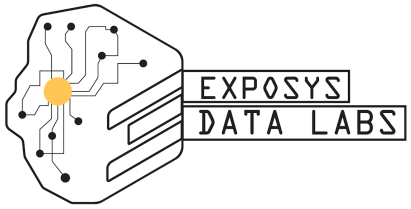
Internship Report
on
Data Science Project

Topic :
Diabetes Disease Prediction

Submitted
by
Richa Jha

Under the Guidance
of
Y .Vishnuvardhan

Performed at



Exposys Data Labs

P.M R. Residency
Ground Floor, No-5/3 Sy. No.10/6-1
Doddaballapur Main Road
Yelahanka Bengaluru, Karnataka 560064

Duration of Internship :
1 Month

Abstract

Diabetes is a most common disease in these world affecting major part of the population. Although it is a global concern but still most part of population are unaware to deal with it. When the level of glucose in our blood increased (often called hyperglycaemia) above the required level then the state is called Diabetes. Inability of pancreas to produce required insulin or inability to use it properly to convert glucose into energy is the cause of Diabetes. Diabetes majorly of 3 types : Diabetes Mellitus, Diabetes Insipidus, Gestational Diabetes. So in this project. It is a Data science project in which we need to predict whether a given data point is diabetic or not. Since it comes under Classification technique which is a type of Supervised Machine Learning, here the model is built using different classification algorithms like K-Nearest Neighbors, and Support Vector Machine. The model built with K-Nearest Neighbors gave high accuracy out of all in predicting whether the given data point is diabetic or not.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 BASIC INFORMATION

Diabetes is a most common disease in these world affecting major part of the population. Although it is a global concern but still most part of population are unaware to deal with it. When the level of glucose in our blood increased (often called hyperglycaemia) above the required level then the state is called Diabetes. Inability of pancreas to produce required insulin or inability to use it properly to convert glucose into energy is the cause of Diabetes.

Diabetes majorly of 3 types :

- Diabetes Mellitus :

This is the most common form of Diabetes which can occur at any age because of the current lifestyle of Humans.
These disease is caused due to inability of functioning of pancreas thus fluctuating the level of glucose in the blood.
Its general symptom can include Weakness, Shaking , sweating , Hunger etc
Diabetes Mellitus types are:

Type 1 Diabetes : It is a state where pancreas are unable to produce enough insulin. It is an Autoimmune disorder which develops rapidly often in younger ages. It immediately need for insulin to be injected in body to make insulin balance stable. It can be deadly if not caught in time. This are not preventable just after diagnose we can make the insulin level stable by insulin therapy.
Symptoms can be : Unexplained Weight loss, Tiredness and weakness, Increased Thirst etc.

Type 2 Diabetes : This is state of body where insulin is produced properly but unable to use it properly and thus makes body insulin resistant. It can prevented through proper diet, exercise and lifestyle. Symptoms can be obesity and weakness etc. It is not genetic disorder so it can come even if it not in hereditary.

- Gestational Diabetes :

It is a temporary state in pregnancy which occur during the gestation period. Develop in 2nd or 3rd trimester which ends after pregnancy. It can be treated through proper diet and exercise. But If proper care not taken can risk to Type 2 Diabetes later in life.

- Diabetes Insipidus :

It is a rare state where there is no role with glucose in blood while this takes place in kidney. Kidney produces a lot of dilute pale urine which does not have glucose.

Symptom can be excessive thirst or hunger , excessive urine , excessive thirst and dehydration.

1.2 DATA SCIENCE

Data Science has an intersection with artificial intelligence but is not a subset of artificial intelligence. This means it is that technique which uses some techniques of artificial intelligence, machine learning, and deep learning. By using this techniques data are first visualized, then builds model, and at last do prediction.

