A. Course Handout (For Students & Faculty)

Institute/School/College Name	Chitkara University Institute of Engineering & Technology		
Department/Centre Name	Department of Computer Science & Engineering		
Programme Name	Bachelor of Engineering- Computer Science & Engineering **(Artificial Intelligence)**		
Course Name	**Data Analytics**	Session	2022-23
Course Code	**AI105**	Semester/Batch	3rd /2022
Lecture/Tutorial (Per Week)	**2-0-4**	Course Credits	**4**
Course Coordinator Name	Dr. Sushil Kumar Narang		

1. Scope & Objective of the Course:

With great amounts of data comes a great need for data analysts. Organizations generate and collect an exponentially growing amount of data: wringing actionable answers and insights out of the chaos is a valuable and in-demand skill set to have. Organizations across industries need these answers and insights to improve the decisions they make. B2B and B2C commerce, health care, manufacturing, and marketing all use data analytics to improve processes and enhance profits.

This course prepares the students for a new career in the high-growth field of data analytics and about the process for planning data analysis solutions and the various data analytic processes that are involved. This course takes you through five key factors that indicate the need for specific AWS services in collecting, processing, analyzing, and presenting your data. This includes learning basic architectures, value propositions, and potential use cases

The core objectives of this course are :

- To gain an immersive understanding of the practices and processes used by a data analyst in their day-to-day job operations
- To teach how to clean and organize data for analysis, and complete analysis and calculations using spreadsheets, SQL and Python programming
- To inculcate key analytical skills (data cleaning, analysis, & visualization) and tools (Python programming, Tableau, Power BI)
- To know how to visualize and present data findings in dashboards, presentations and commonly used visualization platforms.

2. Course Learning Outcomes:

At the end of the course, students will be able to:

CLO01: Develop the ability to build proficiency with statistical analysis of data

CLO02: Apply data science concepts and methods to solve problems in real-world contexts and will communicate these solutions effectively

CLO03: Carry out standard data visualization and formal inference procedures

CLO04: Perform data cleaning, and transform variables to facilitate analysis by integrating data from disparate sources

CLO05: Build and enhance business intelligence capabilities by adapting the appropriate technology and software solutions

CLO-PO Mapping grid |Program outcomes (POs) are available as a part of Academic Program Guide

Course Learning Outcomes	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CLO1	L	L	L	L	L					L	L	M
CLO2	H	M	H	H	M					L	L	M

CLO3	H	M	H	H	M					L	L	M
CLO4	H	M	H	H	M					L	L	M
CLO5	M	M	M	H	M					L	L	M

3. **Recommended Books (Reference Books/Textbooks):**
B01: Jojo Moolayil, "Smarter Decisions: The Intersection of IoT and Data Science", PACKT, 2016.
B02: Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
B03: David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics", EMC 2013
B04: Raj, Pethuru, "Handbook of Research on Cloud Infrastructures for Big Data Analytics", IGI Global

4. **Other readings & relevant websites:**

S.N.	Link of Journals, Magazines, Websites, and Research Papers
1.	https://www.tutorialspoint.com/statistics/index.html
2.	https://iridl.ldeo.columbia.edu/dochelp/StatTutorial/index.html
3.	https://www.khanacademy.org/math/statistics-probability
4.	https://www.analyticsvidhya.com/blog/2021/02/an-intuitive-guide-to-visualization-in-python/
5.	https://www.educba.com/data-science/data-science-tutorials/tableau-tutorial/
6.	https://matplotlib.org/stable/tutorials/index.html
7.	https://seaborn.pydata.org/tutorial.html

5. **Recommended Tools and Platforms**

Python, Jupyter Notebook, Visual Studio Code, Anaconda, Tableau, Microsoft Power BI

6. **Course Plan:**

Lecture Number	Topics	Details
1	Understanding data	Introduction – Types of Data: Numeric – Categorical – Graphical – High Dimensional Data
2	Classification of digital Data	• Structured, Semi-Structured and Un-Structured • Sources of Data
3	Case Studies on Different types of Data	Time Series – Transactional Data – Biological Data – Spatial Data – Social Network Data
4	Data Evolution	Understand issues relating to acquisition, cleaning and loading of data, Data Deluge, Data lake
5-7	Python Numpy	• Arrays, Indexing, Slicing • Different Array Operations • Linear Algebra Operations
8-12	Python Pandas	• Dataframes, Handling of data • Series • Data wrangling, Alignment and Indexing, Handling Missing Data, Data Cleaning • Merging and Joining Dataframes, Grouping • Concayenation and Aggregation • Masking, Performing Mathematical Operations on Data

