



Machine Learning in Predicting Diabetes in the Early Stage

Niranj Rajesh
CSE S7
Roll no:20
Guided by Vishagini V

Content

- Abstract
- Introduction
- Literature Survey
- Data Set
- Existing System
- Proposed System
- Results
- Conclusion
- References

Abstract

- Diabetes is a common disease and its early symptoms are not very noticeable, the conventional and tedious method ends with consulting a doctor or diagnostic center.
- The objective of this report is to develop a system which can perform early prediction of diabetes for a patient with high accuracy.
- We use Machine Learning technique(Random Forest algorithm) to achieve a system that gives the best results for diabetic prediction.
- The Prediction system showed that it is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

Introduction

- Machine learning in healthcare is becoming more widely used and is helping patients and clinicians in many different ways.
- Diabetes can lead to problems affecting almost every body systems. Hence, it is necessary to prevent and diagnose the disease.
- An accurate and timely diagnosis will help the patient prevent it. Since the medical resources are limited and only certain number of patients can be diagnosed by doctors.
- Hence there is a need to make an efficient prediction model, which can save medical resources and help patients make a self-test accurately.



Literature Survey

| Paper Name | Advantages | Disadvantages |
|--|---|---|
| Predicting Diabetes in Healthy Population through Machine Learning. | Uses less features, for type-2 diabetes. | Only has accuracy of 84.1%.Prediction for later stages of diabetes only. |
| Decision Tree and Random Forest Classification Model to Predict Diabetes | A hybrid of decision tree and random forest is used. | Only has accuracy of 72% and 76.5% for (RT and RFC respectively).Uses Plasma glucose as attribute requires medical tests. |
| Diabetes prediction using artificial neural networks el jerjawi | Established model,87% prediction accuracy comparatively higher. | Also uses Plasma Glucose as attribute. Samples do not hold all races. |


Data Set

The data set we used is from UCI machine learning repository.

There are 17 attributes and 520 instances totally, 320 of which is positive cases and 200 samples are negative cases.

We investigate relation between each attribute and diabetes.

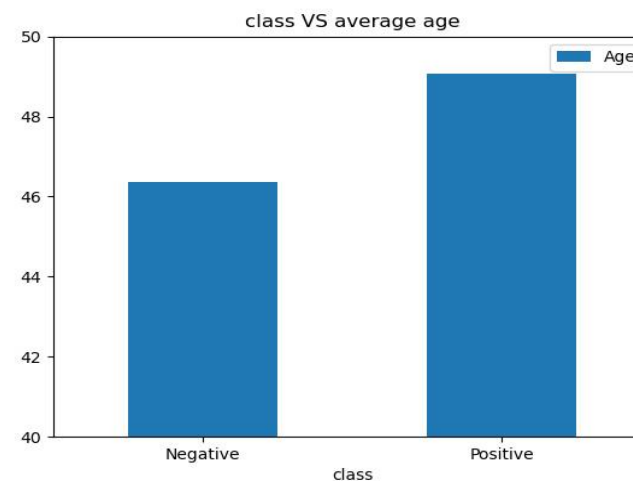
| | | | | | |
|---------------------------|-------------------|------------------------|---------------------|-------------------------|-----------------|
| Age | Polyphagia | Parital paresis | Sex | Muscle stiffness | Polyuria |
| Sudden weight loss | Itching | Obesity | Irritability | Delayed healing | weakness |
| Visual blurring | Alopecia | Genital thrush | Polydipsia | Class | |



Data Set

- Since age is the only numerical value we plot the average age for people with each class.
- For the other variables we calculate morbidity for each class of every attribute.
- As result average age of people with diabetes is higher and symptoms like alopecia and itching imply lower probability.

| | Alopecia | delayed healing | Gender | Itching | muscle stiffness |
|-----|--------------------|-----------------|--------------|-----------------|------------------|
| No | 71.0% | 59.4% | 90.1% | 62.2% | 56.9% |
| Yes | 43.6% | 64.0% | 44.8% | 60.9% | 69.2% |
| | sudden weight loss | Genital thrush | Irritability | partial paresis | Polydipsia |
| No | 43.6% | 58.7% | 53.3% | 43.2% | 33.1% |
| Yes | 86.6% | 71.6% | 87.3% | 85.7% | 96.6% |
| | Obesity | Polyphagia | Polyuria | visual blurring | weakness |
| No | 60.0% | 46.3% | 29.4% | 50.5% | 47.4% |
| Yes | 69.3% | 79.7% | 94.2% | 75.1% | 71.5% |



