EXPOSYS DATA LABS

Domain: DATA SCIENCE

Topic: PREDICTION OF DIABETES

Project by

SHOBHAN AKSHAY GIRIDHARAN

SAMUEL WILLIAM ROBERT

SNEHA JAYWANT SABLE

ABSTRACT

Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are Decision Tree (DT), XG Boost, Support Vector Machine (SVM) and Random Forest (RF). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively. Our Result shows that Random Forest achieved higher accuracy compared to other machine learning techniques.

Keywords: Diabetes, Machine, Learning, Prediction, Dataset

TABLE OF CONTENTS

INTRODUCTION

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin.

According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world.

Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly.

Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction

EXISTING METHOD

Categories of Increased Risk for Diabetes (Pre-diabetes) Recommendations Testing to assess risk for future diabetes in asymptomatic people should be considered in adults of any age who are overweight or obese (BMI ≥25 kg/m2 or ≥23 kg/m2 in Asian Americans) and who have one or more additional risk factors for diabetes. For all patients, particularly those who are overweight or obese, testing should begin at age 45 years. If tests are normal, repeat testing carried out at a minimum of 3-year intervals is reasonable. To test for diabetes, the A1C, FPG, and 2-h PG after 75-g OGTT are appropriate. In patients with diabetes, identify and, if appropriate, treat other CVD risk factors. Testing to detect type 2 diabetes should be considered in children and adolescents who are overweight or obese and who have two or more additional risk factors for diabetes.

There are various causes of type 2 diabetes. Although the specific etiologies are not known, autoimmune destruction of β-cells does not occur, and patients do not have any of the other known causes of diabetes. Most, but not all, patients with type 2 diabetes are obese. Obesity itself causes some degree of insulin resistance. Patients who are not obese by traditional weight criteria may have an increased percentage of body fat distributed predominantly in the abdominal region.

PROPOSED METHOD WITH ARCHITECTURE

Goal of the paper is to investigate for model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

SR NO.	ATTRIBUTES
1	Pregnancy
2	Glucose
3	Blood pressure
4	Skin thickness
5	Insulin
6	BMI
7	Diabetes Pedigree Function
8	Age

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics. Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labelled as 0 means negative means no diabetes and 268 labelled as 1 means positive means diabetic.