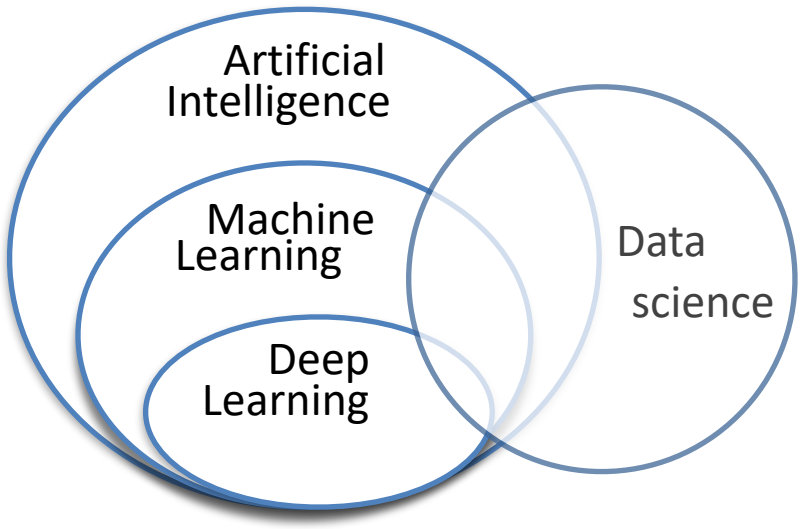


Fig 1.2.1 Intersection of Data Science with AI, ML, DL

It is the art and science of drawing actionable insights from the data.

Applications of Data science are :
- Healthcare
- Telecom
- Bank
- Retail, and etc.

1.3 MACHINE LEARNING

Machine Learning is a subset of Artificial Intelligence. It is a set of algorithms that train on a data set to make predictions or take actions in order to optimize some systems.
It is a technique in which computer program learn from experiences or past experiences. Then specifically related to that it performs the task.
Machine Learning algorithm create model using sample data or training data. Which takes the decision or prediction to perform the task. Then model checks the accuracy of that prediction, if accuracy is acceptable and prediction is correct, then machine learning algorithm gets deploy. But if accuracy is not acceptable then model again get trained until accuracy doesn't get accepted.

In machine learning whatever the processes are involved are similar to data mining and predictive modeling.
- Data mining is the process of finding the patterns from the large dataset in which machine learning, statistics, and database systems are used. It is also known as KDD(Knowledge discovery in databases).
- Predictive Modeling is the process of predicting future outcomes by using data mining and probability.

It uses mathematical module to find out hidden pattern from the data. This solves the real life problem or business problem. Machine Learning do complex task easily. For example face detection, self driving car, Alexa, etc.

Machine learning algorithms has three parts:-

- Input
- Output
- Objective Function or Performance matrix

The steps involved in machine learning are:-

- Problem Identification
- Data Collection and validation
- Model Building
- Feedback

1.2.1 MACHINE LEARNING MODELS

In Machine Learning, techniques and data are more important, but to identify the type of problem (supervised or unsupervised) is equally important.

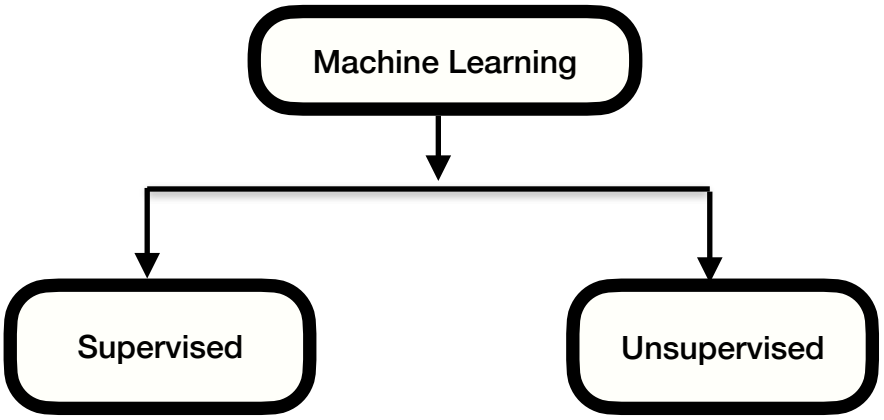


Fig 1.2.1 Types of Machine Learning

Supervised Learning:-

In this input and output are already available. On the basis of this input and output or labeled data model is created and on that model new input are given and checked whether valid output is coming or not. In this model machine knows the features.

