**Diabetes Prediction**

**by using**

**Machine Learning**

A PROJECT REPORT

Submitted by:

**RAKTIMA DATTA**

**HRITWIKA BANIK**

**NATIONAL INSTITUTE OF TECHNOLOGY AGARTALA**

***Under the Project Guidance of***

**Sofikul Mullick**

***At***

## Ardent Computech Pvt. Ltd.

****

**CERTIFICATE**

This is to certify that **Raktima Datta and Hritwika Banik** have successfully completed the project titled ***“Diabetes Prediction by using machine learning”*** under my supervision during the period from January 4, 2021 to January 19, 2021 which is in fulfillment of their training in Data science and machine learning using python.

**Signature of Mentor**

**Sofikul Mullick Ardent Computech Pvt. Ltd**

**Salt Lake, Kolkata**

**ACKNOWLEDGEMENT**

The project ***“Diabetes Prediction by using machine learning”*** would not have been possible without the constant guidance of our guide **Sofikul Mullick ,** Ardent Computech Pvt. Ltd., who guided us throughout this process. We are immensely thankful to him for his valuable ideas on improvement of the project.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sl No** | **Topic** | **Page No** |
| 1. | Abstract | 5 |
| 2. | Introduction | 6 |
| 3. | Problem Definition | 8 |
| 4. | Project Goal | 8 |
| 5. | Methodology | 9 |
| 6. | Project Objective | 11 |
| 7. | Project Workflow | 11 |
| 8. | Project Implementation | 12 |
| 9. | Step-by-Step Working | 14 |
| 10. | Project Limitations | 16 |
| 11. | Future Scope | 17 |
| 12. | Summary | 18 |
| 13. | Bibliography | 19 |

**ABSTRACT**

With the emerging increase of diabetes , there is a strong need for supporting the medical decision making process. In this project we aim to predict whether a person has diabetes or not.

In this project, Logistic Regression is used to predict whether the person is diabetic or not .Logistic Regression are considered as helpful methods for the diagnosis of many diseases.

They, in fact , are probable models which have been proved useful in displaying complex systems and showing the relationships between variables in a graphical way . This project gives us the outcome whether a person is diabetic or not.

**INTRODUCTION**

**Python** is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects

**Machine learning** (**ML**) is the study of computer algorithms that improve automatically through experience.It is seen as a subset of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data”, in order to make predictions or decisions without being explicitly programmed to do so.Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics

Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed.

**Logistic regression** is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, **logistic regression**(or **logit regression**) is estimating  the parameters of a logistic model (a form of binary regression.

**PROJECT DEFINITION**

The main objective of this project is to predict the diabetes. The purposed work focuses on to predict diabetes using probabilistic model of Regression. This integrated technique of classification gives a promising classification results with utmost accuracy rate. For detecting a disease a number of test should be required from the patient. But using data mining technique the number of test should be reduced.This reduced test plays an important role in time and performance. Insulin is one of the most important hormones in the body. It aids the body in converting sugar ,starches and other food items into energy needed for daily life.. This disease is refer to diabetes.The cause of diabetes is mystery,although obesity and lack of exercise appear to possibly play significant roles.

**PROJECT GOAL**

The goal is to predict whether a person has diabetes or not. The goal is achieved by using logistic regression in machine learning.

**METHODOLOGY**

* Data Selection :

Data is the foundation for any machine learning project. The job is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.

* Data Visualisation:

A large amount of information represented in graphic form is easier to understand and analyze. Some companies specify that a data analyst must know how to create slides, diagrams, charts, and templates.

* 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. A specialist also detects outliers — observations that deviate significantly from the rest of distribution.

* Data Splitting:

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

* Model Selection:

After a data scientist has preprocessed the collected data and split it into three subsets, he or she 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 in new data .The purpose of model training is to develop a model.

* Model Evaluation:

The goal of this step is to develop the simplest model able to formulate a target value fast and well enough and check the accuracy .

**PROJECT OBJECTIVE**

The main objective of this is Diabetes Prediction by using machine learning.

**PROJECT WORKFLOW**

This is the detailed work architecture where we are showing the process of Diabetes Prediction using machine learning.

**Data Visualization**

**Model Evaluation**

**Data Splitting**

**Model Training**

**Data Cleaning**

**Data Selection**

**PROJECT IMPLEMENTATION**

* SELECTION OF DATA:

The process of selecting data depends on the type of project we desire to. The data set can be collected from various sources such as a file, database, sensor and many other such sources.

