Machine Learning Project

(2020-2021)

Predicting Type II Diabetes using Medical Records

Final Report

Institute of Engineering and Technology

Submitted by:

Tushar Saxena(181500762)

Vipul(181500801)

Sakshi Bhardwaj(181500608)

Umesh Pratap Singh(181500767)

Submitted To:

Mrs. Pooja Pathak

Department of Computer Science & Technology

Abstract

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic centre and consulting doctor. But the rise in machine learning approaches solves this critical problem. The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore, three machine learning classification algorithms namely Decision Tree, Logistic Regression and Naive Bayes are used in this experiment to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. The performances of all the three algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show Naive Bayes outperforms with the highest accuracy of 74.89% comparatively other algorithms.

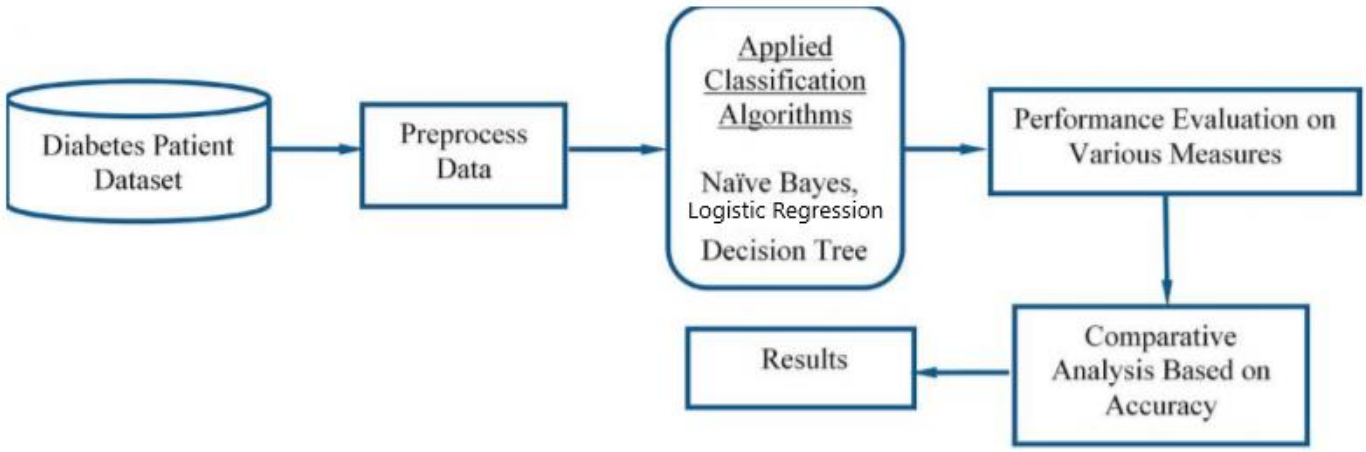
Introduction

Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier. Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if diabetes remains untreated. Some of the severe complications include diabetic ketoacidosis and nonketotic hyperosmolar coma. Diabetes is examined as a vital serious health matter during which the measure of sugar substance cannot be controlled. Diabetes is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. The early identification is the only remedy to stay away from the complications. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms works better in diagnosing different diseases. Data Mining and Machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study. This research work focuses on pregnant women suffering from diabetes. In this work, Naive Bayes, Logistic Regression, and Decision Tree machine learning classification algorithms are used and evaluated on the PIDD dataset to find the prediction of diabetes in a patient. Experimental performance of all the three algorithms are compared on various measures and achieved good accuracy.

Methodology Used

Model Diagram:

Proposed procedure is summarized in figure below in the form of model diagram. The figure shows the flow of the research conducted in constructing the model.

