IJARIIT
International Journal Of Advance Research, Ideas And Innovations In Technology

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 6)
Available online at:https://www.ijariit.com

Data set using Weka tool with data mining techniques

Ashwin R.
win9096@gmail.com
Bannari Amman Institute of Technology, Coimbatore, Tamil Nadu

ABSTRACT

In this paper all the datamining techniques and some of the dataset applications have been illustrated using datamining tool. Similarly Datamining technique is based on the neural network and Genetic algorithm. Here, how the given dataset has been pre-processed using WEKA tool has been illustrated in this paper with of some data pre-processing algorithms and rules.

Keywords*:Data mining, Data pre-processing, classification, clustering, decision tree algorithms, Weka tool*

1. INTRODUCTION

Firstly, we have to know that what is datamining and how it is used in current technology field ? So the best answer from side is that data mining means getting knowledge from the data and it is used in interdisciplinary field brings out the various techniques that have been used in machine learning, pattern recognition , data visualization, collecting statistics and databases .This has an alternative names like knowledge discovery , knowledge extraction, data harvesting, Business intelligence. Let us compare with Bigdata and Data-science. Big data is a huge data that has large number of datasets so it is complex for data pre-processing and inadequate.While data-science is to extract the knowledge from the data.Whereas,Datamining is simple search and Query processing. Then we step into the architecture of data mining in simple ways.

- **User Interface:** It communicates between user and data mining.It visualize the result and perform exploratory on given data or schema.
- **Pattern Evaluation:** It tests the data and evaluates the pattern.
- **Data-Mining Engine:** Perform some of the activities like characterization, classification and prediction.
- **Knowledge Base:** Organise the attributes into the levels of abstraction and user to access the patterns or threshold
- **Data Cleaning:** This requires the source of data for cleaning ,integration and selection.

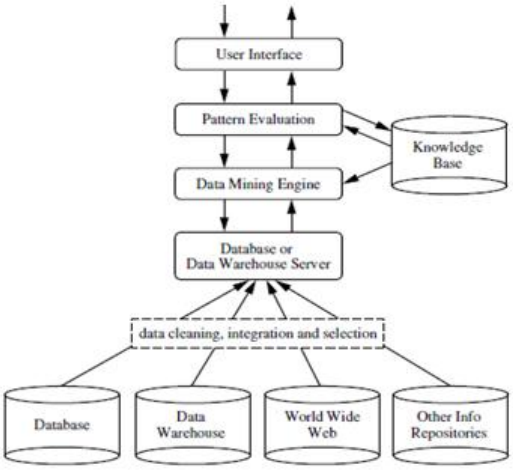


Fig.1: Architecture of Datamining

We clearly understood that some source of data is collected and goes to integration process in terms of datastore and then it is going for data pre-processing well the data is prepared through datamining tool in terms of patterns and symbols it comes to the knowledge to the users in visualization.

2. DATA PREPROCESSING

Data preprocessing is a technique is used to convert the raw material into useful and efficient format.

Steps:

1. **Data cleaning**: It is classified into two types, a.noisy data and b.missing data

a)Noisy data : It is a meaningless data it is later interpreted by systems. This kind of data comes from error in data entry and data collections. This can be overcome by some of the methods like:

Binning Method

This is one of the method is used to sort data to smooth it. The entire data is divided into segments and various methods are followed to overcome the error.

Regression

This method is made the data smooth by fitting it to a regression function. It uses linear regression(single independent variable) and multiple regression (multiple independent variable)

Clustering

This method groups the data and form a cluster. The outliers may be fall outside the clusters.

b) Missing data:

This method arises when there is missing of data. It is classified as follows:

Delete The Data In Tuples

This method follows when there is large number of dataset and multiple values are missing inside the tuple.

Fill The Values

This method follows where you can missing data’s manually by taking mean or some value.

2.**Data Transformation:** It is used to transform the data. The method involves such as:

a) Normalization: The values in the data specifies in the range of some(-1.0 -- 1.0 or 0--1.0)

b) Attribute selection: In this method new attributes are developed with help of given data.

3. **Data Reduction:** Since the datamining handle the large amount of data in order to make the analysis becomes harder and complex in many cases. In order to maintain storage efficiency ,reduce cost and storage, data reduction method is used.

