Diabetes Disease Prediction

EXPOSYS
DATA LABS

Presented
by
Richa Jha

Contents

INTRODUCTION

- 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.
- 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 :

  1) Diabetes Mellitus : Type 1 Diabetes , Type 2 Diabetes

  2) Gestational Diabetes

  3) Diabetes Insipidus

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.

- It is the art and science of drawing actionable insights from the data and these can be used in various industries but using it in Health industry need caution and proper data.

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.
- The steps involved in machine learning are:-
  - Problem Identification
  - Data Collection and validation
  - Model Building
  - Feedback

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

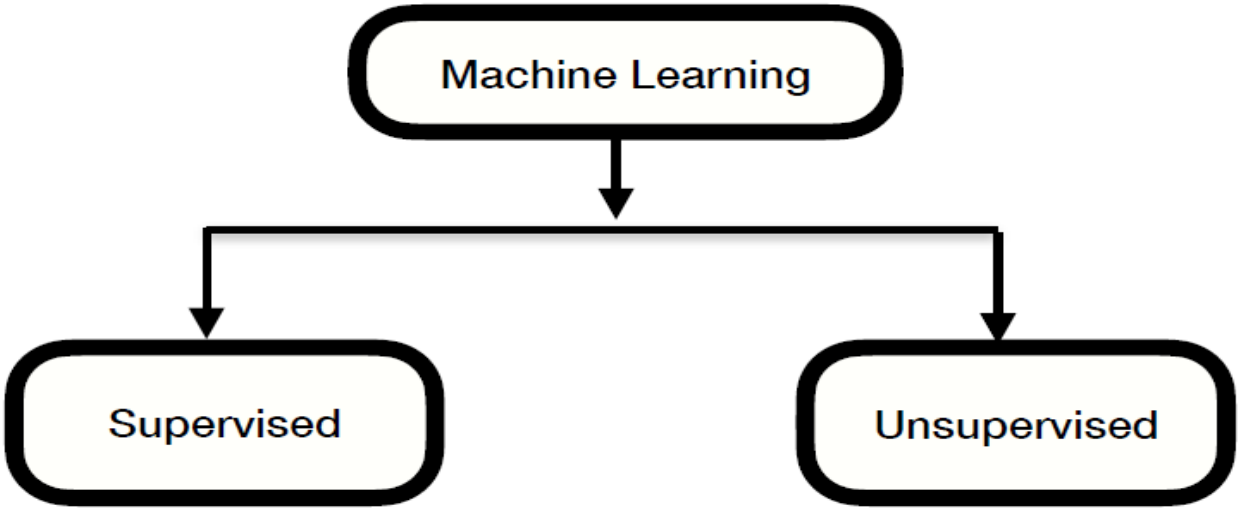

Fig 1.2.1 Types of Machine Learning

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.
- 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.

- **Regression:-**
  - Finds 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 behavior. 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

Proposed Methodology

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 Support Vector Machine gave high accuracy out of all

in predicting whether the given data point is diabetic or not.

Libraries Used :

- **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.

