Name of Institute: Indus Institute of Technology and Engineering
Name of Faculty: Pruthvi Patel

Course code: CE0519/CS0519/IT0519
Course name: Data Science
Pre-requisites: Basic knowledge of Mathematics, Programming, Database
Credit points: 4
Offered Semester: V

Course Coordinator
Full Name: Pruthvi Patel
Department with sitting location: Computer engineering (4th Floor Staff Room)
Telephone:
Email: pruthvipatel.ce@indusuni.ac.in
Consultation times: Everyday 4:00 to 5:00 PM

Course Lecturer
Full Name: Pruthvi Patel
Department with sitting location: Computer engineering (4th Floor Staff Room)
Telephone:
Email: pruthvipatel.ce@indusuni.ac.in
Consultation times: Everyday 4:00 to 5:00 PM

Students will be contacted throughout the Session via Mail with important information relating to this Course.

Course Objectives

By participating in and understanding all facets of this Course a student will:

1) Learn the fundamentals of data analytics and the data science pipeline
2) Learn how to scope the resources required for a data science project
3) Apply principles of Data Science to the analysis of business problems.
4) Apply data mining software to solve real-world problems.
5) Employ cutting edge tools and technologies to analyse Big Data

Course Outcomes (CO)

1) Students will demonstrate knowledge of data analytics.
2) Students will demonstrate the ability to think critically in making decisions based on data.
3) Students will able to interpret data, extract meaningful information, and assess findings.
4) Students will identify and analyze social, legal, and ethical issues in data science.
5) Students will be able to choose and apply tools and methodologies to solve data science tasks.
6) Students will be able to implement statistics and mathematical concepts along with machine learning algorithm.

Course Outline

UNIT-I [12 hours]
Introduction
Defining Data Science , What do data science people do?, Current landscape of perspectives, Data Science in Business, Use Cases for Data Science
UNIT-II [12 hours]
Statistical Inference: Statistical modeling, probability distributions, fitting a model, Intro to R
Descriptive Statistics: Introduction to the course, Descriptive Statistics, Probability Distributions
Inferential Statistics: Inferential Statistics through hypothesis tests, Permutation & Randomization Test
UNIT-III [12 hours]
Machine Learning Introduction and Concepts: Differentiating algorithmic and model based frameworks, Regression: Ordinary Least Squares, Ridge Regression, Lasso Regression, K Nearest Neighbors, Regression & Classification
Supervised Learning with Regression and Classification techniques: Bias-Variance Dichotomy, Model Validation Approaches, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Regression and Classification Trees, Support Vector Machines, Ensemble Methods: Random Forest, Neural Networks, Deep learning
Unsupervised Learning and data modeling: Clustering, Associative Rule Mining, Logical Modeling: Converting a conceptual model to logical model, Integrity constraints, Normalization
UNIT-IV [12 hours]
Data Visualization: Basic principles, ideas and tools for data visualization
Data Science and Ethical Issues: Discussions on privacy, security, ethics

Method of delivery

Chalk and Board, Presentations, Inverse Classroom, AV(Reference Videos), self study material, Practical Demo

Study time

Theory- 3 Hours, Practical- 2 Hours

CO-PO Mapping (PO: Program Outcomes)

CO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO 1	3	1	2	-	-	-	-	-	-	-	-	-
CO 2	1	2	3	-	2	-	-	-	-	-	-	-
CO 3	1	2	3	-	2	-	-	-	-	-	-	-
CO 4	-	3	2	-	2	-	-	-	-	-	-	-
CO 5	-	2	2	-	3	-	-	-	-	-	-	-
CO 6	-	2	2	1	3	-	-	-	2	2	1	-

Blooms Taxonomy and Knowledge retention (For reference)

(Blooms taxonomy has been given for reference)