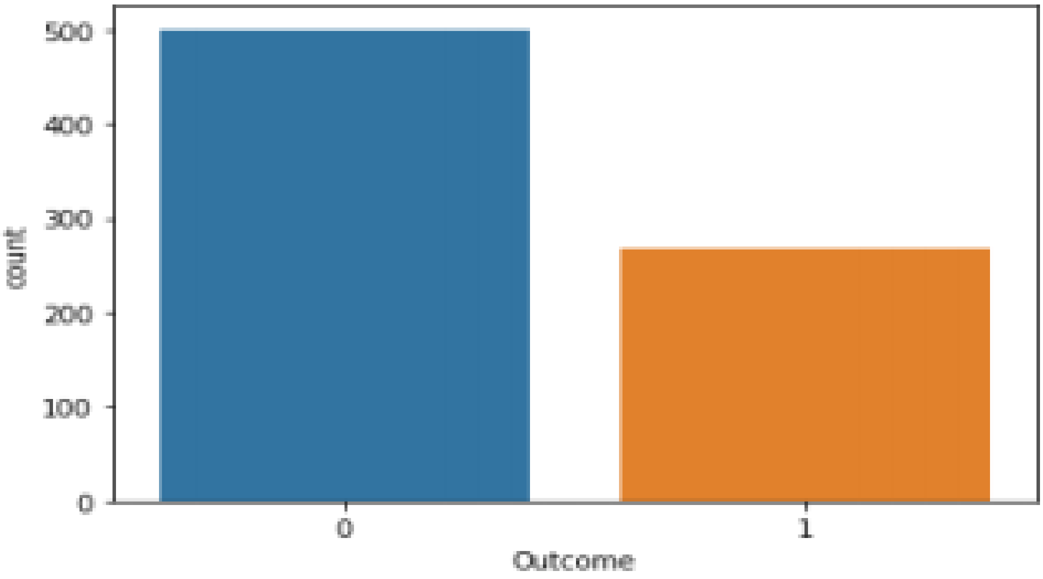

Figure 1: Ratio of Diabetic and Non Diabetic Patient

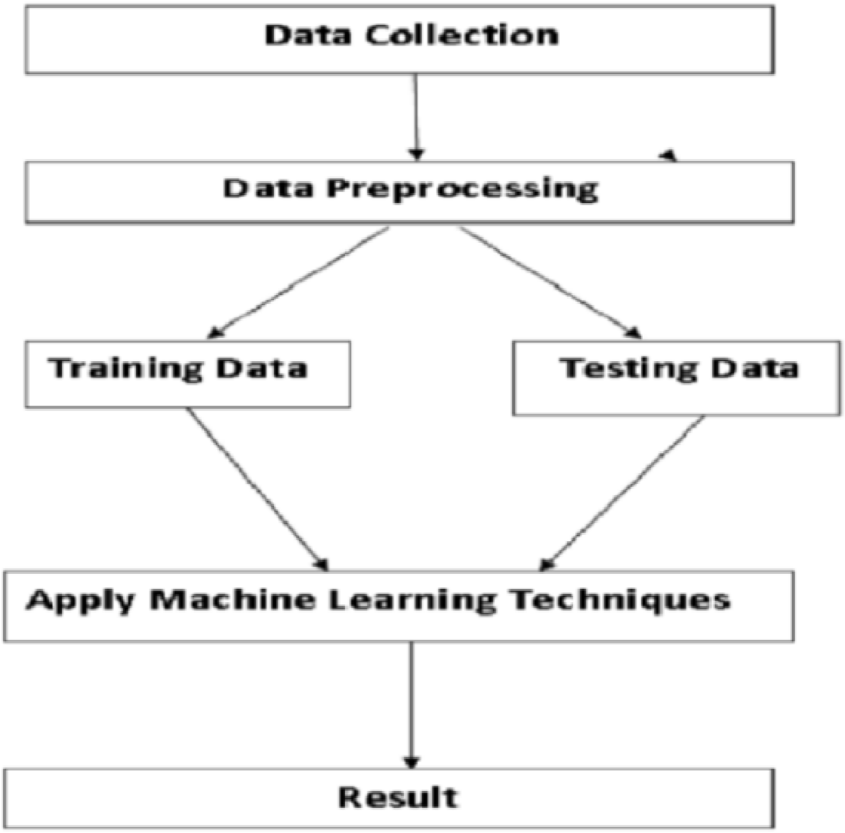

Figure 2: Overview of the Process

Procedure of Proposed Methodology

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. KNearest Neighbor, Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analysing based on various measures conclude the best performing algorithm.

METHODOLOGY

Data Pre-processing- Data pre-processing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data pre-processing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre-processing in two steps.

1). **Missing Values removal**- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.

2). **Splitting of data**- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.

Apply Machine Learning- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyse the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.

IMPLEMENTATION

1) **Random Forest** – It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

- The first step is to select the “R” features from the total features “m” where R<<M.
- Among the “R” features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until ”l” number of nodes has been reached.
- Built forest by repeating steps a to d for “a” number of times to create “n” number of trees.

2) **Decision Tree**- Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc.

Steps for Decision Tree Algorithm-

- Construct tree with nodes as input feature.
- Select feature to predict the output from input feature whose information gain is highest.
- The highest information gain is calculated for each attribute in each node of tree.
- Repeat step 2 to form a sub-tree using the feature which is not used in above node.

3) **XG BOOST –** XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

The algorithm differentiates itself in the following ways:

- A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.
- Portability: Runs smoothly on Windows, Linux, and OS X.
- Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.
- Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.

4) **Support Vector Machine**- Support Vector Machine also known as SVM is a supervised machine learning algorithm. SVM is most popular classification technique. SVM creates a hyper plane that separate two classes. It can create a hyper plane or set of hyper plane in high dimensional space. This hyper plane can be used for classification or regression also. SVM differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyper plane performs the separation to the closest training point of any class.

Algorithm-

- 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. Margin = distance to positive point + Distance to negative point.

LIBRARY AND IDE USED IN THE PROJECT:-

Libraries used in this project were mainly

1. Seaborn
2. Matplotlib
3. Pandas
4. Numpy
5. Sci-kit learning

IDE Used for compiling:- Jupyter Notebook (Python)

CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, XGBoost, Random Forest, Decision Tree classifiers are used. And 76% classification accuracy has been achieved. The Experimental results can be assisted health care to take early prediction and make early decision to cure diabetes and save humans life.

REFERENCES

[1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[2] K. Vijiya Kumar, B. Lavanya, I. Nirmala, S. Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[3] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13