PROJECT REPORT ON

Diabetes Prediction using Machine Learning

SUBMITTED IN PARTIAL FULFILLMENT OF REQUIREMENT FOR THE

MASTER OF SCIENCE OF INFORMATION TECHNOLOGY

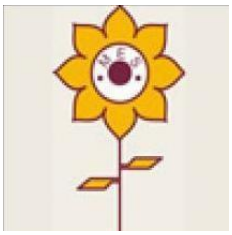
A.Y. 2022-2023

SUBMITTED BY MS. JANHAVI SANTOSH DEVARE

ROLL NO: 9055

UNDER THE GUIDANCE OF

Mr. Abhijeet Salvi



DEPARTMENT OF INFORMATION TECHNOLOGY

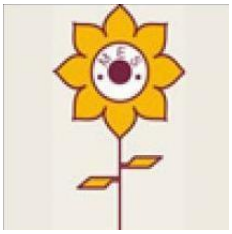
MAHATMA EDUCATION SOCIETY'S

PILLAI'S COLLEGE OF ARTS, COMMERCE & SCIENCE

NEW PANVEL, RAIGAD PIN CODE-410206 MAHARASHTRA

A. Y. 2022-2023

CERTIFICATE



This is to certify that the project entitled, "**DIABETES PREDICTION USING MACHINE LEARNING",** is successfully completed by **Ms. Janhavi Santosh Devare** as per the syllabus and in partial fulfillment for the completion of MSc. in Information Technology. It is also to certify that this is the work of the candidate one during the academic year 2022-2023.

Exam Seat No: 9055

Place: NEW PANVEL

Signature of Guide	**Signature**	**Signature of**
Examiner	**Principal/Co-coordinator**	**External**

Seal

DECLARATION

I hereby declare that the project entitled, **"DIABETES PREDICTION USING MACHINE LEARNING"** done at Pillai's College of Arts, Commerce and Science, has not been in any case duplicated to submit to any other university for the award of any degree. To the best of my knowledge other than me, no one has submitted to any other university.

The project is done in partial fulfillment of the requirement for the award of **MASTER OF SCIENCE (INFORMATION TECHNOLOGY)** to be submitted as a final semester project as part of our curriculum.

JANHAVI SANTOSH DEVARE

INDEX

Chapters		Page No.
Abstract		i
Acknowledgement		ii
List of Figures		iii
List of Tables		iii
Chapter 1	Introduction	1
Chapter 2	Literature Survey	3
2.1	Research Gap	7
2.2	Scope of the Project	7
2.3	Problem Statement	8
Chapter 3	Proposed Methodology	9
3.1	System Architecture	9
3.2	Proposed System	10
3.3	Modeling	14
3.4	Applications	17
Chapter 4	Code Implementation	18
Chapter 5	Result Analysis	28
Chapter 6	Conclusion	29
	Reference	30
	Bibliography	32

ABSTRACT

Diabetes is a serious disease that affects majority of the population. It is caused by increased blood sugar level due to imbalance in insulin production by the body, which leads to variety of disorders like heart failure, blindness, kidney failure, blood pressure etc. which also affect other parts of the body. The diabetic disease cannot be cured but it can be controlled and managed by timely detection. The subject needs to visit a diagnostic center every time, to get their reports and consult a doctor which is time consuming.

Machine learning algorithms provide better results in diabetes detection by constructing models from patients data. The aim of the project is to make an early prediction of diabetes more precisely by using variety of machine learning algorithms and comparing the accuracy among them. The models consider few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. The project incorporates the algorithms like Naive Bayes, K-Nearest Neighbour(KNN), Logistic Regression, Support Vector Machine(SVM) and Random forest for obtaining the early prediction, high accuracy, and performance comparison with the related works.

ACKNOWLEDGEMENT

When I got the opportunity to do a work on **DIABETIES PREDICTION USING MACHINE LEARNING.**

I took this challenge and finished the work with my hard work. We dedicate this page with our profound and earnest thanks to our internal guide **Prof. Sagar Kulkarni** who helped us on numerous aspects of our project and unrelentingly guided and encouraged us the drive to take up and complete this project.

I am also thankful to our Msc. IT Coordinator, **Prof. Abhijeet Salvi** and our principal, **Dr Gajanan Sir** for giving me opportunity and providing the required resources and infrastructure to complete this project.

We are conscious that our project work rather than this acknowledgement will be a more appropriate way to express our gratitude towards the people mentioned above.

Ms. Janhavi Santosh Devare
Class: MSc.IT

List of Figures

Sr. No.	Figure Name	Page No.
Fig No. 3.1	System Architecture	9
Fig No. 3.2	Data Pre-processing	11
Fig No. 3.3	Feature Extraction	12
Fig No.3.4	Train and Test	14
Fig No. 3.5	Support Vector Machine	15
Fig No. 3.6	Logistic Regression	16
Fig No. 4.1	Importing Libraries & Dataset	18
Fig No. 4.2	Pre-processing	19
Fig No. 4.3	Correlation matrix	19
Fig No. 4.4	Heatmap	20
Fig No. 4.5	Train-test split	21
Fig No. 4.6	Training model	21

List of Tables

Sr. No.	Table Name	Page No.
Table No. 2.1	Literature Review Summary	5
Table No. 3.1	Dataset Attributes	10

CHAPTER 1

INTRODUCTION

All around there are numerous ceaseless infections that are boundless in evolved and developing nations. One of such sickness is diabetes. Diabetes is a metabolic issue that causes blood sugar by creating a significant measure of insulin in the human body or by producing a little measure of insulin. Diabetes is perhaps the deadliest sickness on the planet. It is not just a malady yet, also a maker of different sorts of sicknesses like a coronary failure, visual deficiency, kidney ailments and nerve harm and so on.

Subsequently, the identification of such chronic metabolic ailment at a beginning period could help specialists around the globe in forestalling loss of human life. Presently, with the ascent of machine learning. AI, and neural system, and their application in various domains, we may have the option to find new realities from existing well-being-related information indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database. The point of this framework is to make an ML model, which can anticipate with precision the likelihood or the odds of a patient being diabetic. The ordinary distinguishing process for the location of diabetes is that the patient needs to visit a symptomatic focus. One of the key issues of bio-informatics various laboratory tests can entangle the procedure of identification of the disease. This model can foresee whether the patient has the diabetes or not, aiding specialists to ensure that the patient in need of clinical consideration can get it on schedule and also help anticipate the loss of human lives.

DNA makes neural networks the apparent choice. Neural networks use neurons to transmit data across various layers, with each node working on a different weighted parameter to help predict diabetes.

Presently, with the ascent of machine learning, AI and neural system, and their application in various domains, we may have the option to find an answer for this issue. ML strategies and neural systems help scientists to find new realities from existing well-being related information indexes, which may help in ailment supervision and detection. The current work is completed utilizing the Pima Indians Diabetes Database.

