SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

B.Tech., Computer Science and Engineering

<table>
<tr><td>Course Title</td><td colspan="3">Data Science</td><td colspan="3">Course Type</td><td colspan="3">INTEGRATED</td></tr>
<tr><td>Course Code</td><td colspan="3"></td><td colspan="3">Class</td><td colspan="3">IV SEMESTER</td></tr>
<tr><td rowspan="6">Instruction delivery</td><td>Activity</td><td>Credits</td><td>Credit Hours</td><td colspan="4" rowspan="2">Total Number of Classes per Semester</td><td colspan="2" rowspan="2">Assessment in Weightage</td></tr>
<tr><td>Lecture</td><td>3</td><td>45</td></tr>
<tr><td>Tutorial</td><td></td><td></td><td rowspan="3">Theory</td><td rowspan="3">Tutorial</td><td rowspan="3">Practical</td><td rowspan="3">Self-study</td><td rowspan="3">CIE</td><td rowspan="3">SEE</td></tr>
<tr><td>Practical</td><td>1</td><td>15</td></tr>
<tr><td>Self-study</td><td></td><td></td></tr>
<tr><td>Total</td><td>4</td><td>60</td><td>45</td><td>0</td><td>15</td><td>0</td><td>50%</td><td>50%</td></tr>
<tr><td rowspan="9">Names Course Instructors</td><td>Course Lead</td><td colspan="8">Mr. Soumalya Ghosh</td></tr>
<tr><td>Course Coordinator</td><td colspan="8">Mr. Hariprasath K</td></tr>
<tr><td colspan="4">Theory</td><td></td><td colspan="4">Practical</td></tr>
<tr><td colspan="4">Mr. S.Prakash</td><td></td><td colspan="4">Mr. S.Prakash</td></tr>
<tr><td colspan="4">Mr. A.Booblan</td><td></td><td colspan="4">Mr. A.Booblan</td></tr>
<tr><td colspan="4">Mr. Soumalya Ghosh</td><td></td><td colspan="4">Mr. Soumalya Ghosh</td></tr>
<tr><td colspan="4">Mr. Vikash Kumar Mishra</td><td></td><td colspan="4">Mr. Vikash Kumar Mishra</td></tr>
<tr><td colspan="4">Mr. Hariprasath k</td><td></td><td colspan="4">Mr. Hariprasath k</td></tr>
<tr><td colspan="4">Mr. Soumalya Ghosh</td><td></td><td colspan="4">Mr. Soumalya Ghosh</td></tr>
<tr><td colspan="4">Mr. Vishwa Pratap Singh</td><td></td><td colspan="4">Mr. Vishwa Pratap Singh</td></tr>
</table>

COURSE OVERVIEW

Statistical experiment design and analytics are at the heart of data science. Data Science is a branch that deals with the extraction of information from data. It is a relatively new field that combines aspects of statistics, machine learning, and computer science. In this course you will design statistical experiments and analyze the results using modern methods. This course is intended to provide a thorough and sound understanding of the essential theoretical base, an introduction into the important and useful techniques of modelling and also an understanding of the broad applications of Data Science, and apply them to solve some real-world problems.

COURSE OBJECTIVE

- To understand concepts of data science.
- To understand popular tools of data analysis.
- To learn data science in python.
- To apply different visualization Techniques on datasets.
- To implement learning techniques in data science.

COURSE OUTCOME

CO1	Understand introductory knowledge in the essentials of Data Analysis and Data science
CO2	Apply and Implement concepts of data analysis in google sheets and excel; algorithmic principles and Programming knowledge using Python language
CO3	Demonstrate different data virtualization and optimization techniques.
CO4	Apply and Implement ML processing principles using Probability and Statistics**.**

CO No.	**Bloom's Taxonomy Level (BTL)**					
	Remember (L1)	**Understand (L2)**	**Apply (L3)**	**Analyze (L4)**	**Evaluate (L5)**	**Create (L6)**
CO1	×					
CO2	×	×	×	×		×
CO3			×	×		
CO4			×	×	×	×

