Data Mining Lab List

Section I - Preprocessing

1. Perform Imputation on Titanic data set
2. Perform Discretization on Iris dataset
3. Perform continuization on Titanic dataset
4. Perform normalization on Iris dataset
5. Perform Randomization on Iris dataset
6. Perform Remove Sparse on zoo data set
7. Perform Feature Selection on Wine dataset
8. For the dataset wine.csv
    a) Replace missing values by the mean of the values of records having the same class value.
    b) Display the entire data after replacement.
    c) Perform binning (3 bins) for the attribute residual sugar
    d) Remove redundant variables/features having high correlation.
    e) Select important variables/features using Information gain and gain ratio.
    f) Perform normalization [-1,1] on the attribute quality and display the full dataset.
    g) Do a stratified random sampling to draw a sample size of approximately 100 out of the total records.
    h) Split the dataset into 70%training data set and 30% test dataset

9. Use mtcars data set to,
    a) Replace the missing data with the average/median of the feature wt
    b) Transform the numerical variable am to manual-0 and automatic-1.
    c) Transform the numerical variable gear by appending "gear" to the no.of gears given in the feature.
    d) Add a new attribute Engine type based on the condition for the attribute vs (0 = V-shaped, 1 = straight)
    e) Scale the feature disp

Section II - Data Visualization

1. Use car.csv data set to,

a) Plot a bar chart to compare the price of different makes of car.
b) Create a histogram for analyzing make and mileage.
c) Create a histogram for analyzing price. Show a stacked column distribution with respect to Type. Write your inferences for the price of cars with respect to the above variables.
d) Visualize a bar plot for, model Vs door. Write your inferences.
e) Create a boxplot for price w.r.t make
f) Create a violin plot for price w.r.t type.

2. Illustrate the following using diamonds data set

a) Create a histogram of "carat" w.r.t cut
b) Set the bin width of the histogram to 20
c) Make a scatterplot: carat vs price, set the color to clarity
d) Make a scatterplot: carat vs price, set the color to clarity. Also add regression line to the plot
e) For carat vs cut, make a violin and a boxplot.
f) Illustrate Heat map and Venn Diagram using the data set.
g) Illustrate freeviz, linear projection and radviz usig the data set.

Section III - Association Rule Mining

1. Generate association rules using Market Basket Data set and compare the different measures to assess the quality of rules.
2. Generate association rules using the Food Mart Data set and compare the different measures to assess the quality of rules.

Section IV - Classification

Demonstration of classification and prediction techniques – Analysis and evaluation of Model Performance. Explain the evaluative report of the classifier for the generated classifiers.

1. Generate a classifier from Iris dataset using Decision Tree.
2. Generate a classifier using Housing Dataset Decision using Decision Tree and Naïve Bayesian Classifier and compare the results.
3. Generate a classifier in Orange Tool from Titanic dataset using K-Nearest Neighbor and SVM Classification. Compare the models.
4. Generate a classifier for housing dataset using Linear Regression.
5. Generate a classifier for heart disease dataset using Logistic Regression.

Section V - Clustering

1. Demonstration of Clustering Techniques-Analysis and Evaluation of Model Performance on Iris Dataset using K-Means Algorithm
2. Demonstration of Clustering Techniques- Analysis and Evaluation of Model Performance on Housing Dataset using K-Means Algorithm
3. Demonstration of Clustering Techniques Analysis and Evaluation of Model Performance using Course Grades dataset in with Hierarchical Clustering Algorithms

Section VI - Project

i. Abstract
ii. Introduction
iii. Materials and Methods
iv. Data Visualization and Interpretation
v. Preprocessing and Feature Selection
vi. Model Construction (at least 5 models)
vii. Performance and Evaluation of Models
viii. Results and Discussion
ix. Conclusion
x. References