A PROJECT REPORT ON *PREDICTION OF DIABETES* BY USING PROBABILISTIC MODEL OF MACHINE LEARNING

SUBMITTED BY:

Name: **PRATIK KUMAR**

University roll number: **10900116073**

Registration number: **161090110078 (2016-2020)**

Year: **3rd Year**

Department of Computer Science & Engineering
Netaji Subhash Engineering College
Garia, Kolkata – 700152

CONTENTS

1. ACKNOWLEDGEMENT
2. ABSTRACT
3. INTRODUCTION
    a. DATA MINING
    b. DIABETES
        i. TYPE OF DIABETES
            1. TYPE-1
            2. TYPE-2
    c. MACHINE LEARNING
        i. SUPERVISED LEARNING
        ii. SEMI-SUPERVISED LEARNING
        iii. REINFORCEMENT LEARNING
4. PROJECT OBJECTIVE
5. PROBLEM STATEMENT & DESCRIPTION
    a. DATASETS DESCRIPTION
    b. RESULT
6. HARDWARE & SOFTWARE REQUIREMENTS
    a. HARDWARE USED
    b. SOFTWARE USED
7. LOGISTIC REGRESSION ALGORITHM
8. HOW IT WORKS
9. ADVANTAGES OF LOGISTIC REGRESSION ALGORITHM
10. DISADVANTAGES OF LOGISTIC REGRESSION ALGORITHM
11. CODE
12. PLOTS
13. SUMMARY
14. EXPERIMENT RESULTS
15. CONCLUSION
16. BIBLIOGRAPHY

Acknowledgement

The achievement that is associated with the successful completion of any task would be incomplete without mentioning the names of those people whose endless cooperation made it possible. Their constant guidance and encouragement made all our efforts successful.

We take this opportunity to express our deep gratitude towards our project mentor, **Mr. Rictor Bhowmic** for giving such valuable suggestions, guidance and encouragement during the development of this project work.

Last but not the least we are grateful to all the faculty members for their support.

Thank you!

ABSTRACT

The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. With the emerging increase of diabetes, that recently affects around 346 million people, of which more than one-third go undetected in early stage, a strong need for supporting the medical decision-making process is generated.

Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Diabetes is ascribed to the acute conditions under which the production and consumption of insulin is disturbed in the body which consequently leads to the increase of glucose level in the blood.

Using data mining methods to aid people to predict diabetes has gain major popularity. In this project, Logistic Regression is used to predict the persons whether 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 graphic way. The dataset used is Pima Indian Diabetes dataset, which collects the information of persons with and without diabetes.

INTRODUCTION

Diabetes:-Diabetes is a chronic disease that occurs when the human pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces, which leads to an increase in blood glucose levels. Generally a person is considered to be suffering from diabetes, when blood sugar levels are above normal.

Types of Diabetes:-The three main types of diabetes are described below:

Type 1 – In this type of diabetes, the pancreatic cells that produce insulin have been destroyed by the defense system of the body.

Type 2-In this case the various organs of the body become insulin resistant, and this increases the demand for insulin. Gestational diabetes – It is a type of diabetes that tends to occur in pregnant women due to the high sugar levels as the pancreas don’t produce sufficient amount of insulin. Controlling the blood glucose level of diabetic patients and keeping it within the normal range (70 mg/dL -120 mg/dL) is therefore the focal goal of physicians.

MACHINE LEARNING:-

Machine learning is a branch of computer science that consists of algorithms that can learn from data, it provides set of methods that can detect patterns in the data and use the patterns to generate future predictions.

Because of new computing technologies, machine learning today is not like machine learning of the past. With the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not.

Furthermore, predicting the disease early leads to treating the patients before it becomes critical.The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results.

There are three core types of machine learning- supervised learning, unsupervised learning, and reinforcement learning.

1. **Supervised Learning:-**The main goal in supervised learning is to learn a model from labeled training data that allows us to make predictions about unseen or future data. Here, the term **supervised** refers to a set of samples where the desired output signals (labels) are already known.

2. **Semi-supervised Learning:-**It uses both labeled and unlabeled data for training-typically a small amount of labeled data and large amount of unlabeled data(because unlabeled data is less expensive and take less effort to acquire).

3. **Reinforcement Learning:-**This learning used forrobotics,gaming and navigation.With reinforcement learning,the algorithm discovers through trial and error which actions yield the greatest reward.

PROJECT OBJECTIVE:-

- 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 harmones in the body.

- It aids the body in converting sugar ,starches and other food items into energy needed for daily life.