PROGRAM OUTCOMES

PO1	**Engineering Knowledge:** Apply the knowledge of mathematics, science, Engineering fundamentals, and an engineering specialization to the solution of complex engineering problems
PO2	**Problem Analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO3	**Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate

	consideration for the public health and safety, and the cultural, societal, and environmental considerations.
PO4	**Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions
PO5	**Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
PO6	**The Engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
PO7	**Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development
PO8	**Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO9	**Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
PO11	**Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO12	**Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1:	Able to analyze, design and implement sustainable and ethical solutions in the field of computer science.
PSO2:	Able to use problem solving skills to develop efficient algorithmic solutions.

COURSE ARTICULATION MATRIX

CO/PO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PS O 1	PS O 2
CO1	3	3	1											
CO2	2	3	3	1	2									
CO3				2	3				2				2	3
CO4				3	3				3				3	3

COURSE ASSESSMENT

SN o	Assessment Tools	CIE							Total CIE marks	SEE
		QUIZ1 /AAT	CAT 1	QUIZ2 /AAT	CAT 2	LAB	LAB EXAM	Course -based Project		
1	Integrated		A1		A2	A3	A4			
		0	30	0	30	20	20	0	100	100

COURSE CONTENT

THEORY

What is Data science? Data analysis. Statistics fundamentals for data science. Different popular software tools for data science and data analysis: Installation and usage - Use of formulae to calculate the values in excel, statistical operations in excel sheets (Descriptive Statistic, histogram, correlation, moving average, exponential smoothing), Filter and VLOOKUP in excel. Other operations to manipulate and summarize information present in the data - Introduction about NumPy, Different NumPy Operations, Broadcasting with NumPy, Introduction about Pandas, Reading or Loading data into Data frame, Pandas Data Frame Manipulations, Data Loading /Reading in different formats (CSV, Excel, Json, HTML) - Introduction to data Visualizations, Principles Behind Data Visualizations, Histograms-Visualize, Box Plots-Visualize, the Distribution of Continuous Numerical Variables (Bar Plots, Pie Chart, Line Chart). Data Visualization using R- Line Plots and Regression - Supervised machine learning Fundamentals- Regression and Classification. Unsupervised Learning in Python: K- Means Theory/ Implementation, Quantifying K-Means Clustering Performance, Hierarchical Clustering Theory, Principal Component Analysis (PCA) theory / Implementation. Selection criteria for number of clusters choosing - Recent articles on Research journals of Data science and AI. Scopus and SCI indexed resources can be considered.

PRACTICAL
Working with excel, Statistical analysis with excel, Python revision and introduction to NumPy - Implementing Array operations using NumPy -Implementing NumPy Broadcasting - Implementing NumPy string functions - Implementing Data Frames using PANDAS Library - Implement Pandas Library for working with missing values - Implement Matplotlib Library for data visualization - Implement K-means Clustering algorithm - Implement Principal Component Analysis for Dimensionality Reduction - Implementing Regression models.

LESSON PLAN FOR THEORY

Session	Topics	Skills to be Learned
1.	What is Data science? Preparing and gathering data and knowledge, Philosophies of data science	Students gets the knowledge on Overview of the Data Science Process
2.	Data Science Process and Data analysis -	
3.	Statistics fundamentals for data science	
4.	Data all around us: the virtual wilderness, Data wrangling: from capture to domestication, Data science in a big data world	
5.	Different popular software tools for data science	
6.	Different popular software tools for data science	
7.	data analysis: Installation and usage	
8.	data analysis: Installation and usage	
9.	Use of formulae to calculate the values	Students gets the knowledge on Data Analysis with Excel and Google sheets
10.	Use of formulae to calculate the values	
11.	statistical operations in sheets (Descriptive Statistic, histogram, correlation, moving average, exponential smoothing)	
12.	statistical operations in sheets (Descriptive Statistic, histogram, correlation, moving average, exponential smoothing)	
13.	Filter and VLOOKUP	
14.	Other operations to manipulate and summarize information present in the data	
15.	Other operations to manipulate and summarize information present in the data	
16.	Other operations to manipulate and summarize information present in the data	
17.	Introduction about NumPy	

