JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Prediction Of Diabetes Using Machine Learning

Author' Sujith Naidu Boddeda , Sankara Rao T

M.Tech Scholar, Associate Professor

1CSE Dept,Gitam university,Vishakapatnam,AP,India 2. 1CSE Dept.,Gitam university,Vishakapatnam

*Email: sboddeda@gitam.in, sterli@gitam.edu

Abstract: Diabetes mellitus is a chronic disease characterized by hyperglycemia. Diabetes is generally occurred due to genes, bad lifestyle, lack of exercise, overweight. Diabetes is one of the most populated diseases in the world according to WHO. Early prediction of diabetes helps us to reduce the risk factor and helps in improving treatment. The prediction of diabetes is sort of tough because of the smaller number of data availability along outliers within the data sets. In this we are proposing a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection and different Machine Learning (ML) classifiers (k-nearest Neighbor, Decision Trees, Random Forest, Logistic regression, support vector machine). Along with this, we proposed the work of hybrid algorithm. In this we performed hybridization using different combination of algorithms for better accuracy the data set used is Pima Indian Diabetes Data set.

Keywords: Machine learning, Diabetes, SVM, K-Nearest Neighbor, Decision Tree, Random Forest, Logistic regression, PIMA data set, Hybrid algorithm.

I. Introduction

Diabetes is the known deadliest diseases in the world. It just not only effects the sugar level but also effects our heart, eyes, kidney diseases, etc. Generally, the hormone insulin is which moves sugar from the blood into your cells that is stored or used for energy. When a person is affected with diabetes, your body either does not make enough insulin or cannot use the insulin produced by body. It is a problem that occurs when the glucose produced is not properly used by human body. Glucose is the main source of energy for our body cells. The levels of glucose in our body are controlled by a hormone called insulin.

The main objective of this study is to create a model which may analyze or predict the likelihood of diabetes in patients with maximum accuracy.

Here are different types of diabetes: Type 1: It can develop at any age but, occurs most frequently in children and in between teens to adult age mostly. Type 2: It generally occurs in adults and this type of diabetes is the result of diabetic cases today. If you have type 2 diabetes, it means that your body is not able to use insulin properly that is produced. Type 3: It is a type of diabetes that occurs during the time of pregnancy this type of diabetes sometimes can be effected to both the mother and child. Some of the symptoms of diabetes are Abnormal thirst and dry mouth, Sudden weight loss, Frequent urination, Lack of energy, tiredness, Constant hunger, Blurred vision, Bed wetting, Sores that do not heal. The causes of diabetes are overweight, unhealthy lifestyle, obesity etc. Machine learning helps us in making good analysis or research. Machine learning is one of the most essential domains for such predictions

and analysis. According to World Health Organization data, India has the highest number of diabetes cases. Data mining plays a vital role in collection or gathering the data of different kinds and filed in such a way that can be easily accessed. By using the supervised machine learning we train the algorithms with data sets which are used for testing and prediction of data.

II. Related work

Diabetes Prediction Using Different Machine Learning Approaches by Priyanka Sonar and Jaya Malini (2019) they used many machine learning algorithms like decision trees, random forest and even used artificial neural network for prediction and find better accuracy. They used PIMA dataset. Firstly, they found to have many numbers of decision trees to find better accuracy and trained RF and ANN to find the hidden patterns in the data and they obtained 0.75 accuracy. They mentioned that even an unstructured data can also be helpful for making prediction because it helps us learn new thing from the existing data.

Prediction of diabetes in healthy population using machine learning by T. Abbas, Marely Rios (2019) also used Pima Indian data set they used 10-fold cross validation and used support vector machine for prediction in this they used mat lab for machine learning routines and data processing they obtained an accuracy of 0.72 In a possible extension of this study, the prediction models may be applied on other similar datasets that include the OGTT measurements.

Accurate Diabetes Risk Stratification Using Machine Learning by Md. Maniruzzaman, Md. Jahanur Rahman, BenojirAhamad (2018) used classification algorithms for diabetes prediction. Interesting fact here is these performed research in 2018 by using machine learning classification models like naive Bayesian, logistic regression, Decision tree, support vector machine they obtained 0.79 accuracy and mentioned that further, pre-processing techniques may be used to replace meaningless values by mean or median and outliers by mean or median. There are many other techniques of feature extraction, feature selection, and classification, and performances of presented combinations of system could also be compared to the opposite systems.