Types of Diabetes:

1. Type one diabetes outcomes due to the failure of the pancreas to supply enough hypoglycaemic agent. This type was spoken as "insulin-dependent polygenic disease mellitus" (IDDM) or "juvenile diabetes". The reason is unidentified. The type one polygenic disease found in by doctors children beneath twenty years old. People suffer throughout their life because of the type one diabetic and rest on insulin vaccinations. The diabetic patients must often follow workouts and fit regime which are recommended.
2. The type two diabetes starts with hypoglycaemic agent resistance, a situation in which cells fail to respond the hypoglycaemic agents efficiently. The sickness develops due to the absence of hypoglycaemic agent that additionally built. This type was spoken as "non-insulin-dependent polygenic disease mellitus". The usual cause is extreme weight. The quantity of people affected by type two will be enlarged by 2025. The existence of diabetes mellitus is condensed by 3% in rural zone as compared to the urban zone. The prehypertension is joined with bulkiness, obesity and diabetes mellitus. The study found that an individual United Nations agency has traditional vital sign.
3. Type 3 Gestational diabetes occurs when a woman is pregnant and develops high blood sugar levels without a previous history of diabetes. Therefore, it is found that in total 18% of women in pregnancy have diabetes. So in the older age there is a risk of emerging gestational diabetes in pregnancy. Obesity is one of the main reasons for type-2 diabetes. The type-2 polygenic disease are under control by proper workout and taking appropriate regime. When the aldohexose level isn't reduced by the higher strategies then medications are often recommended. The polygenic disease static report says that 29.1 million people of the United States inhabitants has diabetes.

Causes of Diabetes :

Genetic factor are the main cause of diabetes. It is caused by at least two mutant genes in the 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, Coxasackievirus, mumps, hepatitis B virus, and cytomegalovirus increases the risk of developing diabetes.

Data mining and machine learning have been developing, reliable, and supporting tools in the medical domain in recent years. The data mining methods 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. Machine learning uses various algorithms to learn from the past data and make predictions.

CHAPTER 2

LITERATURE SURVEY

The following chapters give an overview of the various methodologies used by various authors for disease prediction using machine learning methodologies. We can observe that there is fine comparison made between 5 major machine learning algorithms whether they are able to predict the presence of the disease with a greater accuracy, achieving optimal performance.

[1] N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology, In this paper, author had used In this paper, SVM with different kernel functions are applied which are linear, radial and polynomial. The SVM with Linear Kernel function produced 98%, SVM with Radial Kernel Produced 95% and SVM with Polynomial kernel produced 90% for the chosen data set. SVM with linear kernel showed the highest accuracy value for the classification of diabetes.

[2] M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), primary aim of our work is to classify gestational diabetic or non-gestational diabetic. Proposed work is based on decision tree j48. In this paper author had got the 91.2% efficiency .

[3] Aeshah Saad Alanazi and Mohd A. Mezher "Using Machine Learning Algorithms For Prediction Of Diabetes Mellitus,"2020 International Conference on Computing and Information Technology, In this paper, author had used two algorithms of machine learning algorithms and these algorithms are Support Vector Machine and Random Forest to predict the diabetes. Author had used a real dataset collected from Security Force Primary Health Care. The proposed model achieved 98% of accuracy, ROC 99%. The result shows that Random Forest algorithm is has better accuracy score when compared to Support Vector Machine.

[4] G. A . Pethunachiyar, "Classification Of Diabetes Patients Using Kernel Based Support Vector Machine," 2020 International Conference on Computer Communication and Informatics (ICCCI - 2020), In this paper, author had used SVM with different kernel methods are used such as linear, radial and polynomial functions. The accuracy of prediction is calculated using confusion matrix. It is observed that the linear based kernel produces 98% accuracy in predicting the diabetes

patients.ROC curve is used to evaluate the different kernel functions in SVM. Here, Sensitivity (True Positive Rate) and Specificity (False Positive Rate) for different parameters is plotted using R too.l

[5] Amani Yahyaoui, et. al, “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep LearningTechniques, “2019 1st International Informatics and Software Engineering”, . In this paper, author had proposed a DSS for diabetes prediction based on Machine Learning (ML) techniques and they compared conventional machine learning with deep learning approaches. For conventional machine learning method, they considered the most commonly used classifiers: Support Vector Machine (SVM) and the Random Forest(RF). On the other hand, for Deep Learning (DL) they employed a fully Convolutional Neural Network (CNN) to predict and detect the diabetes patients. The overall accuracy obtained using DL, SVM and RF was 76.81%, 65.38% and 83.67% respectively. The experimental results show that RF was more effective for diabetes prediction compared to deep learning and SVM methods.

Literature Review Summary

Table No. 2.1 Literature Review Summary

Sr No.	Paper details	Implementation details	Methods	Accuracy	Observation
1.	N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology	In this paper, SVM four different kernel functions which are linear, polynomial, RBF and sigmoid are used for predicting the diabetes.	Support Vector Machine	98%	In this paper, it is found that RBF kernel best performs for diabetes prediction
2.	M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," 2020 3rd	This study focuses on gestational diabetes by applying Decision Tree j48 classifier on dataset. The dataset is analyzed	Decision Tree	91%	Dataset is analyzed using weka tool. Only one machine learning algorithm is used that is

	International Conference on Intelligent Sustainable Systems (ICISS),	using weka tool			decision tree J48
3.	Aeshah Saad Alanazi and Mohd A. Mezher "Using Machine Learning Algorithms For Prediction Of Diabetes Mellitus,"2020 International Conference on Computing and Information Technology	In this paper, two machine learning algorithms for diagnosing and predicting diabetes. The diagnostic of the performance of the algorithms was done by developing and analyzing the Receiver Operating Characteristics Curve (The ROC curve), and generating a confusion matrix.	Support Vector Machine and Random Forest	98%	In this paper, SVM & RF is used, it is found that obtained results that RF algorithm has a prediction accuracy of 98% which is better than SVM
4.	G. A . Pethunachiyar, "Classification Of Diabetes Patients Using Kernel Based Support Vector Machine," 2020 International Conference on Computer Communication and Informatics (ICCCI - 2020),	SVM with different kernel functions such as linear, radial and polynomial functions are applied for predicting the diabetes. SVM with linear kernel showed the highest accuracy value for the classification of diabetes.	Support Vector Machine	91%	In this paper, author has used ROC curve to evaluate the different kernel functions in SVM and SVM with Linear Kernel function produced better accuracy,
5	Amani Yahyaoui, Jawad Rasheed A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques, "2019 1st International Informatics and Software Engineering"	In this paper, study performed a comparative analysis of machine learning and deep learning-based algorithms for prediction of diabetes.	Support vector machine, Random forest, Convolutional neural network	83.67%	It is observed that RF was more effective for classification of the diabetes in all rounds of experiments

2.1 Research Gap :

From the above study of literature review it is observed that there are some gaps that can be improved by adding some functionality.

- The project can be extended to apply feature selection method before training the model.
- Projects contains high false positive that can be improved.
- Implementation of new ML techniques is essential but without the understanding of the existing issues and barrier in diabetes detection, the advancement is not possible.
- There is no interactive tool for users to predict diabetes. Where User can enter the details through the web application.

2.2 Scope of the Project :

The disease diagnosis system will permit end-users to predict heart disease and diabetes.

2.2.1 Growth of AI Systems

Artificial Intelligence is one of the hottest topics today.

The revenue for cognitive and artificial intelligence systems is expected to hit $12.5 billion.