18.	Different NumPy Operations	Students gets the knowledge on Data science in python with NumPy and PANDAS
19.	Broadcasting with NumPy	
20.	Introduction about Pandas	
21.	Different Built-in Pandas Function Reading or Loading data into Data frame	
22.	Pandas Data Frame Manipulations Data Loading /Reading in different formats (CSV,Excel,Json,HTML)	
23.	Data Loading /Reading in different formats (CSV,Excel,Json,HTML) Cont..	
24.	Introduction to data Visualizations	Students aware on various data visualization tools
25.	Principles Behind Data Visualizations	
26.	Histograms-Visualize	
27.	Box Plots-Visualize	
28.	the Distribution of Continuous Numerical Variables (Bar Plots,	
29.	Pie Chart, Line Chart).	
30.	Data Visualization using R- Line Plots and Regression	
31.	Data Visualization using R- Line Plots and Regression	
32.	Supervised machine learning	Student gets introduced to Machine Learning with Data models
33.	Fundamentals- Regression and	
34.	Classification.	
35.	Supervised/Unsupervised Learning in Python:	
36.	K- Means Theory	
37.	Implementation, Quantifying K-Means.	
38.	Clustering Performance Hierarchical Clustering Theory	
39.	Principal Component Analysis (PCA) theory	
40.	PCA - Implementation	
41.	Selection criteria for number of clusters choosing	
42.	Latest papers published	Students gets introduced to recent trends in Data science and tools
43.	Patents filed	
44.	Latest products	
45.	SCI / Scopus Paper discussion	

LESSON PLAN FOR PRACTICAL

S.No	Topics	Skills
1	Working with Formulas and Functions in Excel	Students gets the knowledge on Data Analysis with Excel and Google sheets
2	Working with Filtering Operations, Pivot tables	
3	Charts using Excel – bar chart, pie chart, scatter plot	
4	Demonstrate the Histogram and Descriptive Statistics using Excel	
5	Implement Moving Average and Exponential Smoothing using Excel	
6	Python revision and introduction to NumPy	Student get introduced to NumPy and able to build models using NumPy
7	Implementing Array operations using NumPy	
8	Implementing NumPy Broadcasting	
9	Implementing NumPy string functions -add, multiply, center, split, join	
10	Implementing NumPy string functions - capitalize, lower,	
11	Implementing Data Frames using PANDAS Library	Students gets the knowledge on Data science in python with PANDAS
12	Implement Pandas Library for working with missing values	
13	Implement bar chart using Matplotlib Library for data visualization	Student gets introduced to Machine Learning with Data
14	Implement scatter plot using Matplotlib Library for data visualization	
15	Implement sub plot using Matplotlib Library for data visualization	
16	Implement K-means Clustering algorithm	
17	Implement Principal Component Analysis for Dimensionality Reduction	
18	Implementing Regression models.	

Text Books:

1. Data Science from Scratch: First Principles with Python 1st Edition, by Joel Grus, O'Reilly Publication,2020.
2. James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning with applications in R. Springer, 2013.
3. Han, J., Kamber, M., Pei, J. Data mining concepts and techniques. Morgan Kaufmann, 2011.

Reference Books:

1. T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2e, 2008.
2. Christopher Bishop. Pattern Recognition and Machine Learning. 2e.

List of NPTEL/MOOCS/SWAYAM/Courses/Video:

1. https://nptel.ac.in/courses/110106072/
2. https://nptel.ac.in/courses/110106073
3. https://www.youtube.com/watch?v=j3lgxdylktxv

Webliography:

1. https://www.w3schools.com/datascience/
2. https://www.w3schools.com/python/pandas/default.asp
3. https://www.w3schools.com/excel/index.php
4. https://www.w3schools.com/googlesheets/index.php
5. https://www.geeksforgeeks.org/data-science-tutorial