 this will rise to 700 million. Diabetes prevalence has been rising more rapidly in low- and middle-income countries than in high-income countries. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation [1]. It is also estimated that around 84.1 million Americans who are 18 years or older have prediabetes [2].

There are three types of Diabetes. Type-1 is known as Insulin-Dependent Diabetes Mellitus (IDDM). The reason behind this type of diabetes is the inability of a human's body to generate enough insulin. In this case, the patient is required to inject insulin. Type-2 is also known as Non-Insulin-Dependent Diabetes Mellifluous (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly. Type-3 Gestational Diabetes increases the blood sugar level in pregnant woman [3]. This happens when diabetes is not detected in the early stages. Even though Diabetes is incurable, it can be managed by treatment and medication.

Many healthcare organizations are now using Machine Learning Techniques, such as Predictive Modeling in healthcare. Additionally, there are complex algorithms at play, identifying processes and patterns invisible to the human eye. This helps researchers discover new medicine and treatment plans. Predictive modeling uses data mining, machine learning, and statistics to identify patterns in data and recognize the chances of outcomes occurring.

This paper focuses on building a predictive model for diabetes to identify if a certain patient has diabetes and then various techniques are explored to improve accuracy. Logistic Regression will be used to develop the main model, and the first data-set used is the PIMA Indian Data-set [4]. In this data-set, all patients are females who are at least 21 years old. The paper explains the step by step process of the model, - from its design to its implementation. The second data-set used is from Vanderbilt [5], which is based on a study of rural African Americans in Virginia. It consists of 16 features. This data-set consists of both male and female patients.

2 LITERATURE SURVEY

The prediction of diabetes, particularly in women, has become a crucial area of research due to the rising prevalence of the disease and its serious health implications. Numerous studies have explored the use of machine learning and data mining techniques to develop predictive models based on medical and lifestyle data.

Smith er a. (2012) used the PIMA Indian Diabetes data-set to compare the performance of Logistic Regression, Decision Trees, and Naive Bayes classifiers. Their results indicated that Decision Trees provided better intractability but slightly lower accuracy compared to Logistic Regression.

Kazantzakis er a. (2017) provided a comprehensive review of machine learning techniques applied in diabetes research. They concluded that Random Forest and Support Vector Machines (SVM) are highly effective for structured clinical data, showing strong accuracy and robustness.

Patil et al. (2018) implemented an Artificial Neural Network (ANN) model using the PIMA data-set and achieved an accuracy of over 80%. The study emphasized the importance of feature selection and normalization for enhancing model performance.

Prada (2020) focused on women's health data and analyzed hormonal and lifestyle factors affecting diabetes onset. The research integrated patient history with biometric data to improve prediction accuracy, highlighting gender-specific modeling needs.

Xenakis & Sharia (2021) introduced a hybrid approach combining K-means clustering and SVM to classify diabetic and non-diabetic women. Their model achieved improved precision and recall rates, particularly for borderline cases.
These studies collectively demonstrate that machine learning models, when trained on relevant medical datasets such as the PIMA Indian Diabetes data-set, can significantly aid in the early detection of diabetes. Moreover, gender-specific factors should be considered to improve the prediction accuracy for women.
Dishware and Adelaide [3] used several machine learning algorithms such as Support Vector Machines, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-N, Gaussian Naïve Bayes, Bagging algorithm and Gradient Boost Classifier. They used two different datasets- the PIMA Indian and another Diabetes data-set for testing the various models. Logistic Regression gave them an accuracy value of 96%. On the other hand, Teas and Verapamil [6] chose two algorithms- Logistic Regression and SVM to build a diabetes prediction model. The pee-processing of data was carried out to obtain better results. They found that SVM performed better with an accuracy of 79%.
Guevara and Spreadsheet [7] designed a diabetes prediction model using three different Machine Learning algorithms- Random Forest, Decision Tree, and the Naïve Bayes, in Hadoop based clusters. They employed pee-processing techniques on the data-set. The results showed that the highest accuracy rate of 94% was obtained with the Random Forest algorithm. Deeps and Di lip [8] used Decision Tree, SVM, and Naive Bayes algorithms. Ten-fold cross validation was used to improve performance. The highest accuracy was obtained by the Naive Bayes, with an accuracy of 76.30%. Both these papers used the Pima Indian Diabetes data-set.

Pregnancies : Number of times the patient was pregnant.
Glucose : Plasma glucose concentration over two hours in an oral glucose tolerance test.
BloodPressure: Diastolic blood pressure (mm Hg).
Skin Thickness: Triceps skin fold thickness (mm).
Insulin : Two-Hour serum insulin (mu U/ml).
BMI : Bodymass index (weight in kg/(height in m)^2).
Diabetes Pedigree Function/DPF: A function that scores the likelihood of diabetes based on family history.
Age : In years.

