B.E. COMPUTER SCIENCE & ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)
Outcome Based Education (OBE) and Choice Based Credit System (CBCS)
SEMESTER - V

DATA MINING AND VISUALIZATION LABORATORY

Course Code	S5CCSL01	CIE Marks	50
Teaching Hours/Week (L:T:P)	(0:0:2)	SEE Marks	50
Credits	1	Exam Hours	3
Lecture Hours	-	Practical Hours	28hrs

Course objectives: This course will enable students to:	
1	To understand the basic concepts of data mining
2	To implement Data Mining techniques, their need, scenarios (situations) and scope of their applicability
3	Synthesize the solution through advanced data mining tool to solve data analytics problems

1. Experiment to be conducted using WEKA tool:

	outlook	temperatu	humidity	windy	play
2					
3	sunny	85	85	FALSE	no
4	sunny	80	90	TRUE	no
5	overcast	83	86	FALSE	yes
6	rainy	70	96	FALSE	yes
7	rainy	68	80	FALSE	yes
8	rainy	65	70	TRUE	no
9	overcast	64	65	TRUE	yes
10	sunny	72	95	FALSE	no
11	sunny	69	70	FALSE	yes
12	rainy	75	80	FALSE	yes
13	sunny	75	70	TRUE	yes
14	overcast	72	90	TRUE	yes
15	overcast	81	75	FALSE	yes
16	rainy	71	91	TRUE	no

i) *Preprocess(Data Cleaning, Data Integration, Data Transformation, Data Reduction)* and *Classify (Posteriori and Priori)* panels. Analyze Input and Output Attributes.
ii) Calculate the information of the whole data set on the basis of whether play is held or not.
iii) Draw the histogram to show how the values of the *play* class occurs for each value of the *outlook* attribute
iv) Derive minimum and maximum values, mean, and standard deviation
v) Perform operations such as filter, delete, invert, Pattern, Undo, Edit, search, Select, Conversions etc
vi) Examine the Output , classification error and Kappa statistics
vii) Visualize threshold curve
viii) Apply Logistic Regression model to classify
ix) Measure the log likelihood of the clusters of training data. (Consider large data set.)
x) Derive Information gain
xi) Build Decision Tree on Humidity attribute. Also demonstrate decision tree after analysis of
    a) Sunny and Overcast dataset
    b) Sunny, Overcast and Rainy Data set
xii) Compute Gini Index representing with respect to Temperature, Humidity, and Windy attributes.
xiii) Obtain the Prediction of Play 'Yes' as well as 'No' for an unknown instance
xiv) Apply Naïve Bayes Classifier to the Weather play data set and derive the probability for play no given outlook rainy

	xv) Apply classification on given data. Remove label and then apply clustering. Perform class to cluster evaluation. Apply classification on un labelled dataset by removing play attribute. Prepare the analysis report. **Interpret the results.**
2.	Experiment using WEKA tool. Consider the following data set

No.	eid Numeric	ename Nominal	salary Numeric	exp Numeric	address Nominal
1	101.0	raj	10000.0	4.0	pdtr
2	102.0	ramu	15000.0	5.0	pdtr
3	103.0	anil	12000.0	3.0	kdp
4	104.0	sunil	13000.0	3.0	kdp
5	105.0	rajiv	16000.0	6.0	kdp
6	106.0	sunitha	15000.0	5.0	nlr
7	107.0	kavitha	12000.0	3.0	nlr
8	108.0	suresh	11000.0	5.0	gtr
9	109.0	ravi	12000.0	3.0	gtr
10	110.0	ramana	11000.0	5.0	gtr
11	111.0	ram	12000.0	3.0	kdp
12	112.0	kavya	13000.0	4.0	kdp
13	113.0	navya	14000.0	5.0	kdp

Use the data sources, like ARFF, XML ARFF files. Do the following

a. Classify , Invoke MultiLayerPerception
b. Build neural network GUI as below

ii) Beginning the process of editing the network to add a second hidden layer
iii) The finished network with two hidden layers
iv) Apply Lazy classifier, multi instance classifier
v) Apply any MetaLearning Algorithm
vi) Optimize base classifier's performance
vii) Use clustering algorithm such as Cobweb, and Hierarchical Cluster
viii) Select attribute by specifying an evaluator and a search method
ix) Insert 30 more records in this file. Perform clustering on this new dataset. Identify optimum number of clusters. After identification of optimum number of clusters, prepare clustering on this number.
x) Perform data analysis on the result obtained and prepare an analysis report for the same
xi) Apply Apriori, Interpret the results.
xii) Apply Association rules and interpret the results.
xiii) Apply Association mining with the Apriori alforithm and find the best rules with threshold value of support of 50% and confidence of 70%
xiv) Interpret the results obtained at Summary. Analyse Precision, Recall, F Score values.
xv) Install R package at Weka environment and install rpart package. Implement decision tree.

