Code: CDS3005	Title: Foundations of Data Science	Course Type	LP
		Credits	3

Course Objectives:

- To understand the basic concepts of probability, statistics, matrics, linear algebra, statistical models, feature extractions.
- To understand various data preprocessing techniques.
- To understand various supervised, unsupervised machine learning classification algorithms.
- To understand various supervised, unsupervised machine learning clustering algorithms.
- To understand various data visualization techniques using PowerBI or Tableau.

Course Outcomes:

At the end of the course, students should able to.

CO1: describe concepts of probability, statistics, matrics, linear algebra, statistical models and feature extractions.

CO2: carryout various data preprocessing tasks.

CO3: apply classification machine learning algorithms.

CO4: apply clustering machine learning algorithms.

CO5: apply tools for data visualization.

Student Outcomes (SO): a, b, c, l

a. An ability to apply the knowledge of mathematics, science and computing appropriate to the discipline.

b. An ability to analyze a problem, identify and define the computing requirements appropriate to its solution.

c. An ability to design, implement and evaluate a system / computer- based system, process, componentor program to meet desired needs.

l. An ability to apply mathematical foundations, algorithmic principles and computer science theory in the modelling and design of computer-based systems (CS).

Module No.	Module Description	No.of Hours	SO
1	**Topic:** Introduction: What is Data Science? Big Data and Data Science – Datafication – Current landscape of perspectives – Skill sets needed; Matrices – Matrices to represent relations between data, and necessary linear algebraic operations on matrices -Approximately representing matrices by decompositions (SVD and PCA); Statistics: Descriptive Statistics: distributions and probability – Statistical Inference: Populations and samples – Statistical modeling – probability distributions – fitting a model – Hypothesis Testing – Intro to R/ Python.	9	a, b, c
2	**Topic:** Data preprocessing: Data cleaning – data integration – Data Reduction Data Transformation and Data Discretization. Evaluation of classification methods – Confusion matrix, Students T-tests and ROC curves-Exploratory Data Analysis – Basic tools (plots, graphs and summary statistics) of EDA Philosophy of EDA – The Data Science Process.	9	a, b, c
3	**Topic:** Basic Machine Learning Algorithms: Association Rule mining – Linear Regression- Logistic Regression – Classifiers – k-Nearest Neighbors (k-NN) k-means -Decision tree – Naive Bayes- Ensemble Methods – Random Forest Feature Generation and Feature Selection – Feature Selection algorithms – Filters; Wrappers; Decision Trees; Random Forests.	9	a, b, c

4	**Topic:** Clustering: Choosing distance metrics – Different clustering approaches – hierarchical agglomerative clustering, k-means (Lloyd's algorithm), – DBSCAN – Relative merits of each method – clustering tendency and quality.	9	a, b, c
5	**Topic:** Data Visualization: Basic principles, ideas and tools for data visualization.	9	a, b, c

	Guest Lecture on Contemporary Topics	**2**
	Total Hours:	**45**

Mode of Teaching and Learning: *Flipped Class Room, Activity Based Teaching/Learning, Digital/Computer based models, wherever possible to augment lecture for practice/tutorial and minimum 2 hours lectures by industry experts on contemporary topics*

Mode of Evaluation and assessment:
The assessment and evaluation components may consist of unannounced open book examinations, quizzes, student's portfolio generation and assessment, and any other innovative assessment practices followed by faculty, in addition to the Continuous Assessment Tests and Term End Examinations.

Text Book(s):

1.	Cathy O'Neil and Rachel Schutt, "Doing Data Science, Straight Talk From The Frontline", O'Reilly, 2014.
2	Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques", Third Edition. ISBN 0123814790, 2011.
3	Mohammed J. Zaki and Wagner Miera Jr, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.

Reference Book(s):

1.	Matt Harrison, "Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visualization, O'Reilly, 2016.
2.	Joel Grus, "Data Science from Scratch: First Principles with Python", O'Reilly Media, 2015.
3.	Wes McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", O'Reilly Media, 2012.

Indicative List of Experiments not less than 10

No.	Description of Experiment	SO
1	Programs for represent relations between data, and necessary linear algebraic operations on matrices.	a.b.c
2	Programs to representing matrices by decompositions (SVD and PCA).	a.b.c
3	Programs for Data cleaning, transformation.	a.b.c
4	Program for Linear Regression- Logistic Regression – Classifiers – k-Nearest Neighbors (k-NN), k-means -Decision tree – Naive Bayes- Ensemble Methods – Random Forest.	a.b.c
5	Program to Feature Generation and Feature Selection.	a.b.c
6	Program for hierarchical agglomerative clustering, k-means.	a.b.c
7	Program to demonstrate data visualization tools like PowerBI or Tableau.	a.b.c

Recommendation by the Board of Studies on	
Approval by Academic council on	
Compiled by	**Dr. Pon Harshavardhanan**