INTERNSHIP PROJECT REPORT

DIABETES PREDICTION MODEL

Submitted by

Sanglap Karmakar

USN – 1DB20CS094

Department of Computer Science

Don Bosco Institute of Technology, Bangalore

E-mail: sanglapkarmakar2001@gmail.com

Supervised by

Mr. Aravind Kumar

R & D Engineer

EXPOSYS DATA LABS

E-mail: exposysdatalabs@gmail.com

Date of Submission: 09.09.2023

BANGALORE, KARNATAKA

Acknowledgement

I would like to take this opportunity to express my heartfelt gratitude to all those who contributed to the successful completion of this diabetes prediction model data science project. I have completed this work under the mentorship of Mr. Aravind Kumar, R & D Engineer. I am doing an online internship on Data Science where I have learnt the various Machine Learning Algorithms from my mentor as Course Instructors.

My heartfelt thanks go out to my friends and family for their unwavering encouragement and understanding throughout this project's duration. Your support provided the motivation and strength needed to overcome challenges and remain dedicated to achieving my objectives.

To everyone who played a role in this data science project, your collective efforts have led to the development of a valuable diabetes prediction model. Thank you all for your contributions and support.

Abstract

Diabetes Prediction:

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. This project aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, KNN. This project also aims to propose an effective technique for earlier detection of the diabetes disease using Machine learning algorithms and end to end deployment using flask.

Causes of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes on chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackie virus, mumps, hepatitis B virus, and cytomegalovirus increases the risk of developing diabetes.

Types of Diabetes

Type 1

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90%of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Data mining and machine learning have been developing, reliable, and supporting tools in the medical domain in recent years. The data mining method is used to pre-process and select the relevant features from the healthcare data, and the machine learning method helps automate diabetes prediction. Data mining and machine learning algorithms can help identify the hidden pattern of data using the cutting-edge method; hence, a reliable accuracy decision is possible. Data Mining is a process where several techniques are involved, including machine learning, statistics, and database system to discover a pattern from the massive amount of dataset.

Index

Introduction

Diabetes, a pervasive metabolic disorder, has emerged as a global health challenge affecting millions of lives across the world. The intricate interplay of genetics, lifestyle factors, and environmental influences makes diabetes a complex and multifaceted disease. It not only threatens the well-being of individuals but also places a significant burden on healthcare systems and resources. Early detection and timely intervention are crucial in managing diabetes effectively, preventing complications, and ultimately saving lives.

In recent years, the field of data science has witnessed remarkable advancements, fueled by the exponential growth of data and the development of sophisticated machine learning and artificial intelligence techniques. These advancements offer a promising avenue for addressing the challenges posed by diabetes. By harnessing the power of data science, we can develop predictive models that have the potential to revolutionize the way we approach diabetes diagnosis and management.

This project endeavours to explore the application of data science, specifically machine learning, in the context of diabetes prediction. By leveraging large datasets and cutting-edge algorithms, we aim to create a robust and accurate predictive model capable of assessing an individual's risk of developing diabetes. Such a model has the potential to not only streamline the diagnostic process but also empower healthcare professionals with timely and actionable insights, enabling them to provide personalized care and interventions.

In this journey, we will delve into the intricacies of diabetes, examine the dataset at our disposal, explore feature engineering techniques, and train and evaluate various machine learning models. Our ultimate goal is to develop a predictive model that not only achieves high accuracy but is also interpretable and can be seamlessly integrated into clinical practice. Through this data-driven approach, we hope to contribute to the early detection and effective management of diabetes, thereby improving the quality of life for individuals at risk and reducing the societal burden of this pervasive disease.

