ANNA UNIVERSITY, CHENNAI
UNIVERSITY DEPARTMENTS
REGULATIONS – 2023
FACULTY OF INFORMATION & COMMUNICATION ENGINEERING
RAMANUJAN COMPUTING CENTRE
MINOR DEGREE ON 'DATA SCIENCE'

Courses for minor degree on 'Data Science'

SL. NO	COURSE CODE	COURSE TITLE	CONTACT PERIODS	L	T	P	C
1.	CSM507	Foundations of Data Science with Python	3	3	0	0	**3**
2.	CSM508	Machine Learning for Data Science	3	3	0	0	**3**
3.	CSM509	Data Visualization	3	3	0	0	**3**
4.	CSM510	Data Security & Privacy	3	3	0	0	**3**
5.	CSM511	Big data analytics	3	3	0	0	**3**
6.	CSM512	Exploratory Analysis	3	3	0	0	**3**
Total Credits							**18**

CSM507	FOUNDATIONS OF DATASCIENCE WITH PYTHON	L	T	P	C	
		3	0	0	3	

Course Objectives:

- To understand fundamentals and the process of data science.
- To comprehend different types and representation of data and analyze them.
- To apply inferential techniques to extrapolate information from the available data.
- To utilize the Python libraries for Data Wrangling.
- To interpret data and present it using visualization libraries in Python.

Unit I	INTRODUCTION	9

Data Science: Benefits and uses – facets of data – Data Science Process: Overview – Defining research goals – Retrieving data – Data preparation – Exploratory Data analysis – build the model – presenting findings and building applications

Unit II	DESCRIBING DATA	9

Types of Data – Types of Variables – Basic Statistical descriptions of Data – Describing Data with Tables and Graphs –Describing Data with Averages – Describing Variability – Normal Distributions and Standard (z) Scores

Unit III	PROBABILITY & STATISTICS	9

Probability Review, Joint & Conditional. Review on statistics- Populations and Samples – Sampling Distribution of the mean – Hypothesis testing – Z Test – One-Tailed and Two-Tailed Tests – Estimation – t-Test for one Sample – Analysis of Variance for one factor – Chi-Square Test

Unit IV	PYTHON LIBRARIES FOR DATA WRANGLING	9

Basics of Numpy arrays – aggregations –computations on arrays –comparisons, masks, boolean logic – fancy indexing – structured arrays – Data manipulation with Pandas – data indexing and selection – operating on data – missing data – Hierarchical indexing – combining datasets – aggregation and grouping – pivot tables

Unit V	DATA VISUALIZATION	9

Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn.

Total: 45 Periods

References

1.	David Cielen, Arno D. B. Meysman, and Mohamed Ali, “Introducing Data Science”, Manning Publications, 2016.
2.	Robert S. Witte and John S. Witte, “Statistics”, Eleventh Edition, Wiley Publications, 2017.
3.	Jake VanderPlas, “Python Data Science Handbook”, O’Reilly, 2016.
4.	Avrim Blum, John Hopcroft, Ravindran Kannan, “Foundations of Data Science”, Cambridge Press, 2020

Course Outcomes:

Upon completion of the course, the students will be able to

- Understand data science fundamental and follow the correct process for applying data science.
- Represent and understand data in different formats and analyse it.
- Infer new information from the data using different analysis techniques.
- Gather, collect, and transform raw data into useful formats with Python libraries.
- Apply Python libraries to visualize and study data.

CSM508	MACHINE LEARNING FOR DATA SCIENCE	L	T	P	C	
		3	0	0	3	

Course Objectives:

- To understand the basic concepts of machine learning.
- To understand and build supervised learning models.
- To understand neural network and learn combination of classifiers
- To understand and build unsupervised learning models.
- To design and analysis of probabilistic graphical models

Unit I	INTRODUCTION TO MACHINE LEARNING	9

Machine Learning – Basic concepts – Types of Machine learning – Examples & Applications - Data Pre-processing - Noise Removal – Normalization – Bias & Variance, Review on Probability – Conditional probability – Bayesian conditional probability

Unit II	SUPERVISED LEARNING	9

Linear Regression Models: Multiple regression – Logistic regression, Naïve Bayes classifier, Nearest Neighbour and KNN Algorithm, Decision Trees, Support Vector Machines, Kernel functions

Unit III	NEURAL NETWORKS, ENSEMBLE TECHNIQUES	9

Artificial Neural Network(ANN), perceptron, multilayer perceptron, Back propogation network(BPN) activation functions, gradient descent optimization, error back propagation, Unit saturation (vanishing gradient problem) - ReLU, hyperparameter tuning, batch normalization, regularization, Ensemble Methods – Bagging, Boosting