* VISUALIZATION OF DATA:

Data visualization is the graphical representation of.information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in .

* DATA PRE-PROCESSING :

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

* SELECTION OF DEPENDENT AND INDEPENDENT DATA:

We need to select the dependent and independent data and store them in y and x.

* SPLITTING OF THE DATA:

We train the classifier using ‘**training data set**’, then test the performance of your classifier on unseen ‘**test data set**’.We split the data for training and testing by using the ‘train\_test\_split’ .

* FITTING THE MODEL:

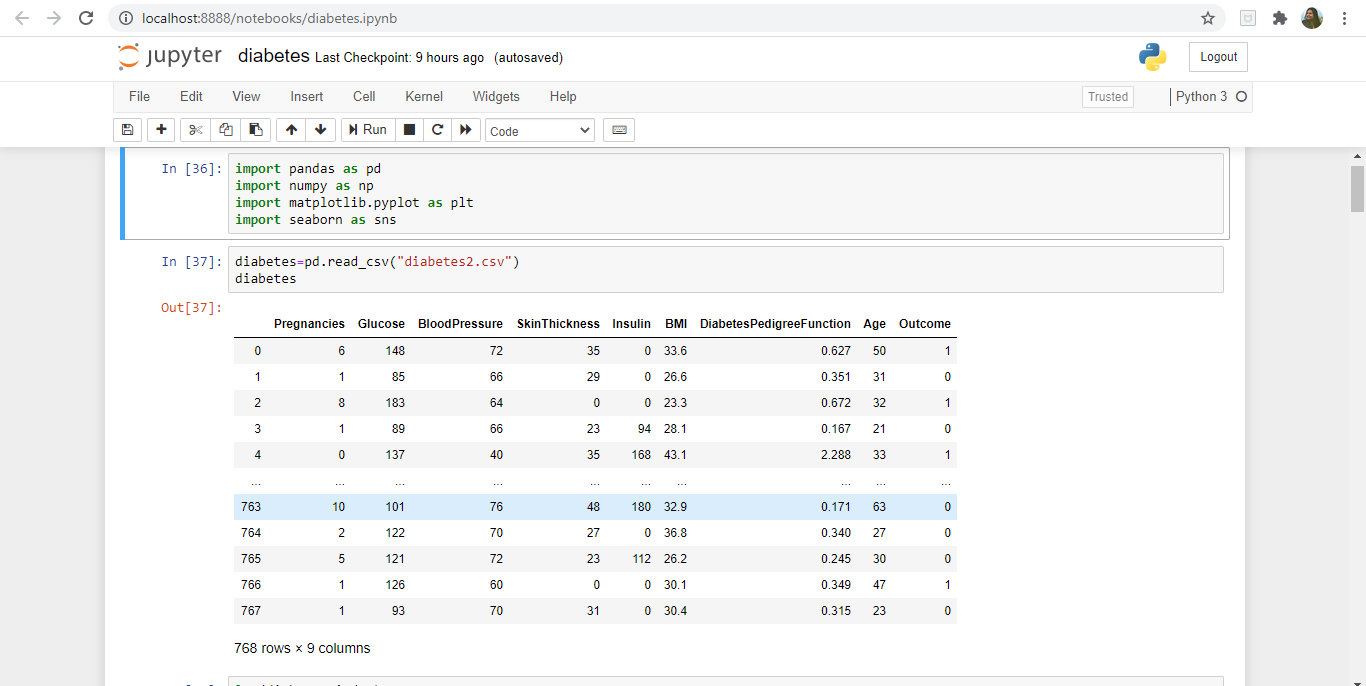
In a data set, a training set is implemented to build up a model.Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model is trained.

* MODEL EVALUATION :

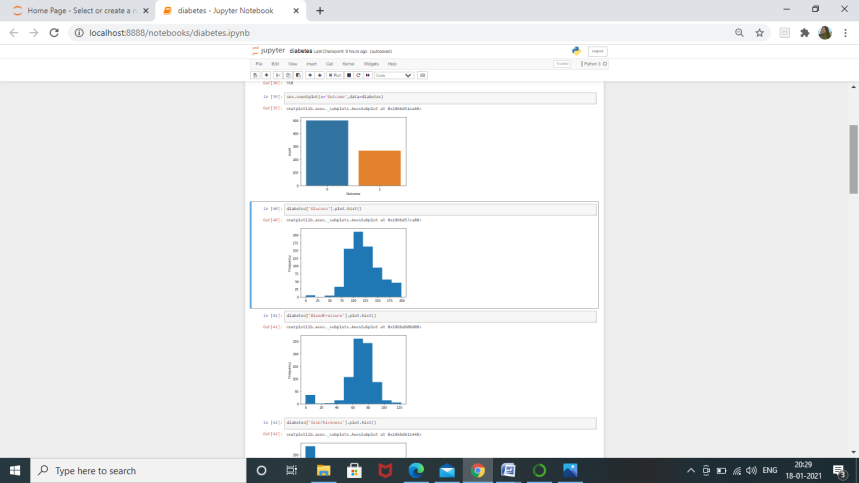
It is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

