Analysis and prediction of diabetes diseases

1 Vijayalakshmi P B

Department of Computer Applications Amal Jyothi College of EngineeringKanjirappally, India [**vijayalakshmipb2021@mca.ajce.in**](mailto:vijayalakshmipb2021@mca.ajce.in)

2 Paulin Paul

Department of Computer Applications Amal Jyothi College of EngineeringKanjirappally, India [**paulinpaul@amaljyothi.ac.in**](mailto:paulinpaul@amaljyothi.ac.in)

***Abstract*-Today, the data mining is popular as an important field in healthcare sector for deeper study of medical data and providing accurate predictions of diseases.Various diseases such as** **stroke, diabetes, cancer, hypothyroid and heart disease, etc are identified using data mining techniques. To predict if the individual is infected by diabetes or not, the required dataset was downloaded. As the number of people affected by diabetes increases day by day this prediction helps to find if the patient is diabetic or not.**

**In machine learning analyzing and summarizing data from different aspects into valuable information is the main point of view. The data from different dimensions are analysed then it categorize the relationships. WEKA is a data analysis tool for machine learning classification. The vital technique with more applications in various fields is called Machine learning. It is used to classify each item in a set of data into one predefined set of classes. This research paper presents the analysis and prediction of diabetes diseases. The proposed work focuses on machine learning techniques and using the WEKA tool.**

I.INTRODUCTION

Waikato University, New Zealand developed WEKA – (Waikato Environment for Knowledge Analysis). It provides different pre-processing, machine learning algorithms, data mining tools. It includes different types of algorithms such as classification algorithm, clustering algorithm, regression algorithm, association rule mining, attribute selection algorithm etc. The data visualization facilities are also available. WEKA is a free software comes under free licence. Data pre-processing, implementation of several Machine Learning algorithms, and visualization tools are available in WEKA so that the machine learning techniques can be developed and we can apply them to real-world data mining problems. The data are stored in arff file format. It is easy to transform the EXCEL file to the ARFF format. An arff file is full of instances.

An arff file by can be created by the use of Notepad or word. The name of the dataset is associated with the relation in the dataset. Attribute informations are also available in weka . The data is with data. WEKA allows different types of formats.

Identification and prediction of diabetes disease at the beginning of the disease is a better method to prevent the disease from causing other deadly diseases. The Diabetes. arff dataset contains different attributes that can be useful for the prediction of diabetes. The ARFF file used in WEKA for diabetes disease prediction contains the details of many patients and their symptoms. Machine learning algorithms and data mining techniques are widely used for prediction and classification. The J48 and Naive Bayesian algorithm is used for classification. These are the best machine learning algorithms.

1. LITERATURE SURVEY

K.Priyadarshini and Dr.I.Lakshmi proposed [1] is a effective methodology by using a Neural Network for the automated detection of Diabetes Diseases. The work [2] is about the diagnoses of diabetes by using ANNs,the preprocessing is performed ,the missing values are also considered and replaced. Through the Modified training set, a better accuracy was achieved with the lesser time required for training the set. Karnika Dwivedi [4] used different datasets like CPCSSN and four machine learning algorithms to predict Diabetes Diseases (DD) in early-stage to save human life from an early death.To predict if the individual is infected by diabetes or not, the required dataset was downloaded. As the number of people affected by diabetes increases day by day this prediction helps to find if the patient is diabetic or not.

In this paper decision tree(J48), K-star etc, are used to predict diabetes, and the results were compared to each other.

