CODE	COURSE NAME	CATEGORY	L	T	P	CREDIT
ITL411	DATA ANALYTICS LAB	PCC	0	0	3	2

Preamble: Data analytics lab is a practical course to supplement the Data analytics theory course. The implementation of machine learning algorithms using R and experimenting with the dynamic, interactive visualization techniques using Tableau will equip the students to pursue careers in the data analytics domain. A familiarization of the popular analytic tools like Hadoop can help in academic projects or to carry out data analysis in new application areas.

Prerequisites:

- ITT201 - Data Structures
- ITT 206 - Database Management Systems
- MAT 208 - Probability, Statistics and Advanced Graph theory
- ITT 306 - Data Science

Course Outcomes: After the completion of the course the student will be able to:

CO No.	Course Outcome (CO)	Bloom's Category Level
CO 1	Solve simple problems of statistical analysis of data using Microsoft Excel	Level 3: Apply
CO 2	Analyze the textual dataand time series data with the data visualization techniques in R	Level 3: Analyze
CO 3	Implement the basic statistical techniques and machine learning algorithms using R	Level 3: Apply
CO 4	Execute HDFS commands and apply Map Reduce technologies associated with big data analytics using HADOOP	Level 3: Apply
CO 5	Analyzereal world data by applying the suitable visualization techniques in Tableau	Level 4: Analyze

Mapping of Course Outcomes with Program Outcomes

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
CO 1	3	3	3	3	3	-	-	-	-	-	-	2
CO 2	3	3	3	3	3	2	2	2	2	2	-	2
CO 3	3	3	3	2	3	-	-	-	-	-	-	2
CO 4	3	3	3	2	3	-	-	-	-	-	-	2
CO 5	3	3	3	3	3	2	2	2	2	2	-	2

3/2/1: High/Medium/Low

Assessment Pattern

Mark distribution

Total Marks	Continuous Internal Evaluation (CIE)	End Semester Examination (ESE)	ESE Duration
150	75	75	2.5 hours

Continuous Internal Evaluation Pattern:

Attendance : 15 marks
Continuous Assessment : 30 marks
Internal Test (Immediately before the second series test) : 30 marks

End Semester Examination Pattern: The following guidelines should be followed regarding award of marks

(a) Preliminary work : 15 Marks
(b) Implementing the work/Conducting the experiment : 10 Marks
(c) Performance, result and inference (usage of equipments and troubleshooting) : 25 Marks
(d) Viva voce : 20 marks
(e) Record : 5 Marks

General instructions: Practical examination to be conducted immediately after the second series test covering entire syllabus given below. Evaluation is a serious process that is to be conducted under the equal responsibility of both the internal and external examiners. The number of candidates evaluated per day should not exceed 20. Students shall be allowed for the University examination only on submitting the duly certified record. The external examiner shall endorse the record.

Sample Course Level Assessment Questions

Course Outcome 1 (CO1):

1. Use Excel's Descriptive Statistics data analysis tool to show the descriptive statistics for the two samples.

Sample 1	Sample 2
28	14
19	18
56	9
23	2
24	26
35	32
99	100
10	62
4	53
67	25
82	84
17	36
33	20
61	46
48	21
85	44
72	65
97	74
12	35
29	15
34	11

2.The given data shows the age of individuals and their average medical expenses per month. Apply linear regression in Excel to draw the regression line and predict the average medical expenses of specific individuals.

Age(X)	Average Amount Spent on Medical Expenses(per month in Rs)(y)
15	100
20	135
25	135
37	150
40	250
45	270
48	290
50	360
55	375
61	400
64	500
67	1000
70	1500

3. Consider the waiting time of the customer at the cash counter of the SBI bank branch during peak hours, which was observed by the cashier. Create a histogram in Excel based on the below data.

Customer Waiting Time (in mins)
2.30
5.00
3.55
2.50
5.10
4.21
3.33
4.10
2.55
5.07
3.45
4.10
5.12

Course Outcome 2 (CO2):

1. Write an R program to perform sentiment analysis using the movie review dataset. (Reference Dataset: https://ai.stanford.edu/~amaas/data/sentiment/)
2. Write an R program to create a corpus of documents and preprocess them in R using stemming, stop word removal, whitespace removal, convert them to lowercase and remove punctuations.
3. Write an R program to create a term document matrix for a corpus in R.
4. Write an R program to find the frequent terms in a document and remove sparse terms in R.
5. Plot a distance versus time scatter plot.
6. Implement the analysis of single variable and multi variable data using histogram, boxplot,whisker plot, barplot and scatter plot (Use the default mtcars dataset in R).
7. Collect data related to user's preferences for products and implement a product recommender system.

Course Outcome 3 (CO3):

1.Given the following data about average rainfall in every month in the year of 2017.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Rainfall (mm)	10	10	10	10	10	560	640	520	320	90	20	10

Calculate Arithmetic, Geometric, Harmonic mean, Median and Mode, First quartile, 56thpercentile for the above data using R.

2.Interpret the data in Anscombe dataset in R with linear regression.