Supervised Learning are,

- Already Tagged Data
- Features and labels present

- There are two types of supervised Learning:-
  - Classification:-
    - Classify which label a given set of features belongs to.
    - This algorithm is used when there are limited number of answers Like, yes or no, 0 or 1, true or false, etc.
      For example, Is it cold? -> Yes or No
    - In these algorithm we have only two outputs. So, when there are only two outputs (only two choices) then that type is called as 2 class classification algorithm. But if there are more than two choices then this type is called as multi class classification.
    - To solve the classification problem the algorithms required are,
      - Logistic regression
      - K-Nearest Neighbors (KNN)
      - Support Vector machine (SVM)
      - Naive Bayes
      - Decision Tree Classification
      - Random Forest Classification
  - Regression:-
    - Find out the value of the label using previous data.
      For example, What will be the temperature tomorrow?
      So the output will be suppose, 28.
    - To solve the regression problem the algorithms required are,
      - Linear Regression
      - Multiple Linear Regression
      - Polynomial Regression
      - Support Vector Regression
      - Decision Tree Regression
      - Random Forest Regression

Unsupervised Learning:-

All a machine knows is the data in front of it. No Features, No Labels. Machine seeing the data for the first time.

In this model,

- Data are not already tagged
- Features and labels not present
- Training not done
- Clustering:- Discover the inherit groupings in the data, such as growing customers by purchasing behaviour. There are two types of algorithm used for clustering and they are,
  - K-Means Clustering
  - Hierarchical Clustering
- Association:- Association rule learning problem, such as people that buy X also tend to buy Y. Algorithm used for association are,
  - Apriori
  - Eclat

To know how a problem cane solved using machine learning models, some of the examples are:-

1. Is this person is diabetic? (Yes or no)

Algorithm used will be Classification Algorithm.

2. Is this A or B?

Algorithm used will be Classification Algorithm

3. Is this weird?

Algorithm used will be Anomaly Detection Algorithm.

4. How many?

Algorithm used will be Regression Algorithms.

5. How is this organised?

Algorithm used will be Clustering Algorithm.

CHAPTER 2

PROPOSED METHODOLOGY

2.1 PROJECT DETAILS

It is a Data science project in which we need to predict whether a given data point is diabetic or not. Since it comes under Classification technique which is a type of Supervised Machine Learning, here the model is built using different classification algorithms like K-Nearest Neighbors, Naive Bayes , and Support Vector Machine. The model built with K-Nearest Neighbors gave high accuracy out of all in predicting whether the given data point is diabetic or not.

Libraries used in this project are:-

- **NumPy :**

- Here, Num means numeric and Py means python.
- It is a scientific computing library for python.
- It supports multi-dimensional array. It is used to represent large number of data in the form of array.
- NumPy Library is used for numeric calculation.
- To store data it uses less memory. It is very convenient and process fast.
- NumPy for Machine Learning:- Any machine learning model can't directly perform operation on images so for that images are converted into the numpy array format then machine learning can perform operations.

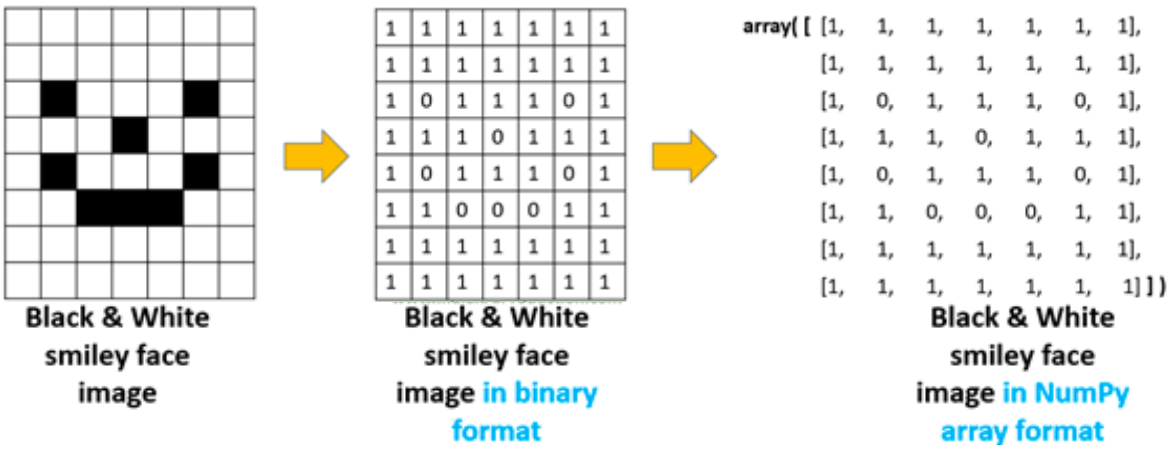


Fig 2.1.1 NumPy for Machine Learning

Here in fig., image is converted to binary format(numeric format) then in NumPy array format. Now machine learning model can perform operation on this format.