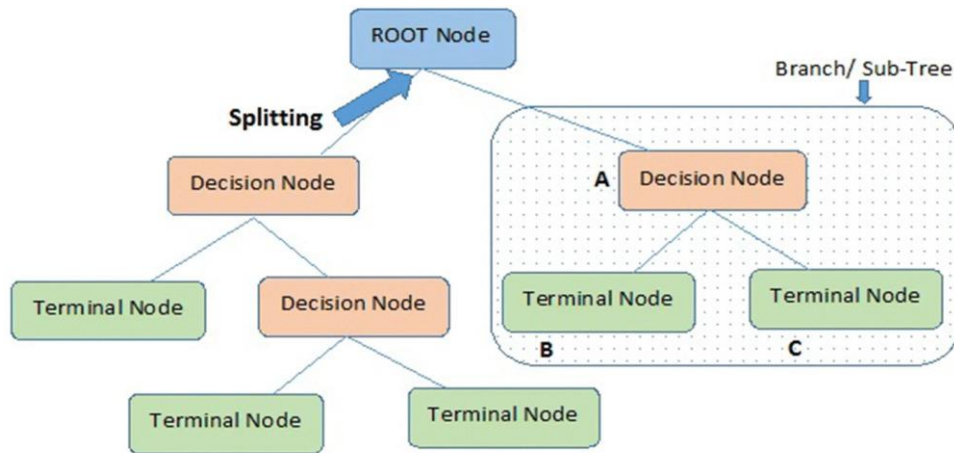
Existing System

Decision Tree Model

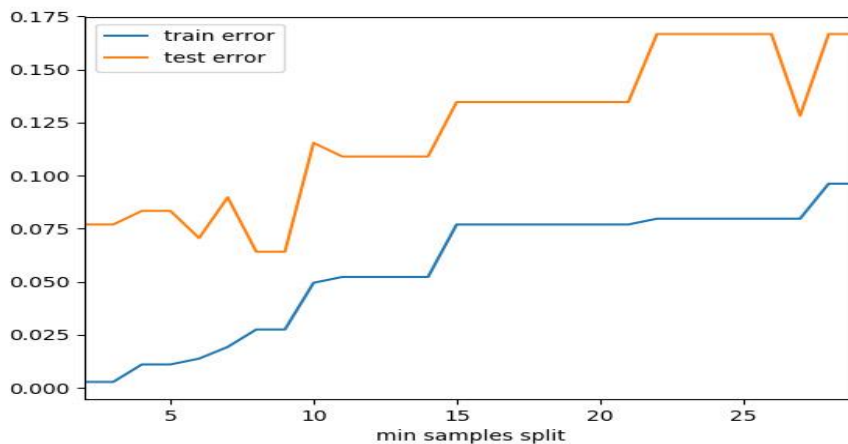
Decision Tree Model

- Decision tree is a tree-like structure to create a model that make predictions based on input variables.
- In this case a Decision tree is created from the given dataset.
- 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.

Decision Tree Model



Note:- A is parent node of B and C.



- There are many ways to choose the best attribute to be as the root node, based on the degree of impurity of the child nodes.
- The Performance measure like Entropy are used.
- These measures are done for all attributes and comparison is done, to select the best split.
- We build the model and calculate train error and test error which is presented in the figure using different sample split.
- The accuracy of the resulting model was 78.1768%.

Decision Tree Model

Disadvantages of the Decision Tree Model.

- Comparatively poor accuracy.
- The decision tree model is exceptionally sensitive to the slightest change in data.
- Overfitting of decision tree leading to inaccurate predictions.
- Decision tree can grow out and be more complex leading to large computational resources and training time.



