Data Science

Minor Degree in "Data Science"

Course Structure						
S. No.	**Course Code**	**Title**	**L**	**T**	**P**	**Credits**
1	DAS-01	Introduction to Data Science	3	0	2	4
2	DAS-02	Introduction to AI and ML	3	0	2	4
3	DAS-03	Computational Data analytics	3	0	2	4
4	DAS-04	Web Data Mining	3	0	0	3
5	DAS-05	Analysing, Visualizing and Applying data science with python	3	0	2	4
TOTAL			**15**	**0**	**8**	**19**

Detailed Syllabus

Course Code	:	DAS-01
Course Title	:	Introduction to Data Science
Number of Credits	:	4 (L: 3; T: 0; P: 2)
Course Category	:	DAS

Coursera Link: https://www.coursera.org/professional-certificates/ibm-data-science

Nptel Link: https://nptel.ac.in/courses/106106212

Course Objective:

- To Provide the knowledge and expertise to become a proficient data scientist;
- Demonstrate an understanding of statistics and machine learning concepts that are vital for data science;
- Produce Python code to statistically analyse a dataset;
- Critically evaluate data visualisations based on their design and use for communicating stories from data;

Course Contents:

Module 1: [7 Lectures]
Introduction to Data Science, Different Sectors using Data science, Purpose and Components of Python in Data Science.

Module 2: [7 Lectures]
Data Analytics Process, Knowledge Check, Exploratory Data Analysis (EDA), EDA- Quantitative technique, EDA- Graphical Technique, Data Analytics Conclusion and Predictions.

Module 3: [11 Lectures]
Feature Generation and Feature Selection (Extracting Meaning from Data)- Motivating application: user (customer) retention- Feature Generation (brainstorming, role of domain expertise, and place for imagination)- Feature Selection algorithms.

Module 4: [10 Lectures]
Data Visualization- Basic principles, ideas and tools for data visualization, Examples of inspiring (industry) projects- Exercise: create your own visualization of a complex dataset.

Module 5: [7 Lectures]
Applications of Data Science, Data Science and Ethical Issues- Discussions on privacy, security, ethics- A look back at Data Science- Next-generation data scientists.

Lab Work:
1. Python Environment setup and Essentials.
2. Mathematical computing with Python (NumPy).
3. Scientific Computing with Python (SciPy).
4. Data Manipulation with Pandas.
5. Prediction using Scikit-Learn
6. Data Visualization in python using matplotlib

Text Books/References:

1. Business Analytics: The Science of Data - Driven Decision Making, U Dinesh Kumar, John Wiley & Sons.
2. Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools, Davy Cielen, John Wiley & Sons.
3. Joel Grus, Data Science from Scratch, Shroff Publisher/O'Reilly Publisher Media
4. Annalyn Ng, Kenneth Soo, Numsense! Data Science for the Layman, Shroff Publisher Publisher
5. Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from The Frontline. O'Reilly Publisher.
6. Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press.
7. Jake VanderPlas, Python Data Science Handbook, Shroff Publisher/O'Reilly Publisher Media.
8. Philipp Janert, Data Analysis with Open Source Tools, Shroff Publisher/O'Reilly Publisher Media.

Course Outcomes: After completion of course, students would be able:

1. To explain how data is collected, managed and stored for data science;
2. To understand the key concepts in data science, including their real-world applications and the toolkit used by data scientists;
3. To implement data collection and management scripts using MongoDB.

Course Code	:	DAS-02
Course Title	:	Introduction to AI and ML
Number of Credits	:	4 (L: 3; T: 0; P: 2)
Course Category	:	DAS

Coursera: https://www.coursera.org/learn/machine-learning

Nptel: https://nptel.ac.in/courses/106106202

Course Objective:

- To understand basics of machine learning in data science.
- To understand various basic machine learning algorithm that can be used with various type of data.

Course Contents:

Module 1: [6 Lectures]
Linear Regression: Basic facts of linear regression, implementation of linear regression, case studies of linear regression using data set

Module 2: [8 Lectures]
Logistic Regression: Basic facts and implementation of logistic regression, solve a case study to predict output using existing data set

Module 3: [11 Lectures]
Clustering and Principle Component Analysis: K means and hierarchical clustering, how to make market strategies using clustering, recommendation and PCA

Module 4: [9 Lectures]
Support Vector Machine: basics of SVM and use it to detect the spam emails and recognize alphabets

Module 5: [8 Lectures]
Model Selection and advanced regression: use of Lasso and Ridge

Lab Work:

1. Use python to predict employee attrition in a firm and help them plan their manpower. (take data set from kaggle).
2. Create customer clusters using different market strategies on a data set.
3. Make a movie recommendation system.
4. Develop a prediction mechanism to predict which employee can go on leave in a company in near future.
5. Recognizing alphabets using SVM.