- **Pandas:**

- Pandas is a powerful python data analysis toolkit.
- Pandas is used for data manipulation, analysis and cleaning.
- Open Source
- For data manipulation it is a fast and efficient data frame.
- It can read different types of formats like, csv, JSON, etc.
- Importance of Pandas:- Suppose in data science projects, to find useful insights or patterns from raw data firstly raw data should be in proper format. This is known as data preprocessing. In data preprocessing panda libraries are used. To prepare and process the data this, library is very fast and efficient. This library have many features like, it can use for handling missing values, reshaping the datasets, etc.

- **SKLearn:**

Using SKLearn for machine learning:-

- Simple and efficient tools for data mining and data analysis.
- Accessible to everybody and reusable in various contexts.
- Built on Numpy, Scipy, and Matplotlib.
- Open source, commercially usable-BSD license.

- **Matplotlib:**

- Matplotlib is a 2D and 3D plotting python library.
- This library is used when there is a bulk of data at that time we can transform that data in a graphical representation and can analyse the data easily.
- It is use to create high quality graph.
- Matplotlib graphs are Histogram, Bar charts, Scatterplots, etc.

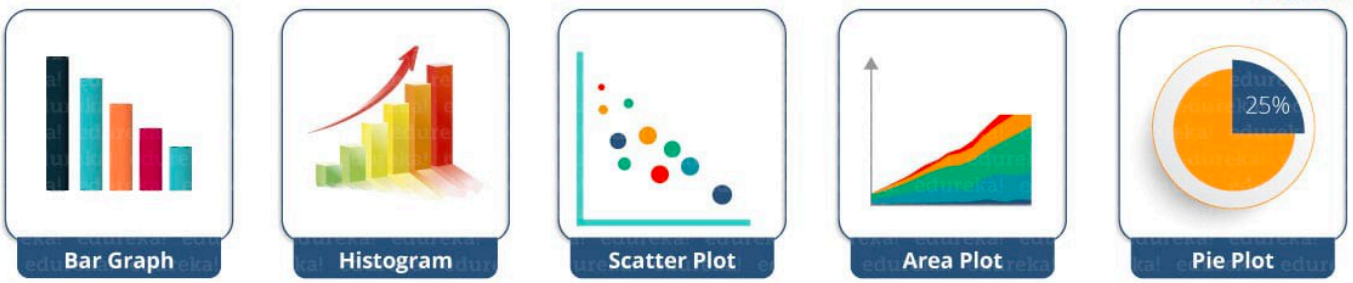


Fig 2.1.2 Matplotlib Plots

- Seaborn:
- It is a python library use for data visualization. This library is build using matplotlib library.
- It is use for statistical library.
- Seaborn graphics are Heatmap, Pair plot, Facet grid, etc.
- Dependency of Seaborn:- Seaborn library is dependent upon Python, NumPy, Pandas, Scipy, Matplotlib.