The J48 method provides effective and better accuracy than the other methods in weka data mining tools. N.Nandhini. [2] used Naive Bayes and J48 data mining algorithms for identifying different type of diabetes diseases. The WEKA tool is used for the classification and found that the Naive Bayes is the best which have an accuracy of 76.96% than the other algorithms which are used. Here the Prediction of diabetes is done using Pima Indian diabetes dataset,It is the best data set than any other datasets that are now available which have 80% and 81% accuracy achieved for data set by using 10-fold cross-validation and by spitting data into 30% testing and 70% training. Balamurugan. S Appavu alias [3] This research study compares different algorithms that are also used to predict diabetes using data mining technology. Also authors classifiers J48 Decision Tree, KNearest Neighbors, and Random Forest, Support Vector Machines to classify patients with diabetes disease. The authors compared five prediction models for predicting diabetes disease using 9 important attributes under different circumstances. Here the results shows that the decision tree J48 classifier have a higher accuracy of 73.82 % than the other five classifiers.Rther than the other classifiers the KNN (k=1) and Random Forest performance are much better and they provide an accuracy of 100%. So from the results we can understand that noisy data from our dataset will provide a good result for our problem. Diabetes Mellitus (DM) is commonly referred to as Diabetes; it is a common disease which can harm the human body and can affect the origin of other harmfull diseases to the humans.A worldwide disease that can affect people in different ways is Diabetes and it is a long-lasting disease.If the body is not producing enough insulin it can cause the formation of diabetes. To maintain the energy level human body require a loat of insulin which is Insulin which is secreted by the pancreas known as the most important hormone in human body.It may produce the symptoms of frequent urination, increased thirst, and hunger. Diabetes can be controlled with the help of insulin injections, a healthy diet, and regular exercise. Diabetes can also lead to other diseases such as blindness, blood pressure, heart diseases, and kidney disease, etc. There are four types of diabetes. Researchers have developed a decision support system for the diagnosis of an illness that makes use of data mining techniques [2]. Three different artificial intelligence classifier algorithms namely Multilayer Perceptron, Naive Bayes Classifier, and J.48 were applied to Diabetes dataset. These classifiers are generally used for the fields of data mining, biomedical engineering, and diagnosing patient's medicine.

1. METHODOLOGY

After opening the WEKA. The following screen WEKA GUI Chooser application will start and you can see the following in “Fig. 1”.



Fig.1. The screen appears after opening WEKA

The following run five different types of applications can be seen in the GUI Chooser

* Explorer
* Experimenter
* Knowledge Flow
* Workbench
* Simple CLI

We need to click on the workbench. Then a window is open as shown in “Fig. 2”.

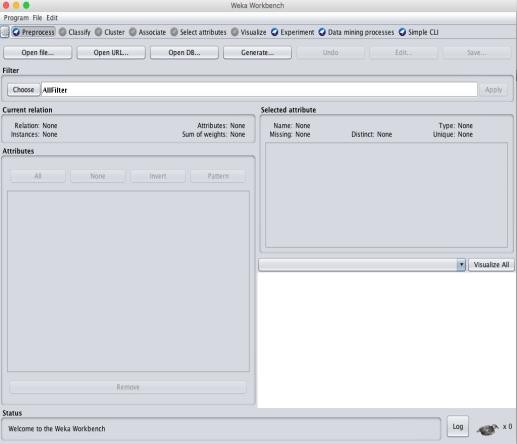


Fig.2. WEKA workbench

Then we open the dataset. The dataset that is used for the analysis and prediction of diabetes disease is an ARFF extension file as shown in “Fig.3”.

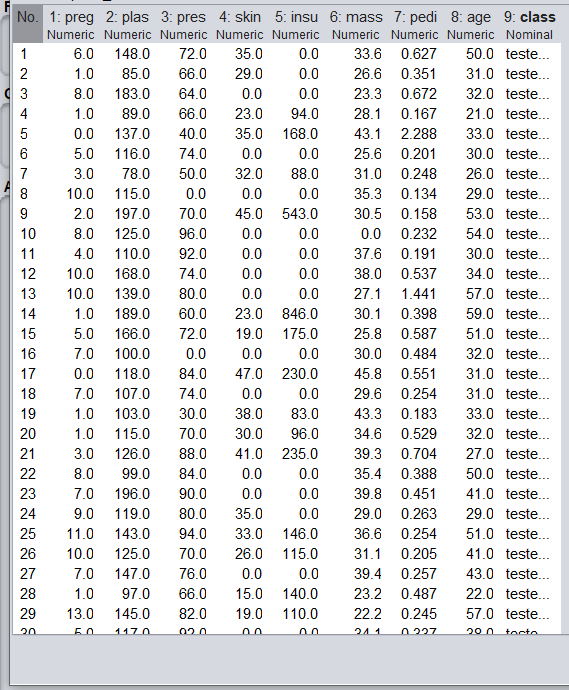


Fig.3. Dataset of diabetes

There are attributes in the dataset. Some of them are:

* Preg
* Plas
* Pres
* Skin
* Insulin
* Mass
* Pedigree
* Age

Data preprocessing is the next level is to do. Data preprocessing transforms the data into a new form that is suitable to perform mining algorithms. 9 attributes are selected to preprocessing data with 768 instances and 0% missing values in our data. Data preprocessing of diabetes data is shown in “Fig.4”.