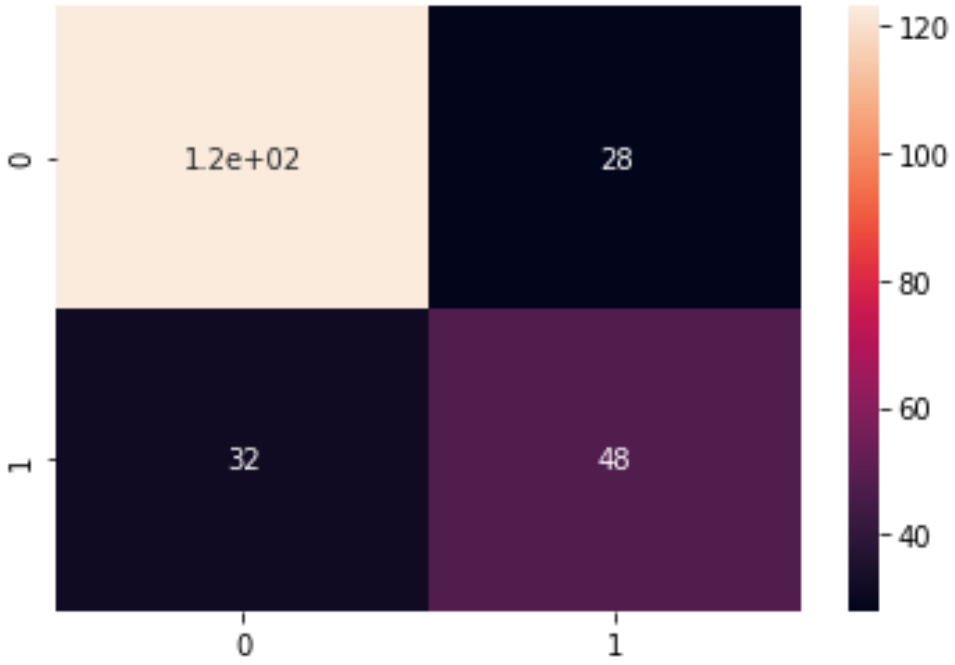


Brief Description of Algorithm used:

Logistic Regression:

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 [1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labelled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labelling; the function that converts log odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probity model; the defining characteristic of the logistic model is that increasing one of the independent

variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.



Naive Bayes Classifier:

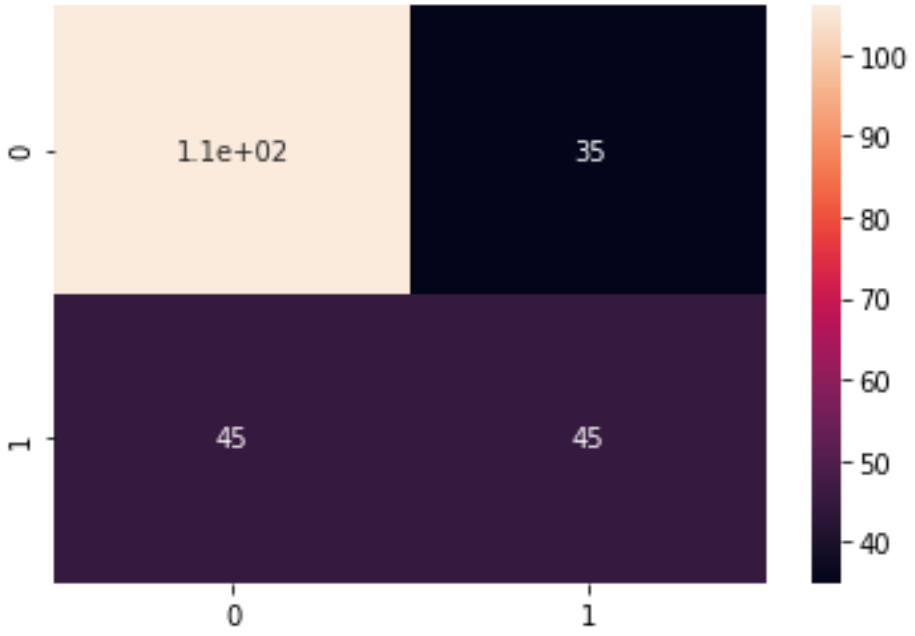
Naive Bayes is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose. It works well for the data with unbalancing problems and missing values. Naive Bayes is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability P (C X) can be calculated from P(C), P(X) and P (X C). Therefore,

P(C X) = (P(X C) P(C))/P(X) .

P(X C) = predictor class's probability.