2.2 PROCESS FLOW

The steps involved in the machine learning process,

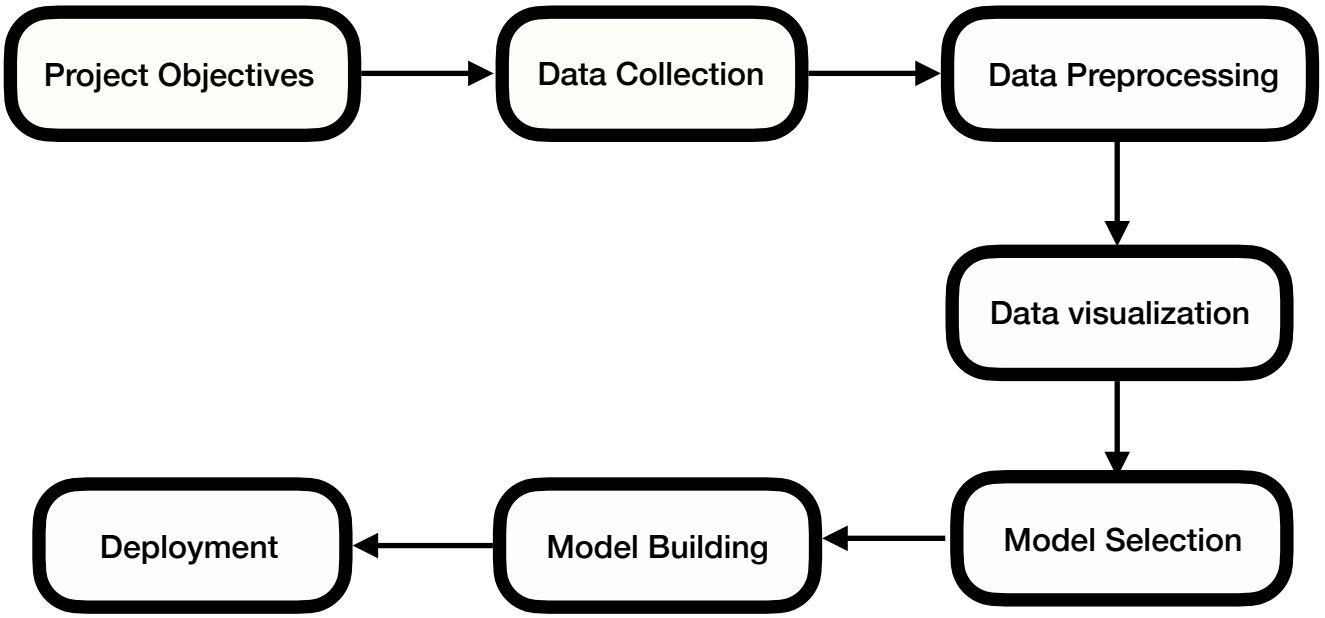


Fig 2.2.1 Process Flow

Step 1: Project Objectives

This is the first step in which business problems and requirements are specified. For example, In this project the objective is, to predict whether the person has Diabetes or not based on various features like Number of Pregnancies, Insulin Level, Age, BMI.

Step 2: Data Collection

In this process, the data required for project are collected or gathered.

There are two types of data collection:-

1. Primary data
2. Secondary data

Primary data :- This are the original data collected through researcher or user who is doing project. This is the first hand source of data collection.

Secondary data :- If the data is collected from a person other than the user then that data is called as secondary data. This type of data are passed through statistical process or some modification and given to user.

Now by using this process data are collected. But in this raw data their are some impurities like,

- Between integer value there are some text values,
- Some values are missing,
- Inconsistency in data i.e data is not in ordered form or not in proper format, and etc.

So because of these impurities or problem work can't be done on this data. Therefore, to overcome from this problem after the data collection process data preprocessing is done.

Step 3: Data Preprocessing

In this process, raw data or original data are purified. For data preprocessing the python libraries like NumPy, Panda, Sklearn, etc are used.

The steps in this process are,

1. Data Cleaning:

   In this step, missing data are filled using nan or 0. Any row who's heading or label is not given are ignored or removed. Any type of irrelevant data are removed. After this, noisy data (meaningless data or corrupted data) are deleted or removed.

2. Data Transformation:

   In this step, feature scaling is done. In this scikit-learn library is used. Normally for solving classification problem data transformation is used.

3. Dimensionality Reduction:

   In this step, dimensions are reduced like 3D data is converted to 2D data. So in this way data are analysed.

Step 4: Data Visualization

Data visualization is a graphical representation of data. Here, data are transformed to piecharts, graphs, bar graphs, histograms, etc. For data visualization python libraries like matplotlib, seaborn, etc are used. Values are created from data by doing Visualization. New patterns are discovered and to find out useful insights. This is like comparative analysis here data are compared and find out the relationships between x(independent variable) and y(dependent variable or target variable) variable.

Step 5: Model Selection

In this step, we have to select best machine learning model according to the data which will give us the accurate result. In simple language, we have to find out the which algorithm should be used for this particular data.

Suppose we have to predict the numeric value, then linear regression algorithm will be used or if we have to do classification then logistic regression algorithm can be used and so on. So basically perfect model or algorithms are selected according to the problem statement.

There are many machine learning models, but in this project we will see classification algorithms. Since this project (diabetes disease prediction) comes under Classification technique which is a type of Supervised Machine Learning, here the model is built using different classification algorithms and some of them are:

- Logistic regression :-

  This algorithm is used to solve the classification problems. The nature of the response variable is categorical. It helps to calculate the possibility of a particular event taking place.