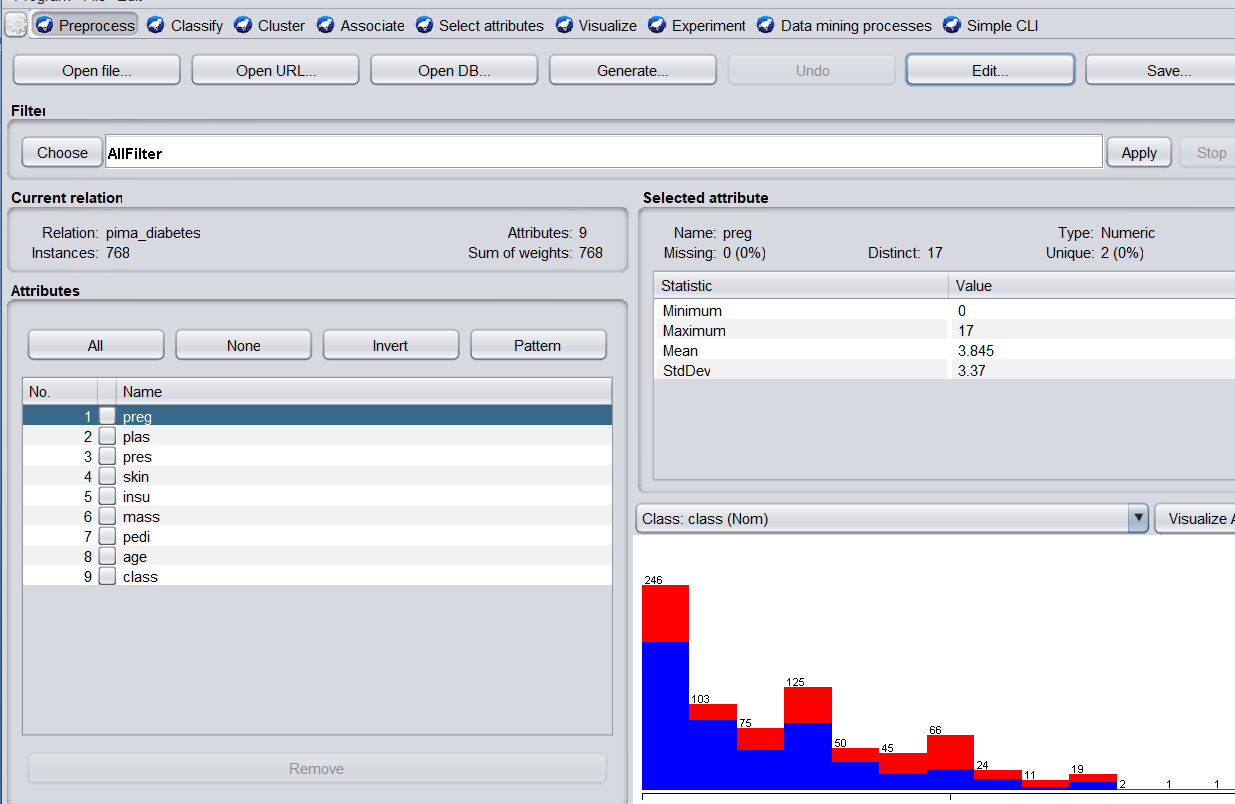


Fig.4. diabetic dataset preprocessing

The visualization of all attributes is shown in “Fig.4.1”.