Unit IV	UNSUPERVISORY & REINFORCEMENT LEARNING	9

Clustering – Distance Function, Minimum, maximum & average connection, Hierarchical Clustering, agglomerative – K Means clustering, Self-organizing Map, Reinforcement Learning overview

Unit V	GRAPHICAL MODELS & DIMENSION REDUCTION	9

Directed Graphical Models, Bayesian Networks, Markov Models, Hidden Markov Models, Inference- Learning Generalization, Dimension reduction-Curse of Dimensinality, PCA

Total: 45 Periods

COURSE OUTCOMES:	
At the end of this course, the students will be able to:	
CO1:	Explain the basic concepts of machine learning.
CO2:	Construct supervised learning models.
CO3:	Construct unsupervised learning algorithms.
CO4:	Evaluate and compare different models
CO5:	Design of experiments using machine learning

References:	
1.	Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, Fourth Edition, 2020.
2.	Stephen Marsland, "Machine Learning: An Algorithmic Perspective, "Second Edition", CRC Press, 2014
3.	Sridhar S & Vijayalakshmi M, “Machine Learning”, Oxford University Press, 2021
4.	Tom Mitchell, "Machine Learning", McGraw Hill, 3rd Edition, 1997.
5.	Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, "Foundations of Machine Learning", Second Edition, MIT Press, . 2018.
6.	Ian Goodfellow, Yoshua Bengio, Aaron Courville, "Deep Learning", MIT Press, 2016
7.	Sebastain Raschka, Vahid Mirjalili , "Python Machine Learning", Packt publishing, 3rd Edition, 2019.
8.	Francois Chollet, “Deep Learning with Python”, Second Edition, Manning Publications, 2021.

	DATA VISUALIZATION	L	T	P	C	Credits
		3	0	0	3	3

Course Objectives:

- To understand the fundamentals of data visualization.
- To know the working principles of various information visualization depth tools.
- To acquire knowledge about the issues in data representation.
- To visualize the Data using tools Tableau
- To gain skill in designing real time interactive information visualization system.

Unit I	**INTRODUCTION**	**9**

Context of data visualization – Definition, Methodology, Visualization design objectives. Key Factors – Purpose, visualization function and tone, visualization design options – Data representation, Data Presentation, Seven stages of data visualization, widgets, data visualization tools. Mapping - Time Series - Connections and Correlations - Scatterplot Maps - Trees, Hierarchies, and Recursion - Networks and Graphs

Unit II	**VISUALIZATION TECHNIQUES FOR TIME-SERIES, TREES & GRAPHS**	**9**

Mapping - Time series - Connections and correlations – Indicator-Area chart-Pivot table-Scatter charts, Scatter maps - Tree maps, Space filling and non-space filling methods-Hierarchies and Recursion - Networks and Graphs-Displaying Arbitrary Graphs-node link graph-Matrix representation for graphs- Info graphics

Unit III	**TEXT AND DOCUMENT VISUALIZATION**	**9**

Acquiring data, - Where to Find Data, Tools for Acquiring Data from the Internet, Locating Files for Use with Processing, Loading Text Data, Dealing with Files and Folders, Listing Files in a Folder ,Asynchronous Image Downloads, Web Techniques, Parsing data - Levels of Effort, Tools for Gathering Clues, Text Markup Languages, Regular Expressions, Grammars and BNF Notation, Compressed Data, Vectors and Geometry, Binary Data Formats, Advanced Detective Work.

Unit IV	**INTERACTIVE DATA VISUALIZATION**	**9**

Drawing with data – Scales – Axes – Updates, Transition and Motion – Interactivity - Layouts – Geomapping – Exporting, Framework – D3.js, Tableau
Dashboards

Unit V	**SECURITY IN DATA VISUALIZATION**	**9**

Port scan visualization - Vulnerability assessment and exploitation - Firewall log visualization - Intrusion detection log visualization -Attacking and defending visualization systems – Creating secured visualization system..

Total: 45 Periods

Course Outcomes:

Upon completion of the course, the students will be able to

- Apply mathematics and basic science knowledge for designing information visualizing System.
- Collect data ethically and solve engineering problem in visualizing the information.
- Implement algorithms and techniques for interactive information visualization.
- Conduct experiments by applying various modern visualization tool and solve the
- space layout problem.
- Analyze and design system to visualize multidisciplinary multivariate Data individually or in teams.

Develop a cost effective and a scalable information visualization system.