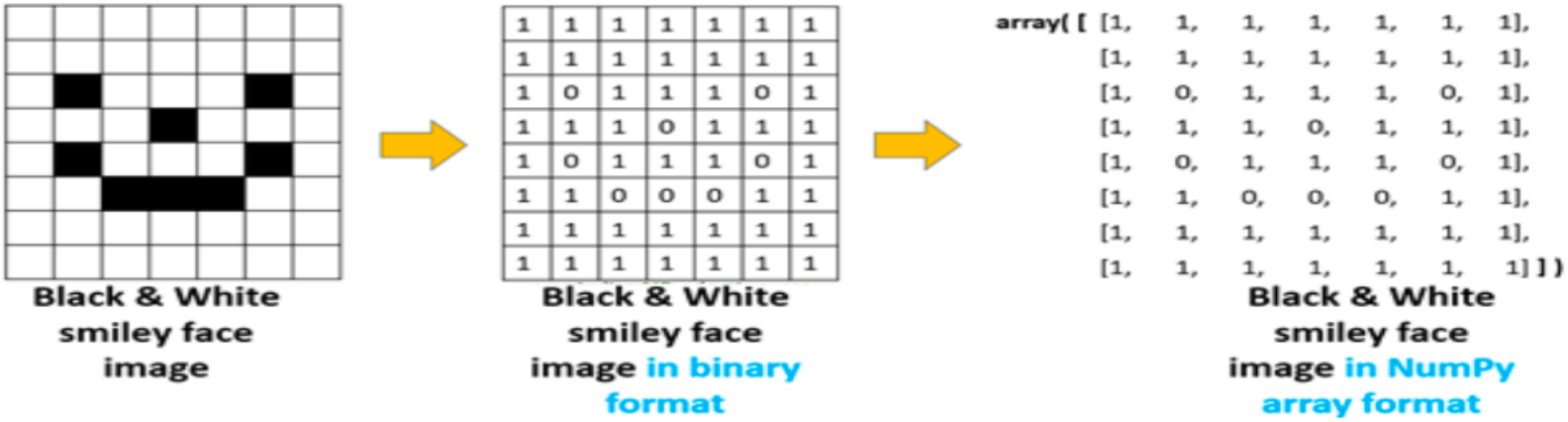

Fig 2.1.1 NumPy for Machine Learning

- **• Pandas:**

- Pandas is a powerful python data analysis toolkit.

- Pandas is used for data manipulation, analysis and cleaning.

- **• SKLearn:**

Using SKLearn for machine learning:-

- Simple and efficient tools for data mining and data analysis. –

-Accessible to everybody and reusable in various contexts.

- **• Matplotlib:**

- 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.