3.	Consider the data set given below

Relation: employee

No.	age Nominal	income Nominal	stud Nominal	creditrate Nominal	buyscomp Nominal
1	L20	high	no	fair	yes
2	20-40	low	yes	fair	yes
3	G40	medium	yes	fair	yes
4	L20	low	no	fair	no
5	G40	high	no	excellent	yes
6	L20	low	yes	fair	yes
7	20-40	high	yes	excellent	no
8	G40	low	no	fair	yes
9	L20	high	yes	excellent	yes
10	G40	high	no	fair	yes
11	L20	low	yes	excellent	no
12	G40	high	yes	excellent	no
13	20-40	medium	yes	excellent	yes
14	L20	medium	yes	fair	yes
15	G40	high	yes	excellent	yes

a. Load ARFF file and explore knowledge flow interface
b. Configure the data source , check the status area after executing the configuration
c. Perform operations such as Attribute Selection, Filter, Classify, Data Sink, Visualization and Evaluation
d. Apply incremental learning and analyze the result
e. Do clustering : use generator properties, two clustering schemes, and result panel
f. Generate Confusion Matrix and Interpret the results
g. Construct Decision tree
h. Perform Linear Regresssion and Analyze , Validate and Visualize the data Apply Association mining with the Apriori alforithm and find the best rules with threshold value of support of 50% and confidence of 70%
i. Interpret the results obtained at Summary. Analyse Precision, Recall, F Score values.

I. <u>Consider glass data set.</u>

i) How many attributes are there in the dataset? What are their names? What is the class attribute? Run the classification algorithm IBk (weka.classifiers.lazy.IBk). Use cross-validation to test its performance, leaving the number of folds at the default value of 10.

ii) What is the accuracy of IBk (given in the Classifier Output box)? Run IBk again, but increase the number of neighboring instances to k = 5 by entering this value in the KNN field. Use cross-validation as the evaluation method.

iii) What is the accuracy of IBk with five neighboring instances (k = 5)?

iv) Obtain best accuracy higher than the accuracy obtained on the full dataset. Verify ,Is this best accuracy an unbiased estimate of accuracy on future data?

v) Record the cross-validated accuracy estimate of IBk for 10 different percentages of class noise and neighborhood sizes

vi) Analyze, What is the effect of increasing the amount of class noise?

vii) Analyze, What is the effect of altering the value of *k*?

viii) Verify the amount of training data

Additional: For both problems defined under I. and II. Use R package installed at Weka environment, derive decision tree. Interpret the results.

4. Set up SQL database and insert sample data consisting of customer data. Integrate SQL database with Weka.

j. Perform data preprocessing, classification, clustering tasks on customer data.
k. Apply filters and interpret the results
l. Connect WEKA to a relational database using JDBC.
m. Retrieve customer data directly using SQL queries.
n. Load data into WEKA's Explorer interface.
o. Apply classification algorithms (e.g., J48, Naive Bayes) to predict customer spending behavior.
p. Use clustering (e.g., k-means) for market segmentation.
q. Generate evaluation metrics (accuracy, precision, recall).