  It takes features and training data (labels). It fits a linear model. Instead of giving the result, it gives the logistic of the result.

  Logistic regression is a sigmoid function and shape of the curve is S.
  The formula for Sigmoid curve is,

$$Y = 1 / 1+e^{-x}$$

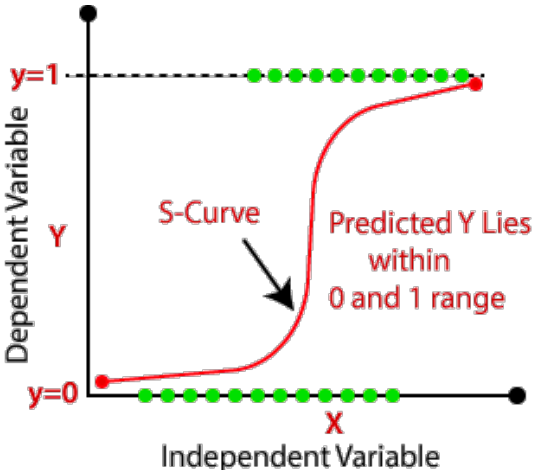


Fig 2.2.2 Logistic regression

  Applications of Logistic regression are weather prediction, disease prediction, image categorisation, and etc.

- K-Nearest Neighbors (KNN) :-

  It is a supervised learning algorithm that can perform both classification and regression task using k (numbers) neighbours (instances). In this,

firstly train sets fit into the model and then input the test set into the trained model to obtain the required results.

- Support Vector machine (SVM):

  It is a supervised learning algorithm that can perform both classification and regression task. Support vector machine makes sure that when hyperpath plane is created then two margin line is also created. This margin lines are created, have some distance so that it is easily linearly separable for both the classification points.

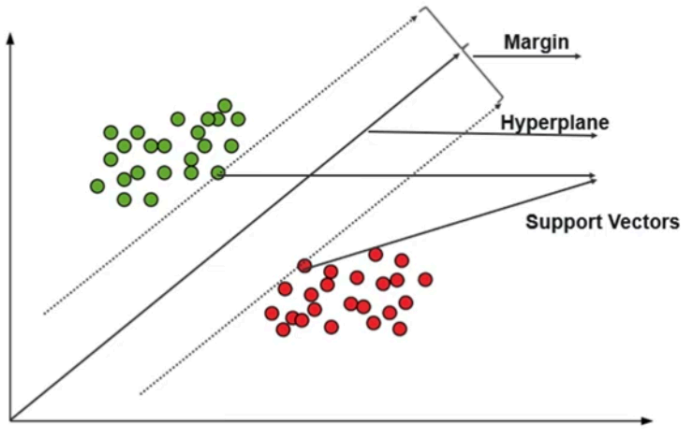


Fig 2.2.3 SVM

The main goal of the SVM algorithm is to develop the best decision boundary that can segregate n-dimensional space into classes. So that the new data points can be easily put in the correct category – Hyperplane.

- Naive Bayes:
  - It is a supervised learning algorithm in which features are independent.
  - In this, labels are already assigned and feature detection is defined in advance.
  - Classifiers are created by training them on labeled data.
  - It can solve both continuous and categorical value attributes.
  - It is a process of two step,
    - Model Construction and
    - Model Usage

- It is a classifier which uses Bayes Theorem. Bayes theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes theorem is stated mathematically as the following equation,

$$P(A/B) = P(B|A)\ P(A)/P(B)$$

Step 6: Model building

In this step, by using algorithm data are split into training and testing sets like 70-30, 80-30, and etc.

Suppose data are split into 70-30 then this means that,

- 70% part of the data are for training phase in which model get trained and
- remaining 30% data are for testing phase.

After this process, prediction is done from that model. If prediction is not accurate then again model gets trained and testing is done again.

This happen again and again until accurate prediction is not calculated. And in this way model improve itself according to desired output. Now model is ready for prediction and for solving problem.

Step 7: Model Deployment

This is the last stage of the machine learning life cycle. Machine learning model are deployed into production environment for taking decision.

CHAPTER 3

IMPLEMENTATION

3.1 PROBLEM STATEMENT:

The goal of the project is to,

1. Prepare the data-set using several methods to train the model.
2. Build a model which can give high accuracy of predicting the disease.