2.2.2 Availability of Doctors and Chatbots

Other than disease diagnosis, artificial intelligence can be used to streamline and optimize the clinical process. There is only one doctor for over 1600 patients in India .AI health assistants can help in covering large part of clinical and outpatient services freeing up doctor's time to attend more critical cases. Chatbots like "Your.MD" can assist patients by understanding patients' symptoms and suggest easy-to-understand medical information about their condition. Other assistants like "Ada" integrated with"Amazon Alexa" provides a detailed symptom assessment report and also provides an option to contact a real doctor. Such assistants make use of Natural Language Processing and Deep Learning to understand the user and generate suggestions.

2.2.3 Internet of Things (IoT), Healthcare and Machine Learning

Increasing use of Internet of Things has promising benefits in healthcare. Dynamically collecting patient data using remote sensors can help in early detection of health problems and aid in preventive care.

2.3 Problem Statement

The primary goal is to develop a prediction engine which will allow the users to checkwhether they have diabetes or heart disease sitting at home. The user need not visit thedoctor unless he has diabetes or heart disease, for further treatment. The prediction engine requires a large dataset and efficient machine learning algorithms to predict the presence of the disease. Pre-processing the dataset to train the machine learning models, removing redundant, null, or invalid data for optimal performance of the prediction engine.

Doctors rely on common knowledge for treatment. When common knowledge is lacking, studies are summarized after some number of cases have been studied. But this process takes time, whereas if machine learning is used, the patterns can be identified earlier. For using machine learning, a huge amount of data is required. There is very limited amount of data available depending on the disease. Also, the number of samples having no diseases is very high compared to the number of samples having the disease.This project is about performing case study to compare the performance of various machine learning algorithms to help identify such patterns in (i) and to create a platform for easier data sharing and collaboration.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 System Architecture

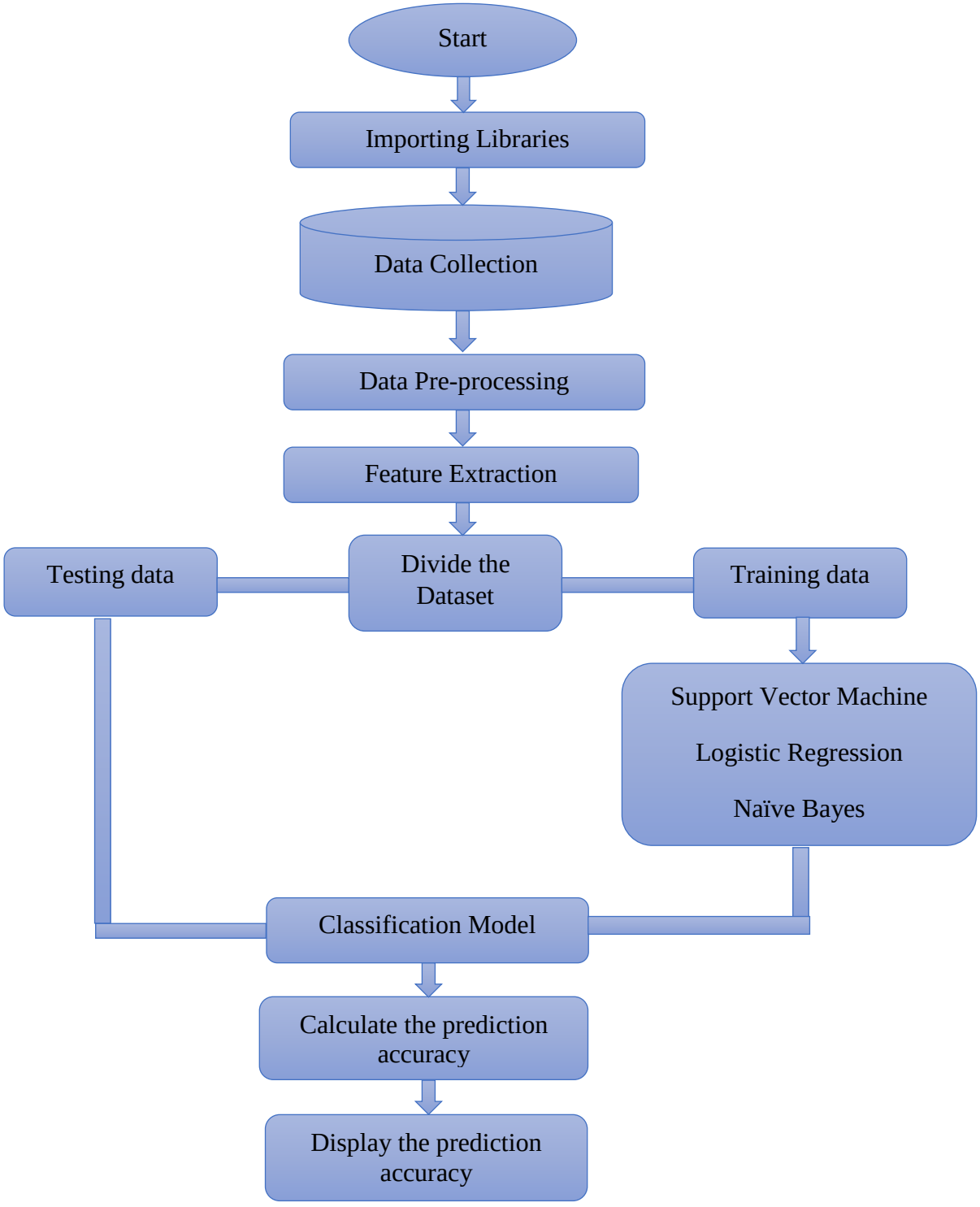


Fig No. 3.1 System Architecture

3.2 Proposed System :

1] Data Collection :

Data collection is the first step of of any machine learning project. The goal of this step is to identify and obtain all data-related problems. In this step, it identify the different data sources, as data can be collected from various sources such as files, database, internet, or mobile devices. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

This step includes the below tasks:

- Identify various data sources
- Collect data
- Integrate the data obtained from different sources

In this project I will be using PIMA Indian dataset for diabetes prediction which I have taken from www.kaggle.com.

PIMA Indian dataset has the following attributes:

Table No. 3.1 Dataset Attributes

Sr.No.	Attributes	Description of attributes	Range
1.	Pregnancy	Number of times a participant is pregnant	0–17
2.	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test	0–199
3.	Diastolic Blood pressure	It consists of Diastolic blood pressure (when blood exerts into arteries between heart)(mm Hg)	0–122
4.	Skin Thickness	Triceps skinfold thickness (mm).It concluded by the collagen content	0–99
5.	Serum Insulin	2-Hour serum insulin (mu U/ml)	0–846
6.	BMI	Body mass index (weight in kg/(height in m)^2)	0–67.1
7.	Diabetes pedigree Function	An appealing attributed used in diabetes prognosis	0.078–2.42
8.	Age	Age of participants	21–81
9.	Outcome	Diabetes class variable, Yes represent the patient is diabetic and no represent patient is not diabetic	Yes/No

2] Data Pre-processing :

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. The purpose of pre-processing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

