**DIABETIC PREDICTION USING MACHINE LEARNING AND PYTHON PROGRAMMING LANGUAGE**

Submitted By

Harshit Kumar

Dhiraj Kumar

Dinkar Kumar chaubey

Danish Shafi

Project Guide : Sofikul mallick



Start Date : 18th August 2020

End Date : 19th September 2020

CERTIFICATE

This is to certify that Harshit Kumar , Dhiraj Kumar, Dinkar Kumar chaubey and Danish Shafi successfully completed the project titled "Diabetic Prediction using Machine learning and python programming language " under my supervision during the period from 18 August 2020 to 19 September 2020 which is in fulfillment of their training in “Diabetic prediction using machine learning and python programming language .

Signature of the Supervisor

Date :

CONTENTS:-

1. Acknowledgment

2. Abstract

3. Introduction

3.1 problem statement

3.2 goal of competition

4. Required statistical concepts

5. Methodology

5.1 SVM

5.2 decision tree

5.3 KNN

5.4 logical Regression

5.5 Random Forest

6.Data Set

7. Project Implementation

7.1 Library(Pandas)

7.2 Library(Matplotlib)

7.3 Library (Seaborn)

7.4 Library (Sklearn)

8.Result

9.Conclusion

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 **Sofikul Mallick**, for giving such valuable suggestions, guidance and encouragement during the development of this project work.

ABSTRACT

The report presents the three tasks completed during Pandemic situation at Ardent Computech Pvt. Ltd. through online class Which are listed below:

1. Understand of the Problem objective & business implication

2. Understanding the data & build the model .

3. Evaluation of the model.

All these tasks have been completed successfully and results were according to Expectations. All the tasks were need very systematic approach, starting from the behavior of the data to the application of the algorithm and till evaluation of the model. The most challenging task was the domain knowledge, to understand the behavior of the data. Once the data has been prepared, we applied statistical algorithm for model building. It is one of the major area and really need very fundamental and conceptual knowledge of Advanced Statistics.

Introduction

**Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can access data and use it learn for themselves. **Python** offers concise and readable code. While complex algorithms and versatile workflows stand behind **machine learning** and AI, **Python's** simplicity allows developers to write reliable systems. Developers get to put all their effort into solving an ML problem instead of focusing on the technical nuances of the language.

Problem Statement

The datasets consists of several medical predictor variables and one target variable i.e CV(class variable). Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old .

Goal of Competition :

We need to build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not?

Required statistical concepts for this project:

Five **machine** getting to know **algorithms** namely SVM, selection Tree, NaiveBayes , Logistic Regression and KNN are **used** to hit upon **diabetes**. This may be capable of predict the chance degrees of **diabetes** and gives the first-class getting to know set of rules **with** better accuracy comparatively different **algorithms**.Diabetes isn't a hereditary disorder however heterogeneous group of disorder which could ultimately result in an boom of glucose within the blood and lack of glucose inside the urine. Diabetes is typically resulting from genetics, way of life and surroundings. Eating an dangerous weight loss plan, being overweight play role in developing the diabetes. High blood sugar tiers can also result in kidney diseases, coronary heart illnesses. The excess of sugar in the blood can harm the tiny blood vessels in your frame.

METHODOLOGY

Algorithms used

1.SVM

2 Decision Tree

3 KNN

4 Logistic Regression

5 Random Forest

Svm :

SVM is the standard supervised learning algorithm. It does the complex data transformations and separates the data based on the outputs and it can be used for both classifications and regression challenges. In SVM there are different hyper planes which divides the data .wehave to select the hyper plane which divides the class better. To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin. If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to select the class which has the high margin. The performance of SVM algorithm version for prediction of diabetes the use of confusion matrix is as follows..

|  |  |
| --- | --- |
| Tested Positive | 70 |
| Tested Negative | 141 |

DECISION TREE :

Decision tree set of rules is a supervised studying algorithm. It is used to remedy the type problems. In this algorithm whole facts is represented inside the shape of tree in which every leaf is corresponds to the class label and attribute are corresponds to inner node of the tree. The fundamental venture is to discover the foundation node in each stage. The overall performance of choice Tree set of rules model for prediction of diabetes is as follows

|  |  |
| --- | --- |
| Tested Positive | 79 |
| Tested negative | 113 |

KNN:

KNN is the one of the best set of rules used in gadget getting to know. In this set of rules the complete dataset is sorted. The data is divided into classes, if other data is want to classify then it finds the neighbours of that element based on the majority number of votes for the label . Initialize the data it calculates the distance between the classes and finding neighbours and voting for labels. The overall performance of KNN set of rules version for prediction of diabetes is as follows.

|  |  |
| --- | --- |
| Tested Positive | 44 |
| Tested Negative | 148 |

Logistic Regression :

Logistic regression is a machine getting to know classifier. This set of rules is used to split the observations for discrete classes. The outputs given by using the logistic regression is based totally on the opportunity feature. It uses the fee function that's known as as ―sigma‖ characteristic. Sigma function is more complex than the normal linear function. Logistic regression limit the cost function value between 0 to 1.

Random forest :

Random forest is a supervised system getting to know set of rules. It's also used to remedy classification and regression additionally. In this algorithm it consists of the trees. The number of tree structures present in the data is directly proportional to the accuracy of the result. Each internal node within the tree corresponds to an attribute and every leaf node represents class label.

|  |  |
| --- | --- |
| Total positive | 68 |
| Total negative | 124 |

Confusion matrix for Random forest

Data set

The statistics set is taken from Pima Indians Dataset Database (PIDD) that's to be had at UCI device studying. The statistics set has many impartial variables along with Glucose, Blood pressure, skin Thickness, BMI etc. Records set is trained to get the accurate end result and similarly it is tested.

|  |  |  |
| --- | --- | --- |
| Data base | No. of traits | No. of occurance |
| PIDD | 8 | 769 |

Accuracy is measured by the formula given below

ACCURACY = TP + TN / TP + FP + TN + FN Where, TP: True Positive TN: True Negative FP: False Positive FN: False Negative.

Result :

In the given Table we can see the algorithm which gives the better accuracy on the data set.

|  |  |  |
| --- | --- | --- |
| Algorithms | Accuracy | Misclassification |
| SVM | 73.43 | 26.5 |
| Decision Tree | 72.91 | 27.08 |
| Random Forest | 74.4 | 25.5 |
| KNN | 71.3 | 28.64 |
| Logistic Regression | 72.39 | 27.60 |

Conclusion

At last by using all these five machine learning algorithms we had measured different parameters within the dataset and we had came through better accuracy rate with random forest with nearly 75%. This work can be extended by adding any other algorithm which can give better accuracy than random forest…