Existing Method

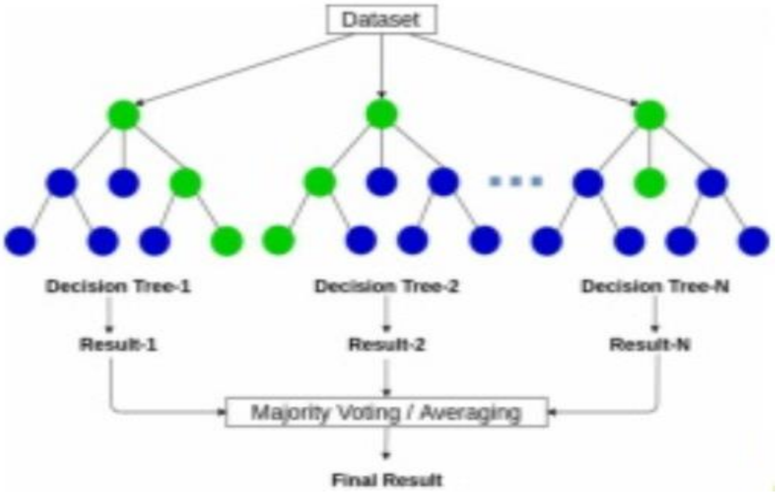
Gradient boosting

Gradient boosting is a powerful ensemble machine learning algorithm. It’s popular for structured predictive modelling problems, such as classification and regression on tabular data, and is often the main algorithm or one of the main algorithms used in winning solutions to machine learning competitions, like those on Kaggle. There are many implementations of gradient boosting available, including standard implementations in SciPy and efficient third-party libraries. Each uses a different interface and even different names for the algorithm.

Random Forest Classifier

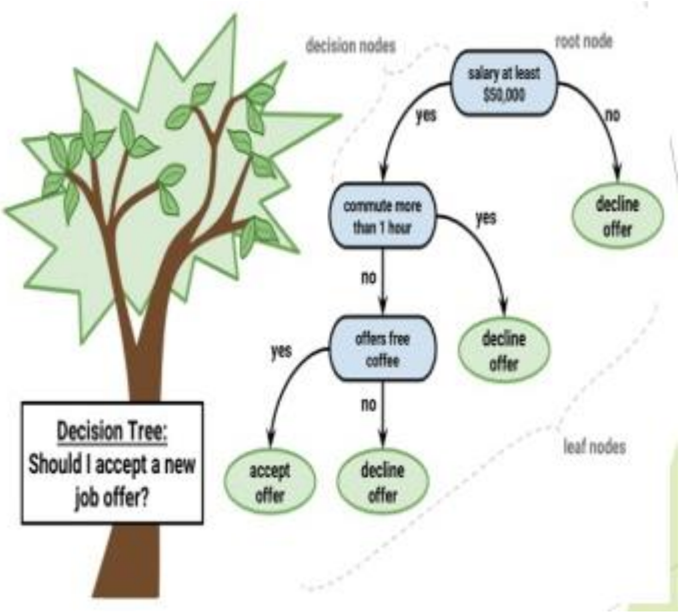
The Random Forest Classifier Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is one of the widely used algorithms, which perform well with any kind of dataset, be it classification or regression.

It learning, which is a process of combining multiple classifiers to solve a complex problem, and at the end, the results are either made an average of all the classifiers or mode of all the classifiers. is based on the concept of ensemble.



Decision Tree Decision tree

Decision Tree Decision tree, as the name suggests, creates a branch of nodes. Where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and the last nodes are termed as the leaf nodes. Leaf node means there cannot be any nodes attached to them, and each leaf node (terminal node) holds a class label. The decision tree is one of the most popular algorithms in machine learning, it can be sued for both classification and regression.