Again, performed research in 2020 by using ensemble methods such as Ada boost and random forest multi-layer perceptron which is often used to find the hidden layers and correlation of the variable and is often used in supervised machine learning and for performed for better accuracy with different models leaving a scope again mentioning that different combinations of algorithms can also be tried.

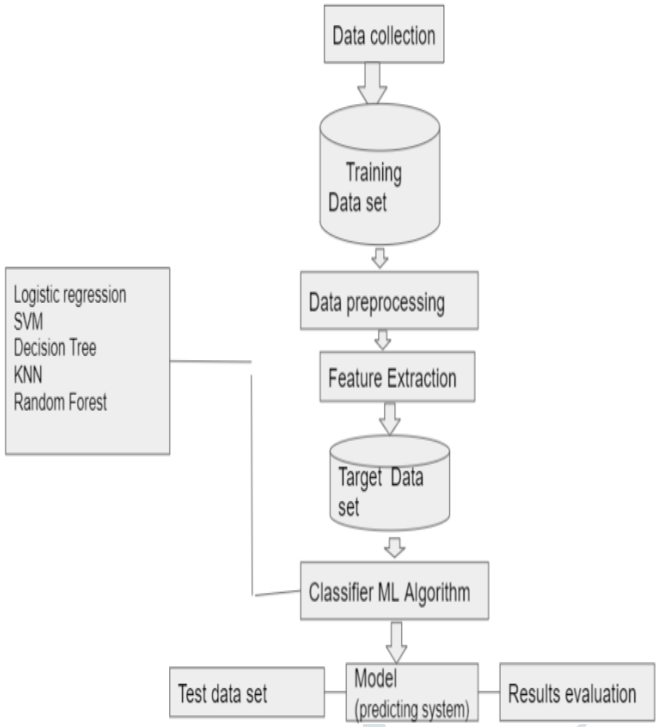
Diabetes Prediction Using Ensemble of Different Machine Learning Classifiers by Md. Kamrul Hasan, Md. Ashraful Alam (2020) used both classification and ensemble models Ada boost, XG boost, multilayer perceptron which is often used to find the hidden layers and correlation of the variable and is often used in supervised machine learning for diabetes prediction and also used k-fold cross validation and mentioned that The correlation based attribute selection can improve the correlation between attribute and target outcome.

Prediction and diagnosis of future diabetes risk: a machine learning approach by Ritu Chauhan, Ashish Kumar Mourya they used Gradient Boosting, Logistic Regression and Naive Bayes. They obtained certain accuracy and in this they mentioned that we can perform the same algorithms on different data sets and compare the accuracy and mentioned that trying on large data sets help to give us better performances.

By making a study of research papers some of the machine learning classification models like logistic regression, support vector machine, decision tree, naive Bayesian, random forest, k-nearest neighbor had been implemented. Along with this on making a proper study we proposed hybrid algorithm.

Our main objective is to analyze diabetes dataset and study whether it is possible to predict diabetes from medical test results or not. To investigate on some existing models to identify its advantages for detection of diabetes. To design new model using Machine learning approach and compare with the existing model.

III. Basic concept



In this we collected our data set from Kaggle that is PIMA Indian dataset. This data set is used in different research not only for diabetes prediction but also helpful for kidney related disease predictions. This data set contains 768 patient records of PIMA Indian dataset with 9 attributes. Description of the data set is mentioned in a tabular form below.

Initially we pre-process the data by removing unwanted data and outlier rejection is done all the noise is removed We used different classification algorithms for our diabetes prediction.

We used different algorithms for experimental study.

Algorithms used:

- 1.Random Forest
- 2.Logistic Regression
- 3.SVM
- 4.Decision tress
- 5.Naive Bayesian
- 6.K Nearest Neighbor

III.I Random Forest

Random forest is one among the supervised machine learning algorithms. This method is employed for both classification and regression. It is a collection of trees and during the training and testing the average prediction of individual trees are taken.

In random forest we take our data set and divide it into different subsets. Each data in the subset is different from one another and then the trees are trained to their subsets then from each tree the output is predicted by most votes. Use random subset of features to split the tree. All the outputs of the tree are individually collected. The increase in number of trees helps us to give a good prediction. Each tree in random forest moves to its largest extension without using pruning initially.