Text Books/References:

1. Machine Learning using Python , U Dinesh Kumar and Manaranjan Pradhan, John Wiley & Sons.
2. Advanced Data Analytics Using Python: With Machine Learning, Deep Learning by By Sayan Mukhopadhyay, Apress.
3. Practical Data Mining" by Monte F. Hancock, Auerbach Publication.
4. "Machine Learning for Absolute Beginners: A Plain English Introduction (Second Edition)" by Oliver Theobald.
5. Practical Data Science with R, Nina Zumel, John Wiley & Sons.
6. Python for Data Science for Dummies, John Paul Mueller, Luca Massaron, John Wiley & Sons.
7. Big Data and Analytics, Seema Acharya and Subhashini Chellappan, Wiley Publication.

Course Outcomes: After completion of course, students would be able:

1. To explain how data is collected, managed and stored for data science;
2. To use various type of Machine learning model
3. To implement various ML algorithms on data models

Course Code	:	DAS-03
Course Title	:	Computational Data Analytics
Number of Credits	:	4 (L: 3; T: 0; P: 2)
Course Category	:	DAS

Coursera: https://www.coursera.org/specializations/statistics

Nptel: https://nptel.ac.in/courses/111106154

Course Objective:

- To learn how to think about your study system and research question of interest in a systematic way in order to design an efficient sampling and experimental research program.
- To understand how to analyze collected data to derive the most information possible about your research questions.

Course Contents:

Module 1: [6 Lectures]
Introduction to R Computing language. Best practices in executing Reproducible Research in data science, Sampling and Simulation. Descriptive statistics, and the creation of good observational sampling designs.

Module 2: [8 Lectures]
Data visualization, Data import and visualization, Introduction to various plots

Module 3: [10 Lectures]
Frequentist Hypothesis Testing, Z-Tests, Power Analysis

Module 4: [10 Lectures]
Linear regression, diagnostics, visualization, Likelihoodist Inference, Fitting a line with Likelihood, Model Selection with one predictor

Module 5: [8 Lectures]
Bayesian Inference, Fitting a line with Bayesian techniques, Multiple Regression and Interaction Effects, Information Theoretic Approaches

Lab Work:
1. To give a basic insight of R and its various libraries.
2. Libraries in R. R as a Data Importing Tool, Dplyr. Forcats.
3. Simulation and Frequentist Hypothesis testing, Simulation and Power.
4. Bayesian computation in R, Fitting a line with Bayesian techniques.

Text Books/References:

1. Practical Data Science with R, Nina Zumel, John Wiley & Sons.
2. N. C. Das, Experimental Designs in Data Science with Least Resources, Shroff Publisher Publisher.

3. Hadley Wickham, Garret Grolemund, *R for Data Science*, Shroff Publisher/O'Reilly Publisher Publisher
4. Benjamin M. Bolker. *Ecological Models and Data in R*. Princeton University Press, 2008. ISBN 978-0-691-12522-0.
5. John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage Publications, Thousand Oaks, CA, USA, second edition, 2011. ISBN 978-1-4129-7514-8.

Course Outcomes: After completion of course, students would be able to:

1. Explain how data is collected, managed and stored for data science;
2. When to use which type of Machine learning model.
3. Implement various ML algorithms on data models.

Course Code	:	DAS-04
Course Title	:	Web Data Mining
Number of Credits	:	3 (L: 3; T: 0; P: 0)
Course Category	:	DAS

Coursera: https://www.coursera.org/specializations/data-mining

Nptel: https://nptel.ac.in/courses/106105239

Course Objective:

- To learn how to extract data from the Web.
- To understand how to analyze collected data to derive the most information

Course Contents:

Module 1: [6 Lectures]

Introduction to internet and WWW, Data Mining Foundations, Association Rules and Sequential Patterns, Basic Concepts of Association Rules, Apriori Algorithm, Frequent Itemset Generation, Association Rule Generation, Data Formats for Association Rule Mining, Mining with multiple minimum supports, Extended Model, Mining Algorithm, Rule Generation

Module 2: [8 Lectures]

Mining Class Association Rules, Basic Concepts of Sequential Patterns, Mining Sequential Patterns on GSP, Mining Sequential Patterns on Prefix Span, Generating Rules from Sequential Patterns

Module 3: [10 Lectures]

Concepts of Information Retrieval, IR Methods, Boolean Model, Vector Space Model and Statistical Language Model, Relevance Feedback, Evaluation Measures, Text and Web Page Pre-processing, Stopword Removal, Stemming, Web Page Preprocessing, Duplicate Detection, Inverted Index and Its Compression, Inverted Index, Search using Inverted Index, Index Construction, Index Compression, Latent Semantic Indexing, Singular Value Decomposition, Query and Retrieval, Web Search, Meta Search, Web Spamming.