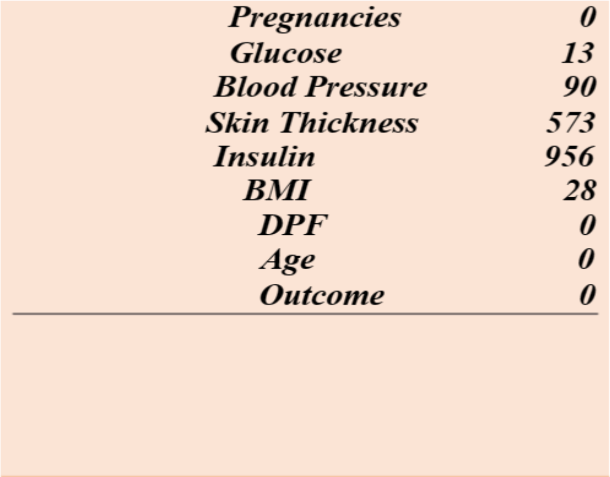


Proposed Methods

I] Dataset collection – It includes data collection and understanding the data to study the hidden patterns and trends which helps to predict and evaluating the results. Dataset carries1405 rows i.e., total number of data and 10 columns i.e., total number of features. Features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age.

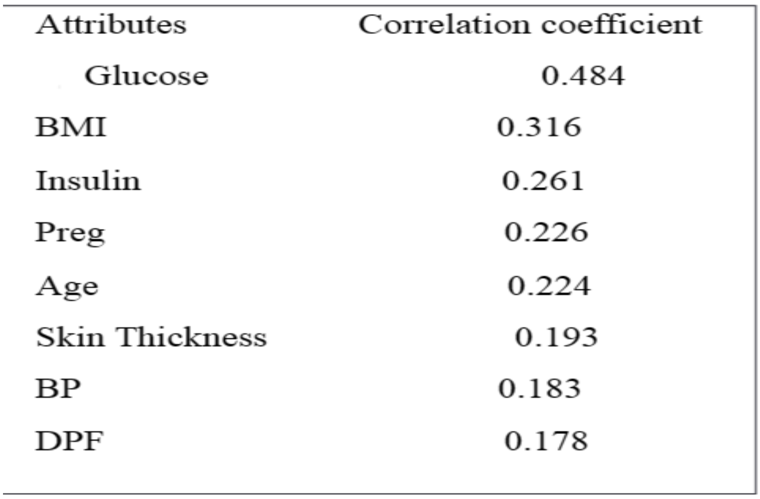
II) Data Pre-processing: This phase of model handles inconsistent data in order to get more accurate and precise results like in this dataset Id is inconsistent so we dropped the feature. This dataset doesn’t contain missing values. So, we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then data was scaled using Standard Scaler. Since there were a smaller number of features and important for prediction so no feature selection was done.

III]Missing value identification: Using the Panda library and scikit-learn, we got the missing values in the datasets, shown in Table 2. We replaced the missing value with the corresponding mean value.



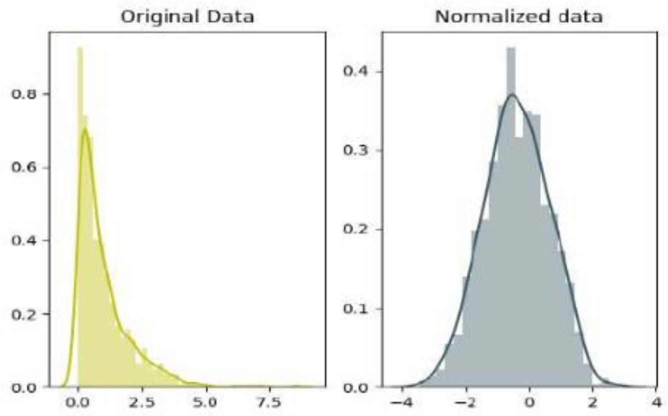
Pregnancies	*0*
Glucose	*13*
Blood Pressure	*90*
Skin Thickness	*573*
Insulin	*956*
BMI	*28*
DPF	*0*
Age	*0*
Outcome	*0*

IV] Feature selection: Pearson’s correlation method is a popular method to find the most relevant attributes/features. The correlation coefficient is calculated in this method, which correlates with the output and input attributes. The coefficient value remains in the range between −1 and 1. The value above 0.5 and below −0.5 indicates a notable correlation, and the zero value means no correlation.