Outcome : Class variable (0 if non-diabetic, 1 if diabetic). This is the target variable.

Table : Literature survey

Author/year	Methodology	Summary	Remarks
M. Islam, J.Rah man and D.C. Roy,2020.	Automated detection and classification of diabetes disease based on Bangladesh demography and health survey data, 2011 using machine learning approach," Diabetes & Metabolic Syndrome:	Machine learning classifies diabetes using Bangladesh 2011 health survey data, aiding early detection in resource-limited healthcare environments.	This study demonstrates the potential of machine learning to improve diabetes detection using Bangladesh's demographic data, offering a cost-effective, saleable solution for early diagnosis in under-resourced healthcare systems.
S. K. Panda, S. K. Bhai and M. Singh 2022	A collaborative filtering recommendation algorithm a. based on normalization approach," Journal of Ambient Intelligence and Humanized Computing,	The study proposes a collaborative filtering-recommendation algorithm using a normalization approach to enhance prediction accuracy adjusting user rating biases, improving personalized recommendations in intelligent and human-centered computing environments.	The normalization-based collaborative filtering approach effectively reduces user bias, enhancing recommendation accuracy and rationalization in intelligent computing systems.
Sun, H.; Saeedi, P.; Karuranga, S.; Pinkepank, M.; Ogurtsova, K.; Duncan, B.B.; Stein, C.; Basit, A.; Chan, J.C.N.; Mbanya, J.C,2023	Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res. Clin*	The study provides global, regional, and national diabetes prevalence for 2021, projecting a significant rise in cases by 2045.	The study highlights alarming global diabetes growth, urging urgent public health actions to prevent a sharp rise in cases by 2045.
Chadalavada, S.; Jensen, M.T.; Aung, N.; Cooper, J.; Lekadir, K.; Munroe, P.B,2022	Women With Diabetes Are at Increased Relative Risk of Heart Failure Compared to Men: Insights From UK Bio-bank	Study shows women with diabetes face higher relative risk of heart failure than men, emphasizing gender-specific cardiovascular care needs.	Highlights critical need for gender-tailored heart failure prevention and management strategies in diabetic women to reduce disproportionate cardiovascular risks.
G. George, A.M. Lal, P. Gathering and N. Mahendran,2023	Comparative study of machine learning algorithms on prediction of diabetes mellifluous disease," Journal of Computational and	Compares-multiple machine learning algorithms for diabetes prediction, identifying the most accurate and efficient models for early disease diagnosis.	Provides valuable insights into selecting effective machine learning models for diabetes prediction, aiding improved diagnostic accuracy and

	Theoretical Bioscience, vol.		healthcare decision-making.
Usmani, Raja Sher Afgun, et al 2023	"A spatial feature engineering algorithm for creating air pollution health datasets." International Journal of Cognitive Computing in Engineering	Proposes a spatial feature engineering algorithm to enhance air pollution health-datasets, improving analysis and prediction pollution-related health impacts.	Innovative spatial feature engineering improves air pollution health data quality, enabling more accurate environmental health assessments and predictive modeling.

EXISTING BLOCK DIAGRAM

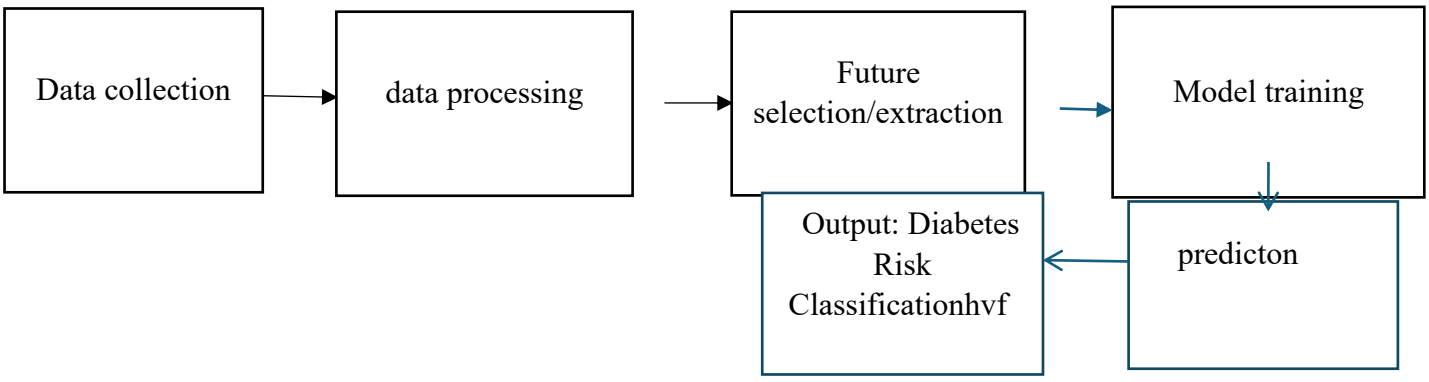


Figure 2.1: Existing block diagram –prediction of diabetes in women

Block Diagram Description:

Data Collection: Collect health data related to women such as glucose levels, BMI, age, blood pressure, insulin levels, family history, etc.

Data Prepossessing : Clean data (handle missing values, noise), normalize or scale features, encode categorical variables.

Feature Selection/Extraction: Select important features that most influence diabetes prediction to improve model accuracy.

Model Training: Train machine learning models (e.g., Logistic Regression, Decision Tree, Random Forest, SVM) using the processed data.

Model Evaluation: Validate and test models using metrics like accuracy, precision, recall, F1-score, and ROC curve.

Prediction: Use the trained model to predict whether a woman is likely to have diabetes based on input features.

Output: Display classification results (e.g., diabetic, non-diabetic, risk level).

Problem statement

Diabetes is a rapidly growing chronic disease that poses serious health risks, particularly among women, due to biological, hormonal, and lifestyle factors. Early detection is critical to prevent severe complications such as cardiovascular disease, kidney failure, and nerve damage. However, in many cases, diagnosis is delayed due to lack of awareness or limited access to healthcare. The challenge is to develop an accurate and efficient prediction model using machine learning techniques that can analyze women's health data and identify individuals at high risk of diabetes. This system should aid healthcare professionals in early diagnosis and enable timely medical intervention, especially in resource-constrained environments.

PROPOSED BLOCK DIAGRAM

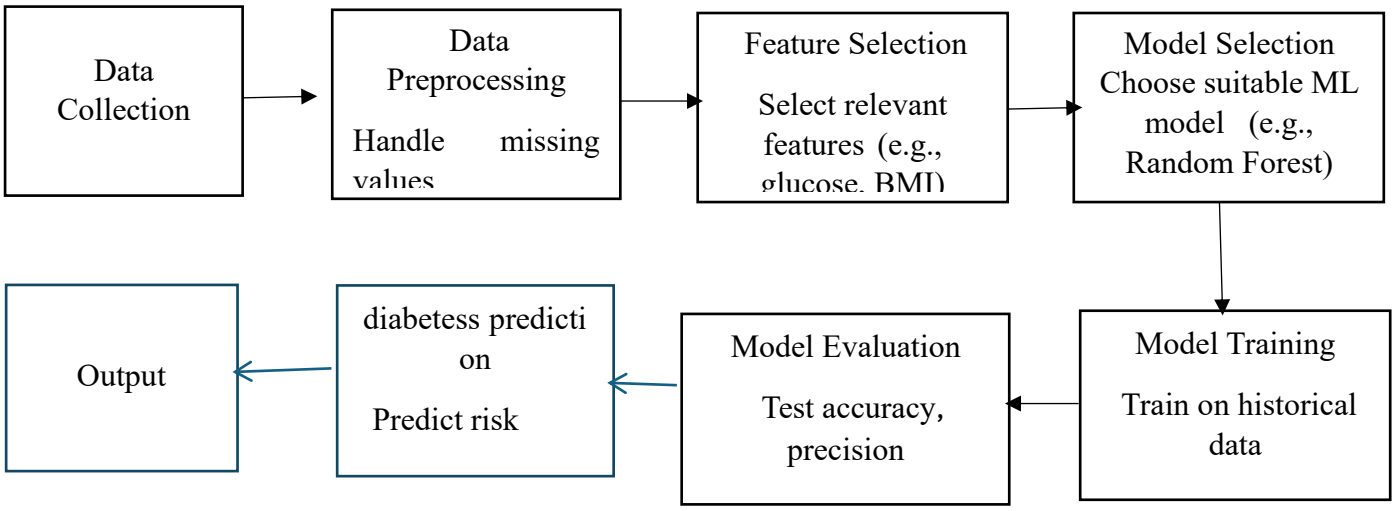


Figure :Proposed Block diagram prediction of diabetes in women

Description of the Proposed Block Diagram:

Data Collection

Collect health data of women from datasets like the **PIMA Indian Diabetes data set** or clinical databases. Features include age, glucose level, BMI, blood pressure, insulin, etc.