- However if the body does not produce a properly used insulin the redundant amount of sugar will be driven out by urination

- This disease is refer to diabetes.The cause of diabetes is mystery,although obesity and lack of exercise appear to possibly play significant roles.

- In early the ability to diagnose diabetes plays an important role for the patient’s treatment process.

Problem Statement and Description :

Prediction of diabetes using Logistic Regression : To identify whether a given person in dataset will be diabetic, non diabetic or pre-diabetic will be done on basis of attribute values.

Dataset contains all the details of person like fast gtt value, casual gtt value, number of time pregnant, diastolic blood pressure (mmhg), triceps skin fold thickness(mm), serum insulin(µU/ml), body mass index (kg/m), diabetes pedigree function and age of person.

Attributes like fast gtt, casual gtt, diastolic blood pressure values exceeding a specific value may contribute to identify whether a person is diabetic, non diabetic or prediabetic.

The aim of prediction of diabetes is to make aware people about diabetes and what it takes to treat it and gives the power to control. It makes necessary chances to improve lifestyle. classification evaluation for the prediction performance of Bayesian algorithms to predict diabetes. The proposed Logistic Regression will predict the persons having diabetic or not.

HARDWARE & SOFTWARE REQUIREMENTS

HARDWARE USED

1. Intel Core i3 (2nd generation, 2.4 GHz, Cache 3M)
2. 2gb DDR3 Ram
3. Hard Disk
4. Intel HD Graphics

SOFTWARE USED

1.windows
2.Anaconda

The Logistic Regression Algorithm

Logistic Regression is one of the most used Machine Learning algorithms for binary classification. It is a simple Algorithm that you can use as a performance baseline, it is easy to implement and it will do well enough in many tasks. Therefore every Machine Learning engineer should be familiar with its concepts. The building block concepts of Logistic Regression can also be helpful in deep learning while building neural networks.

Like many other machine learning techniques, it is borrowed from the field of statistics and despite its name, it is not an algorithm for regression problems, where you want to predict a continuous outcome. Instead, Logistic Regression is the go-to method for binary classification. It gives you a discrete binary outcome between 0 and 1. To say it in simpler words, it’s outcome is either one thing or another.

A simple example of a Logistic Regression problem would be an algorithm used for diabetes detection that takes an input and should tell if a patient has diabetes (1) or not (0).

How it works

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using it’s underlying logistic function.

These probabilities must then be transformed into binary values in order to actually make a prediction.

Advantages of logistic regression:

- it is more robust: the independent variables don’t have to be normally distributed, or have equal variance in each group.
- It does not assume a linear relationship between the IV and DV.
- It may handle nonlinear effects.
- You can add explicit interaction and power terms.
- The DV need not be normally distributed.
- There is no homogeneity of variance assumption.
- Normally distributed error terms are not assumed.
- It does not require that the independents be interval.
- It does not require that the independents be unbounded.

Disadvantages of logistic regression:

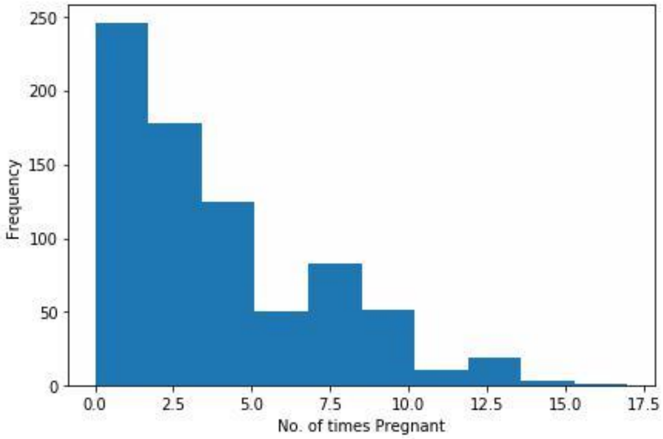
- Identifying Independent Variables
- Limited Outcome Variables
- Independent Observations Required
- Overfitting the Model