Attributes	Correlation coefficient
Glucose	0.484
BMI	0.316
Insulin	0.261
Preg	0.226
Age	0.224
Skin Thickness	0.193
BP	0.183
DPF	0.178

V] Scaling and Normalization: We performed feature scaling by normalizing the data from 0 to 1 range, which boosted the algorithm's calculation speed. scaling means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points are, like support vector machines (SVM) or k-nearest neighbours (KNN).With these algorithms, a change of "1" in any numeric feature is given the same importance.



VI] Splitting of data: After data cleaning and pre-processing, the dataset becomes ready to train and test. In the train/split method, we split the dataset randomly into the training and testing set. For Training we took 1600 sample and for testing we took 400 sample.

VII] Design and implementation of classification model: In this research work, comprehensive studies are done by applying different ML classification techniques like DT, KNN, RF, NB, LR, SVM.

VIII] Machine learning classifier: We have developed a model using Machine learning Technique. Used different classifier and ensemble techniques to predict diabetes dataset. We have applied SVM, LR, DT and RF Machine learning classifier to analyse the performance by finding accuracy of each classifier All the classifiers are implemented using scikit-learn libraries in python. The implemented classification algorithms are described in next section.

Methodology

Random forest algorithms have three main hyper-parameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

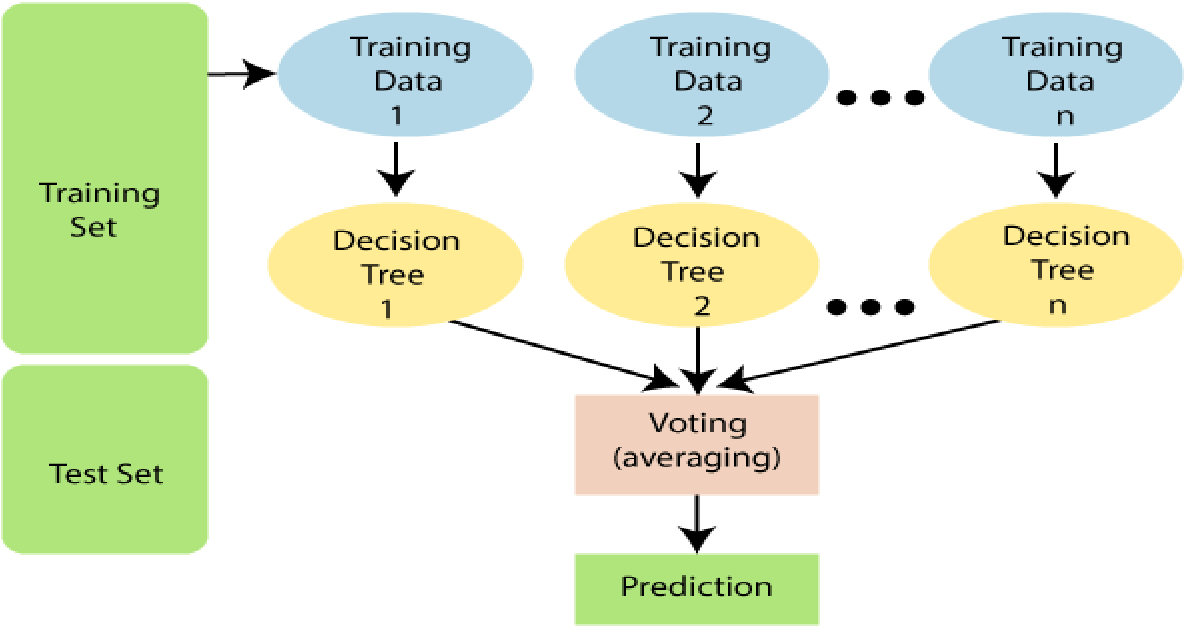
The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we'll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority vote—i.e. the most frequent categorical variable—will yield the predicted class. Finally, the oob sample is then used for cross-validation, finalizing that prediction.

