18CSE355T

DATA MINING AND ANALYTICS

DIABETES PREDICTION ANALYSIS

PROJECT REPORT

Submitted

UppuManikanta[RA201100301112]

Allen Jerome[RA2011003011167]

MAY 2023

AIM:

To make diabetes prediction analysis using WEKA tool.

OBJECTIVE:

Diabetes is one of the major international health problems. World Health Organization report says that around 422 million people have diabetes worldwide. Data mining plays a huge role in predicting diabetes in the healthcare industry. There are many algorithms developed for prediction of diabetes. But most of the algorithms failed in case of the accuracy estimation. Also, there is a need to automate the overall process of diabetes prediction. This automation of diabetic database helps in identification of impact of diabetes on various human organs. More the accuracy of prediction, more the chances of accurate severity estimation. Therefore this project concentrated on providing different prediction methods of diabetes.

PROBLEM STATEMENT:

The objective of this project is to build a model that can predict whether a person is likely to have diabetes or not based on certain clinical parameters. Our task is to train a model using this dataset to accurately predict whether a person has diabetes or not based on these features. The model should be able to generalize well to new and unseen data and achieve high accuracy and precision in its predictions.

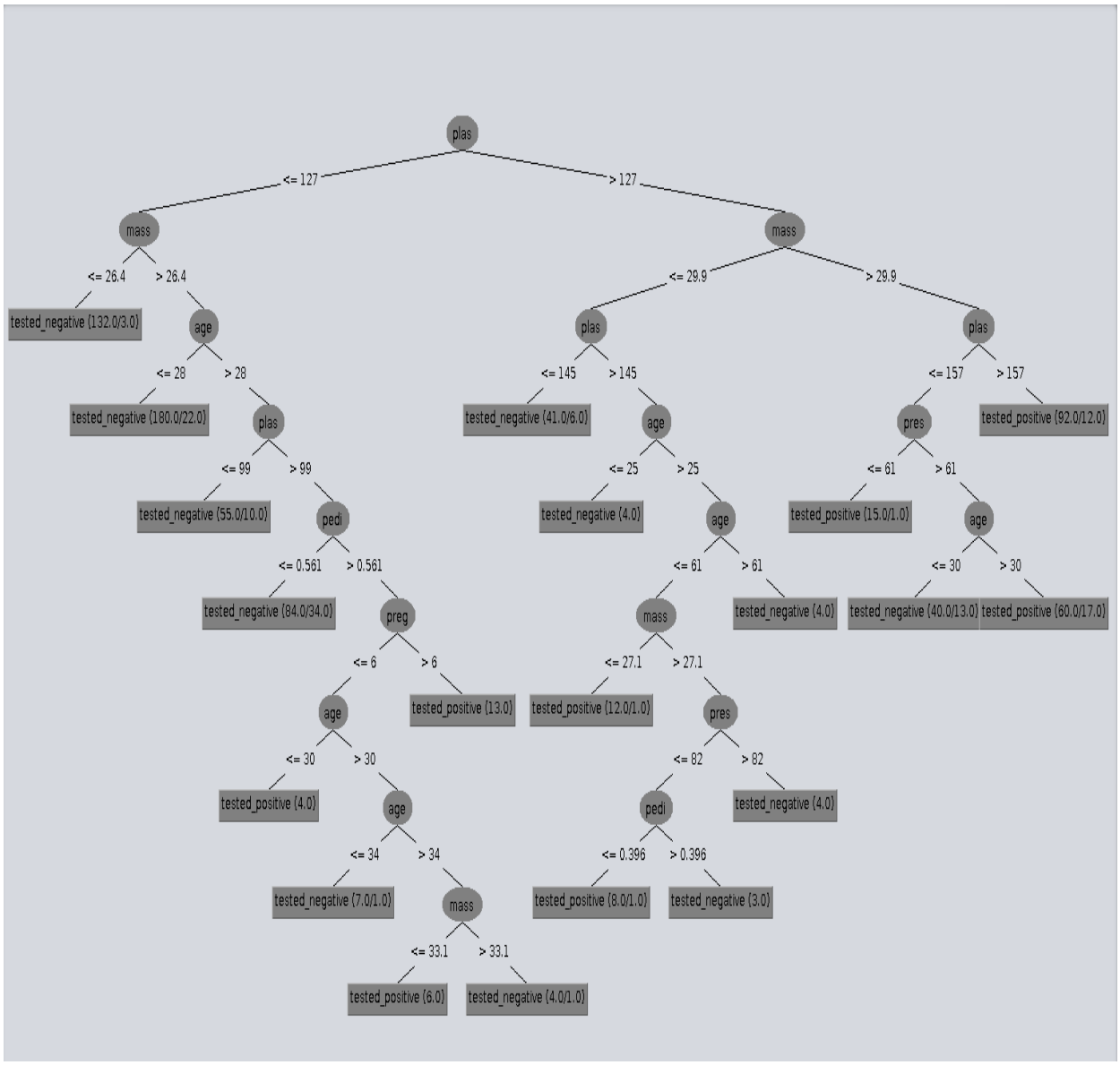
DATASET USED:

The study implemented in WEKA and dataset of Sylhet Diabetes Hospital have been used. The following classification algorithms were used in the study:

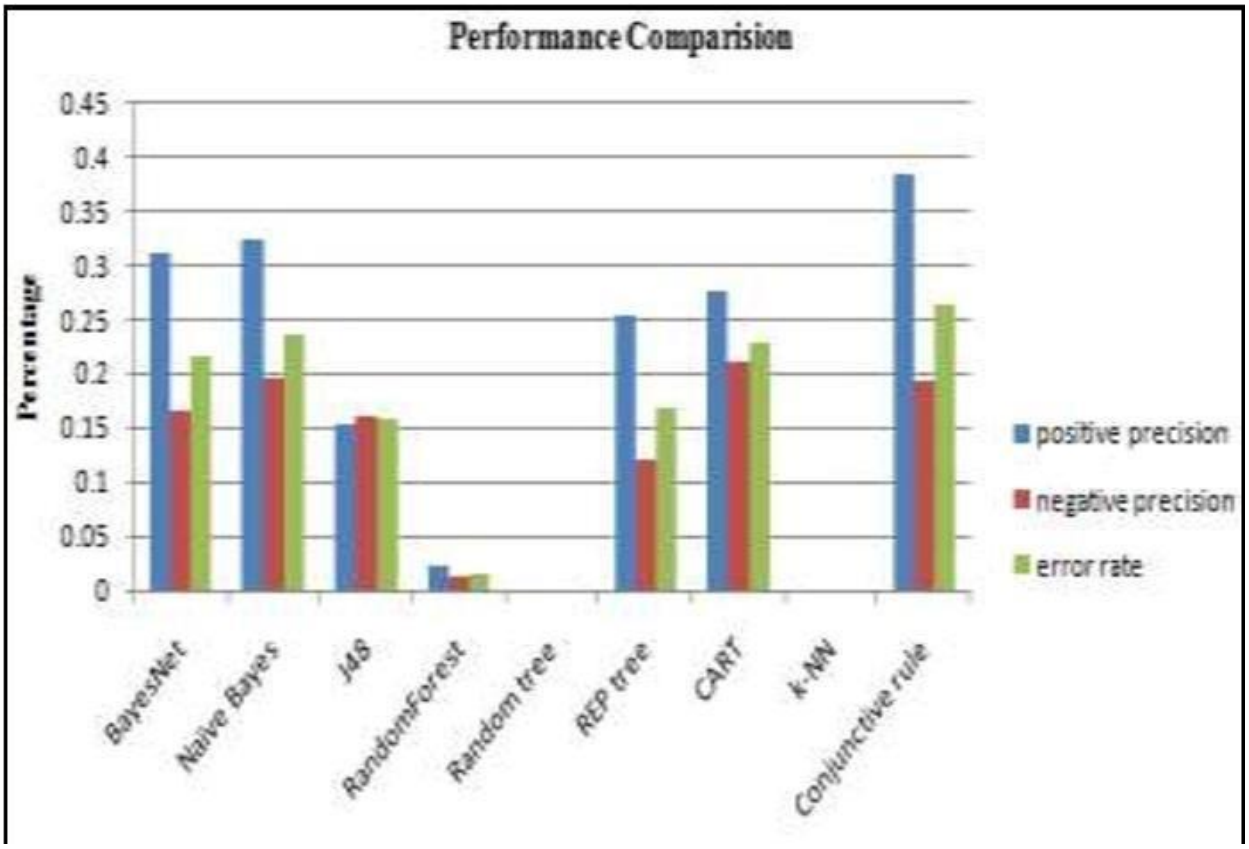
- Bayes Network
- Naïve Bayes
- Decision tree (J48)
- Random tree
- Random forest
- kNN
- Support Vector Machines

Attribute Number	Attribute Name	Answer	
1	Age	20-65	
2	Sex	1. Male	2.Female
3	Polyuria	1.Yes	2.No
4	Polydipsia	1.Yes	2.No.
5	Sudden weight loss	1.Yes	2.No.
6	Weakness	1.Yes	2.No.
7	Polyphagia	1.Yes	2.No.
8	Genital thrush	1.Yes	2.No.
9	Visual blurring	1.Yes	2.No.
10	Itching	1.Yes	2.No.
11	Irritability	1.Yes	2.No.
12	Delayed healing	1.Yes	2.No.
13	Partial paresis	1.Yes	2.No.
14	Muscle stiffness	1.Yes	2.No.
15	Alopecia	1.Yes	2.No.
16	Obesity	1.Yes	2.No.
17	Class	1.Positive,	2.Negative.

DECISION TREE:



RESULT ANALYSIS:



Attribute Rank	Attribute Nominal	Attribute
0.3623	3	Polyuria
0.3619	4	Polydipsia
0.172	2	Gender
0.1518	5	Sudden weight loss
0.1467	13	Partial paresis
0.0912	11	Irritability
0.0883	7	Polyphagia
0.0551	15	Alopecia
0.047	9	Visual blurring

Gain Ratio is used for reducing the bias resulting effect which causes from the use of information gain. The information gain measure is biased toward tests with many outcomes. This situation means that, Gain Ratio prefers to select attributes which have a large number of values. Gain Ratio is for adjusting the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

Gain Ratio = Information Gain / Split Information

CONCLUSION:

Applying data mining techniques on medical datasets is an interesting topic for researchers as there are lots of health issues and cases to investigate. On the other hand, classification is a very useful technique for knowledge discovery because it can accurately and efficiently classifies the data. As a result, k-NN is an effective classification technique to identifying diabetes disease and helps to medical workers for faster decision making about this disease, based on the attributes.