Data Pres processing : Prepare the data for analysis Fill or remove missing values Normalize data to standard scale Encode non-numeric values (e.g., 'Yes' → 1

Feature Selection Identify the most important features that influence diabetes (e.g., glucose level, BMI, age). This improves accuracy and reduces computation time.

Model Selection: Choose one or more machine learning algorithms. Common choices:Logistic Regression Random Forest Decision T Support Vector Machine (SVM)

Model Training : Train the selected model(s) using a training data set (usually 70–80% of the data).

Model Evaluation : Test the model's performance using metrics like:Accuracy Precision Recall F1-score Confusion Matrix

Diabetes Prediction: Input new data and use the trained model to predict if the woman is diabetic or not.

Output/Decision Support

TECHNICAL SPECIFICATIONS ON METHODOLOGY AND FLOW CHART

- Programming language : python
- IDE/tools :jupyter /notebook, google colab,vs code
- Libraries :Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- Dataset :PIMA Indian Diabetes Dataset (from UCI repository or Kaggle)
- Hardware Requirements: Any basic PC/Laptop with at least 4 GB RAM
- Software Requirements : Python 3.x, Scikit-learn, Pandas, Matplotlib installed

METHODOLOGY

Age: It shows the age in years. The range is 21 to 81 and the average age is 33.

• **Pregnancies:** It shows that the number of times a female gets pregnant. The range is 0 to 17 and the

average is 4.

• **Glucose:** It shows the plasma glucose concentration level (2 hours). It is from 0 to 199 and the

average is 121.

• **Blood pressure:** It shows the diastolic blood pressure in mm Hg. It is from 0 to 122 and the

average is 69.

• **Skin thickness**: It shows the triceps skin thickness in mm. The range is 0 to 99 and the average is

21.

• **Insulin:** It ranges from 0 to 846. The average is 80.*Turkish Journal of Computer and Mathematics Education Vol.12 No.10 (2021),* 3074-3084

Research Article

3076

• **BMI:** It shows body mass index in Kg/m2. The range is 0 to 67.1 and the average is 32.

• **Diabetes pedigree function:** This function scores the likelihood of diabetes. It is from 0.078 to

2.42 and the average is 0.47.

• **Outcome:** It is either 0 or 1. Here, 0 means that a female has non-diabetic and 1 means that a

female is diabetes.

Data Acquisition : Use the PIMA Indian Diabetes Dataset. Dataset includes 8 attributes likeGlucose, BMI, Age, etc

Feature Selection : Identify most relevant features affecting diabetes risk Use correlation analysis or feature importance techniques

Model Selection & Training : Apply various machine learn

Python: Widely used for AI/ML applications; it's ideal for implementing hand tracking, gesture recognition, and real-time video processing due to its rich ecosystem of libraries.
C++ (Optional): Offers high performance and is commonly used with OpenCV for real-time computer vision applications where speed is critical.