Proposed System

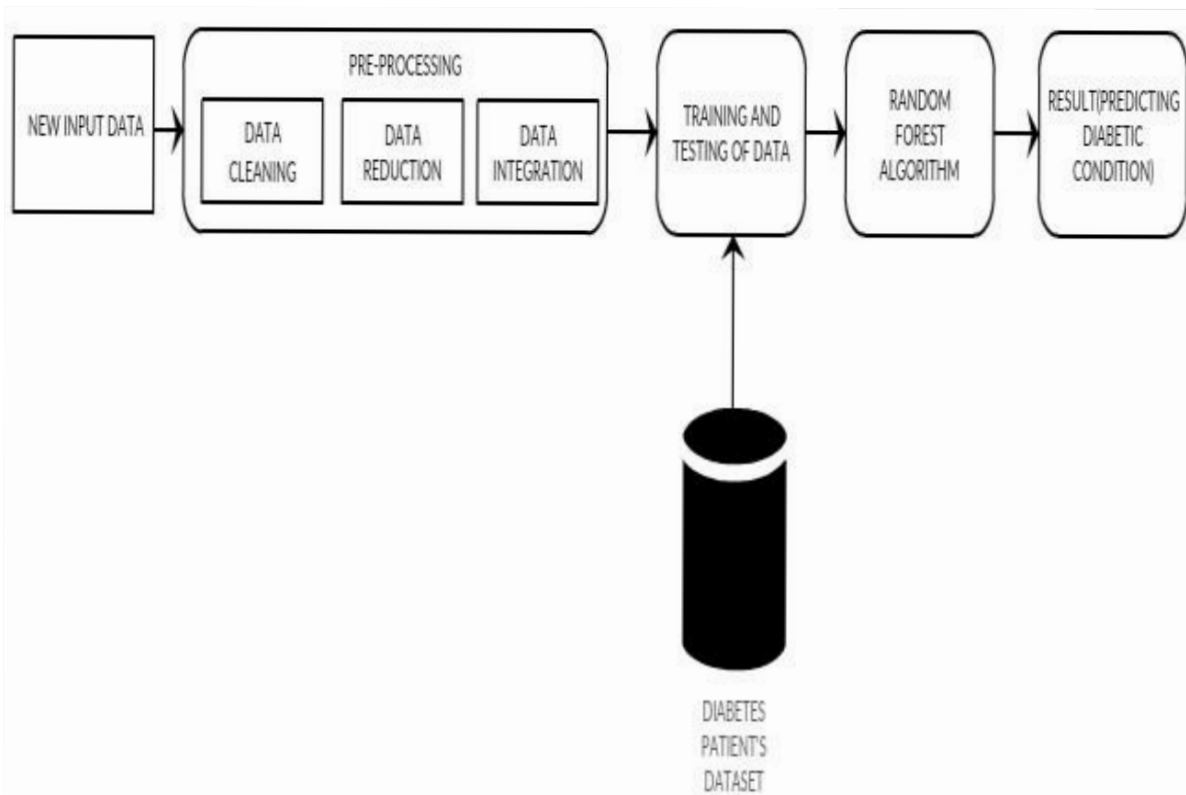
Random Forest Model

Proposed System

Why Random Forest Model over the Decision Tree Model?

- Due to the instability of single decision tree, it comes out the ensemble learning which combines multiple models to improve the prediction accuracy overall and reduce the variance.
- Random Forest Technique is designed to increase the accuracy of the decision tree.
- A set of decision trees are grown and each tree votes for the most popular class, then these votes are integrated and a class is predicted.

System Architecture



- The datasets are collected from the database.
- The data was obtained from UCI learning repository.
- The New Input Data is given and goes through various phases.
- In the next step the new input data will be pre-processed which will include data cleaning, integration and transformation.
- Before that data is finally tested with the model.

Data Pre-processing

Data pre-processing is a vital step in data discovery methodology. Since most of the healthcare information contain missing value, inconsistency and duplications.

- a) **Data cleaning:** It is the tactic of detection and correcting corrupt or inaccurate records from a record ,table, set or data. eg: drop the duplicates, check for null values and replace with mean values.
- b) **Data Integration:** It is the technique that involves combining data from multiple heterogenous data sources to provide a unified view.
- c) **Data Reduction:** It is the transformation of numerical and alphabetical digital data into a corrected, ordered and simplified type.