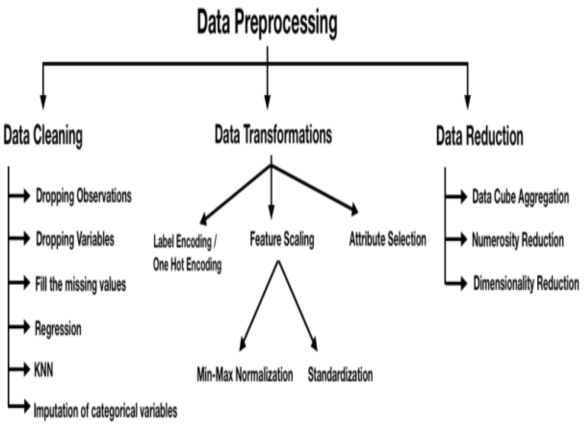


Fig.2:Data preprocessing inData mining

3. WEKA TOOL

Data mining is a process of extracting useful information from large amount of dataset. Data mining is also used in many health care applications to get the better quality of service and to reduce the medicine effect. Hospitals have to reduce the cost for treatment.Health care data is huge and enormous.It consist of patient data,organization data and some medicine data.Treatment has records the millions of patient in data mining techniques answers the essential and needed questions related to health care.Data mining techniques are useful to collect the medicinal data in terms of association rule mining for searching the frequent patterns, classification, prediction and clustering.This techniques are more useful in predicting the heart diseases,breast cancer, diabetes and current situation i.e, covid-19 and etc.In this paper illustrates the diabetes dataset using datamining tool.

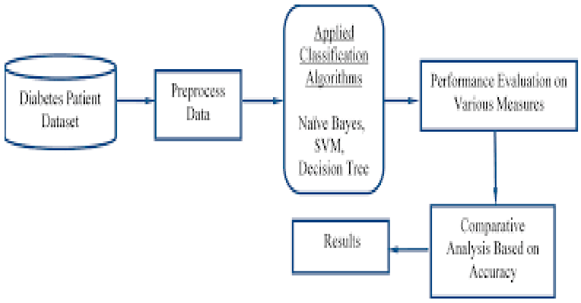
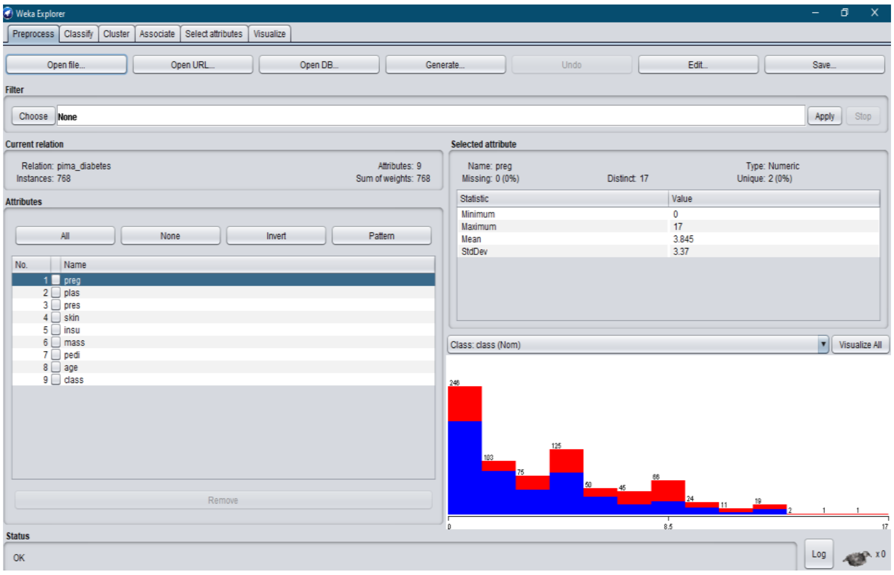

Diabetes Patient Dataset
Preprocess Data
Applied Classification Algorithms
Naive Bayes, SVM, Decision Tree
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Results

Fig. 3: Block Diagram Diabetes Dataset

STEP-1 You have to download the dataset from internet .arff extension (Attribute -relation file format).