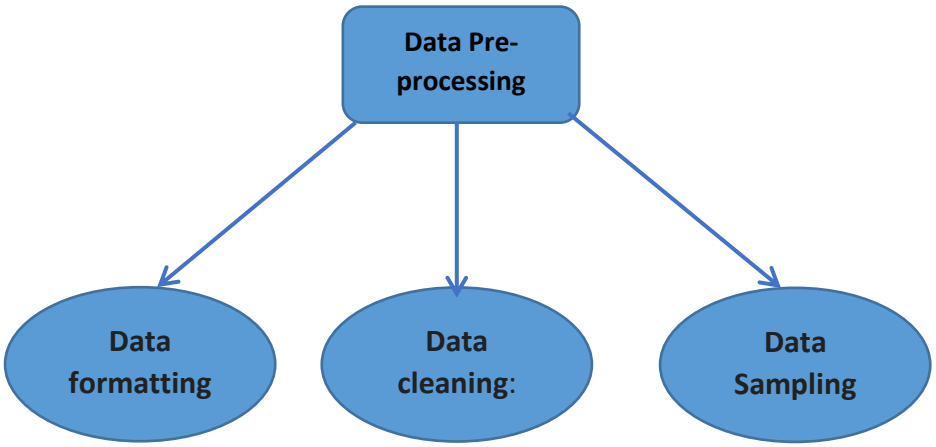


Fig No. 3.2: Data Pre-processing

Data formatting: The importance of data formatting grows when data is acquired from various sources by different people. The first task is to standardize record formats. Then checks whether variables representing each attribute are recorded in the same way. Titles of products and services, prices, date formats, and addresses are examples of variables. The principle of data consistency also applies to attributes represented by numeric ranges.

Data cleaning: This set of procedures allows for removing noise and fixing inconsistencies in data. A data scientist can fill in missing data using imputation techniques, e.g. substituting missing values with mean attributes. A specialist also detects outliers — observations that deviate significantly from the rest of distribution. If an outlier indicates erroneous data, a data scientist deletes or corrects them if possible.This stage also includes removing incomplete and

useless data objects.

Data sampling: Big datasets require more time and computational power for analysis. If a dataset is too large, applying data sampling is the way to go. This technique is used to select a smaller but representative data sample to build and run modelsmuch faster, and at the same time to produce accurate outcomes.

Data Transformation: Data transformation is the process of converting data from one format or structure intoanother format or structure. Data transformation is critical to activities such as data integration and data management. Data transformation can include a range of activities: you might convert data types, cleanse data by removing nulls or duplicate data, enrichthe data, or perform aggregations, depending on the needs of your project.

Scaling: Data may have numeric attributes (features) that span different ranges, for example, millimetres, meters, and kilometres. Scaling is about converting these attributes so that they will have the same scale, such as between 0 and 1, or 1 and 10 for the smallest and biggest value for an attribute.

3] Feature extraction:

Feature selection is one of the important concepts of machine learning, which highly impacts the performance of the model. As machine learning works on the concept of "Garbage In Garbage Out", so we always need to input the most appropriate and relevant dataset to the model in order to get a better result. A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection. Some of the popular methods of feature selection methods are chi square test, information gain, variance threshold.

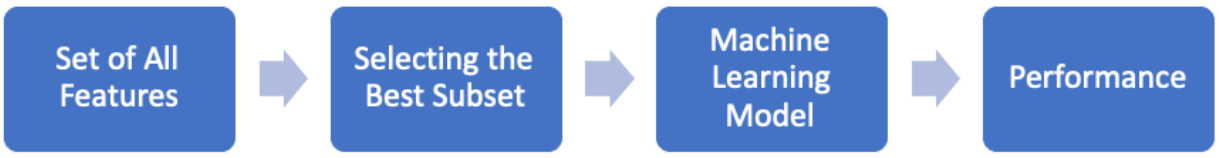


Fig No. 3.3: Feature Extraction

4]Training and testing the Dataset :

After preparing the dataset in full informative(removing unnecessary information)e format the dataset is divided into two major parts: i] Training data, ii] Testing data

Training Data: The training data is the biggest subset of the original dataset, which usually 70% or 75% of whole dataset, which is used to train or fit the machine learning model. Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task. The training data varies depending on whether we are using supervised learning or unsupervised learning algorithms.

In this project, I am applying 3 machine learning algorithm for training and building the model, which are Support vector machine, Logistic regression naïve bayes.

Validation data: During training, validation data infuses new data into the model that it hasn't evaluated before. Validation data provides the first test against unseen data, allowing us to evaluate how well the model makes predictions based on the new data. It is not always necessary to use validation data, but it can provide some helpful information to optimize hyperparameters, which influence how the model assesses data.

Testing Data: Once we train the model with the training dataset and assess the model performance using the validation data set., it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another subset of original data, which is independent of the training dataset. However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. One can split the data into a 70:20:10 ratio.

At this stage, I will check and compare the testing accuracy with the training accuracy, which means how accurate the model is with the test dataset against the training dataset.

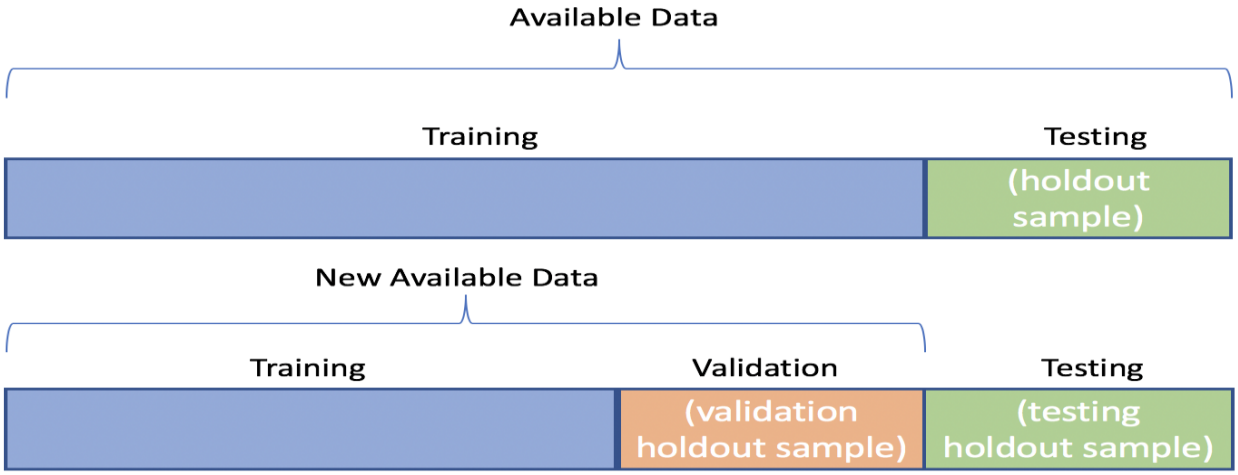


Fig No. 3.4: Training and Testing

3.3. Modelling

After pre-processing the collected data and split it into three subsets, we can proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data — an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Applied Machine learning algorithms :

Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm used in finding solutions to regression and classification problems. SVM classification performs the transformation of the training data into a high dimension that is then classified according to the optimality of the separation hyperplanes boundaries. . The distance between the classes is referred to as margin. SVM algorithms finds a margin such that its distance is maximum. The higher the margin, the better the classification accuracy can be obtained for the classifier. Numerous researchers have used SVM in studies relating to diabetes classification activities.