The basic Random Forest procedure may not work well in situations where there are a large number of features but only a small proportion of these features are informative with respect to sample classification. This can be addressed by encouraging the procedure to focus mainly on features and trees that are informative. Some methods for accomplishing this are:

- Pre-filtering: Eliminate features that are mostly just noise.

- Enriched Random Forest (ERF): Use weighted random sampling instead of simple random sampling at each node of each tree, giving greater weight to features that appear to be more informative.

- Tree Weighted Random Forest (TWRF): Weight trees so that trees exhibiting better accuracy are assigned higher weights.

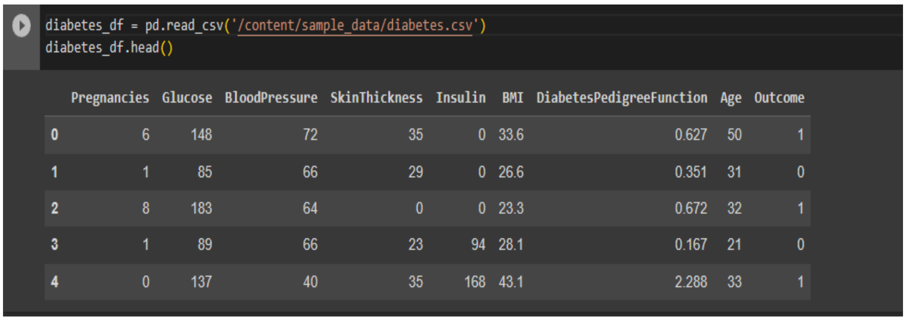

Training Set
Test Set
Training Data 1
Training Data 2
Training Data n
Decision Tree 1
Decision Tree 2
Decision Tree n
Voting (averaging)
Prediction

Implementation

Data Preprocessing

Data Loading

We began by loading the diabetes dataset using the Pandas library in Python:

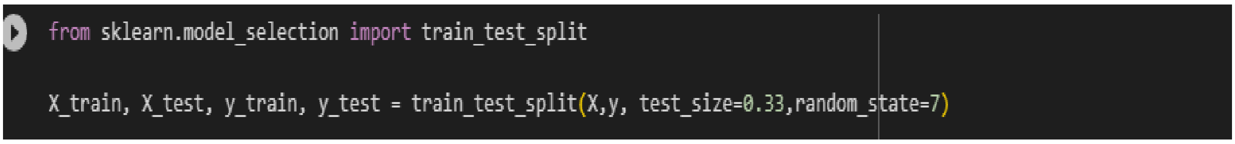


```
diabetes_df = pd.read_csv('/content/sample_data/diabetes.csv')
diabetes_df.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Data Splitting

We divided the data into training and testing sets:

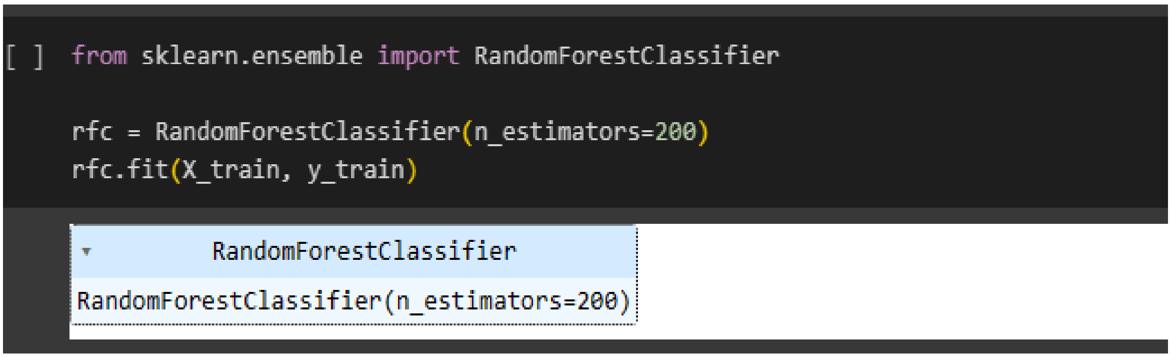


