DEPARTMENT OF COMPUTER SCIENCE
RAJAGIRI COLLEGE OF SOCIAL SCIENCES
(Autonomous)

M.Sc. COMPUTER SCIENCE
(Data Analytics)

DATA MINING LAB
CSDA 207
LAB RECORD

NAME : BALU S UNNY

SEMESTER : SECOND

REGISTER NO : 2217013

DEPARTMENT OF COMPUTER SCIENCE
RAJAGIRI COLLEGE OF SOCIAL SCIENCES
(Autonomous)

M.Sc. COMPUTER SCIENCE (Data Analytics)

CERTIFICATE

NAME : BALU S UNNY

SEMESTER : SECOND

REGISTER NO : 2217013

Certified that this is a bonafide record of work done by ***BALU S UNNY*** *MSCCS2211 in the Software Laboratory of Rajagiri Department of Computer Science, Kalamassery.*

Sunu Mary Abraham
Faculty in Charge

Dr. Bindiya M Varghese
Dean, Computer Science

Internal Examiner

External Examiner

Place : Kalamassery
Date :

Table of Contents

SECTION	TOPICS	PAGE NO
Section 1 A	**Preprocessing (Orange Tool)**	
1	Perform Imputation on dataset Heart Disease	1
2	Perform Discretization on Iris dataset	4
3	Perform continuization on Titanic dataset	7
4	Perform normalization on Iris dataset	10
5	Perform Randomization on Iris dataset	13
6	Perform Remove Sparse on zoo data set	16
7	Perform Feature Selection on Wine dataset	18
8	Perform Feature Selection on Lenses dataset	20
Section 1 B	**Preprocessing (Orange Tool)**	
	Dataset student.csv	
1	Replace missing values by the mean of the values of records having same class value. Display the entire data after replacement.	22
2	Perform binning(3 bins) for the attribute Announcements View.	24
3	Remove redundant variables/features having high corelation.	26
4	Select important variables/features using Information gain and gain ratio.	28
5	Perform normalization [-1,1] on the attribute raised hands.	29
6	Do a stratified random sampling to draw a sample size of approximately 100 out of the total records.	30
7	Partition the data into 2 data sets(60:40) using random partitioning.	32
	Dataset mtcars.csv	
1	Replace the missing data with the average/median of the feature wt	34
2	Transform the numerical variable am to manual-0 and automatic-1.	36
3	Transform the numerical variable gear by appending “gear” to the no. of gears given in the feature.	38
4	Add a new attribute Engine type based on the condition for the attribute vs (0 = V-shaped, 1 = straight)	40
5	Scale the feature disp	42

6	Split the dataset into 70%training data set and 30% test dataset	44
Section 1 C	**Preprocessing (WEKATool)**	
1	Perform Feature Selection on Wine dataset (correlation)	45
2	Perform Feature Selection on Lenses dataset(Information Gain and Gain Ratio)	46
Section 2 A	**Data Visualization (Orange Tool)**	
	Dataset car.csv	
1	Plot a bar chart to compare the price of different makes of car.	48
2	Create a histogram for analysing city mileage.	50
3	Create a histogram for analysing price. Show a stacked column distribution with respect to fuel type. Similarly create a histogram for price w.r.t body style and price w.r.t engine location. Write your inferences for price of cars w.r.t the above variables.	52
4	Visualize a bar plot for engine size Vs make. Similarly visualize a bar plot for city mpgvs fuel type and write your inferences.	55
5	Create a scatter plot for price, vs engine size, w.r.t num_of_cylinders(color), aspiration(shape), wheel_base(size).	57
6	Create a boxplot for price w.r.t body_styles.	59
7	Create a violin plot for price w.r.t aspiration.	61
8	Illustrate sieve diagram and mosaic display for city_mpg vs highway_mpg	63
	Dataset diamonds.csv	
1	Create a histogram of “carat” w.r.t cut	65
2	Set the bin width of the histogram to 20	67
3	Make a scatterplot: carat vs price, set the color to clarity	69
4	Make a scatterplot: carat vs price, set the color to clarity. Also add regression line to the plot	71
5	For carat vs cut, make a violin and a boxplot.	73
6	Illustrate Heat map and Venn Diagram using the data set.	75
7	Illustrate freeviz, linear projection and radviz usig the data set.	77
Section 2 B	**Data Visualization (Weka)**	

1	Give a visualization of the distribution of Iris dataset w.r.t all the features	78
2	Display the plot matrix for the Iris data set	79
Section 3 A	**Association Rule Mining**	
1	Generate association rules using Market Basket Data set in Orange Tool. Compare the different measures to assess the quality of rules.	80
2	Generate association rules using Food mart Data set in Orange Tool. Compare the different measures to assess the quality of rules.	81
3	Generate association rules using supermarket Data set in WEKA using Apriori algorithm.	84
4	Generate association rules using supermarket Data set in WEKA using FP – growth Algorithm	85
Section 3 B	**Association Rule Mining**	
1	Generate association rules using Lenses Data set in Orange Tool	86
2	Generate association rules using Lenses Data set in WEKA using Apriori algorithm.	88
3	Generate association rules using Lenses Data set in WEKA using FP – growth Algorithm	90
Section 4	**Classification**	
1	Generate a classifier in Orange Tool from Iris dataset using Decision Tree.	91
2	Generate a classifier in Orange Tool from Titanic dataset using Decision Tree	92
3	Generate a classifier in Weka Tool from Pima- Diabetes dataset using Decision Tree.	93
4	Generate a classifier in Weka Tool from contact lenses dataset using Decision Tree.	94
5	Generate a classifier in WEKA using Housing Dataset Decision using Decision Tree and Naïve Bayesian Classifier and compare the results.	96
6	Generate a classifier in Orange Tool from Iris dataset using K-Nearest Neighbour and SVM Classification. Compare the models.	97
7	Generate a classifier in WEKA using Iris dataset with K-Nearest Neighbour Classification and SVM Classification. Compare the models.	98
8	Generate a classifier in Orange Tool /WEKA for diabetes dataset using Linear Regression.	99
9	Generate a classifier in orange Tool/WEKA for heart disease dataset using Logistic Regression.	100

10	Generate at least 5 classifier models in Weka /Orange for the titanic data set provided. Do the necessary preprocessing before training the model? Evaluate the models generated and choose the best model	101
Section 5	**Clustering**	
1	Demonstration of Clustering Techniques-Analysis and Evaluation of Model Performance on Iris Dataset in Orange Tool using K-Means Algorithm	102
2	Demonstration of Clustering Techniques- Analysis and Evaluation of Model Performance on Housing Dataset in Orange Tool using K-Means Algorithm	103
3	Demonstration of Clustering Techniques- Analysis and Evaluation of Model Performance on Housing Dataset in WEKA Tool with K-Means Algorithm	104
4	Demonstration of Clustering Techniques Analysis and Evaluation of Model Performance using Course Grades dataset in Orange Tool with Hierarchical Clustering Algorithms	106
5	Demonstration of Clustering Techniques - Analysis and Evaluation of Model Performance using Iris dataset in WEKA with Hierarchical Clustering Algorithms	108
Section 6	Project Report	110