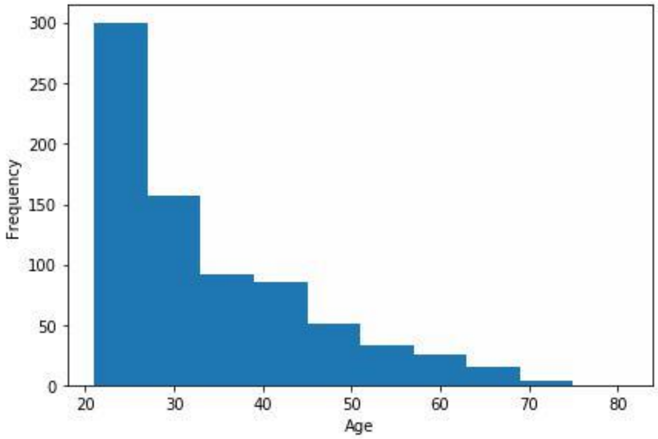
Code:

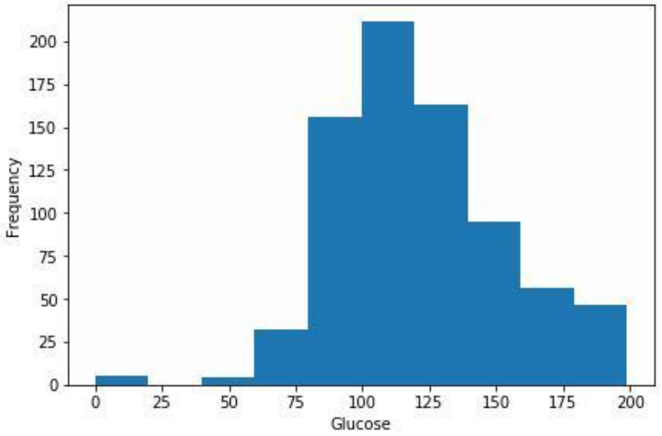
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score
from sklearn.model_selection import learning_curve

def plot_learning_curve(estimator, title, X, y, train_sizes, ylim=None, cv=None,
                        n_jobs=None):

    plt.figure()
    plt.title(title)
    if ylim is not None:
        plt.ylim(*ylim)
    plt.xlabel("Training examples")
    plt.ylabel("Score")
    train_sizes, train_scores, test_scores = learning_curve( estimator,
        X, y, cv=cv, n_jobs=n_jobs, train_sizes=train_sizes)
    train_scores_mean = np.mean(train_scores, axis=1)
    train_scores_std = np.std(train_scores, axis=1)
    test_scores_mean = np.mean(test_scores, axis=1)
    test_scores_std = np.std(test_scores, axis=1)
    plt.grid()

    plt.fill_between(train_sizes, train_scores_mean -
                     train_scores_std, train_scores_mean +
                     train_scores_std, alpha=0.1, color="r")
    plt.fill_between(train_sizes, test_scores_mean - test_scores_std,
                     test_scores_mean + test_scores_std, alpha=0.1, color="g")
    plt.plot(train_sizes, train_scores_mean, 'o-',
             color="r", label="Training score")
    plt.plot(train_sizes, test_scores_mean, 'o-',
             color="g", label="Cross-validation score")

    plt.legend(loc="best")
    return plt


df = pd.read_csv('/content/diabetes.csv')
X = df.iloc[:, : -1]
y = df.iloc[:, -1]
```

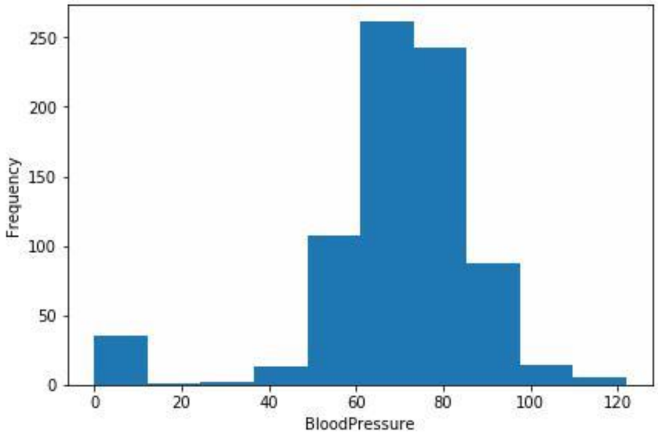
```
#Plotting
plt.figure(figsize=(15,10))
plt.subplot(2, 2, 1)
plt.hist(df['Pregnancies'], bins = 10)
plt.xlabel('No. of times Pregnant')
plt.ylabel('Frequency')

plt.subplot(2, 2, 2)
plt.hist(df['Age'], bins = 10)
plt.xlabel('Age')
plt.ylabel('Frequency')

plt.subplot(2, 2, 3)
plt.hist(df['Glucose'], bins = 10)
plt.xlabel('Glucose')
plt.ylabel('Frequency')

plt.subplot(2, 2, 4)
plt.hist(df['BloodPressure'], bins = 10)
plt.xlabel('BloodPressure')
plt.ylabel('Frequency')