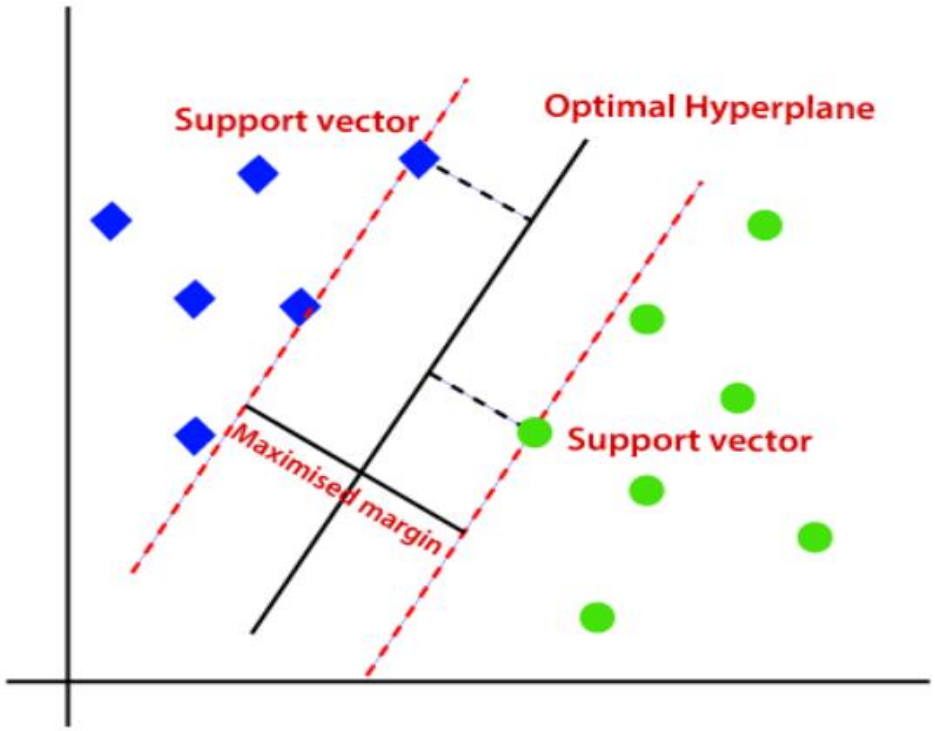


Fig No. 3.5: Support Vector Machine

Logistic Regression: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

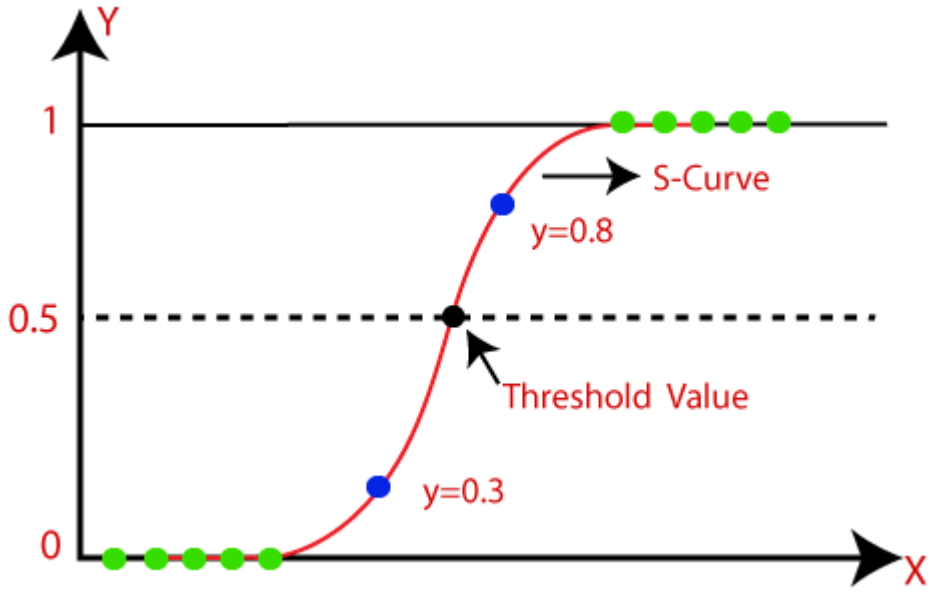


Fig No. 3.6: Logistic Regression

Naïve Bayes: Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in *text classification* that includes a high-dimensional training dataset.Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Model Deployment:

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. Itis one of the last step of the project. Lastly, the algorithms which will having the best accuracy will be used for prediction of diabetes.

3.4 Applications of the project

- Prediction of diabetes at an early stage can lead to improved treatment.
- The user need not visit the doctor unless he/she has the diabetes for further treatment.
- System will provide the interactive tool for the users that will be easy and convenient for the users.
- The model can serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist.
- General physicians can utilize this tool for initial diagnosis of cardio-patients.

CHAPTER 4

4.1 CODE IMPLEMENTATION

Model Building :

Importing Liabraries

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading Dataset

In [5]:
```
data= pd.read_csv('C:/Users/Rupesh/Desktop/MscIT_2/diabetes.csv')
data
```

Out[5]:

	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
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0

Fig No. 4.1: Importing Liabraries & Dataset

Checking for missing values

In [7]:
```
data.isnull()
sns.heatmap(data.isnull())
```

Out[7]: <AxesSubplot:>

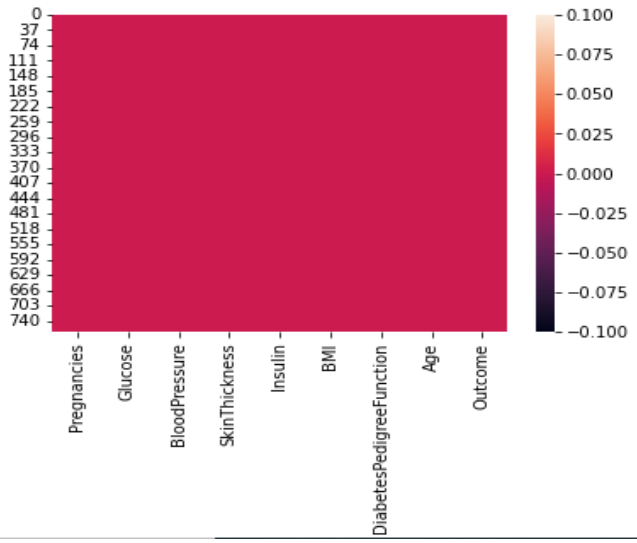


Fig No. 4.2: Preprocessing

Co-relation matrix

In [8]:
```
correlation= data.corr()
print(correlation)
```

```
                          Pregnancies   Glucose  BloodPressure  SkinThickness  \
Pregnancies                  1.000000  0.129459       0.141282      -0.081672
Glucose                      0.129459  1.000000       0.152590       0.057328
BloodPressure                0.141282  0.152590       1.000000       0.207371
SkinThickness               -0.081672  0.057328       0.207371       1.000000
Insulin                     -0.073535  0.331357       0.088933       0.436783
BMI                          0.017683  0.221071       0.281805       0.392573
DiabetesPedigreeFunction    -0.033523  0.137337       0.041265       0.183928
Age                          0.544341  0.263514       0.239528      -0.113970
Outcome                      0.221898  0.466581       0.065068       0.074752

                            Insulin       BMI  DiabetesPedigreeFunction  \
Pregnancies               -0.073535  0.017683                 -0.033523
Glucose                    0.331357  0.221071                  0.137337
BloodPressure              0.088933  0.281805                  0.041265
SkinThickness              0.436783  0.392573                  0.183928
Insulin                    1.000000  0.197859                  0.185071
BMI                        0.197859  1.000000                  0.140647
DiabetesPedigreeFunction   0.185071  0.140647                  1.000000
Age                       -0.042163  0.036242                  0.033561
Outcome                    0.130548  0.292695                  0.173844
```