```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.33,random_state=7)
```

Model Selection and Training

We experimented with various machine learning algorithms to select the best-performing one. Here's an example of training a logistic regression model:



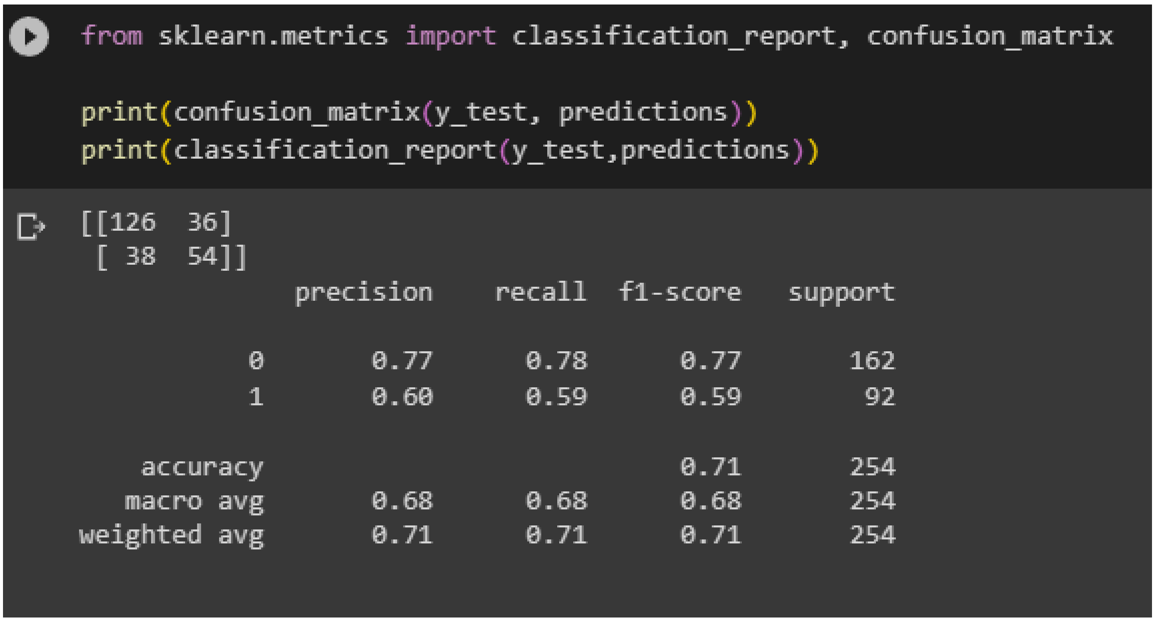
```
from sklearn.ensemble import RandomForestClassifier

rfc = RandomForestClassifier(n_estimators=200)
rfc.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(n_estimators=200)
```

Model Evaluation

We evaluated the model using a range of metrics:



```
from sklearn.metrics import classification_report, confusion_matrix

print(confusion_matrix(y_test, predictions))
print(classification_report(y_test,predictions))
```

```
[[126  36]
 [ 38  54]]
              precision    recall  f1-score   support

           0       0.77      0.78      0.77       162
           1       0.60      0.59      0.59        92

    accuracy                           0.71       254
   macro avg       0.68      0.68      0.68       254
weighted avg       0.71      0.71      0.71       254
```

Conclusion

The implementation of our Diabetes Prediction Model involved data preprocessing, model selection, training, evaluation, and hyper parameter tuning. Our model shows promising results, and it can be a valuable tool for early diabetes risk assessment and intervention in clinical settings.

Remember to adapt this section to your specific project and include any additional details or code snippets relevant to your implementation process. Additionally, provide context and explanations to ensure the reader understands the steps you took to build and evaluate your model.