13-16	Access and combine data from CSV, JSON, logs, APIs, and databases	• Using Pandas to access different sources of Data
17-19	Using SQL with Databases	• DDL, DML, Select and Joins
20-23	Advanced Operations using Pandas	• Statistical Functions • Descriptive Statistics • Working with Text Data • Time Delta • Basic Data Visualization using pandas
24-28	Matplotlib, Seaborn, Cufflinks	• Different Plot types • Scatter, bar, histogram, box, pie, violin • Subplots, axis and figures • Text, labels and annotations • Colormaps • Plotting with Seaborn • Plotting Categorical and Continuous Data • Visualizing Regression models • Plotting interactive plots using Cufflinks
29-33	Using Power BI for Visualization	• Connecting to different data sources • Data types • Working with Meta Data • Calculations • Purpose of Data Analysis Expresssions (DAX) • DAX operators & Functions • Power Query • Different Charts & Reports • Exploring Data geographically • Building dashboard to see insights
34-35	Time-Series Analysis	• Understanding, Trend, Seasonality and residuals, Moving Averages, Expnential Weighted Moving Averages, Autocorrelation, Autoregression models, ARIMA, SARIMAX
ST-1 (Syllabus covered from Lecture 1 to 30)		
36-38	Statistical Learning	• Important statistical concepts used in data science • Difference between population and sample, Types of variables, Measures of central tendency • Measures of variability, Coefficient of variance, Skewness and Kurtosis • Exploratory data analysis: Missing value analysis • The correction matrix, Outlier detection analysis • Inferential Statistics: Normal distribution
39-44	Test hypotheses, Parametric and Non-parametric tests	• Central limit theorem, Confidence interval • T-test, Type I and II errors, Student's T distribution • Non-Parametric Tests: Sign Test • Wilcoxon's Signed Rank Test • Mann-Whitney test

		• Kolmorogov-Smirnov test
45-47	Understanding Regression	Linear and Non-linear Regression
48-50	ANOVA	• One-way and Two-way Analysis of Variance • R square, Correlation and causation • dependent and independent variables- Case Studies
51-53	Identification of regression problems	Case Studies
54-57	Identifications of Classification problems	Case Studies
58-60	Identification of Clustering Problems	Case Studies
ST-2 (Syllabus covered from Lecture)		
END TERM – FULL SYLLABUS		

7. Delivery/Instructional Resources

Lecture No.	Topics	PPT *(Link of ppts on the central server)*	Industry Expert Session *(If **yes**: link of ppts on the central server)*	Web References	Audio-Video
1	Understanding data			https://www.sqlshack.com/introduction-to-data-science-data-understanding-and-preparation/ https://ibm-cloud-architecture.github.io/refarch-data-ai-analytics/preparation/data-understanding/	
2	Classification of digital Data			https://ibm-cloud-architecture.github.io/refarch-data-ai-analytics/preparation/data-understanding/	https://www.youtube.com/watch?v=mm2A5tKVIpg
3	Case Studies on Different types of Data			https://data-flair.training/blogs/big-data-case-studies/	

4	Data Evolution			https://www.kdnuggets.com/2014/06/data-lakes-vs-data-warehouses.html	https://www.youtube.com/watch?v=E49BFhThC3U
5-7	Python Numpy			https://numpy.org/doc/stable/user/index.html#user	https://www.youtube.com/watch?v=j31ah5Qa4QI
8-12	Python Pandas			https://pandas.pydata.org/docs/user_guide/index.html	https://www.youtube.com/watch?v=UB3DE5Bgfx4
13-16	Access and combine data from CSV, JSON, logs, APIs, and databases			https://pandas.pydata.org/docs/user_guide/io.html	https://www.youtube.com/watch?v=GFBxxxjAzaU
17-19	Using SQL with Databases			https://www.sqltutorial.org/	https://www.youtube.com/watch?v=zbMHLJ0dY4w
20-23	Advanced Operations using Pandas			https://www.kdnuggets.com/2019/10/5-advanced-features-pandas.html	https://www.youtube.com/watch?v=DUgd48QYmfI https://www.youtube.com/watch?v=RlliVeig3hc
24-28	Matplotlib, Seaborn, Cufflinks			https://matplotlib.org/stable/plot_types/index.html https://seaborn.pydata.org/tutorial.html https://www.analyticsvidhya.com/blog/2021/06/advanced-python-data-visualization-libraries-plotly/	https://www.youtube.com/watch?v=3Xc3CA655Y4 https://www.youtube.com/watch?v=ooqXQ37XHMM https://www.youtube.com/watch?v=7n5GzKuvPsw
29-33	Using Power BI for Visualization			https://www.tutorialspoint.com/power_bi/index.htm	https://www.youtube.com/watch?v=NalazxBo-90 https://www.youtube.com/watch?v=3u7MQz1EyPY