3.2 REQUIREMENTS:

- SOFTWARE:
    - Anaconda navigator's Jupiter notebook
- WEBSITE:
    - Here, Dataset is taken from kaggle site (https://www.kaggle.com/uciml/pima-indians-diabetes-database).
- LIBRARIES:
    - Pandas,
    - Numpy,
    - Matplotlib,
    - Seaborn,
    - Scikit Learn

3.3 IMPLEMENTATION STEPS:

Step by Step implementation of the project:-

Step 1:-

Import all the required libraries such as numpy, pandas, etc.

For example,

```
import numpy as np
```

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

Fig. 3.3.1 Libraries

This is the syntax one should import libraries like this. But this libraries should be installed first then import the libraries.

Step 2:-
Second step is to read the dataset and stored it into new variable.
Here, Dataset is taken from kaggle site (https://www.kaggle.com/uciml/pima-indians-diabetes-database).

In this, datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.
The following is the description of the 6 variables:

1. Pregnancies: It is no of pregnancies the patient has had which is in terms of numerical values.
2. Glucose: It is blood glucose levels.
3. Blood Pressure: It's Diastolic Blood pressure.
4. Skin Thickness: Triceps skin fold thickness in mm
5. Insulin: It’s the insulin levels in the patient blood.
6. BMI: Body Mass Index i.e., (weight in kg/(height in m)^2).
7. Diabetes Pedigree Function.
8. Age: In years.

9. **Outcome:** In 0 or 1 where 1 represents presents of diabetes and 0 shows there is no diabetes.

To read any dataset we need pandas library. In this project, dataset is in the csv format. Here csv means comma separated value and to read the data of csv format we use,

read_csv()

For example,

```
In [3]: # Lets read the datset

        data = pd.read_csv('diabetes.csv')
```

Fig 3.3.2 Screenshot of code

Pandas library are used to read any format of the dataset. Suppose if dataset is in excel form then,

read_excel().

Now after reading the dataset new variable will be assigned in which this dataset will be stored.

For example, in this project "data" variable is assigned,

data = pd.read_csv()

We assign this variable so that whenever we have to use the dataset we can call this variable.

Step 3:-

Next step is to check the shape of the dataset or dataframe.

For example,

Shape of the dataset : (768, 9)

This means, in dataset there are 768 rows and 9 columns.

Step 4:-

Now next step is to check if there is any missing values present in the dataset. Here there is no missing value present so we can move further.

Step 5:-

Now in this step we will perform the data visualization process. And will perform training and testing.

Step 6:-

Now Next step is to select the model. So, in this project we used the algorithms like K Neighbors Classifier, and Support Vector Classifier.

Step 7:-

After model building next step is to check the highest accuracy among all the algorithms used.

Step8:-

After finding highest accuracy, check whether the model selected is correct or not.

```
In [35]:
#Checking whether the model selected is correct or not

input_data=(4,110,92,0,0,37.6,0.191,30)


input_data_as_numpy_array=np.asarray(input_data)
input_data_reshaped=input_data_as_numpy_array.reshape(1,-1)


std_data=scaler.transform(input_data_reshaped)
print(std_data)

prediction=knn_classifier.predict(std_data)
print(prediction)

if (prediction[0]==0):
     print('the person is not diabetic')
else:
    print('the person is diabetic')
```

```
[[ 0.04601433 -0.34096773  1.18359575 -1.28821221 -0.69289057  0.71168975
  -0.84827977 -0.27575966]]
[0]
the person is not diabetic
```

Fig 3.3.3 Checking the Model

CHAPTER 4

CONCLUSION

In this project, all the steps were performed from Data collecting to Model deployment. In this, classification Learning algorithms and different libraries are used. During Model evaluation, machine learning algorithms on the basis of accuracy score were compared and finds the best one. So, the best accuracy was given by K- Nearest Neighbors (KNN). This model is perfectly predicting the person suffering from diabetes disease.

References

I. https://www.kaggle.com/uciml/pima-indians-diabetes-database

II. Vapnik, V..Statistical learning theory. 1998 (Vol. 3). .New York, NY: Wiley， 1998:Chapter 10-11, pp.401-492

III. Research on Diabetes Prediction Method Based on Machine Learning To cite this article: Jingyu Xue et al 2020 J. Phys.: Conf. Ser. 1684 012062