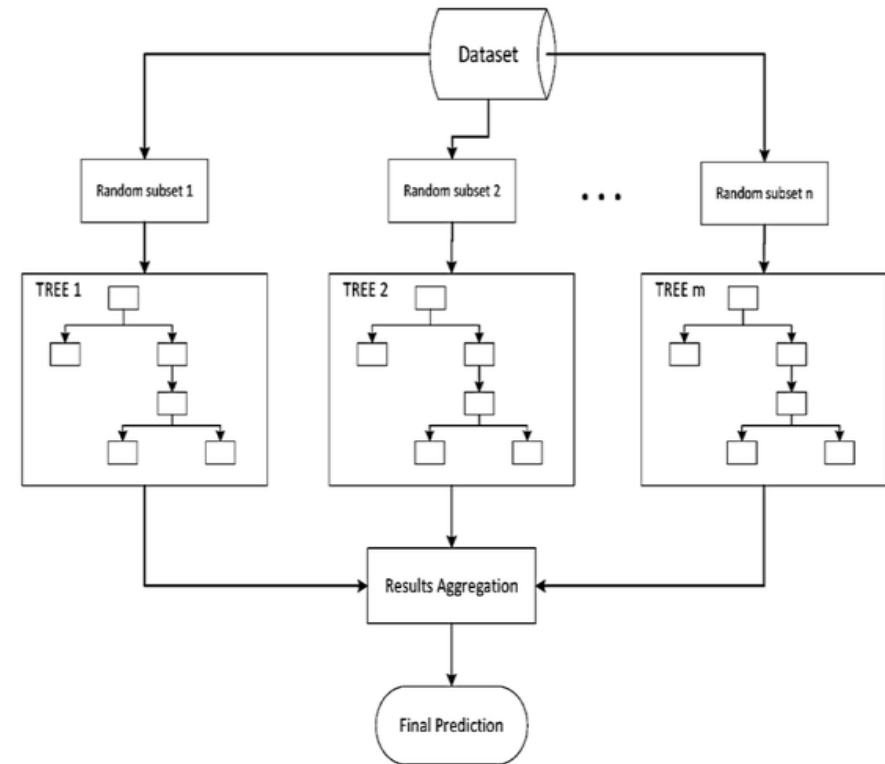


Algorithm

- The first step of random forest is to select the “R” features from the total features “m” where $R \ll m$.
- Among the “R” features, the node using the best split point.
- Split the node into daughter nodes using the best split .
- Repeat the steps until certain number of nodes has been reached.
- Built forest by repeating the steps for certain number of times to create “n” number of trees.

Random Forest Model

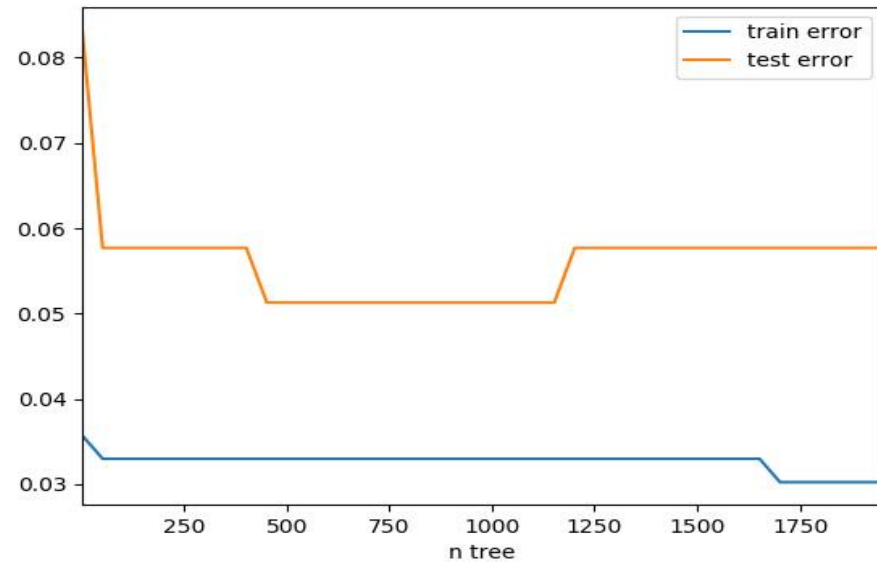
- After taking a glance and use each decision tree created to predict the result.
- Calculate the votes for each predicted target and admit the high voted predicted target as a result of the ultimate prediction.
- **Basically the Random Forest output will be the average value of output of all the decision trees.**