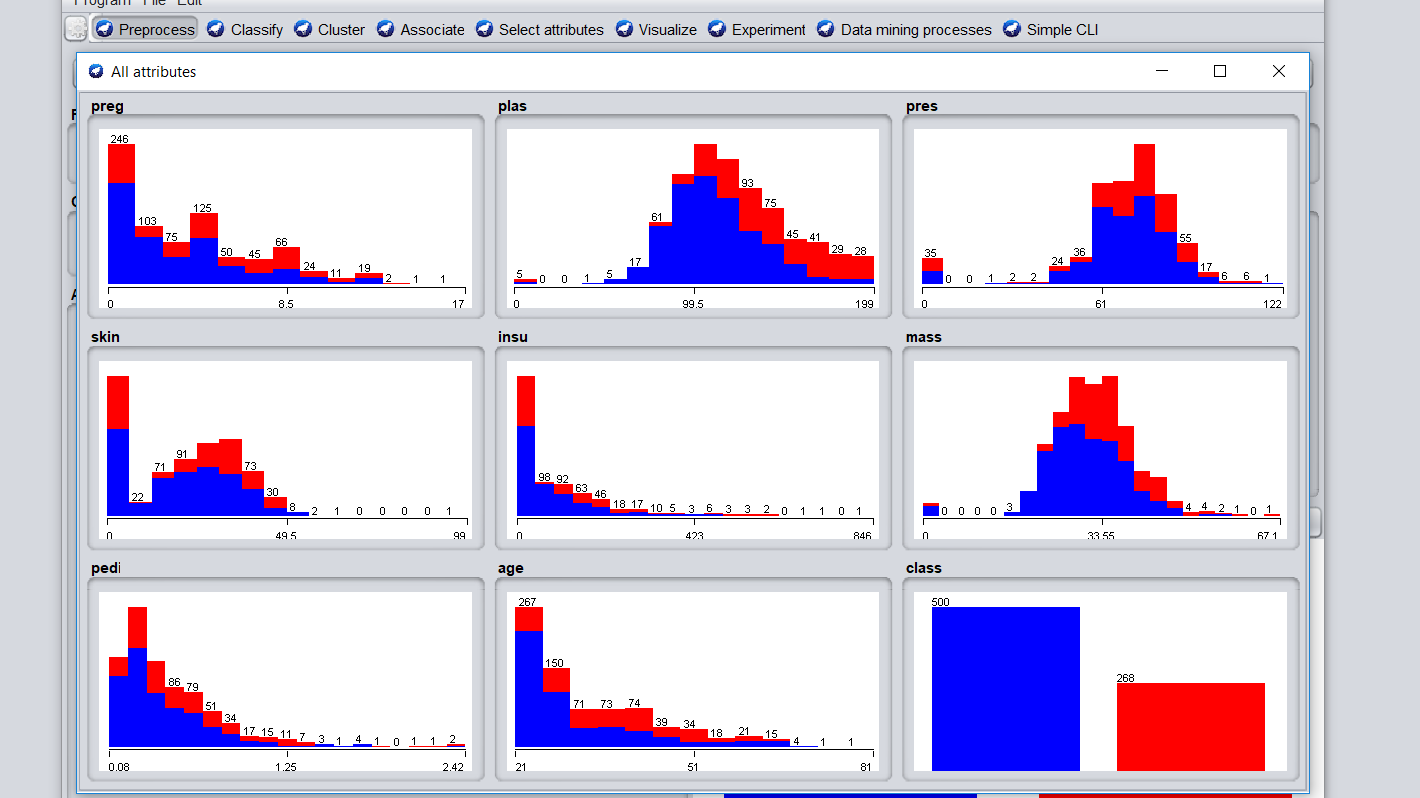


Fig.4.1. Visualization of all attributes

Then classify the data using the J48 algorithm. J48 is a decision-tree-based algorithm. It is an algorithm in machine learning for classifying data categorically and continuously. It measures improved performance and produces a higher rate of accuracy [3].