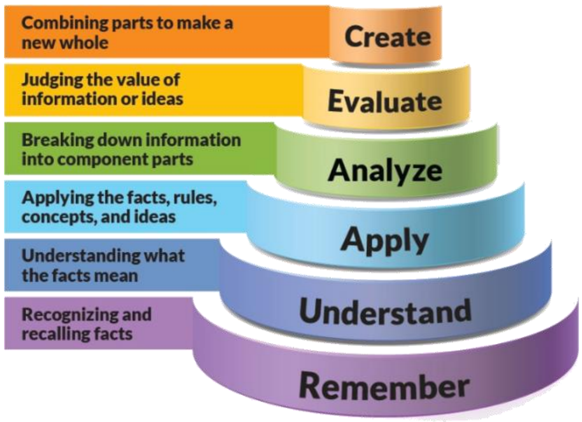


Figure 1: Blooms Taxonomy

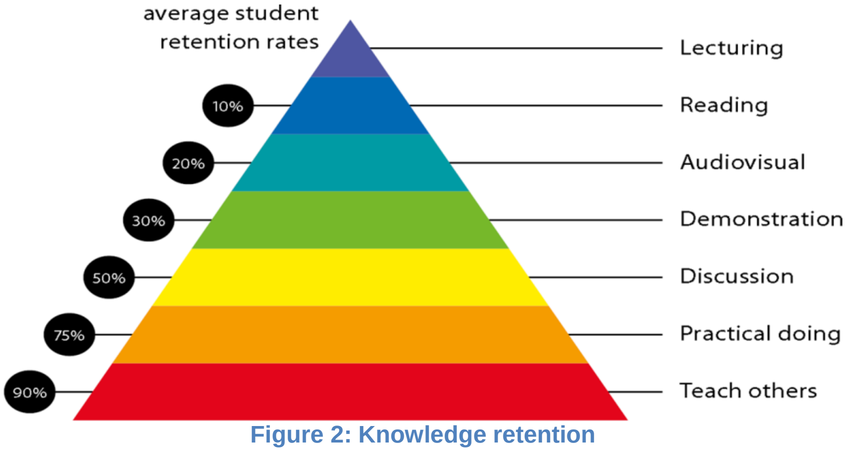


Figure 2: Knowledge retention

Graduate Qualities and Capabilities covered

(Qualities graduates harness crediting this Course)

General Graduate Qualities	Specific Department of _______Graduate Capabilities
Informed Have a sound knowledge of an area of study or profession and understand its current issues, locally and internationally. Know how to apply this knowledge. Understand how an area of study has developed and how it relates to other areas.	**1 Professional knowledge, grounding & awareness**
Independent learners Engage with new ideas and ways of thinking and critically analyze issues. Seek to extend knowledge through ongoing research, enquiry and reflection. Find and evaluate information, using a	**2 Information literacy, gathering & processing**

variety of sources and technologies. Acknowledge the work and ideas of others.	
Problem solvers Take on challenges and opportunities. Apply creative, logical and critical thinking skills to respond effectively. Make and implement decisions. Be flexible, thorough, innovative and aim for high standards.	**4 Problem solving skills**
Effective communicators Articulate ideas and convey them effectively using a range of media. Work collaboratively and engage with people in different settings. Recognize how culture can shape communication.	**5 Written communication** **6 Oral communication** **7 Teamwork**
Responsible Understand how decisions can affect others and make ethically informed choices. Appreciate and respect diversity. Act with integrity as part of local, national, global and professional communities.	**10 Sustainability, societal & environmental impact**

Practical work:

No.	Practical
1.	Introduction to R tool for data analytics science:
2.	Descriptive statistics in R
3.	Reading and writing different types of datasets
4.	Visualizations
5.	Correlation and covariance
6.	Regression model
7.	Multiple regression model
8.	Regression model for prediction
9.	Classification model
10.	Clustering model
11.	Getting started with jupyter notebooks/google collaborator
12.	Libraries: NumPy, ScikitLearn, Pandas,Matplotlib
13.	Clustering and classification
14.	Decision Tree and Random Forest
15.	Neural Networks
16.	Data analysis and visualization with python
17.	Advanced Exercises: Digit Classification Regression Models Prediction of Airbnb Renting Price

Lecture/tutorial times

Branch	Day	Time
5CE	Monday, Tuesday, Wednesday	10:00 AM to 11:00 AM
5CS	Monday, Tuesday	12:20 PM to 1:20 PM
5CS	Friday	11:10 PM to 12:10 PM
5 IT	Wednesday, Thursday, Friday	2:00 PM to 3:00 PM

Attendance Requirements

The University norms states that it is the responsibility of students to attend all lectures, tutorials, seminars and practical work as stipulated in the Course outline. Minimum attendance requirement as per university norms is compulsory for being eligible for mid and end semester examinations.