	r. Optionally automate data refresh using scripts or the WEKA KnowledgeFlow interface. Validate and analyse the result Additional: Derive Fact Table, Star schema, Snowflake schema and do the comparison
5.	Using Tableau do the following. Consider Titanic Data Set. Perform the following using Tableau platform: i. Perform Calculations: i) Calculate the survival rate. Hint: you can use the following formula SUM(IIF([Survived] = 1, 1, 0)) / COUNT([PassengerId]) ii) Calculate the average iii) Calculate the total fare 2. Perform Group Operations: i) Create a group to categorize passengers by age ii) Create a group to categorize passengers by fare iii) Create a set to include only passengers who survived 3. Perform Set Operations: i) Create a set to include only passengers who survived ii) Create a set to include only passengers who did not survive iii)Create a set to include only passengers who traveled in first class 4. Create Dashboard i). Survival Rate by Age Group: Create a bar chart to display the survival rate by age group. ii). Average Fare by Fare Group: Create a bar chart to display the average fare by fare group. iii). Survivors by Class: Create a pie chart to display the number of survivors by class 5. Add Additional Visualizations 6. Combine Dashboards
6.	Using Tableau platform, do the following. Consider Brazilian E-Commerce Public Dataset by Olist from Kaggle. Perform the following using Tableau platform: i) Create a dashboard that displays the total sales, average sales price, and sales quantity for each state in Brazil, filtered by city, with custom geocoding and map layers. Generate mobile responsive dashboard ii) Calculate Total Sales per State, Average Sales Price per State, Sales quantity per state, Forecast sales by region iii) Generate map view iv) Interpret the KPI card v) Demonstrate Advanced Mapping technique vi) Use custom TopoJSON map of Brazil with state and city boundaries. Add heatmap to show density of sales vii) Demonstrate slicers. Add dynamic slicers. Sync slicers across multiple pages Additional: Enable **data export** by filtered view and configure **email alerts** or subscriptions when sales in any region drop below a threshold
7.	Consider Sales Data set. Perform the following using powerBI platform. 1. Create tables with i) manufacturing, sum of sales, columns. ii) Product Category, sum of profits etc iii) Find Answers for the following: a) Which Manufacturer has the highest Sales? b) Which Product Category has the lowest Profit value? c) Which Channel has the highest Cost of Sales? d) Which Manufacturer has the highest Profit? e) Which Promotion Name has the highest Sales? f) Which Product Sub Category has the highest Profit? iv) Perform cross flitering between tables v) Create matrix visualization for product category , Region, sum of sales

	vi) Create a card visualtion for total sales, Total Profits, Avg Profits, Highest Profits, Lower Profits etc. and apply formatting 2. Apply appropriate visualization. Create calculations for Visualization i) Calculate the difference between sum of Sales and Sum of Profits ii) Calcualte Profit Ratio. (Hint: Sum of Profit/Sum of Sales). Also demonstrate using in built funciton, e.g., DIVIDE, AVERAGE etc iii) Calcualte the average sales per product for each of manufacturers. (Hint: Sum of sales/Count of Product Name) iv) Calcualte the Percentage of grand total using inbuit function 3. Apply Filters and Slicers 4. Build Graphs. Draw Trend Analysis Graph. Show trends and forecasting. For example: sum of sales per year, month, quarter etc 5. Create Interactive Dashboard i) Create an Interactive Report The Sales Manager would like you to please create the following interactive report in Power BI Desktop: a) Create a heading - Sales Report b) Create the following Card visualizations: i. Total Sales ii. Total Profit iii. Average Sales iv. Number of Products v. Create a Area graph displaying Sales by Year and Quarter vi. Create a Column graph displaying Profit by Product Category
8.	Consider Sales Data set. Integrate R environment to power BI platform and perform the following i) Create Smart Narrative ii) Using R packages, Build ML based Data Model iii) Perform Visulaization and build report
9.	Consider Sales Data set. Integrate R environment to power BI platform and perform the following i) Create Q&A Visuals ii) Using R packages, Build ML based classification Model iii) Perform Visulaization and build report
10.	Consider Sales Data set. Integrate R environment to power BI platform and perform the following i) Create decomposition tree ii) Using R packages, Build ML based Clustering Model iii) Perform Visulaization and build report
11.	OPEN ENDED MICRO PROJECT

Course Outcomes: On Successful completion of this course, students will be able to	
1	To understand the need for Data Mining and advantages to the business world
2	To get a clear idea of various classes of Data Mining techniques, their need, scenarios (situations) and scope of their applicability
3	To learn the algorithms used for various type of Data Mining problems
4	To understand how to explore and communicate data using data visualization techniques using data analytics tools such as PowerBI and Tableau
5	To derive solutions where business intelligence analytics is applicable.

Course Articulation Matrix (CO-PO and CO_PSO MAPPING)

Course Outcomes	Program Outcomes											PSOs		
	1	2	3	4	5	6	7	8	9	10	11	1	2	3
CO1			2											2
CO2			2											2
CO3			2											2

CO4			2										2
Overall CO			**2**										2

Program Articulation Matrix:

Course Outcomes	**Program Outcomes**											**PSOs**		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**1**	**2**	**3**
			2											**2**