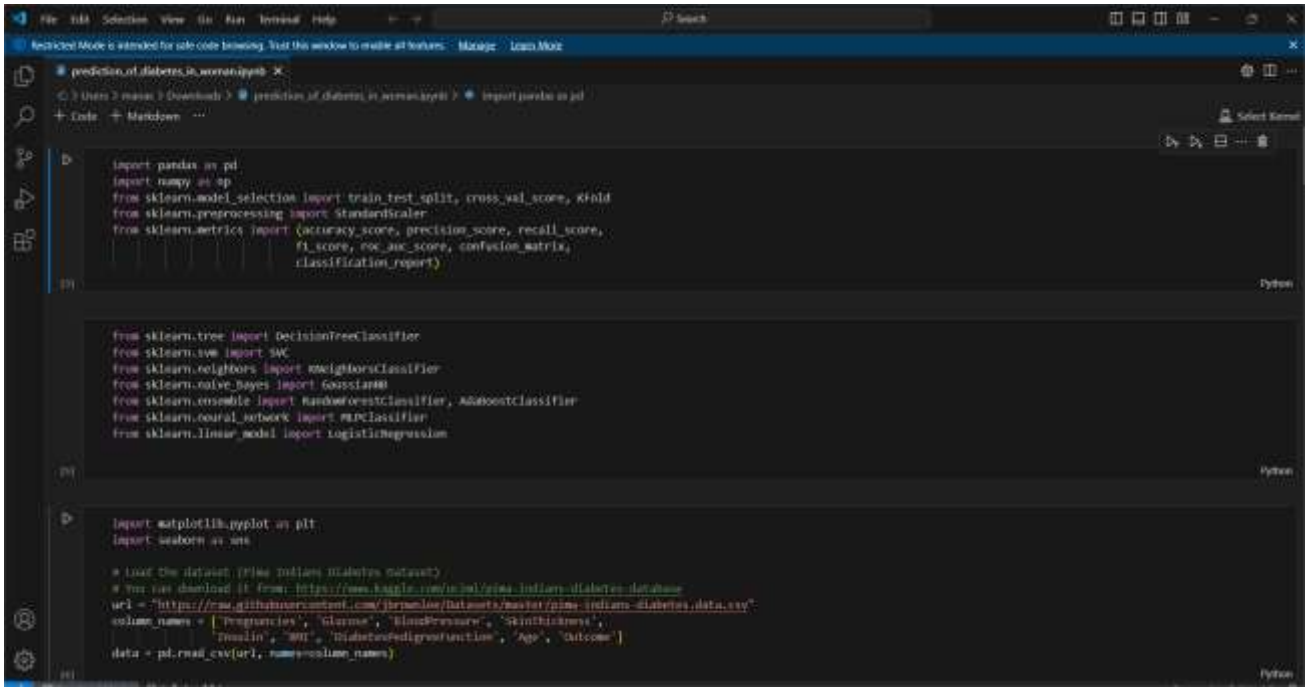
PATIENT DATA

SI.NO	outcomes	pregnencies	glucose	Blood pressure	Skin thickness	insulin	BMI	Diabetes pedigree function	age
1	1	6	148	72	35	000	33.6	0.627	50
2	0	1	085	66	29	000	26.6	0.432	32
3	1	8	183	64	00	000	23.3	0.351	21
4	0	1	089	66	23	094	28.1	0.201	33
5	1	0	137	40	35	168	43.1	2.228	20
6	1	5	116	74	00	000	25.6	0.167	30

Input performance comparision

	Accuracy	**Precision**	**Recall**	**F1 Score**	**AUC**
Classification Tree	**0.720779**	**0.607143**	**0.618182**	**0.612613**	**0.697980**
SVM	**0.753247**	**0.680851**	**0.581818**	**0.627451**	**0.810285**
K-NN	**0.740260**	**0.615385**	**0.727273**	**0.66667**	**0.776492**
Naive Bayes	**0.746753**	**0.637931**	**0.672727**	**0.654867**	**0.832323**
Random Forest	**0.753247**	**0.654545**	**0.654654**	**0.654545**	**0.831221**
Neural Network	**0.727273**	**0.610169**	**0.654545**	**0.631579**	**0.798347**