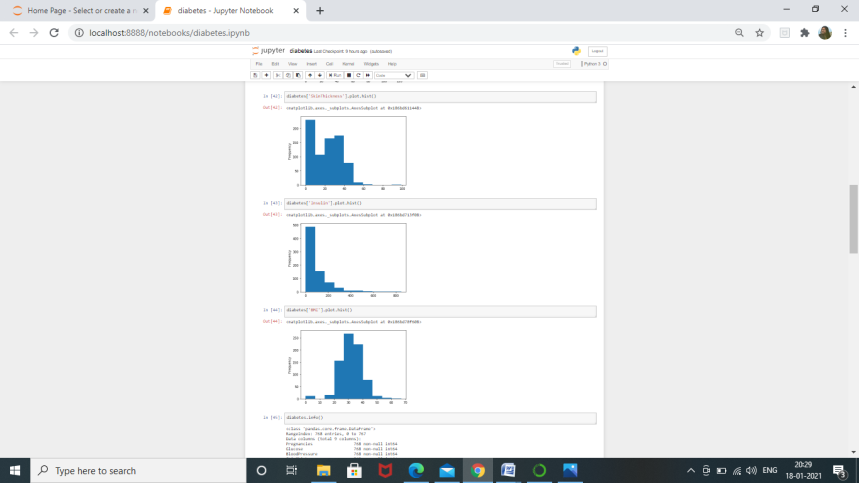
**STEP BY STEP WORKING:**

* SELECTION OF DATA:

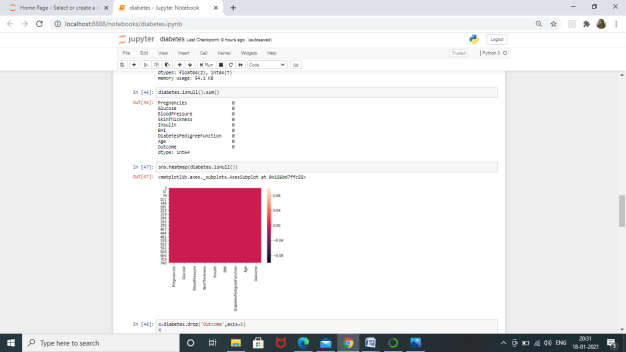
****

* VISUALIZATION OF DATA:

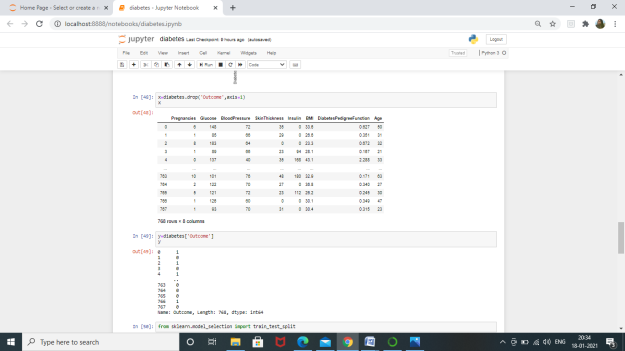




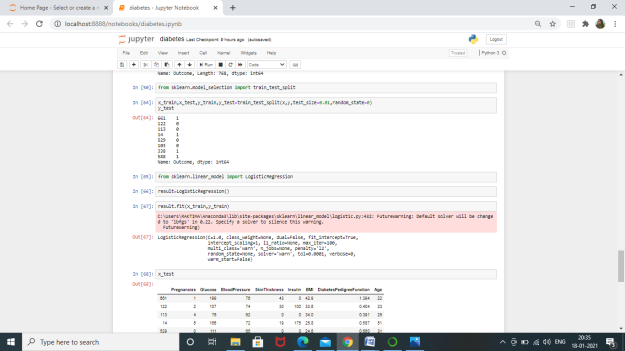
* CLEANING DATA:



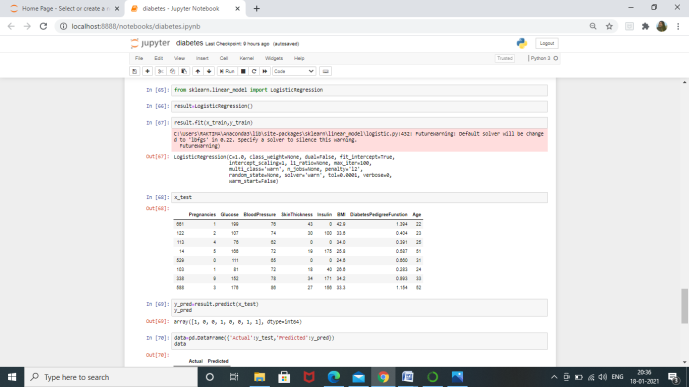
* DEPENDENT AND INDEPENDENT DATA:



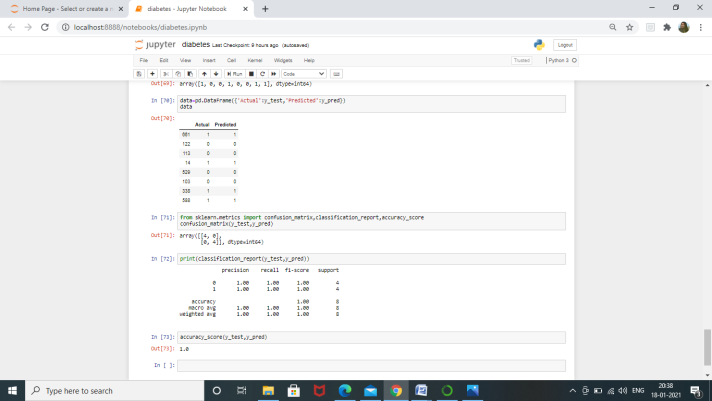
* SPLITTING THE DATA:



* MODEL TRAINING:



* MODEL EVALUATION:



**PROJECT LIMITATIONS:**

* We worked on the **backend part of the system** thus there is no frontend work associated which can result in a more realistic look and focus on user experience.
* While using dataset greater than 1gb , the project won’t work properly.

**FUTURE SCOPE**

* Since here only the backend part of the system is built, we can create a custom frontend which can result in a more realistic look and focus on user experience.
* We can work on the project and use it for greater size dataset .

**SUMMARY**

We load previous datasets to the system.Visualization of data is done to properly know about the data . Data pre-processing is not required as there was no Nan. Following operations are performed on the dataset after that. User input data to the system in order to diagnose whether he has the disease or not. Building model using Logistic Regression Algorithm and train the data set. Test the dataset using model. Get the evaluation result. Get the predicted voting from all classifiers and gives the diagnostic result.

# BIBLIOGRAPHY

* [*https://www.wikipedia.org/*](https://www.wikipedia.org/)
* <https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/>
* [*h ttps://stackify.com/aws-lambda-with-python-a-complete-getting-started-gui d e/*](https://stackify.com/aws-lambda-with-python-a-complete-getting-started-guide/)
* [*https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94*](https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94)

**CODE**

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**diabetes=pd.read\_csv('diabetes2.csv')**

**diabetes**

**len(diabetes.index)**

**sns.countplot(x='Outcome',data=diabetes)**

**diabetes['Glucose'].plot.hist()**

**diabetes['BloodPressure'].plot.hist()**

**diabetes['SkinThickness'].plot.hist()**

**diabetes['Insulin'].plot.hist()**

**diabetes['BMI'].plot.hist()**

**diabetes.info()**

**diabetes.isnull().sum()**

**sns.heatmap(diabetes.isnull())**

**x=diabetes.drop('Outcome',axis=1)**

**x**

**y=diabetes['Outcome']**

**y**

**from sklearn.model\_selection import train\_test\_split**

**x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.01,random\_state=0)**

**y\_test**

**from sklearn.linear\_model import LogisticRegression**

**result=LogisticRegression()**

**result.fit(x\_train,y\_train)**

**x\_test**

**y\_pred=result.predict(x\_test)**

**y\_pred**

**data=pd.DataFrame({'Actual':y\_test,'Predicted':y\_pred})**

**data**

**from sklearn.metrics import confusion\_matrix,classification\_report,accuracy\_score**

**confusion\_matrix(y\_test,y\_pred)**

**print(classification\_report(y\_test,y\_pred))**

**accuracy\_score(y\_test,y\_pred)**