Filter: None

Current relation

Relation: pima_diabetes Attributes: 9
Instances: 768 Sum of weights: 768

Selected attribute

Name: preg Type: Numeric
Missing: 0 (0%) Distinct: 17 Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

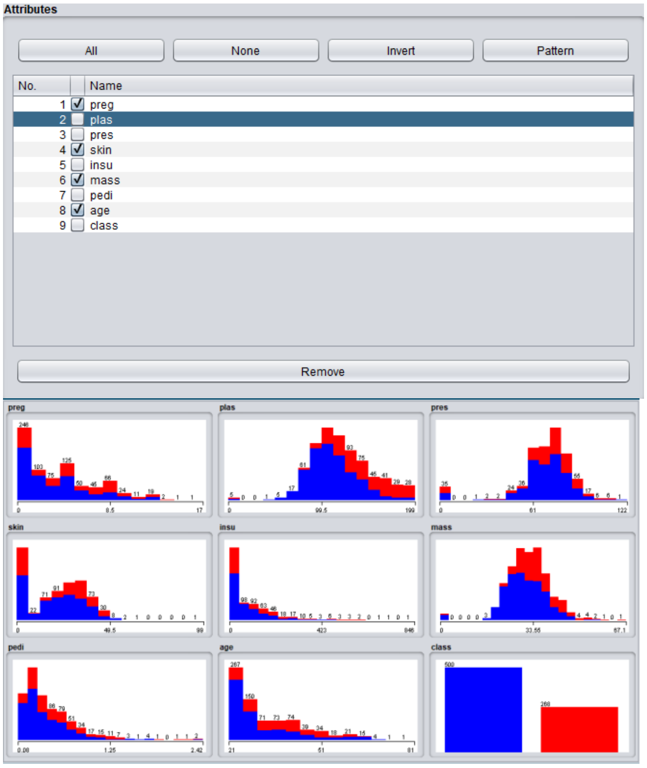
Class: class (Nom)

Attributes

No.		Name
1	□	preg
2	□	plas
3	□	pres
4	□	skin
5	□	insu
6	□	mass
7	□	pedi
8	□	age
9	□	class

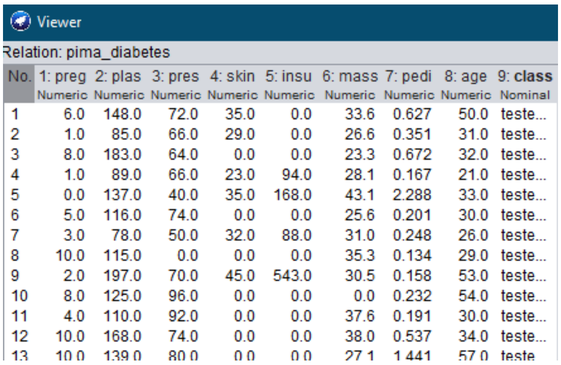
Status: OK

STEP -2 Select the Pre-process tab and Based on the selected attribute result is displayed as graph on the visualization



No.		Name
1	☑	preg
2	□	plas
3	□	pres
4	☑	skin
5	□	insu
6	☑	mass
7	□	pedi
8	☑	age
9	□	class

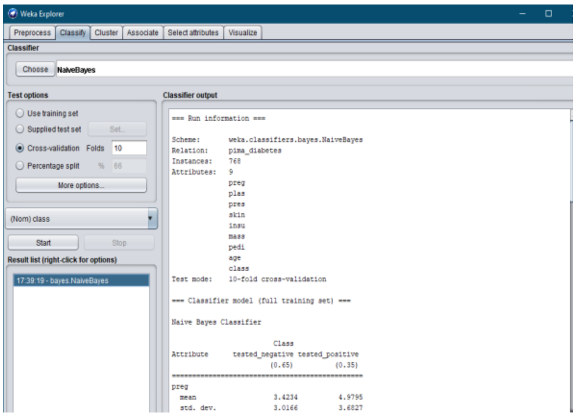
STEP -3 Click edit button on the same section. On this option we can able to edit the dataset such as add, delete and modify the columns and data present in it.



Relation: pima_diabetes

No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	teste...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	teste...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	teste...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	teste...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	teste...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	teste...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	teste...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	teste...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	teste...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	teste...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	teste...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	teste...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	teste...

STEP -4 Next, Click on Classify tab under the drop down box select any one option for your convenient result. I prefer Naïve bayes. For this dataset.