Fig No. 4.3: Correlation matrix

	Age	Outcome
Pregnancies	0.544341	0.221898
Glucose	0.263514	0.466581
BloodPressure	0.239528	0.065068
SkinThickness	-0.113970	0.074752
Insulin	-0.042163	0.130548
BMI	0.036242	0.292695
DiabetesPedigreeFunction	0.033561	0.173844
Age	1.000000	0.238356
Outcome	0.238356	1.000000

```
In [9]: sns.heatmap(correlation)
```

```
Out[9]: <AxesSubplot:>
```

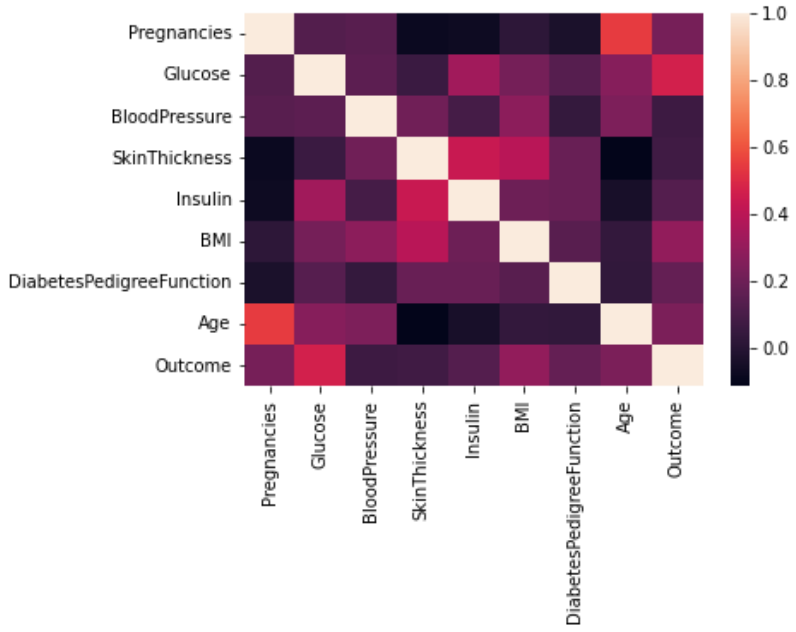


Fig No. 4.4: Heatmap

Train Test split

In [10]:
```
x= data.drop("Outcome", axis=1)
y= data["Outcome"]
y
```

Out[10]:
```
0      1
1      0
2      1
3      0
4      1
      ..
763    0
764    0
765    0
766    1
767    0
Name: Outcome, Length: 768, dtype: int64
```

In [12]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size=0.20)
```

Fig No. 4.5: TrainTest Split

Training the model

In [13]:
```
from sklearn.linear_model import LogisticRegression
model= LogisticRegression()
model.fit(x_train, y_train)
```

```
C:\Users\Rupesh\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:814: ConvergenceWarning: lbfgs failed
rge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[13]:
```
LogisticRegression()
```

In [14]:
```
from sklearn.metrics import accuracy_score
prediction= model.predict(x_test)
print(prediction)
```

```
[1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 1 1 0 0 1 0 0 0 1 1 1 0 1 0 1 0 0 0 0 0
 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1
 0 1 0 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 0
 0 0 1 0 0 0 0 0 1 0 1 0 0 1 1 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1
 1 1 1 0 0 1]
```

```
 1 1 1 0 0 1]
```

In [15]:
```
accuracy= accuracy_score(prediction, y_test)
print(accuracy)
```

```
0.7792207792207793
```

Fig No. 4.6: Training Model

Deploying Model:

Model deployment is done by using PyCharm IDE (Integrated Development Environment) and Django framework. Django is a high-level python web framework that encourages rapid development and clean, pragmatic design.

To start the project in PyCharm IDE following command is written in the terminal of IDE:

→django-admin startproject DiabetesProject

After passing this command some basic files creates automatically such as __init__.py, settings.py, urls.py, manage.py, etc.

To run the project following command is written in terminal :

→ python manage.py runserver

After passing this command local host link is generated http://127.0.0.1:8000/ and on clicking that web browser is opens to show the output.

Home.html

```
{% load static %}

<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Home</title>
    <style type="text/css">
    div{
            color:black;
        }
        h1{
            color: 'white';
            font-family: arial, sans-serif;
            font-size: 60px;
            font-weight: bold;
            margin-top: 200px;
        }
        h2{
            color: 'white';
            font-family: arial, sans-serif;
            font-size: 15px;
            font-weight: bold;
            margin-top: 400px;
        }
        body {
            background-image: url("{% static
'/DiabetesProject/images/Image.jpg' %}"));
            background-repeat: no-repeat;
            background-attachment: fixed;
            background-size: cover;
            background-position: center;
            background-color:skyblue;
```

```
        }
        input[type=submit]{
            background-color: blue;
            border: 2px;
            color: white;
            padding: 16px 32px;
            cursor: pointer;
            margin-top: 15px;
        }
    </style>
</head>
<body>
<div align="center">
    <h1>Welcome to Diabetes Prediction System</h1>
    <form action="predict">
        <input type="submit"value="Lets get started">
    </form>
</div>

</body>
</html>
```

Predict.html

```
{% load static %}

<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Prediction Page</title>
    <style>
        body {
            background-color: skyblue;
            background-image: url("{% static
'/DiabetesProject/images/Image2.jpg' %}");
            background-attachment: fixed;
            background-size: cover;
        }
        .main{
            position: fixed;
            top: 140px;
            left: 410px;
            width: 550px;
            background-color:lightpink;
            border-radius: 10px;
            align-items: center;
            padding: 5%;
        }
        h1{
            color: #0086b3;
            font-size: 30px;
            font-weight: bold;
        }
        input[type=submit]{
            background-color:blue;
            border: 2px;
```

```
            color: white;
            padding: 8px 16px;
            cursor: pointer;
            margin-top: 15px;
        }
    </style>
</head>
<body>
    <div align="center" class="main">
        <h1>Please enter the following information :</h1>
        <form action="result">
            <table>
                <tr>
                    <td align="right">Pregnancies: </td>
                    <td align="left"><input type="text" name="n1"></td>
                </tr>
                <tr>
                    <td align="right">Glucose: </td>
                    <td align="left"><input type="text" name="n2"></td>
                </tr>
                <tr>
                    <td align="right">Blood Pressure: </td>
                    <td align="left"><input type="text" name="n3"></td>
                </tr>
                <tr>
                    <td align="right">Skin Thickness: </td>
                    <td align="left"><input type="text" name="n4"></td>
                </tr>
                <tr>
                    <td align="right">Insulin: </td>
                    <td align="left"><input type="text" name="n5"></td>
                </tr>
                <tr>
                    <td align="right">BMI: </td>
                    <td align="left"><input type="text" name="n6"></td>
                </tr>
                <tr>
                    <td align="right">Diabetes Pedigree Function: </td>
                    <td align="left"><input type="text" name="n7"></td>
                </tr>
                <tr>
                    <td align="right">Age: </td>
                    <td align="left"><input type="text" name="n8"></td>
                </tr>
            </table>
            <input type="submit">
        </form>
        Result: {{result2}}
    </div>
</body>
</html>
```

