Subject Code	Subject Name (Lab oriented Theory Courses)	Category	L	T	P	C
AI19542	DATA SCIENCE USING R	PC	3	0	2	4

Objectives:	
●	To analyze data by applying basic data science techniques.
●	To understand basic constructs of R.
●	To learn and applying basic classification techniques.
●	To learn various black box techniques of classification, market basket analysis and clustering.
●	To evaluate performance of the models.

UNIT-I	**R DATA STRUCTURES**	**9**
Introduction – Managing and understanding data – Console input and output – Data Types – operators – Functions - R Data Structures – Vectors – Factors –Lists – Data Frames – Matrices and arrays – import and export files – Exploring and understanding data – Visualization – Categorical variables exploration – Relations between variables. (T1: Chapter – 1 & 2)		
UNIT-II	**CLASSIFICATION METHODS**	**9**
Classification – Lazy Learner **-** K-Nearest Neighbor – diagnosing breast cancer with kNN algorithm – Probabilistic Learner – Naïve Bayes – filtering mobile phone spam with naïve bayes algorithm – Divide and Conquer **-** Decision Trees and Rules – Understanding decision trees – identifying risky bank loan using C5.0 – Understanding classification rules –.identifying poisonous mushrooms with rule learners. (T1: Chapter – 3, 4 & 5)		
UNIT-III	**REGRESSION AND BLACK BOX METHODS**	**9**
Forecasting numerical data – Understanding regression – predicting medical expenses using linear regression **-** Understanding regression trees and model trees – estimating the quality of wines with regression trees and model trees – Neural Networks and SVM – Understanding neural networks – modeling the strength of concrete with ANNs – Understanding Support Vector Machines – performance OCR with SVMs. (T1: Chapter – 6 & 7)		
UNIT-IV	**PATTERNS AND CLUSTERING**	**9**
Finding Patterns – Market Basket Analysis using Association Rules – Understanding association rules – identifying frequently purchased groceries with association rules – Finding groups of data – Clustering with K-Means – Understanding clustering – Finding teen market segment using k-means clustering. (T1: Chapter – 8 & 9)		
UNIT-V	**EVALUATING MODEL PERFORMANCE**	**9**
Measuring performance for classifier – Beyond Accuracy – Kappa – Sensitivity and Specificity – Precision and recall – F-Measure – Visualization with ROC Curve – Estimate future performance – Improving Model Performance – Improving model performance with meta learners. (T1: Chapter – 10 & 11)		
	Contact Hours	**: 45**

List of Experiments	
1.	Basics of R – data types, vectors, factors, list and data frames.
2.	Program to implement Breast Cancer with kNN.
3.	Program to implement Filtering Mobile phone spam using Naïve Bayes
4.	Program to implement Risky Bank Loans using Decision Trees
5.	Program to implement Predict medical Expense with Linear Regression.
6.	Program to implement Modeling strength of concrete.
7.	Program to implement Identification of frequently Purchased groceries with Apriori algorithm.
8.	Program to implement Finding Teen Segments of Market.
9.	Program to implement Tuning stock models for better performance.
	Contact Hours : 30
	Total Contact Hours : 75

Course Outcomes: On completion of the course, the students will be able to	
●	Understand the application and uses of data science techniques.
●	Apply basic constructs of R.
●	Apply data science by various classification techniques.
●	Apply market basket analysis and clustering techniques.
●	Evaluate the performance of the models built and fine tune the models to improve them.

Text Books:	
1	Brett Lantz , “Machine Learning with R”, ISBN 978-1-78216-214-8, 2019, Packt Publishing.
2	Beginning R: The Statistical Programming Language‖ , Mark Gardener, Wrox Wiley Publication, First Edition, 2012

Reference Books:	
1	Nina Zumel, John Mount, ―Practical Data Science with R‖, Manning Publications, 2014
2	W. N. Venables, D. M. Smith and the R Core Team, ―An Introduction to R‖, 2013
3	Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, ―Practical Data Science Cookbook‖, Packt Publishing Ltd., 2014

Web link:

1. http://www.johndcook.com/R_language_for_programmers.html

CO - PO – PSO matrices of course

PO/PSO CO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO 3
AI19542.1	2	2	2	1	1	-	-	-	-	-	1	2	2	3	2
AI19542.2	2	2	2	1	1	-	-	-	-	-	2	2	2	3	3
AI19542.3	2	2	2	2	2	-	-	-	-	-	2	3	3	3	3
AI19542.4	2	2	2	2	2	-	-	-	-	-	2	3	3	3	3
AI19542.5	2	2	2	2	2	-	-	-	-	-	2	3	3	3	3
Average	2	2	2	1.6	1.6	-	-	-	-	-	1.8	2.6	2.6	3	2.8

Correlation levels 1, 2 or 3 are as defined below:
1: Slight (Low) 2: Moderate (Medium) 3: Substantial (High)
No correlation: “-”