Then the output is predicted by the majority votes from all number of trees. There are different parameters in decision trees are tree depth, number of trees, number of features, number of threshold values etc. If there are any imbalanced tree, we need to be careful because highly imbalanced trees negatively affect predictions.

Confusion matrix for random forest

Tested positive	24
Tested Negative	76

III.II Logistic regression

Logistic Regression is additionally a supervised machine learning classification algorithm. Whenever we want to predict the outcome in binary form, we use logistic regression. They will be continuous or discrete. Logistic regression helps us to find or differentiate between the given data values into separate categories. It classifies the info in two ways which suggests only in 0 and 1 which suggests it classifies whether patient is diabetic or negative. The main moto of logistic regression is to describe the connection between target data and given variables. Logistic regression may be supported on rectilinear model. It uses sigmoid function for predicting the positive or negative class.

Sigmoid function P = 1/1+e – (a + b x) Here P = probability, a and b = parameter of Model.

Confusion matrix for logistic regression

Tested positive	32
Tested Negative	68

III.III Support vector Machine

Support vector machine is also a supervised machine learning algorithm used. This is mainly used for data transformation or separation based on outputs. This can be used for both classification and regression.in this we have a hyper plane that divides our class and calculate the distance between the planes which is called as margin. If the distance between them is low then we may have a chance of misconception so, we should select our classes with high margin for better performance.

In general, we can use logistic regression for out prediction but when we have high dimensional data or more missing values then the model fit becomes difficult. As SVM is a model free method it does not require any assumptions or distributions by the depending and independent variables. Each data point is individually represented as a n-dimensional vector. In this we create n-1 dimensional hyperplanes that separates two classes that are having maximum distance and data point on both sides. We use nonlinear, Kernels to transform into a multidimensional space. There are different types of kernels can be used in this model. Some of them are Sigmoid kernel, Gaussian kernel, polynomial kernel etc.

Confusion matrix for SVM

Tested positive	24
Tested Negative	76

III.IV Decision Tree

It is a supervised machine learning algorithm. This is represented in shape of tree where every node corresponds to each class label given. It gives us a statistical probability where each branch of the tree represents a possible decision, we get a Boolean type of outcome from every internal node the tree.

In decision tree we need to avoid overfitting of data, if there is any noise or the data set is small, we get problems in predicting. To overcome that reduced error pruning is used so that each node can be replaced by the most important attribute that helps in predicting. We make use of continues valued attribute. It is simple and fast process. It gives us a statistical probability where each branch of the tree represents a possible decision, we get a Boolean type of outcome from every internal node of the tree.

Confusion matrix for Decision Tree

Tested positive	34
Tested Negative	56

III.V Naive Bayesian

Naive Bayesian is a supervised machine learning algorithm. It is one among the foremost used classification algorithms. This makes predictions depending on different attributes. It makes assumptions of the features that are independent to each other. It is a very fast used algorithm that predicts the class of a test dataset.

Advantage of naive Bayesian is it is simple and easy to implement. More accuracy in result due to higher value of probability the drawbacks Strong assumption on the shape of data distribution. Some Loss of accuracy.

Smoothing technique is used in naive Bayesian algorithm. It is a technique that is used to remove the zero values in the data.

It is based on conditional probability it is one of the best algorithms to be used in classification problems. It is used in problems such as when the data is highly imbalanced naive Bayesian can be used. Even in case of any noise or missing value in the data also this algorithm will be helpful in dealing with such cases.

Confusion matrix for naive Bayesian

Tested positive	28
Tested Negative	72

III.VI K-Nearest Neighbor

KNN is a supervised machine learning algorithm. It is an algorithm it stores all the data in the data set with a set of rules. These are divided into classes for the rest of the data we need to find the nearest neighbor of our classes. For finding we consider most votes of the label.

Here, K= Number of nearby neighbors, it is always a positive integer. Neighbor value is chosen from set of class. Closeness is mainly defined in terms of Euclidean distance. The Euclidean distance between two points P and Q i.e. P (p1,p2, . Pn) and Q (q1, q2,..qn) is defined by the equation.

KNN stores all cases and helps us to classify depending on the similarities of the attributes taken. How it differentiates two things? KNN checks for the nearest points that have common features by using out the common features it classifies between two cases. It

also classifies new classes basing on the features of the data taken. It is a parameter that refers to number of nearest neighbors it includes by taking the majority voting process. It classifies new data point depending on its nearest neighbor.