References

1	Robert Spence, “Information Visualization An Introduction”, Third Edition, Pearson Education, 2014.
2	Colin Ware, “Information Visualization Perception for Design”, Third edition, Margon Kaufmann Publishers, 2012.
3	Robert Spence, “Information Visualization Design for Interaction”, Second Edition, Pearson Education, 2006.
4	Benjamin B. Bederson and Ben shneiderman, “The Craft of Information Visualization”, Morgan Kaufmann Publishers, 2003.
5	Thomas strothotte, “Computational Visualization: Graphics, Abstraction and Interactivity”, Springer, 1998.
6	Matthew O. Ward, George Grinstein, Daniel Keim, “Interactive Data Visualization: Foundation, Techniques and Applications”, Second Edition, A. K. Peters/CRC Press, 2015.
7	Joerg Osarek, “Virtual Reality Analytics”, Gordon’s Arcade, 2016.

	DATA SECURITY AND PRIVACY	L	T	P	C	Credits
		3	0	0	3	3

Unit I	ATTACKS AND PRIVACY	9

Attacks: Analysing common attack vectors – Data Security – Probabilistic reasoning about attacks – Data security mitigations. Privacy aware Machine learning and Data Science: Privacy preserving techniques in ML - Open-source libraries for PPML – Architecting privacy in Data and ML projects

Unit II	ENCRYPTED COMPUTATION	9

Encrypted computation – Types of encrypted computation: Secure Multi-party computation – Homomorphic encryption. Real-world encrypted computation: Private set intersection – Private join and compute – Secure Aggregation – Encrypted Machine Learning. PSI and Moose

Unit III	DATA GOVERNANCE AND PRIVACY APPROACHES	9

Data Governance – Identifying sensitive data – Documenting data for use - Basic Privacy – Anonymization – Differential privacy – Privacy loss – Differential privacy with Laplace mechanism – Gaussian noise for differential privacy – Sensitivity and Privacy units – k-Anonymity – Building Privacy into Data Pipelines

Unit IV	FEDERATED LEARNING AND DATA SCIENCE	9

Distributed data – Distributed Optimization - Federated learning – Architecting federated systems – Open-source federated libraries – Federated data science

Unit V	LEGALITY OF PRIVACY	9

GDPR – CCPA – HIPAA - LGPD - PIPL- Internal policies and contracts – Adhering to contract agreements and law – Interpreting Data protection regulations - Data governance 2.0 - Indian Data Protection Framework - Use case analysis

Total: 45 Periods

Course Outcomes:

- Gain knowledge on the nature of attacks and threats and security management goals and framework
- Knowledge on the landscape of hacking and defense mechanisms
- Able to differentiate and integrate strategies for data security and protecting critical infrastructure
- Able to understand policies to mitigate data security breaching
- Knowledge on IT Act, and amendments, copy rights, IPR and cyber law to deal with offenses..

References

1.	Katharine Jarmul, Practical Data Privacy, O'Reilly Media, Inc, 2023
2.	David Evans, Vladimir Kolesnikov and Mike Rosulek, A Pragmatic Introduction to Secure Multi-Party Computation, NOW Publishers, 2022 (Free access at https://securecomputation.org/)
3.	William Stallings, Cryptography and Network Security - Principles and Practice, Seventh Edition, Pearson, 2017
4.	Indian Data Protection Framework https://www.meity.gov.in/data-protection-framework

	BIG DATA ANALYTICS	L	T	P	C	Credits
		3	0	2	3	3

Course Objectives:

- To understand big data.
- To learn and use NoSQL big data management.
- To learn mapreduce analytics using Hadoop and related tools.
- To work with map reduce applications
- To understand the usage of Hadoop related tools for Big Data Analytics.

Unit I	UNDERSTANDING BIG DATA	5

Introduction to big data – convergence of key trends – unstructured data – industry examples of big data – web analytics – big data applications – big data technologies – introduction to Hadoop – open source technologies – cloud and big data – mobile business intelligence – Crowd sourcing analytics – inter and trans firewall analytics.

Unit II	NOSQL DATA MANAGEMENT	7

Introduction to NoSQL – aggregate data models – key-value and document data models – relationships – graph databases – schemaless databases – materialized views – distribution models – master-slave replication – consistency – Cassandra – Cassandra data model – Cassandra examples – Cassandra clients.

Unit III	MAP REDUCE APPLICATIONS	6

MapReduce workflows – unit tests with MRUnit – test data and local tests – anatomy of MapReduce job run – classic Map-reduce – YARN – failures in classic Map-reduce and YARN – job scheduling – shuffle and sort – task execution – MapReduce types – input formats – output formats.

Unit IV	BASICS OF HADOOP	6

Data format – analyzing data with Hadoop – scaling out – Hadoop streaming – Haddop pipes design of Haddop distributed file system (HDFS) – HDFS concepts – Java interface – data flow – Hadoop I/O – data integrity – compression – serialization – Avro – file-based data structures – Cassandra – Hadoop integration.