Views.py

```
from django.shortcuts import render
import pandas as pd
import numpy as np
```

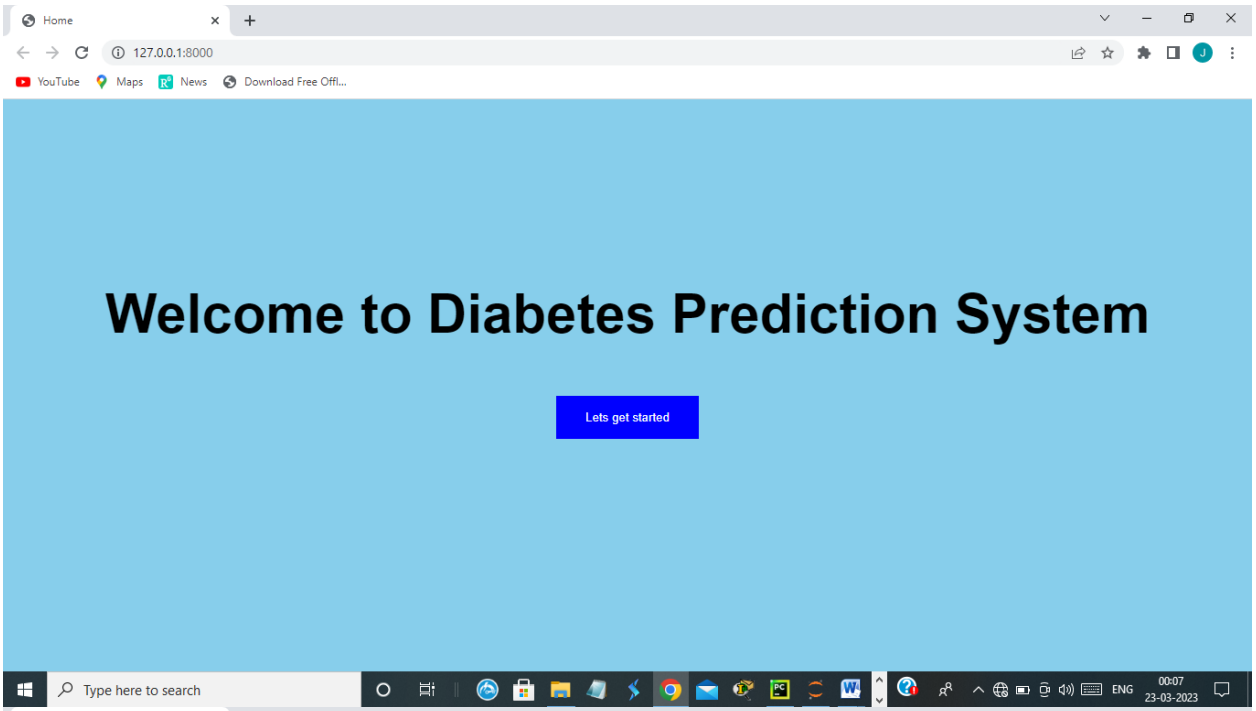
```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

def home(request):
    return render(request, 'home.html')
def predict(request):
    return render(request, 'predict.html')
def result(request):
    data = pd.read_csv(r'C:\Users\Rupesh\PycharmProjects\diabetes.csv')

    x = data.drop("Outcome", axis=1)
    y = data["Outcome"]
    x train, x test, y train, y test = train test split(x, y, test size=0.20)

    model = LogisticRegression()
    model.fit(x_train, y_train)

    val1 = float(request.GET['n1'])
    val2 = float(request.GET['n2'])
    val3 = float(request.GET['n3'])
    val4 = float(request.GET['n4'])
    val5 = float(request.GET['n5'])
    val6 = float(request.GET['n6'])
    val7 = float(request.GET['n7'])
    val8 = float(request.GET['n8'])

    pred = model.predict([[val1, val2, val3, val4, val5, val6, val7, val8]])

    result1 = ""
    if pred==[1]:
        result1 = "Positive"
    else:
        result1 = "Negative"

    return render(request, 'predict.html', {"result2":result1})
```