Fig 2.1.2 Matplotlib Plots

Process Flow

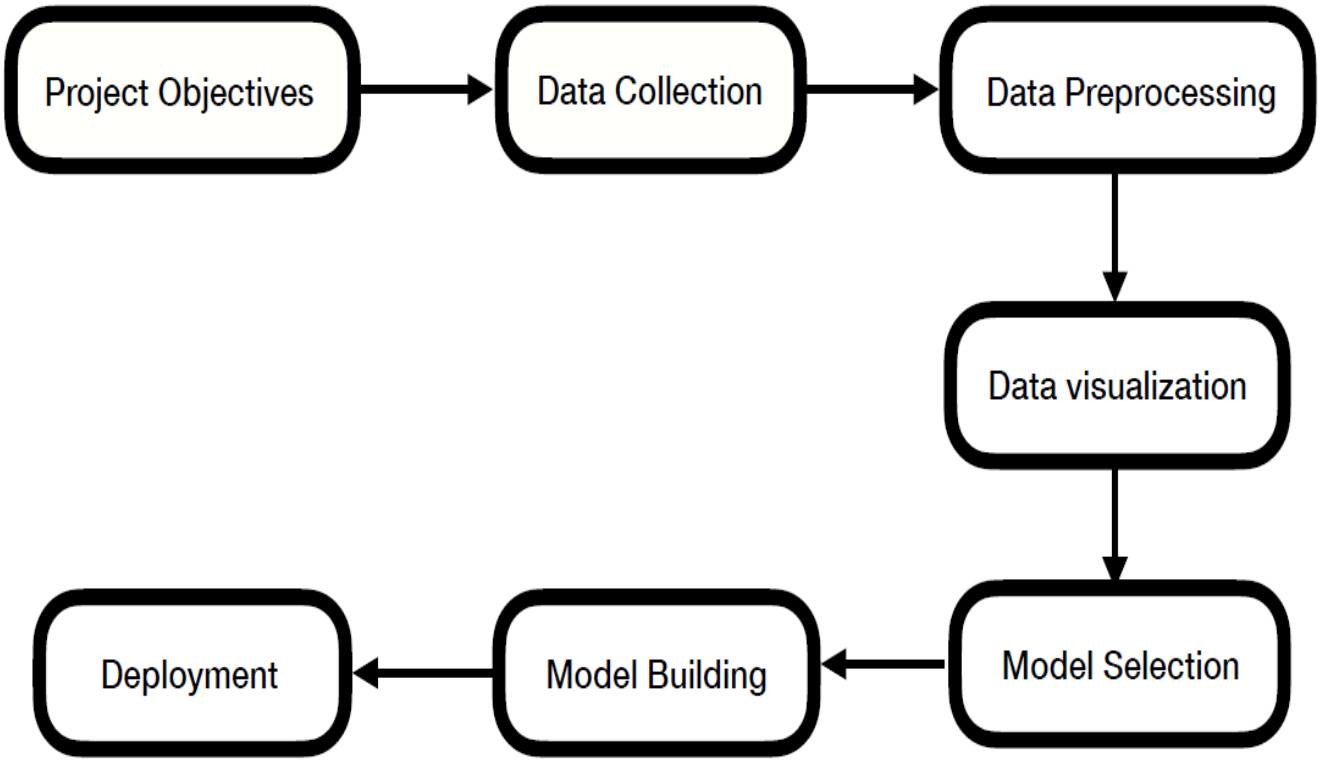

Fig 2.2.1 Process Flow

- **Step 1: Project Objectives :**

  This is the first step in which business problems and requirements are specified.

- **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

- **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.

- **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.

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.

- 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.

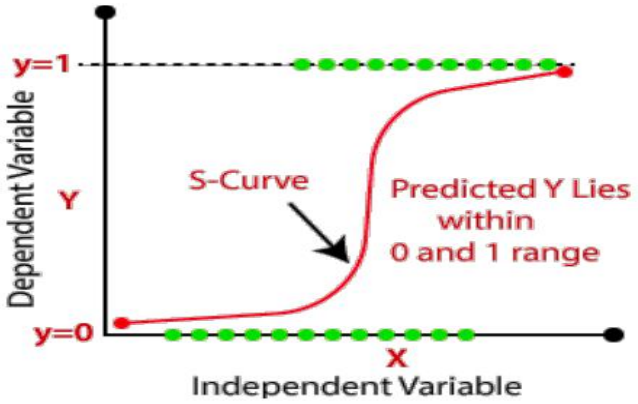

Fig 2.2.2 Logistic regression

- K-Nearest Neighbors (KNN) :-

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

- 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.

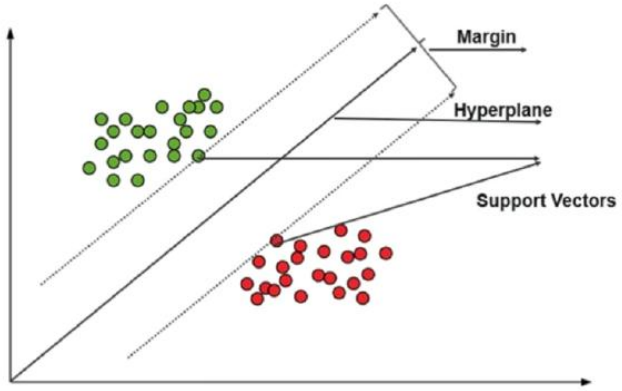

Fig 2.2.3 SVM

- **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.

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

- **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.

IMPLEMENTATION

- **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.

- **Requirements:**

  • **SOFTWARE:** Anaconda navigator's Jupiter notebook

  •**WEBSITE :** Here Dataset is taken from Kaggle site

  • **LIBRARIES**: Pandas, Numpy, Matplotlib, Seaborn, Scikit Learn

- **Implementation Steps :**
  - Step 1:- Import all the required libraries such as numpy, pandas, etc.
  - Step 2:- Second step is to read the dataset and stored it into new variable.
  - Step 3:- Next step is to check the shape of the dataset or dataframe.
  - 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, 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.

CONCLUSION

In this project, all the steps performed from Data collecting to Model deployment. In this project, classification Learning algorithms and different libraries are used. During Model evaluation, we compare machine learning algorithms on the basis of accuracy score and find the best one. So, the best accuracy was given by K-Nearest Neighbors (KNN). 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

THANK YOU!