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state
= 0)

#Random forest
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier()
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
acc = round(accuracy_score(y_test, y_pred), 5)
print("Random Forest Classifier accuracy:", acc)

plot_learning_curve(classifier, "Learning Curve", X_train, y_train,[1,10, 20, 50,
100, 200, 300, 400], (0.7, 1.01))

plt.show()
```

Plots:

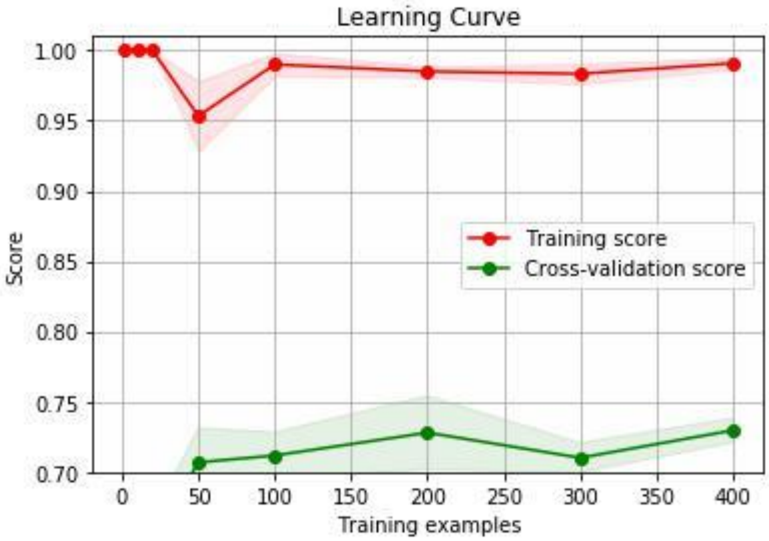








Learning Curve:



Summary:

- Load previous datasets to the system.
- Data pre-processing has done using integrating WEKA tool.
- Following operations are performed on the dataset after that.
    a. Replace Missing values.

    b. Normalization of values.

- 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.

EXPERIMENT RESULTS

Datasets Description:

The dataset that is taken for this work is collected from "Pima Indians Diabetes Database" obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of 768 records of patient data. Here 80% of the data is taken for training and remaining 20% is taken for testing.

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour postload plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care).

Results:

The problem of work is about predicting whether a person is diabetic or non diabetic in a dataset by applying Logistic Regression . This problem is solved using the primary attribute .

The dataset variables which are used for prediction of diabetes are fast plasma glucose concentration in an oral glucose tolerance test ,casual plasma glucose tolerance test and diastolic blood pressure (mmHg) is decision variable.

On validating the test records we obtain accuracy 65%. Where the TP = 86, TN = 30, FP = 29 and FN = 24. From the above values we obtained precision and recall as 0.75324% and 0.75%. Finally we estimate the F measure as 0.75%.

CONCLUSION:

Lately, medical machine learning has gained in interest by the scientific and research communities. Diabetes is considered as the world's fastest-growing chronic disease. It needs continuous self-management and control to maintain blood glucose level within the normal range, in order to prevent complications and prevent diabetic events.

Diabetic is a condition that occurs when blood glucose is too low. The occurrence of diabetic may result in seizures, unconsciousness, and possibly permanent brain damage or death. We proposed a model in predicting diabetes by applying data mining technique.

Diabetes mellitus is a chronic disease and a major public health challenge worldwide. Using data mining methods to aid people to predict diabetes has gain major popularity. In this Logistic Regression is proposed to predict the persons whether diabetic or not.

Results have been obtained. For future work, more input features can be used, e.g. exercise, heart rate, and metabolism rate. In addition, drawn blood samples of plasma insulin are needed to compare with the simulated values. Moreover, we recommend the proposed models to be tested on a larger dataset.

LIMITATIONS:

The logistic regression model has been used for prediction of diabetes. In this model, we have achieved accuracy level of approximate 75% , using different model could help in increasing the accuracy level.

If dataset could be more refined, the accuracy level could have increased.

FUTURE SCOPE:

This project has accuracy of 75.324% and it can be increased by using different approximations in future.

BIBLIOGRAPHY:

□https://towardsdatascience.com/the-logistic-regression-algorithm-75fe48e21cfa

□https://www.tutorialspoint.com/r/index.htm

□https://www.r-bloggers.com/

□https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36?gi=990a4b600b62