34-38	Statistical Learning			https://realpython.com/python-statistics/	https://www.youtube.com/watch?v=mQ-3KwrBIN0
39-44	Test hypotheses, Parametric and Non-parametric tests			https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/	https://www.youtube.com/watch?v=IcLSKko2tsg
45-47	Understanding Regression			https://statisticsbyjim.com/regression/regression-tutorial-analysis-examples/	https://www.youtube.com/watch?v=nk2CQITm_eo
48-50	ANOVA			https://www.reneshbedre.com/blog/anova.html	https://www.youtube.com/watch?v=ITf4vHhyGpc https://www.youtube.com/watch?v=QOl0_OdvbdE
51-53	Identification of regression problems			https://users.stat.ufl.edu/~winner/cases.html	https://www.youtube.com/watch?v=HgfHefwK7VQ&t=268s
54-57	Identifications of Classification problems			https://www.techguruspeaks.com/case-study-classification/	https://www.youtube.com/watch?v=T5zJHhTO1FA
58-60	Identification of Clustering Problems			http://ucanalytics.com/blogs/customer-segmentation-cluster-analysis-telecom-case-study-example/	https://www.youtube.com/watch?v=lc7MLQpjqZ8

8. Action plan for different types of learners

Slow Learners	Average Learners	Fast Learners
• Multiple Remedial Extra Classes • Encouragement for improvement using Peer Tutoring	• Doubt-sessions • Pre-coded algorithms to illustrate concepts and notions • E-notes and E-exercises to read in addition to pedagogic material	• More Practice assignments on real life problems • Engaging students to hold hands of slow learners by creating a Peer Tutoring Group • Participation in Hackathons, competitions.

9. Evaluation Scheme & Components:

Evaluation Component	Type of Component	No. of Assessments	Weightage of Component	Mode of Assessment
Component 1	Subjective Test/Sessional Tests (STs)	02*	40%	Offline/Online
Component 2	End Term Examinations	01	60%	Offline/Online
Total		**100%**		

*Out of 02 STs, the ERP system automatically picks the average of 02 STs Marks for evaluation of the STs as final marks.

10. Details of Evaluation Components:

Evaluation Component	Description	Syllabus Covered (%)	Timeline of Examination	Weightage (%)
Component 1	ST 01	Up to 36%	4th April, 2022	40%
	ST 02	37% - 100%	23rd May, 2022	
Component 2	End Term Examination*	100%	At the End of the Semester	60%
Total				100%

*As per Academic Guidelines minimum 75% attendance is required to become eligible for appearing in the End Semester Examination.

Evaluation Components

Type of Assessment	Timeline of Conduct	Total Marks	Question Paper Format			
			1 Mark MCQ	2 Mark MCQ	5 Mark Question	10 Mark Algorithm/Case Study
Sessional Test 1	4th April, 2022	40	5	5	1	2
Sessional Test 2	23rd May, 2022	40	5	5	1	2
End Term Examination		60	10	5	4	2

B. Syllabus of the Course

Subject: Applied Soft Computing	Subject Code: AI102

S.N.	Topic (s)	No. of Lectures	Weightage %
1	Understanding data: Introduction – Types of Data: Numeric – Categorical – Graphical – High Dimensional Data – Classification of digital Data: Structured, Semi-Structured and Un-Structured - Example, Applications, Sources of Data: Time Series –	4	10%

	Transactional Data – Biological Data – Spatial Data – Social Network Data – Data Evolution, Understand issues relating to acquisition, cleaning and loading of data, Data Deluge, Data lake		
2	Python Numpy, Python Pandas: Dataframes, Handling of data, Data wrangling, Alignment and Indexing, Handling Missing Data, Data Cleaning, Merging and Joining Dataframes, Grouping, Masking, Performing Mathematical Operations on Data, Access and combine data from CSV, JSON, logs, APIs, and databases	14	25%
3	Data Visualization using pandas, Matplotlib, Seaborn, Using Tableau for Visualization, Connecting to different data sources, Data types, Working with Meta Data, Calculations, Purpose of Data Analysis Expresssions (DAX), DAX operators & Functions, Power Query, Different Charts & Reports, Exploring Data geographically, Building dashboard to see insights	10	20%
4	Important statistical concepts used in data science, Difference between population and sample, Types of variables, Measures of central tendency, Measures of variability, Coefficient of variance, Skewness and Kurtosis Exploratory data analysis: Missing value analysis, The correction matrix, Outlier detection analysis	5	8%
5	Inferential Statistics: Normal distribution, Test hypotheses, Parametric and Non-parametric tests, Central limit theorem, Confidence interval, T-test, Type I and II errors, Student's T distribution, Non-Parametric Tests: Sign Test, Wilcoxon's Signed Rank Test, Mann-Whitney test, Kolmorogov-Smirnov test	6	10%
6	Regression, ANOVA(One-way and Two-way Analysis of Variance) ,R-square, Correlation and causation, Introduction to classification problems, Identification of a regression problem, dependent and independent variables, Clustering Problems Different Case Studies	16	27%

This document is approved by:

Designation	**Name**	**Signature**
Course Coordinator	Dr. Sushil Kumar Narang	
Program In-charge	Dr. Kamal Deep Garg	
Cluster Dean	Dr. Sushil Kumar Narang	
Date (DD/MM/YYYY)		