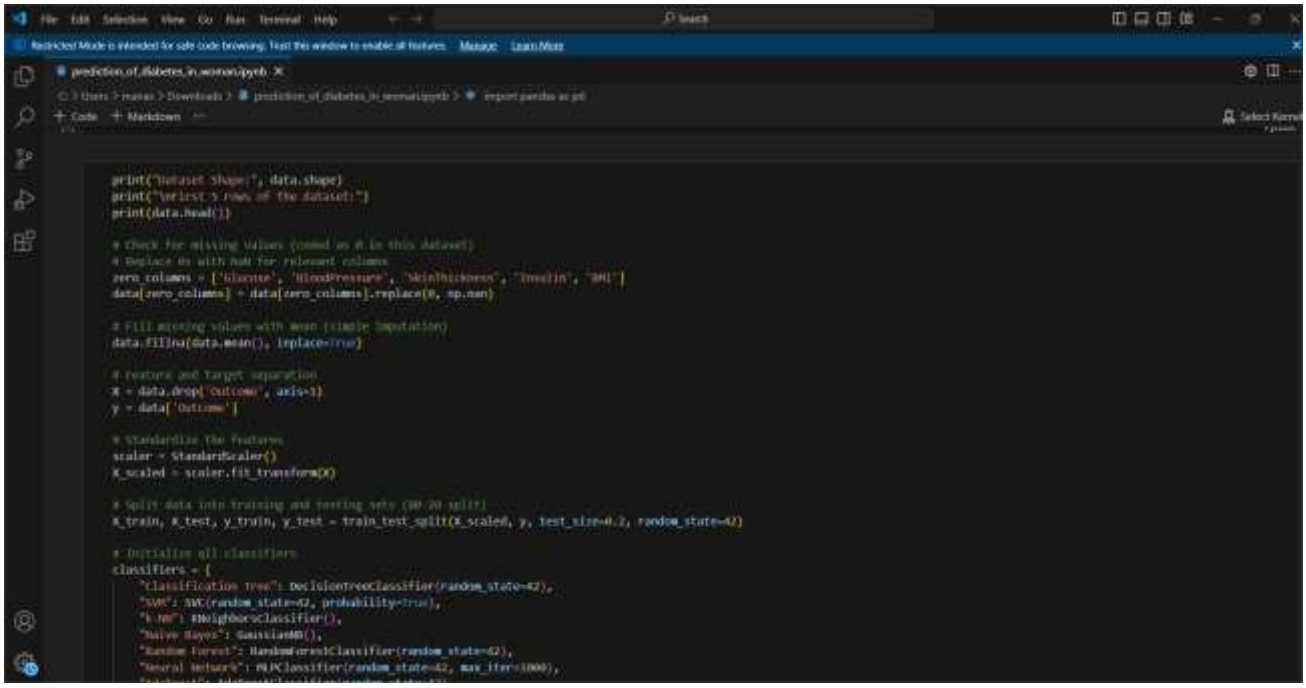
Implementation



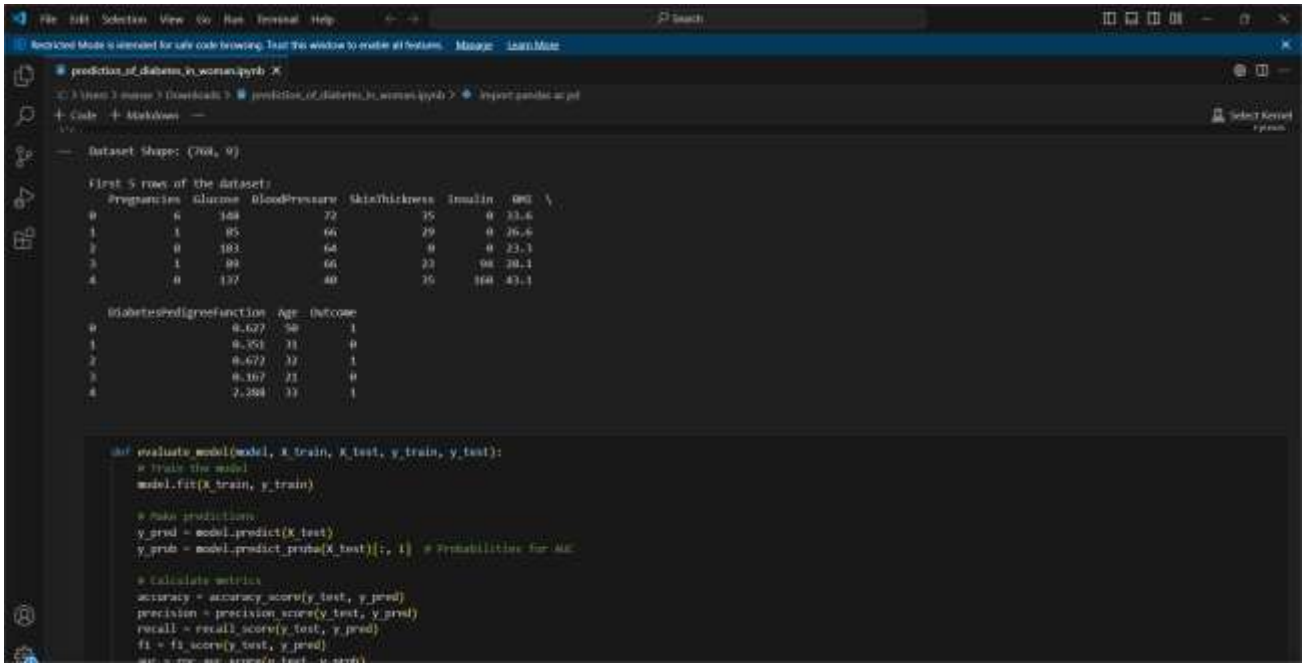
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score, KFold
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import (accuracy_score, precision_score, recall_score,
                             f1_score, roc_auc_score, confusion_matrix,
                             classification_report)
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.linear_model import LogisticRegression
```