Confusion matrix for KNN

Tested positive	26
Tested Negative	73

IV. Data pre-processing

Data mining is usually done for collecting or gathering the required data. Most of the health care related data obtained usually have missing values, unwanted data and other impurities so, in order to change the quality of data the data pre-processing is done. To make sure that machine learning models work effectively by using our data set this process is important. We are using PIMA Indian dataset obtained from Kaggle.

This pre-processing is done in two steps:

Removing the missing values- Removing all the values that are zero(0). Being null valued can not possible for prediction. So, all these values are removed. If any data have maximum of missing values then that particular entire row is eliminated because, the excess of missing values leads to miss classification.

Removing of all the values contain zero helps us in featuring the subsets, which reduces data and so as we can work faster. By using this feature selection all outliers are removed, all the unwanted data or noise is eliminated from your main data. The data removed will not be completed deleted from the system because the unwanted data removed from this can be useful for other predictions so the data will be just remove and it will be formed as a sub set to it.

Splitting of data- After cleaning the info, the entire data gets normalized in training and testing the model. Once we split the info we initially train all the algorithms on the training data set and leave test data set aside. The train and test sets are employed by both classification and regression problems. Initially the subset should fit the model then predictions are made and compared with the prevailing and proposed algorithms counting on their outcomes. Train data set fit the model and therefore test data set is employed for evaluating the model.

IV.1 Data set

We used PIMA Indian Diabetes data set taken from Kaggle. By making use of different parameters, we performed different operations on our data set. we trained our dataset to different algorithms to get the accuracy results.

Accuracy is measured by the formula:

Accuracy = TP+TN/TP+TN+FP+FN

V.Proposed Work

we are proposing the hybridization. In this we performed hybridization on different combination of algorithms in order to obtain better accuracy. Hybrid algorithm means combining or combination of two different algorithms. In this the features of two algorithms are combined to give better performance.

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

Step1: Import required libraries, Import diabetes dataset.

Step2: Missing data is removed.

Step3: Dataset is split into two as Training set and Test set.

Step4: Select the hybrid algorithm that is you can select any two of the machine learning algorithms (K- Nearest Neighbour, Support Vector Machine, Decision Tree, Logistic regression, Random Forest).

Step5: Classifier model is built 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 analyzing based on various measures conclude the best performing algorithm

Hybrid algorithm	Accuracy
Logistic regression and SVM	0.79
Naive Bayesian and SVM	0.80
KNN and Logistic regression	0.81
Random forest and KNN	0.80
KNN and SVM	0.85

In our proposed methodology many combinations were tried but some of them were mentioned below:

VI.Result Analysis

Comparison of six different algorithms is done along with the proposed work. The accuracy of the algorithms are calculated using confusion matrix.

Accuracy is measured by the formula:

Accuracy = TP+TN/TP+TN+FP+FN

ML method used	Accuracy
Logistic Regression	0.77
Support Vector Machine	0.78
Decision Tree	0.72
Random Forest	0.74
K-Nearest Neighbor	0.77
Proposed work (Hybrid algorithm)	0.85

VII.Conclusion

Diabetes prediction is one of the challenging tasks. In this we initially worked on the existing method to check the performances as there is ray to hope we worked to combine two algorithms so as we may obtain better accuracy.

Using PIMA Indian dataset after pre- processing the data, we trained the data on different models so as to check the performances.

we used six different algorithms and highest obtained 0.78 accuracy. For obtaining better accuracy we proposed hybrid algorithm by using different combinations of algorithms we found that by using KNN and SVM hybrid algorithm we obtained a result of 0.85.

VIII. References

Accurate Diabetes Risk Stratification Using Machine Learning (Springier 2018)

Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes (IEEE 2020)

Early Prediction of Diabetes Using Machine Learning (IEEE 2020)

Prediction of Diabetes Using Machine Learning Algorithms in Healthcare (IEEE 2018)

Diabetes prediction Using Different Machine Learning Approaches (IEEE 2019)

Predicting Diabetes in Healthy Population through Machine Learning (IEEE 2019)

Prediction and diagnosis of future diabetes risk: a machine learning approach (IEEE 2019)

Classification and prediction of diabetes disease using machine learning paradigm (Springier 2020)

Diabetes Prediction Using Ensemble of Different Machine Learning Classifiers (IEEE 2020)

Diabetes Mellitus Prediction and Classifier Comparative Study (IEEE 2020)

Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison with Classical Time-Series Models (IEEE 2020)