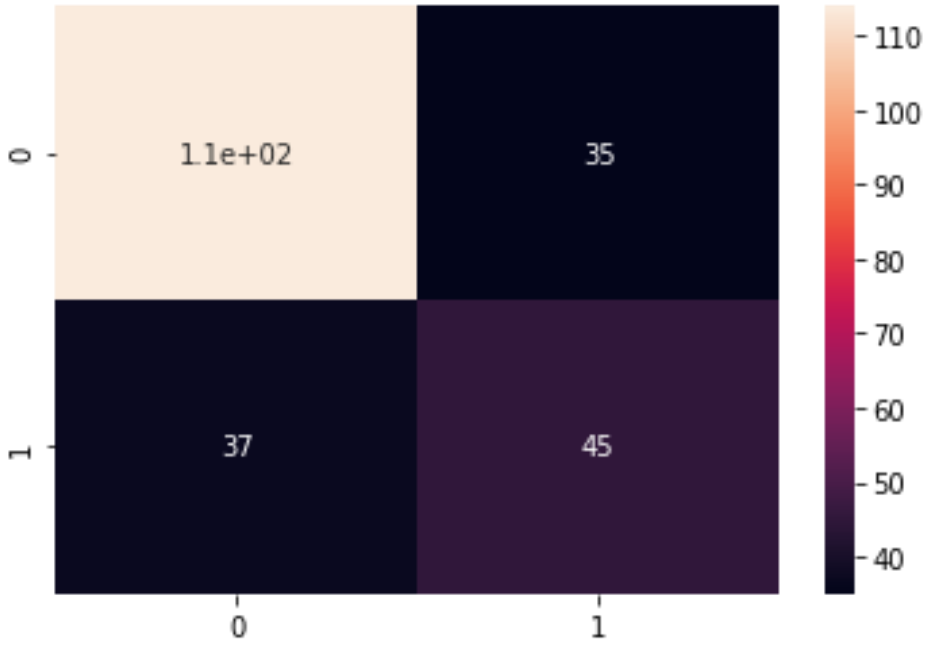
P(C) = class C's probability being true.

P(X) =predictor's prior probability.



Decision Tree Classifier:

Decision Tree is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes.



Dataset Used

In this work WEKA tool, is used for performing the experiment. WEKA is a software which is designed in the country New Zealand by University of Waikato, which includes a collection of various machine learning methods for data classification, clustering, regression, visualization etc. One of the biggest advantages of using WEKA is that it can be personalized according to the requirements. The main aim of this study is the prediction of the patient affected by diabetes using the WEKA tool by using the medical database PIDD. Table-4 shows a brief description of the dataset.

Database	No. of Attributes	No. of Instances
PIDD	8	768

PIDD-Pima Indians Diabetes

Dataset The proposed methodology is evaluated on Diabetes Dataset namely (PIDD) , which is taken from UCI Repository. This dataset comprises of medical detail of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class ’0’ treated as tested negative for diabetes and value of another class ’1’ is treated as tested positive for diabetes. Dataset description is defined by Table-4 and the Table-5 represents Attributes descriptions.

Accuracy Measures

Naive Bayes, Logistic Regression and Decision Tree algorithms are used in this research work. Experiments are performed using internal cross-validation 10-folds. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve) measures are used for the classification of this work. Table-5 defines accuracy measures below:

Table 5. Accuracy Measures

Measures	Definitions	Formula
1. Accuracy (A)	Accuracy determines the accuracy of the algorithm in predicting instances.	A=(TP+TN) / (Total no of samples)
2. Precision (P)	Classifier¢¢s correctness/accuracy is measured by Precision.	P = TP / (TP+ FP)
3. Recall (R)	To measure the classifier¢¢s completeness or sensitivity, Recall is used.	R =TP / (TP+FN)
4. F-Measure	F-Measure is the weighted average of precision and recall.	F=2*(P*R) / (P+R)
5. ROC	ROC (Receiver Operating Curve) curves are used to compare the usefulness of tests.	

Table 6. Comparative Performance of Classification Algorithms on Various Measures.

Classification Algo.	Precision	Recall	F-Measure	Accuracy%
Logistic Regression	0.63	0.6	0.61	74.02%
Naïve Bayes	0.63	0.6625	0.6463	74.89%
Decision Tree	0.51	0.6	0.55	66.6%

Results

Table-6 represents different performance values of all classification algorithms calculated on various measures. From Table-6 it is analysed that Naive Bayes showing the maximum accuracy. So the Naive Bayes machine learning classifier can predict the chances of diabetes with more accuracy as compared to other classifiers.

Notebook Link: https://github.com/Tusharsaxena3112/Ml-Diabetes-Project/blob/master/Diabetes/Diabetes_Prediction.ipynb.

Conclusion

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of disease like diabetes. During this work, three machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on Pima Indians Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 74.89 % using the Naive Bayes classification algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved forth automation of diabetes analysis including some other machine learning algorithms.

References

1. https://towardsdatascience.com/the-complete-guide-to-classification-in-python-b0e34c92e455.

2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232260/#:~:text=(2015)%20proposed%20a%20machine%20learning,and%20they%20get%20preferable%20results.&text=So%20in%20this%20study%2C%20we,network%20to%20predict%20the%20diabetes.

3. https://towardsdatascience.com/building-a-machine-learning-classifier-model-for-diabetes-4fca624daed0.