NaiveBayes

Cross-validation Folds 10

Percentage split % 66

(Nom) class

17:39:19 - bayes.NaiveBayes

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     pima_diabetes
Instances:    768
Attributes:   9
              preg
              plas
              pres
              skin
              insu
              mass
              pedi
              age
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                          Class
Attribute      tested_negative tested_positive
                        (0.65)          (0.35)
===============================================
preg
  mean                  3.4234          4.9795
  std. dev.             3.0166          3.6827
```

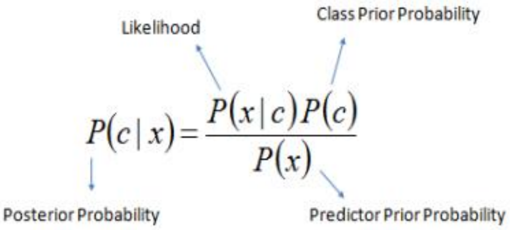
4. NAÏVE BAYES

a) P(c|x) is the posterior probability of class (target) given predictor (attribute).

b) P(c) is the prior probability of class.

c) P(x|c) is the likelihood which is the probability of predictor given class.

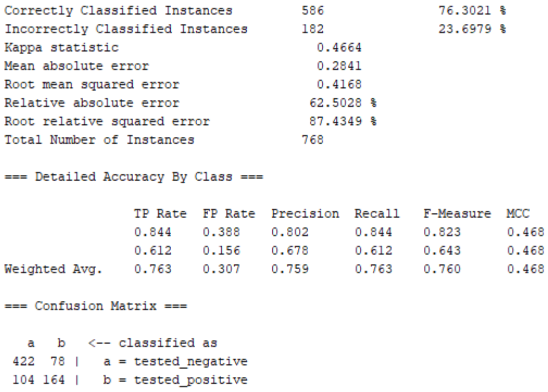
d) P(x) is the prior probability of predictor.



$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

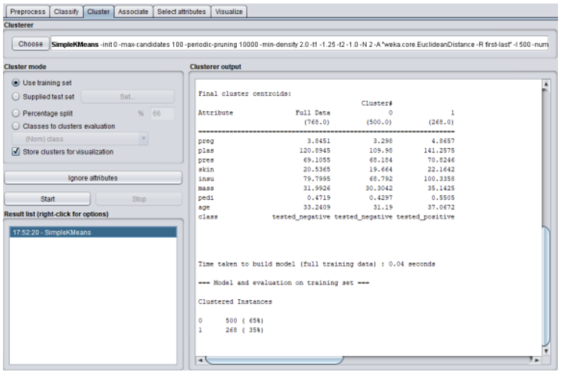
$$P(c\mid \mathrm{X}) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

STEP-5 Result under Naïve bayes classification.



```
Correctly Classified Instances         586               76.3021 %
Incorrectly Classified Instances       182               23.6979 %
Kappa statistic                          0.4664
Mean absolute error                      0.2841
Root mean squared error                  0.4168
Relative absolute error                 62.5028 %
Root relative squared error             87.4349 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.844    0.388    0.802      0.844    0.823      0.468
                 0.612    0.156    0.678      0.612    0.643      0.468
Weighted Avg.    0.763    0.307    0.759      0.763    0.760      0.468

=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

STEP -6 Click on next cluster tab. Select K-means cluster algorithm for this dataset for more accuracy.



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num

Percentage split % 66

(Nom) class

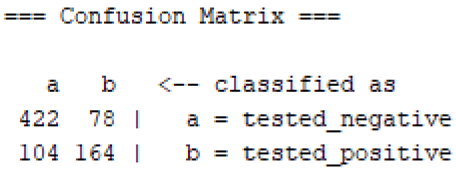
17:52:20 - SimpleKMeans

```
Final cluster centroids:
                                        Cluster#
Attribute                  Full Data           0               1
                             (768.0)     (500.0)         (268.0)
=================================================================
preg                          3.8451       3.298          4.8657
plas                        120.8945      109.98        141.2575
pres                         69.1055      68.184         70.8246
skin                         20.5365      19.664         22.1642
insu                         79.7995      68.792        100.3358
mass                         31.9926     30.3042         35.1425
pedi                          0.4719      0.4297          0.5505
age                          33.2409       31.19         37.0672
class                tested_negative tested_negative tested_positive




Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      500 ( 65%)
1      268 ( 35%)
```

5. RESULT AND CONCLUSION

In this study, the accuracy of datamining techniques is compared. The main goal is to have the higher accuracy. These metrics can be originated from the confusion matrix and it can be easily converted to TP(True-Positive) and FP(Fales-Positive).

Confusion Matrix for Naïve Bayes



```
=== Confusion Matrix ===

   a   b   <-- classified as
 422  78 |   a = tested_negative
 104 164 |   b = tested_positive
```

In this paper, we have discussed some of the effective techniques that can be used for diabetes classification ad accuracy based on the classifier algorithm.An important thing in data mining and machine learning is to build more precise and computationally efficient classifiers for health care and medical applications. The performance of Naïve Bayes shows high level of accuracy compare with other classifiers.

6. REFERENCES

[1] Ian and Eibe Frank, Data Mining, Practical Machine Learning Tools ,Morgan Kaufan,1999.
[2] M. Craven and J. Shavlik, “Learning rules using ANN.