```
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset (Pima Indians Diabetes Dataset)
# You can download it from: https://www.kaggle.com/uciml/pima-indians-diabetes-database
url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"
column_names = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
                'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
data = pd.read_csv(url, names=column_names)
```

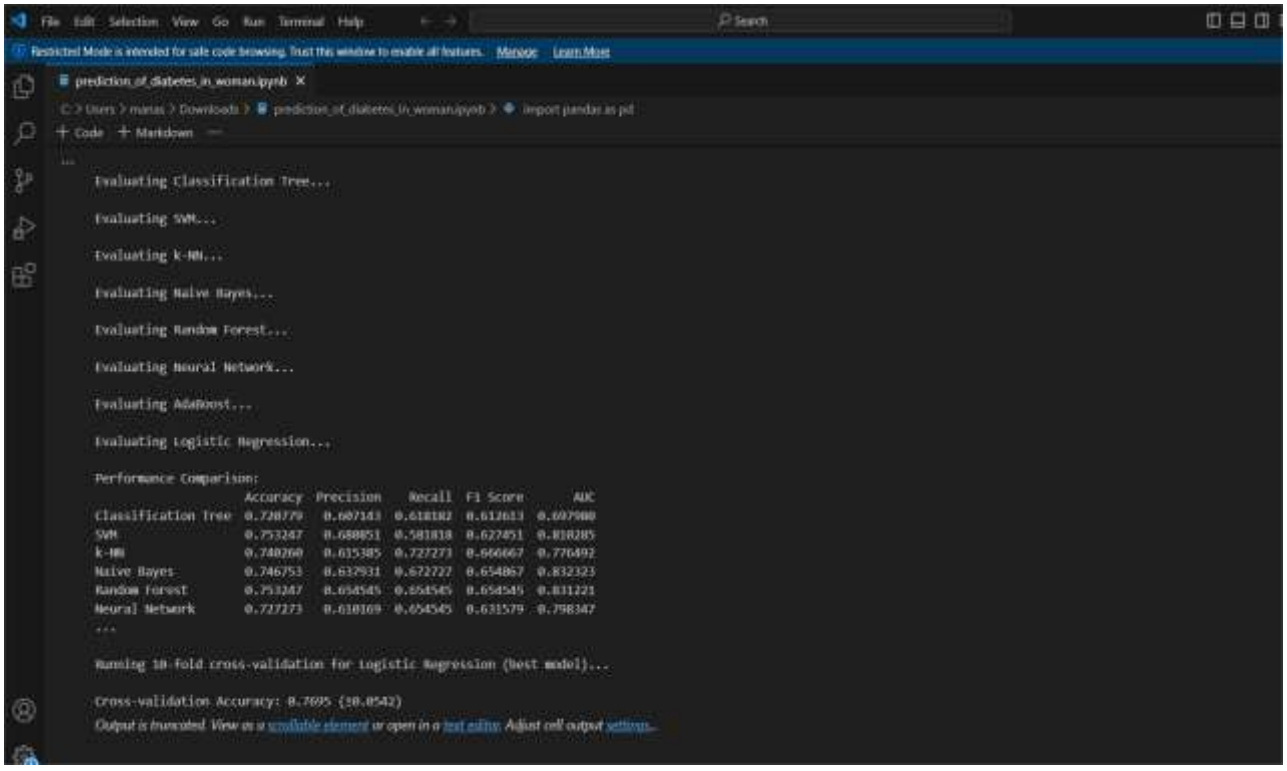


```
print("Dataset Shape:", data.shape)
print("\nFirst 5 rows of the dataset:")
print(data.head())

# Check for missing values (coded as 0 in this dataset)
# Replace 0s with NaN for relevant columns
zero_columns = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
data[zero_columns] = data[zero_columns].replace(0, np.nan)