3. Use logistic regression to find the best predictor variables for customer churn prediction.

4. Using decision trees, predict whether to play golf given factors such as weather outlook, temperature, humidity, and wind.

5. Group 620 high school seniors based on their grades in three subject areas: English, mathematics, and science with K-means clustering method.

Course Outcome 4 (CO4):

1. Write a map reduce program to count the words.
   (https://www.kaggle.com/rtatman/english-word-frequency)
2. Write a map reduce program to mine weather data (Data available at https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all)
3. Write a map reduce program to analyze web log files (https://www.microsoft.com/en-in/download/details.aspx?id=37003)
4. Write HDFS commands to:
   - Create a directory on HDFS in home directory.
   - Create two directories in a single command in home directory.
   - List the directories created in HDFS.
   - Create a sample text file in any of the directories created above.
   - Copy file/files from local file system to one of the directories created on HDFS.
   - Verify the file upload.
   - Copy a file from HDFS to local file system
   - Copy the file from one directory to another directory in HDFS.
   - Move the file from one directory to another directory in HDFS.
   - Copy a file from/To Local file system to HDFS.
   - Display last few lines from the file in HDFS.
   - Display the size of the file in KB and MB in the HDFS.
   - Append a file from Local File system to file in HDFS
   - Merge two file contents (in HDFS) in to one file (in Local file system)
   - Copy one directory structure to another.
   - Set the replication to the file created to 4
   - Remove a file from the directory in HDFS.
   - Remove a directory in HDFS.

Course Outcome 5 (CO5):

1. Using visualization with Tableau , analyze a Superstore data to identify prospective regions for its expansion.
2. There Are Three Customer Segments in the Superstore Dataset. What Percent of the Total Profits Are Associated with the Corporate Segment? Visualize using Tableau
3. Calculate the “average delay to ship using Tableau.”The data set considered should have information regarding order date and ship date for four different regions.

LIST OF EXPERIMENTS

Data Analysis using EXCEL

1.Descriptive Statistics*
2.Linear Regression*
3.Histogram*

R programming

1. Basic Concepts of R - Data structures , Control flow , Functions, Packages*
2 Data reshaping and merging using R *
3.Text Data Analysis using appropriate datasets.*
4. Data Visualisation in R (Scatter plot, Histogram, Box and Whisker, Dot plots,ggplot package).*
5.Exercises to implement Time series Analysis using R.
6.Exercises to create Dashboard, analytics report for a dataset.
7.Recommender systems like product recommendation or movie recommendation

Machine Learning algorithms using R

1.Statistics using R – Mean, Mode, median*
2.Linear Regression and logistic regression*
3.Decision Tree based Classification*
4.K-Means Clustering*
5.SVM classification
6.Neural Network based classification
7. Principal Component Analysis

Big Data Tools and Techniques

1. Installation and configuration of Hadoop*
2. Manipulation of HDFS files using commands*
3. Implementation of Map Reduce programs *
4.Interactive Data Visualization with Tableau Public*
5. Installing and configuring Hive and implementing partitioning and bucketing in Hive
6. Exercises to implement map reduce in MongoDB

*(Note: * marked experiments are mandatory.)*

References

1. Joseph Schmuller. Statistical Analysis with Excel For Dummies (4th. edn.)2016.

2. https://cran.r-project.org/manuals.html

3. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. Wiley Publishing.(1st. ed.). 2015.

4. https://bradleyboehmke.github.io/HOML/index.html

5.https://hadoop.apache.org/

6.https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

7. https://hive.apache.org/

8. https://docs.mongodb.com/manual/installation/

9. https://www.tableau.com/community/academic

ITQ413	SEMINAR	CATEGORY	L	T	P	CREDIT
		PWS	0	0	3	2

Preamble: The course 'Seminar' is intended to enable a B.Tech graduate to read, understand, present and prepare report about an academic document. The learner shall search in the literature including peer reviewed journals, conference, books, project reports etc., and identify an appropriate paper/thesis/report in her/his area of interest, in consultation with her/his seminar guide. This course can help the learner to experience how a presentation can be made about a selected academic document and also empower her/him to prepare a technical report.

Course Objectives:

- To do literature survey in a selected area ofstudy.
- To understand an academic document from the literate and to give a presentation about it.
- To prepare a technical report.

Course Outcomes [COs] : After successful completion of the course, the students will be able to:

CO1	Identify academic documents from the literature which are related to her/his areas of interest (Cognitive knowledge level: **Apply**).
CO2	Read and apprehend an academic document from the literature which is related to her/ his areas of interest (Cognitive knowledge level: **Analyze**).
CO3	Prepare a presentation about an academic document (Cognitive knowledge level: **Create**).
CO4	Give a presentation about an academic document (Cognitive knowledge level: **Apply**).
CO5	Prepare a technical report (Cognitive knowledge level: **Create**).

Mapping of course outcomes with program outcomes:

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	2	2	1	1		2	1					3
CO2	3	3	2	3		2	1					3
CO3	3	2			3			1		2		3
CO4	3				2			1		3		3
CO5	3	3	3	3	2	2		2		3		3