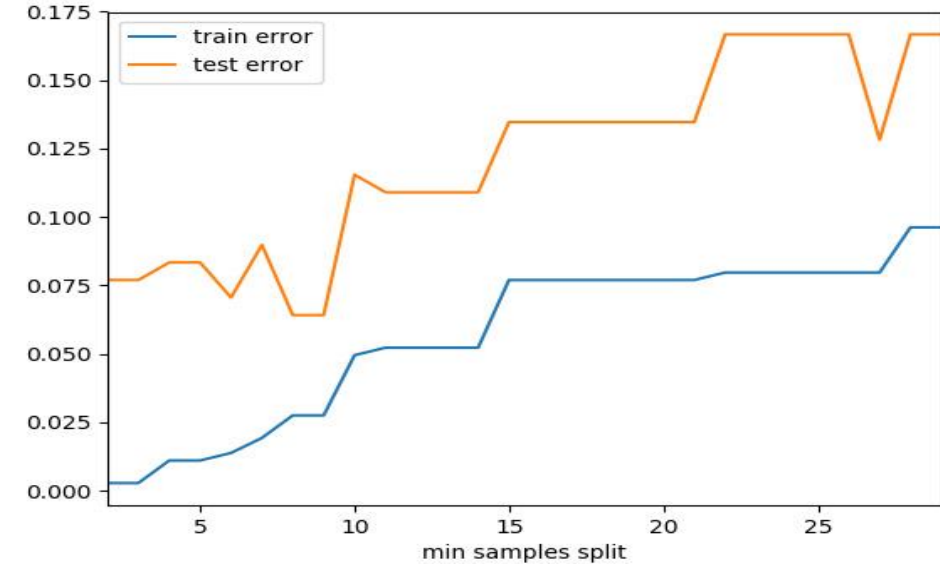
Result

Compared with single decision tree, the testing accuracy of random forest, which is 94.9%, increased near 16.8 percent.

Random Forest Model



Decision Tree Model



Result

- The positive true for positive prediction was 95.70% for test case.
- The negative true for negative prediction was 93.80% for test case.
- Comparatively the results were much better than the decision tree model.
- The testing error keeps stable after falling, which means its not an overfitting model.



Conclusion

- An efficient diabetes prediction model will help doctors make accurate diagnoses and help patients get timely treatment.
- Throughout this work we studied and evaluated the Random Forest algorithm and it has higher accuracy of the other model(Decision tree algorithm).
- The aim of this project is to develop a system which can perform early prediction of diabetes with higher accuracy and provide advance support for predicting the accuracy rate of diabetes.

References

- [1] Fikirte Girma Wolde Michael, Sumitra Menaria, “ Prediction of Diabetes using Data Mining Techniques, Dept.of Computer Science and Engineering, Sumitra.Menaria@paruluniversity.ac.in, Proceedings of the 2nd International conference on Trends in Electronics and Informatics(ICOEI 2018).
- [2] El_Jerjawi, Nesreen Samer & Abu-Naser, Samy S. (2018). Diabetes Prediction Using Artificial Neural Network. International Journal of Advanced Science and Technology 121:54-64
- [3] Deepti Sisodiaa, Dilip Singh Sisodiab,” Prediction Diabetes and Classification Algorithm” A National Institute of Technology, G.E Road, Raipur and 492001, India, International Conference on Computational Intelligence and Data Science.
- [4] A.M.Rajeswari, M.Sumaiya Sidhika, M.Kalaivani C.Deisy,” Prediction of Pre-Diabetes using Fuzzy Logic Based Association Classification”, Thiagarajar College of Engineering, Madurai, India Proceedings of the (ICICCT 2018), cdcse@tce.edu.