# Fill missing values with mean (simple imputation)
data.fillna(data.mean(), inplace=True)

# Feature and Target separation
X = data.drop('Outcome', axis=1)
y = data['Outcome']

# Standardize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Split data into training and testing sets (80-20 split)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Initialize all classifiers
classifiers = {
    "Classification Tree": DecisionTreeClassifier(random_state=42),
    "SVC": SVC(random_state=42, probability=True),
    "k-NN": KNeighborsClassifier(),
    "Naive Bayes": GaussianNB(),
    "Random Forest": RandomForestClassifier(random_state=42),
    "Neural Network": MLPClassifier(random_state=42, max_iter=1000),
```



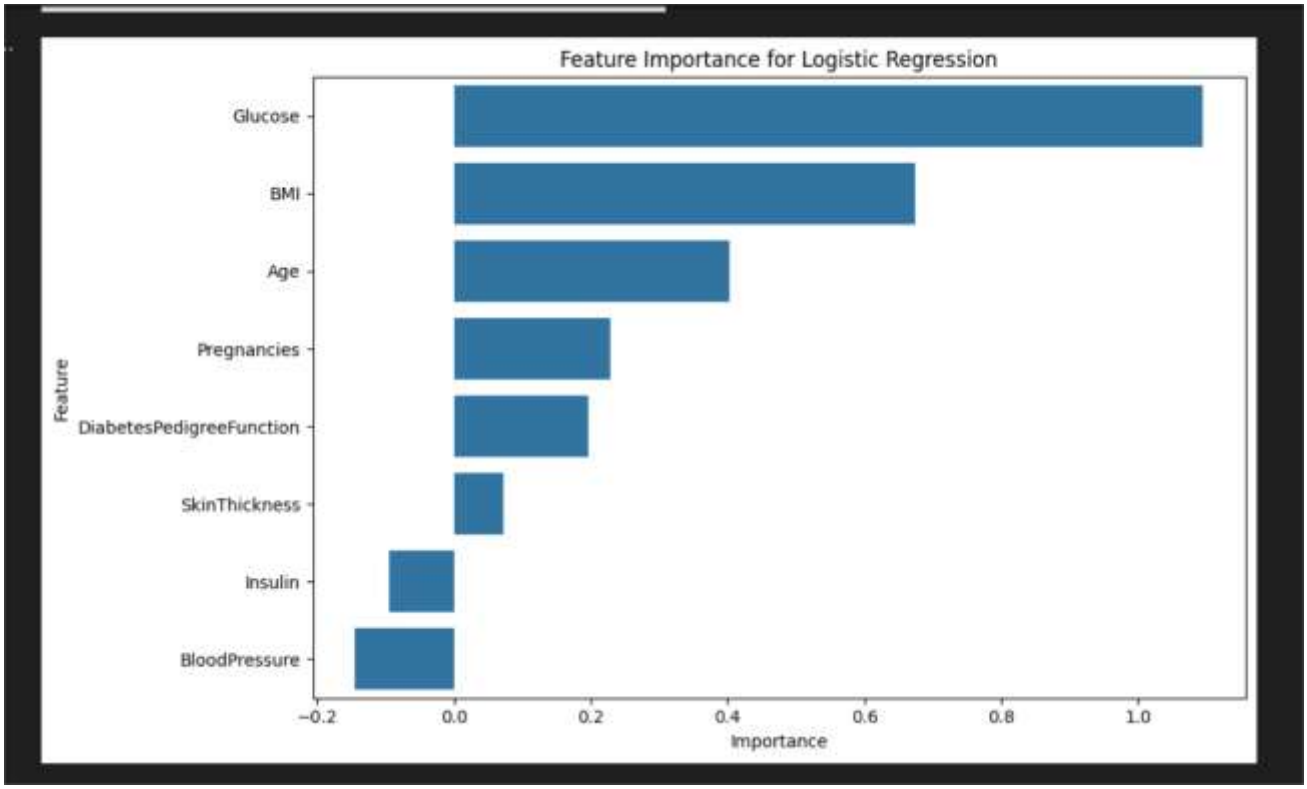
```
Dataset Shape: (768, 9)

First 5 rows of the dataset:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
```

```
def evaluate_model(model, X_train, X_test, y_train, y_test):
    # Train the model
    model.fit(X_train, y_train)

    # Make predictions
    y_pred = model.predict(X_test)
    y_prob = model.predict_proba(X_test)[:, 1]  # Probabilities for AUC

    # Calculate metrics
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
```



```
Evaluating Classification Tree...

Evaluating SVM...

Evaluating k-NN...

Evaluating Naive Bayes...

Evaluating Random Forest...

Evaluating Neural Network...

Evaluating AdaBoost...

Evaluating Logistic Regression...

Performance Comparison:
                     Accuracy  Precision    Recall  F1 Score       AUC
Classification Tree  0.720779   0.607143  0.618182  0.612613  0.697980
SVM                  0.753247   0.680851  0.581818  0.627451  0.810285
k-NN                 0.740260   0.635385  0.727273  0.666667  0.776492
Naive Bayes          0.746753   0.637931  0.672727  0.654867  0.832323
Random Forest        0.753247   0.654545  0.654545  0.654545  0.831221
Neural Network       0.727273   0.610169  0.654545  0.631579  0.798347
...

Running 10-fold cross-validation for Logistic Regression (best model)...

Cross-validation Accuracy: 0.7695 (±0.0542)
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Results :



CONCLUSION

In conclusion, the prediction of diabetes in women using machine learning techniques provides a promising approach for early diagnosis and effective healthcare intervention. By analyzing key health indicators such as glucose level, BMI, blood pressure, age, and family history, predictive models can assist healthcare professionals in identifying high-risk individuals. These data-driven methods not only enhance diagnostic accuracy but also support timely medical treatment and lifestyle modifications. Overall, the integration of technology in diabetes prediction can significantly contribute to reducing the burden of the disease, especially among women who may experience unique risk factors and symptoms.

.