**LAB EXPERIMENT NO. 01**

**Aim:** Perform data Pre-processing task using Weka data mining tool

**Theory:**

WEKA - an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems

**Tasks performed through Weka:**

Preprocessing:

Classification:

Clustering:

Association Rule:

Select Attributes:

Visualization:

A picture containing text, screenshot, indoor

Description automatically generated

Weka GUI

**Preprocessing activities to be observed in Weka:**

1. **Visualization:** Visualize scatter plot for all the attributes from dataset selected from Weka. Determine correlation if any using these plots for different datasets

A picture containing text, screenshot, indoor

Description automatically generated

Weka on loading **Prima Diabetes dataset**

Graphical user interface, application, Word

Description automatically generated

Visualize all – Distribution Plot

Text

Description automatically generated with medium confidence

Correlation between the features

Thus, upon performing **data visualization** we observed :

1. The data distribution with respect to each features and the skewness of the data with respect to that feature.
2. **Scatter plot** between the features. For e.g in the Prima Diabetes dataset:
   * Age vs Pregnancy have no co-relation.
   * Plasma vs insurance were positively correlated
   * Mass vs skin was positively correlated.
3. **Select Attributes:** Apply suitable feature selection filter like GainRatio etc to choose relevant attributes from the list of attributes. Observe the ranks / priority provided by the filter.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Using attribute evaluator as InfoGainAttributeEval on Ranker Search method in the Select Attribute tab we got the order and values of the most important attributes through entropy which further is used for clusters/classification. Thus, from this we know that attribute ‘plas’ , ‘mass’ then ‘age’ and so on hold importance while clustering the instances.

1. **Preprocessing:**
   1. **Visualize All:** Select this button to visualize histograms of all attributes.

Graphical user interface, application, Word

Description automatically generated

* 1. **Filter:** Choose Discretization under Unsupervised and Supervised methods. Observe the discretization and the outliers.

Graphical user interface

Description automatically generated

Filter – Supervised Discretize

Graphical user interface

Description automatically generated

Filter – Unsupervised Discretize

* 1. **IQR:** Observe the IQR values for a selected attribute. Observe the outlier and extreme values

Graphical user interface, application

Description automatically generated

Filter – IOR

Thus, by using IOR filter we can look at the outliers i.e. those values which are outside the 1.5\*IOR range. Data cleaning is necessary as these extreme outliers do affect the model accuracy.

Graphical user interface, application

Description automatically generated

IOR 🡪 detectPerAttribute – True

Teams

Description automatically generated

Visualize all – Outliers and Extreme Values

* 1. **Removethevalue:** Remove instances with outlier values and show the screenshots of dataset before and after the removal.

Application, Teams

Description automatically generated

Before Removal of Outliers

Graphical user interface, application, Teams

Description automatically generated

After Removal of Outliers

1. **Classification:** Perform NB and Random Forest classification

Text

Description automatically generated

A picture containing table

Description automatically generated

Naïve Bayes on pima diabetes

Text

Description automatically generated

A picture containing text

Description automatically generated

Random Forest Classifier

1. **Clustering:** Perform kmeans, hierarchical clustering and explain the output

Graphical user interface, application

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

SimpleKmeans

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

**Hierarchical Clustering**

In clustering the dataset by **KMeans Clustering**, we classified the dataset into 2 clusters:

Cluster 0 – tested\_negative

Cluster 1 – tested\_positive

Out of which a total of 515 instances we classified into Cluster0 and 253 instances were classified into Cluster 1. But out of 515 instances in the Cluster 0, 135 instances were wrongly classified as those were labelled tested\_positive in the actual dataset. Similarly in the cluster1 120 instances were wrongly classified as those were labelled as tested\_negative in the original dataset.

Thus, overall the incorrectly clustered instances were 255 misclassified instances which results to 33.203% error.

Text, letter

Description automatically generated

In clustering the dataset by **Hierarchical Clustering**, we classified the dataset into 2 clusters:

Cluster 0 – tested\_negative

Cluster 1 – tested\_positive

Out of which a total of 767 instances we classified into Cluster0 and 1 data instance was classified into Cluster 1. But out of 767 instances in the Cluster 0, 267 instances were wrongly classified as those were labelled tested\_positive in the actual dataset. In the cluster1, 0 instances were wrongly classified.

Thus, overall the incorrectly clustered instances were 255 misclassified instances which results to 34.7656% error.

Text, letter

Description automatically generated

1. **Association rule mining:** Perform apriori algo and show the rules created

A screenshot of a computer

Description automatically generated with medium confidence

Supermarket dataset

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Association Rules generated by Apriori Analysis

Text

Description automatically generated

**Conclusion:**

In this experiment we learnt to about Weka tool used for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools. Using data visualization we can see the correlation between the features and also the distribution of data for each feature thereby, performing necessary pre-processing needed, like transformation to normalize the skewness in the data. Other pre-processing being data cleaning like removing outliers (i.e those values outside the 1.5IQR range). Also, through Select Attribute we got the preference of each attribute with respect to InfoGain. Then I performed Naives Bayes and Random Forest Classification on the dataset using cross validation. I also performed Clustering Algorithms like Simple KMeans and Hierarchical Clustering and observed the numbers of data instances were correctly and wrongly classified into the new clusters. Finally, performed Association mining using Apriori Algorithm and displayed the association rules.