Details of referencing system to be used in written work

Text books

1. Data Science from Scratch, Steven Cooper, 2018
2. Introduction to Data Science, Laura Igual, Santi segui, Springer, 2017.
3. Applied statistics and probability for engineers, Montgomery, Douglas C., George C. Runger, John Wiley & Sons, 2010
4. Doing Data Science, Straight Talk From The Frontline., Cathy O'Neil and Rachel Schutt , O'Reilly. 2014

Additional Materials

1. Machine Learning: A Probabilistic Perspective. Kevin P. Murphy.
2. Mining of Massive Datasets , Anand Rajaraman and Jeffrey David Ullman 2012

ASSESSMENT GUIDELINES

Your final course mark will be calculated from the following:

CIE (60 Marks) Mid Semester Exam- 40 Marks Internal Evaluation- 20 Marks Case study – 10 Marks Presentation – 5 Marks Attendance – 5 Marks ESE (40 Marks)

SUPPLEMENTARY ASSESSMENT

Students who receive an overall mark less than 40% in mid semester or end semester will be considered for supplementary assessment in the respective components (i.e mid semester or end semester) of semester concerned. Students must make themselves available during the supplementary examination period to take up the respective components (mid semester or end semester) and need to obtain the required minimum 40% marks to clear the concerned components.

Practical Work Report/Laboratory Report:

A report on the practical work is due the subsequent week after completion of the class by each group.

Late Work

Late assignments will not be accepted without supporting documentation. Late submission of the reports will result in a deduction of -% of the maximum mark per calendar day

Format

All assignments must be presented in a neat, legible format with all information sources correctly referenced. **Assignment material handed in throughout the session that is not neat and legible will not be marked and will be returned to the student.**

Retention of Written Work

Written assessment work will be retained by the Course coordinator/lecturer for two weeks after marking to be collected by the students.

University and Faculty Policies

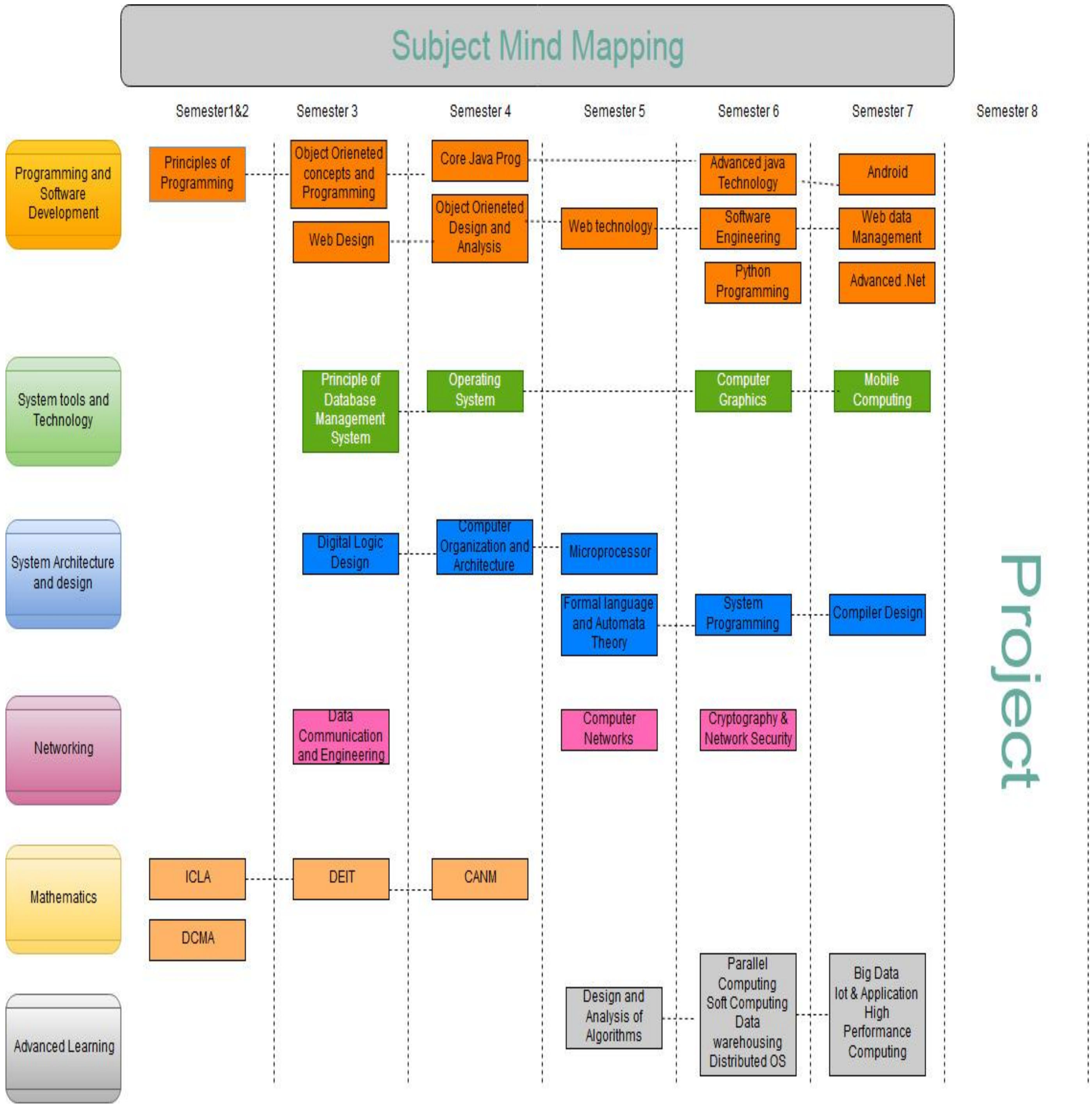
Students should make themselves aware of the University and/or Faculty Policies regarding plagiarism, special consideration, supplementary examinations and other educational issues and student matters.