1. *Algorithm J48:*

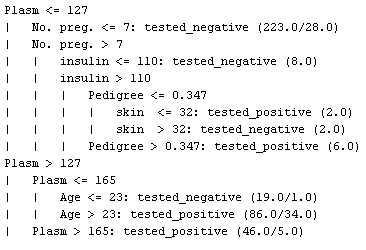


Figure 4.2. J48-Pruned tree

J48 ignores the missing value during the time of building the tree. The J48 algorithm produces the following output shown in “Fig.5”. From the data set J48 builds the decision tree, then it examines results from choosing different attribute for splitting the data. The algorithm continues to run over the smaller subsets. If all instances in a subset belong to the same class the splitting procedure stops. At last the leaf node is created in a decision tree telling to choose that class.

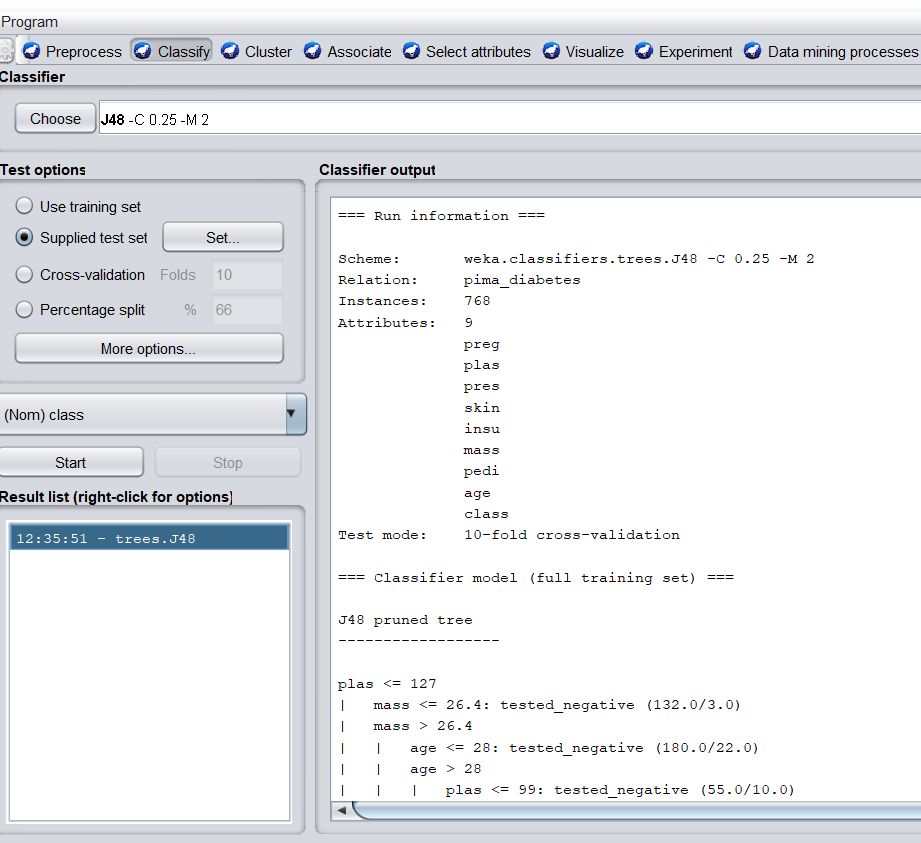


Fig.5. The output of the J48 algorithm

The tree formed using the J48 algorithm is shown in “Fig.5.1”. The problem of identifying to which of a set of categories a new observation belongs is called classification and it is done based on a training set of data containing observations whose category membership is known [4]. To classify the data samples into known classes The training examples are used.