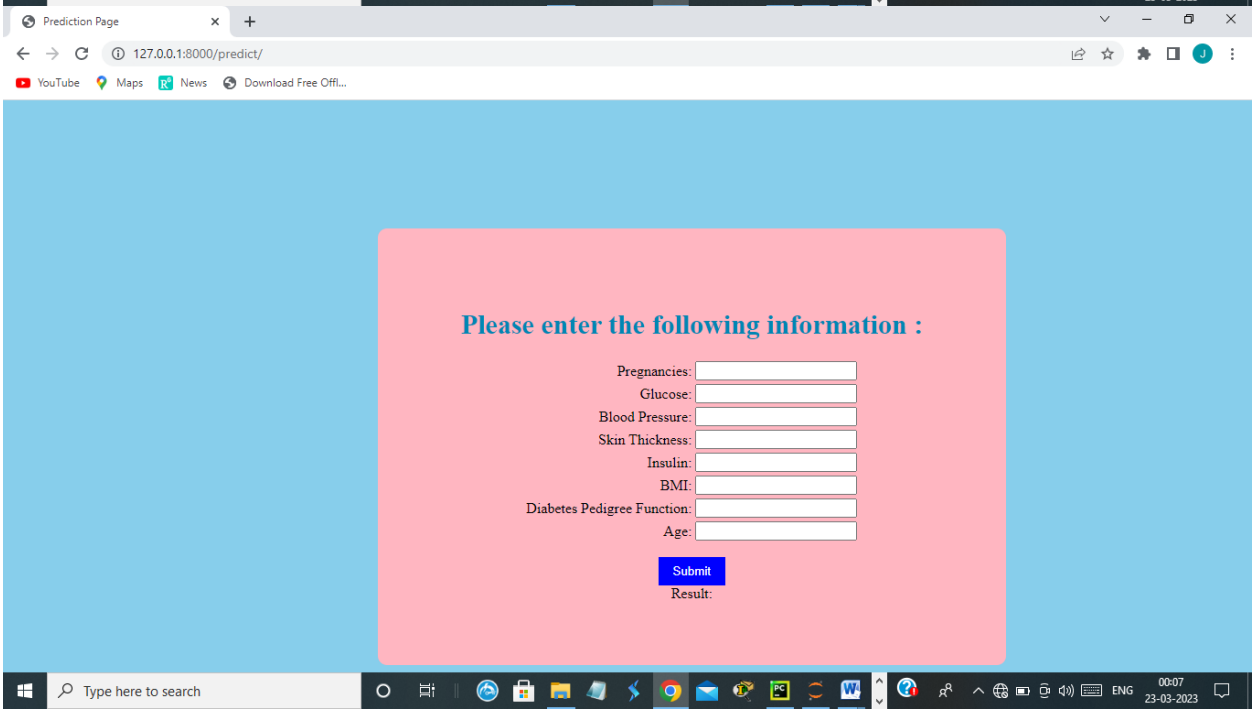
urls.py

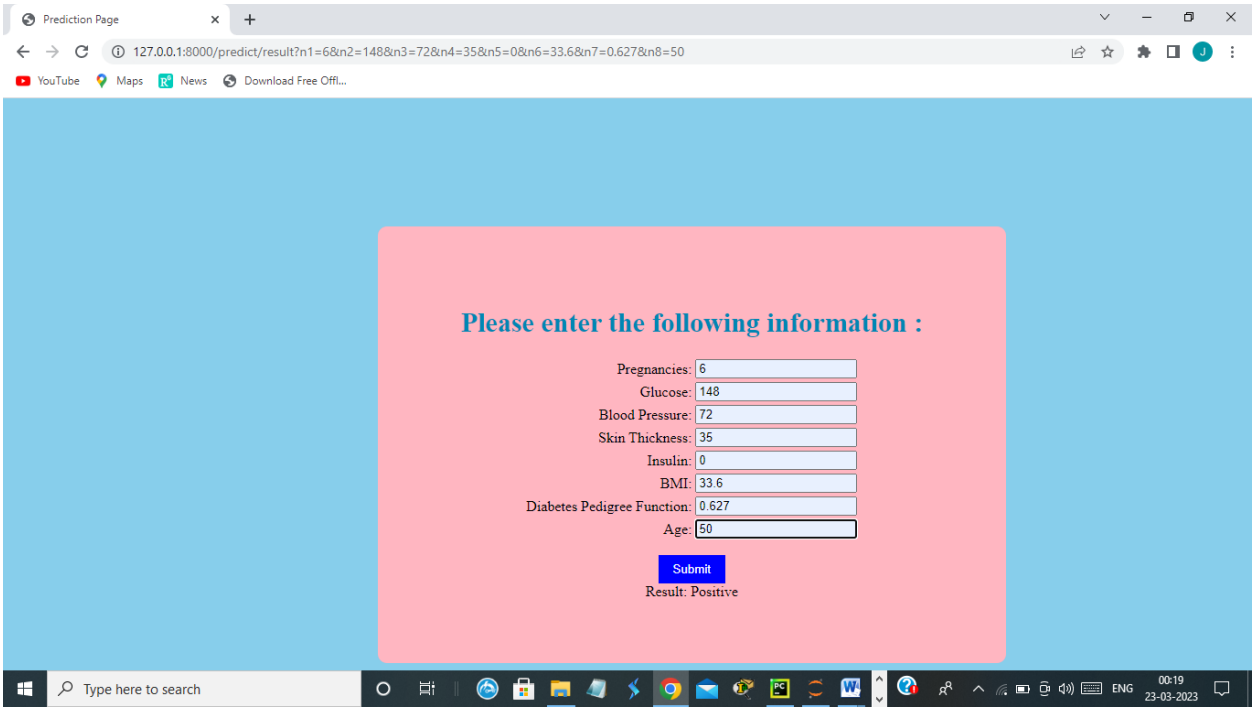
```
"""DiabetesProject URL Configuration

The `urlpatterns` list routes URLs to views. For more information please see:
    https://docs.djangoproject.com/en/4.1/topics/http/urls/
Examples:
Function views
    1. Add an import:  from my_app import views
    2. Add a URL to urlpatterns:  path('', views.home, name='home')
Class-based views
    1. Add an import:  from other_app.views import Home
    2. Add a URL to urlpatterns:  path('', Home.as view(), name='home')
Including another URLconf
    1. Import the include() function: from django.urls import include, path
    2. Add a URL to urlpatterns:  path('blog/', include('blog.urls'))
"""
from django.contrib import admin
```

```
from django.urls import path
from . import views

urlpatterns = [
    path('admin/', admin.site.urls),
    path("", views.home),
    path("predict/", views.predict),
    path("predict/result", views.result),
]
```

Output :






Prediction Page
127.0.0.1:8000/predict/result?n1=6&n2=148&n3=72&n4=35&n5=0&n6=33.6&n7=0.627&n8=50
YouTube
Maps
News
Download Free Offl...
Please enter the following information :
Pregnancies: 6
Glucose: 148
Blood Pressure: 72
Skin Thickness: 35
Insulin: 0
BMI: 33.6
Diabetes Pedigree Function: 0.627
Age: 50
Submit
Result: Positive
Type here to search
ENG
00:19
23-03-2023

CHAPTER 5

RESULT ANALYSIS

Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods. The proposed approach uses various classification and ensemble learning method in which SVM, Logistic regression and Naïve bayes are used. The project predict the diabetes of person based on the relevant medical details that are collected using a Web application. When the user enters all the relevant medical data required in the online Web application, this data is then passed on the trained model for it to make predictions whether the person is diabetic or non-diabetic. After comparing the three different algorithms it is found that logistic regression is providing better accuracy than others. So while deploying the model logistic regression algorithm is used.

CHAPTER 6

CONCLUSION

Computational intelligence has been the potential of expediting early diagnosis of diabetes. Machine learning algorithms such as logistic regression, decision tree, KNN, SVM, naïve bayes, etc are usable in the diagnosis of diabetes. In this project, I have analyzed the predictive performance of three algorithms namely SVM, logistic regression, naïve bayes on diabetes prediction using the kaggle diabetes dataset. Feature extraction was conducted by employing the Pearson's correlation to find the relationship between the class label and other features of the diabetes disease. The feature extraction helped to identify the relevant features and improve the accuracy of the models by using only relevant features in the training. Using the web application, the person with and without diabetes can be predicted. User need to enter the details through the web application.

Further the work can be extended to compare the performance metrics such as precision, recall and F-measure and can be compared with other existing techniques. In future we can improve the feature extraction step by applying deep feature extraction approach and for obtaining a better fitting model to improve the prediction accuracy and we can provide diet plans to the patients according to the diabetes type.

REFERENCES

[1]S. Alanazi and M. A. Mezher, "Using Machine Learning Algorithms For Prediction Of Diabetes Mellitus," *2020 International Conference on Computing and Information Technology*

[2]P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*

[3]N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," *2020 4th International Conference on Electronics, Communication and Aerospace Technology*

*[4]*M. Posonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*

[5]Amani Yahyaoui, Jawad Rasheed,"A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques", "2019 1st International Informatics and Software Engineering"

[6]P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.

[7] Mitushi Soni , Dr. Sunita Varma, 2020, Diabetes Prediction using Machine Learning Techniques, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 09 (September 2020),

[8]https://www.researchgate.net/publication/257334197_Comparative_analysis_of_machine_learning_techniques_in_prognosis_of_type_II_diabetes

[9]https://www.researchgate.net/publication/343536932_A_Review_of_Diabetic_Prediction_Using_Machine_Learning_Techniques

[10]https://www.researchgate.net/publication/336218749_Performance_Evaluation_of_Machine_Learning_Models_for_Diabetes_Prediction

Bibliography

https://ieeexplore.ieee.org/document/9441935

https://www.sciencedirect.com/science/article/pii/S1877050920300557

https://www.sciencedirect.com/science/article/abs/pii/S175199182100019X

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3368308

https://www.researchgate.net/profile/Damian-Mingle/publication/315640596_Predicting_Diabetic_Readmission_Rates_Moving_Beyond_HbA1c/links/58d69922a6fdcc1bae8704c4/Predicting-Diabetic-Readmission-Rates-Moving-Beyond-HbA1c.pdf

https://ieeexplore.ieee.org/abstract/document/9579869