Module 4: [10 Lectures]

Link Analysis, Social Network Analysis, Co-Citation and Bibliographic Coupling, Page Rank Algorithm, HITS Algorithm, CommModuley Discovery, Problem Definition, Bipartite Core CommModuleies, Maximum Flow CommModuleies, Email CommModuleies, Web Crawling, A Basic Crawler Algorithm – Breadth First Crawlers, Preferential Crawlers, Implementation Issues – Fetching, Parsing, Stopword Removal, Link Extraction, Spider Traps, Page Repository, Universal Crawlers, Focused Crawlers, Topical Crawlers, Crawler Ethics and Conflicts.

Module 5: [8 Lectures]

Opinion Mining, Sentiment Classification, Classification based on Sentiment Phrases, Classification Using Text Classification Methods, Feature based Opinion Mining and Summarization, Problem Definition, Object feature extraction, Comparative Sentence and Relation Mining, Opinion Search and Opinion Spam. Web Usage Mining, Data Collection and Preprocessing, Sources and Types of Data, Key Elements of Web Usage Data Preprocessing, Data Modeling for Web Usage Mining, Discovery and Analysis of Web

Usage Patterns, Session and Visitor Analysis, Cluster Analysis and Visitor Segmentation, Association and Correlation Analysis, Analysis of Sequential and Navigation Patterns.

Text Books/References:

1. Mining the Web: Discovering Knowledge from Hypertext Data, Soumen Chakrabarti, Morgan Kaufmann Publishers.
2. Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer Publications, 2011.
3. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, Elsevier Publications 2010.
4. Anthony Scime, Web Mining: Applications and Techniques, 2005.
5. Kowalski, Gerald, Mark T Maybury: Information Retrieval Systems: Theory and Implementation, Kluwer Academic Press, 1997.
6. Mathew Russell, Mining the Social Web 2nd Edition, Shroff Publisher/O'Reilly Publisher Publication.
7. Data Mining and Data Warehousing Principles and Practical Techniques, Parteek Bhatia, Cambridge University Press.

Course Outcomes: After completion of course, students would be able:

1. To explain how data is can be collected from the Web.
2. To extract data and information from the webpages.
3. To make decision based on the data collected.

Course Code	:	DAS-05
Course Title	:	Analysing, Visualizing and Applying data science with python
Number of Credits	:	4 (L: 3; T: 0; P: 2)
Course Category	:	DAS

Coursera: https://www.coursera.org/specializations/data-science-python
Nptel: https://nptel.ac.in/courses/106106179

Course Objective:

- To understand and use all the tools and libraries of python for data science.

Course Contents:

Module 1: [6 Lectures]
Data Analysis libraries: will learn to use Pandas DataFrames, Numpy multi-dimentional arrays, and SciPy libraries to work with a various dataset.

Module 2: [8 Lectures]
Pandas, an open-source library, and we will use it to load, manipulate, analyze, and visualize various datasets.

Module 3: [10 Lectures]
Scikit-learn, and we will use some of its machine learning algorithms to build smart models and make predictions, various parameters that can be used to compare various parameters.

Module 4: [10 Lectures]
Descriptive Statistics, Basic of Grouping, ANOVA, Correlation, Polynomial Regression and Pipelines, R-squared and MSE for In-Sample Evaluation, Prediction and Decision Making

Module 5: [10 Lectures]
Grid Search, Model Refinement, Binning, Indicator variables

Lab Work:

1. Demonstrate knowledge of Data Science and Machine Learning.
2. Explore New York City - 311 Complaints and Housing datasets.
3. Analyze and Visualize data using Python.
4. Perform feature engineering exercise using Python.

Text Books/References:

1. Data Visualization with Python and JavaScript, Kyran Dale, Shroff Publisher/O'Reilly Publisher Publication.
2. Data Science Using Python and R by Chantal D. Larose and Daniel T. Larose
3. Python for Data Science and Visualization -Beginners to Pro, Udemy.

Course Outcomes: After completion of course, students would:

1. To explain how data is can be collected from the Web.
2. To extract data and information from the webpages.
3. To make decision based on the data collected.