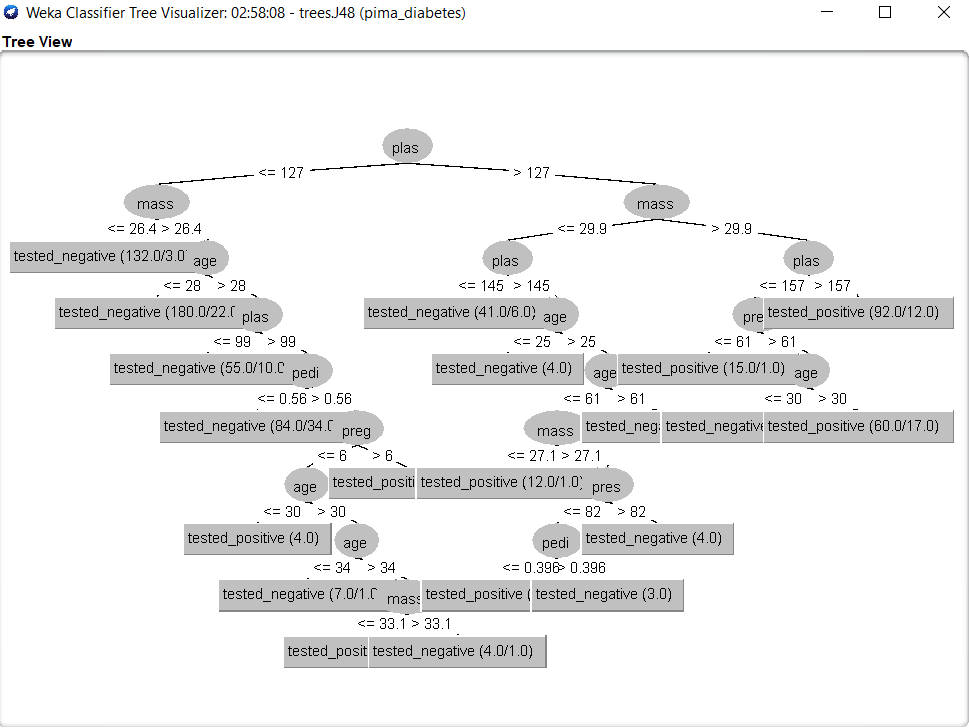


Fig.5.1. Tree of J48 algorithm

1. *Naive Bayesian Classifier*

The result of the Naive Bayes Classifier is shown in “Fig.6”. The Naive Bayes classifier is a classification algorithm based on Bayes Theorem. It is especially fit when the dimensionality of input is high. It is work based on the following Bayesian equation:

*P (T  S) = [ P (S  T) \* P (T)]  P(S)*

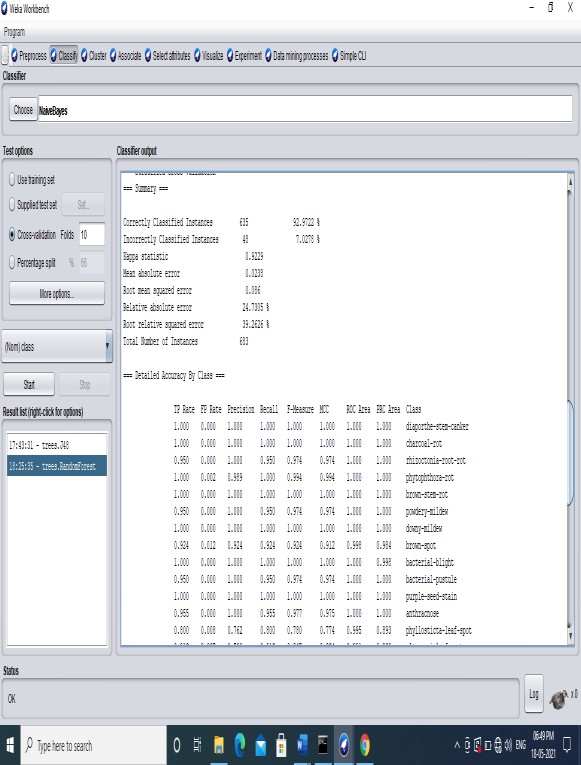


Fig.6 Result of Naive Bayes classifier

1. RESULT

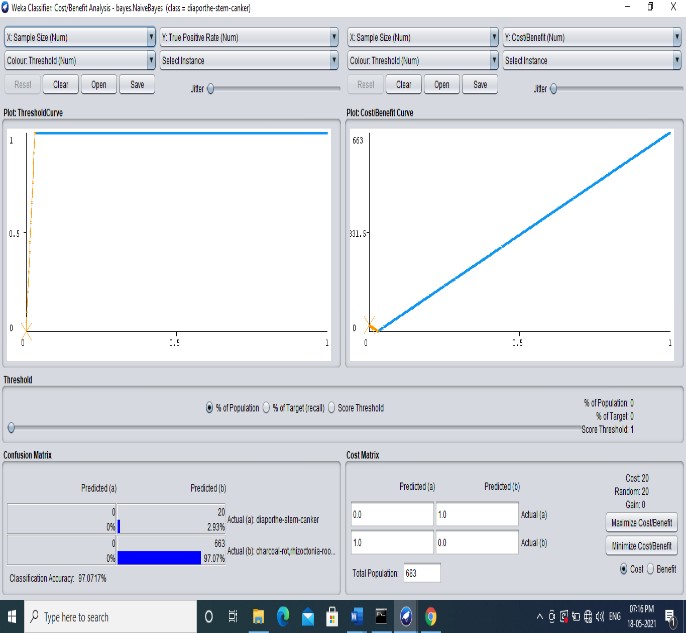
Classification Algorithm Execution

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No** | **Classification technique** | **Accuracy** | **Time** |
| 1 | NAÏVE BAYES | 77.80 | 0 |
| 2 | K-Star | 71.17 | 0 |
| 3 | Random Tree | 75 | 0.02 |
| 4 | OneR | 75.76 | 0 |
| 5 | ZeroR | 66.83 | 0 |
| 6 | Decision Table | 77.8 | 0.06 |
| 7 | Multilayer Perceptron | 74.23 | 1.09 |
| 8 | Simple Logistic | 77.8 | 0.31 |
| 9 | SimpleCart | 76.02 | 0.09 |
| 10 | Hyper Pipes | 66.32 | 0 |
| 11 | JRip | 77.2 | 0.05 |
| 12 | J48 | 79.33 | 0 |

Fig.7 Classification Algorithm Execution

J48 builds the decision tree, then it examines results from choosing different attribute for splitting the data. The algorithm continues to run over the smaller subsets. If all instances in a subset belong to the same class the splitting procedure stops. At last the leaf node is created in a decision tree telling to choose that class.

.

REFERENCES

* 1. K.Priyadarshini, Dr.I.Lakshmi.2017. A Survey on Prediction of Diabetes Using Data Mining Technique from International Journal of Innovative Research in Science, Engineering and Technology, ISSN(Online): 2319-8753 Sonali Vyas, Rajeev Ranjan, Navdeep Singh, Arohan Mathur.2019. Review of Predictive Analysis Techniques for Analysis Diabetes Risk
  2. Dr. K. Thangadurai, N.Nandhini.2016. Comparison of data mining algorithms for prediction and diagnosis of diabetes mellitus from International Journal of Scientific & Engineering Research, Volume 7, Issue 5, ISSN 2229-5518
  3. M Nirmala Devi, Balamurugan.S Appavu alias, U.VSwathi.2013.An amalgam KNN to predict Diabetes Mellitus
  4. Karnika Dwivedi, Dr. Hari Om Sharan.2018. Review on Detection of Diabetes using Data Mining Technology from International Journal of Engineering and Technical Research (IJETR).

Fig.8.Visualization results

V.CONCLUSION

Identification and prediction of diabetes disease at the beginning of the disease is a better method to prevent the disease from causing other deadly diseases. The Diabetes. arff dataset contains different attributes that can be useful for the prediction of diabetes. The ARFF file used in WEKA for diabetes disease prediction contains the details of many patients and their symptoms. Machine learning algorithms and data mining techniques are widely used for prediction and classification. In this research data mining technique applied to classify Diabetes Clinical data and predict the likelihood of a patient being affected with Diabetes or not. Different classification algorithms are applied to Pima Indians Diabetes Database and the result obtained is tabulated in the table. This research can be extended by applying association mining. This subset of the dataset is converted into the required form. This work extends to utilize the implementation of the different datasets.

Comparing the accuracy rate, we can conclude that the Naive Bayes algorithm is more efficient than the J48 algorithm

VI.FUTURE WORK

In this paper, we analyze and predict diabetes disease. Early detection of the disease will help us to prevent the disease. So, the future work we can do is can suggest the proper medicine for each disease that is predicted by the J48 and Naive Bayesian algorithm. For this medicine suggestion, we will use another dataset containing attributes like name of the disease, name of medicine