Plagiarism - Plagiarism is not acceptable and may result in the imposition of severe penalties. Plagiarism is the use of another person's work, or idea, as if it is his or her own - if you have any doubts at all on what constitutes plagiarism, please consult your Course coordinator or lecturer. Plagiarism will be penalized severely.

Do not copy the work of other students.
Do not share your work with other students (except where required for a group activity or assessment

Course schedule (subject to change)

	Week #	Topic & contents	CO Addressed	Teaching Learning Activity (TLA)
	Week 1	Defining Data Science , What do data science people do?, Current landscape of perspectives,	CO1, CO2	Presentations
	Week 2	Data Science in Business, Use Cases for Data Science	CO1, CO6	Inverse Classroom (case Study)
	Week 3	Statistical modeling, probability distributions, fitting a model	CO3, CO6	AV, Presentations, Chalk and Board
	Week 4	Intro to R and Python, Python Libraries	CO5, CO6	Self Study Materials
	Week 5	Introduction to the course, Descriptive Statistics, Probability Distributions	CO3, CO6	Presentations, Chalk and Board
	Week 6	Inferential Statistics through hypothesis tests, Permutation & Randomization Test	CO6	Presentations, Chalk and Board
	Week 7	Differentiating algorithmic and model based frameworks, Regression: Ordinary Least Squares, Ridge Regression, Lasso Regression	CO5, CO6	Practical Demo, AV, Presentations
	Week 8	K Nearest Neighbors, Regression & Classification	CO3, CO5	Practical Demo, Presentations,
	Week 9	Bias-Variance Dichotomy, Model Validation Approaches, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis	CO1, CO3	Practical Demo, AV, Presentations,
	Week 10	Regression and Classification Trees, Support Vector Machines, Ensemble Methods: Random Forest, Neural Networks, Deep learning	CO3, CO5	Practical Demo, AV, Presentations,
	Week 11	Clustering, Associative Rule Mining, Logical Modelling : Converting a conceptual model to logical model , Integrity constraints, Normalization	CO3, CO5	Practical Demo, Presentation
	Week 12	Basic principles, ideas and tools for data visualization, Discussions on privacy, security, ethics	CO1, CO4	Practical Demo, AV, Presentations



Subject Mind Mapping
Semester1&2
Semester 3
Semester 4
Semester 5
Semester 6
Semester 7
Semester 8
Programming and Software Development
Principles of Programming
Object Orieneted concepts and Programming
Web Design
Core Java Prog
Object Orieneted Design and Analysis
Web technology
Advanced java Technology
Software Engineering
Python Programming
Android
Web data Management
Advanced .Net
System tools and Technology
Principle of Database Management System
Operating System
Computer Graphics
Mobile Computing
System Architecture and design
Digital Logic Design
Computer Organization and Architecture
Microprocessor
Formal language and Automata Theory
System Programming
Compiler Design
Networking
Data Communication and Engineering
Computer Networks
Cryptography & Network Security
Mathematics
ICLA
DCMA
DEIT
CANM
Advanced Learning
Design and Analysis of Algorithms
Parallel Computing
Soft Computing
Data warehousing
Distributed OS
Big Data
Iot & Application
High Performance Computing
Project