Unit V	HADOOP RELATED TOOLS	6

Hbase – data model and implementations – Hbase clients – Hbase examples – praxis.
Pig – Grunt – pig data model – Pig Latin – developing and testing Pig Latin scripts.
Hive – data types and file formats – HiveQL data definition – HiveQL data manipulation – HiveQL queries.

Total: 30 Periods

Course Outcomes:
After the completion of this course, students will be able to:

- Describe the big data and use cases from selected business domains.
- Explain NoSQL big data management.
- Install, configure and run Hadoop and HDFS.
- Perform map-reduce analytics using Hadoop.
- Use Hadoop-related tools such as HBase, Cassandra, Pig and Hive for big data analytics.

References

1	Michael Minelli, Michael Chambers, and AmbigDhiraj, “Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today’s Businesses”, Wiley, 2013.
2	Eric Sammer, “Hadoop Operations”, O’Reilley, 2012.
3	Sadalage, Pramod J. “NoSQL distilled”, 2013
4	E. Capriolo, D. Wrampler, and J. Rutherglen, “Programming Hive”, O’Reilley, 2012.
5	Lars George, “HBase: The Definitive Guide: O’Reilley, 2011.
6	Eben Hewitt, “Cassandra: The Definitive Guide: O’Reilley, 2010.
7	Alan Gates, “Programming Pig”, O’Reilley, 2011.

	EXPLORATORY DATA ANALYSIS	L	T	P	C	Credits
		3	0	0	3	3

Course Objectives:

- To outline an overview of exploratory data analysis.
- To implement data visualization using Matplotlib.
- To perform univariate data exploration and analysis.
- To apply bivariate data exploration and analysis.
- To use Data exploration and visualization techniques for multivariate and time series data.

Unit I	**EXPLORATORY DATA ANALYSIS**	**9**

EDA fundamentals – Understanding data science - Significance of EDA – Making sense of data – Comparing EDA with classical and Bayesian analysis – Software tools for EDA – Visual Aids for EDA – Data transformation techniques-merging database, reshaping and pivoting, Transformation techniques.

Unit II	**EDA USING PYTHON**	**9**

Data Manipulation using Pandas – Pandas Objects – Data Indexing and Selection – Operating on Data – Handling Missing Data – Hierarchical Indexing – Combining datasets – Concat, Append, Merge and Join – Aggregation and grouping – Pivot Tables – Vectorized String Operations.

Unit III	**UNIVARIATE ANALYSIS**	**9**

Introduction to Single Variable: Distribution Variables – Numerical Summaries of Level and Spread – Scaling and Standardizing – Inequality.

Unit IV	**BIVARIATE ANALYSIS**	**9**

Relationships between Two Variables – Percentage Tables – Analysis Contingency Tables – Handling Several Batches – Scatterplots and Resistant Lines.

Unit V	**MULTIVARIATE AND TIME SERIES ANALYSIS**	**9**

Introducing a Third Variable – Causal Explanations – Three-Variable Contingency Tables and Beyond – Fundamentals of TSA – Chracteristics of time series data – Data Cleaning – Time-based indexing – Visualizing – Grouping – Resampling.

Total: 45 Periods

COURSE OUTCOMES:	
At the end of this course, the students will be able to:	
CO1:	Understand the fundamentals of exploratory data analysis.
CO2:	Implement the data Visualization using Matplotlib.
CO3:	Perform univariate data exploration and analysis.
CO4:	Apply bivariate data exploration and analysis.
CO5:	Use Data exploration and visualization techniques for multivariate and time series data.

Total: 60 Periods

REFERENCES:	
1	Suresh Kumar Mukhiya, Usman Ahmed, “Hands-On Exploratory Data Analysis with Python”, Packt Publishing, 2020. (Unit 1)
2	Jake Vander Plas, “Python Data Science Handbook: Essential Tools for Working with Data”. First Edition, O Reilly, 2017. (Unit 2)
3	Catherine Mars, Jane Elliott, “Exploring Data: An Introduction to Data Analysis for Social Scientists”, Wiley Publications, 2nd Edition, 2008. (Unit 3,4,5)
4	Eric Pimpler, Data Visualization and Exploration with R, GeoSpatial Training service, 2017.
5	Claus O. Wilke, “Fundamentals of Data Visualization”, O’reilly Publications, 2019.
6	Matthew O. Ward, Georges Grinstein, Daniel Keim, “Interactive Data Visualization: Foundations, Techniques, and Applications”, 2nd Edition, CRC press, 2015.