| B. Tech. Minor in Data Science | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sl. No. | Course code | Course Title | Teaching Dept. | Teaching Hours/Week | | | Examination | | | | Credits |
| | | | | Theory Lecture | Tutorial | Practical | Duration in hr | CIE Marks | SEE Marks | Total Marks | |
| | | | | L | T | P | | | | | |
| 1. | IS1603–1<br>IS1604-1 | Basic Python Programming<br>Data Analysis with Excel* | IS | 1 | 0 | 2 | 3 | 50 | 50 | 100 | 2 |
| 2. | IS2601-1 | Analysing, Visualizing and Applying Data Science with Python | IS | 1 | 0 | 2 | 3 | 50 | 50 | 100 | 2 |
| 3. | IS2003-1 | Introduction to ML | IS | 3 | 0 | 2 | 3 | 50 | 50 | 100 | 4 |
| 4. | IS2004-1 | Computational Data analytics | IS | 3 | 0 | 2 | 3 | 50 | 50 | 100 | 4 |
| 5. | IS1104-1 | Web Data Mining | IS | 3 | 0 | 0 | 3 | 50 | 50 | 100 | 3 |
| 6. | IS1105-1 | Introduction to Big Data | IS | 3 | 0 | 0 | 3 | 50 | 50 | 100 | 3 |
| | | **TOTAL** | | **14** | **0** | **8** | **18** | **300** | **300** | **600** | **18** |

*For the students who have studied Python Programming in their previous semesters.

## Detailed Syllabus

| BASIC PYTHON PROGRAMMING | | | |
|---|---|---|---|
| Course Code | IS1603-1 | CIE Marks | 50 |
| L:T:P | 1:0:2 | SEE Marks | 50 |
| Number of Teaching Hours | 13+26 | Credits | 02 |

**Course Objective:**

1. Construct Python programs using data types and looping.
2. Make use of python operators for manipulating lists, dictionaries and files.
3. Design object-oriented Python programs using classes and objects
4. Design useful stand-alone and GUI applications in Python

### Unit 1

**Introduction to python:** The concept of data types: variables, assignments; immutable variables. Numerical types; arithmetic operators and expressions; comments in the program; Conditions, Boolean logic, logical operators; ranges.

**Control statements**: if-else, loops (for, while); short-circuit (lazy) evaluation.

**String manipulations**: subscript operator, indexing, slicing a string; strings and number system: converting strings to numbers and vice versa, Binary, octal, hexadecimal numbers.

**Lists, tuples, and dictionaries**: basic list operators, replacing, inserting, removing an element; searching and sorting lists.

**Dictionaries**: dictionary literals, adding and removing keys, accessing and replacing values; traversing dictionaries.

15 Hours

**Unit 2**

Text files: manipulating files and directories, OS and sys modules; text files: reading/writing text and numbers from/to a file; creating and reading a formatted file (csv or tab-separated).

File Handling: Reading From Text Files, Writing to Text Files, Seeking Within Files:

**Functions:** Design with functions: hiding redundancy, complexity; arguments and return values; formal vs actual arguments, named arguments. Program structure and design. Recursive functions.

**Classes and OOP**: Classes, objects, attributes and methods; defining classes; design with classes, data modeling; persistent storage of objects, inheritance

15 Hours

**Unit 3**

Graphical user interfaces: Event-driven programming paradigm, Creating simple GUI Buttons, labels, entry fields, dialogs.

9 Hours

Lab Work:

1. Python Environment setup and Essentials.
2. Looping constructs
3. String manipulation
4. File Handling
5. Create GUI

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

1. **Make use of** with the basics of Python programming like data types and looping
2. **Apply** string manipulation, operator overloading concepts in programming.
3. **Make use of** lists, dictionaries and files
4. **Develop** object-oriented Python programs using classes and objects

**Text Books/References:**


1. "**The Fundamentals of Python: First Programs**", Kenneth A. Lambert, 2011, Cengage Learning, ISBN: 978-1111822705
2. "**The Fundamentals of Python: First Programs**", Kenneth A. Lambert, 2011, Cengage Learning, ISBN: 978-1111822705.
3. **"Python Cookbook"**, David M. Baezly O'Reilly Media. 3 edition Joel Grus, Data Science from Scratch, Shroff Publisher/O'Reilly Publisher Media

**E-Resources**

1. For Introduction to Python
https://www.codecademy.com/learn/python
2. For Tkinter www.learnpython.org/

*****

| DATA ANALYSIS WITH EXCEL | | | |
|---|---|---|---|
| Course Code | IS1604-1 | CIE Marks | 50 |
| L:T:P | 1:0:2 | SEE Marks | 50 |
| Number of Teaching Hours | 13+26 | Credits | 02 |

**Course Objective**

1. To learn the basic functionalities of Excel
2. To create and apply formulae in Excel
3. To use Excel for data analysis

**Unit 1**

**Introduction to Excel:** About Excel & Microsoft, Uses of Excel, Excel software, Spreadsheet window pane, Title Bar, Menu Bar, Standard Toolbar, Formatting Toolbar, the Ribbon, File Tab and Backstage View, Formula Bar, Workbook Window, Status Bar, Task Pane, Workbook & sheets

**Columns & Rows:** Selecting Columns & Rows, Changing Column Width & Row Height, Autofitting Columns & Rows, Hiding/Unhiding Columns & Rows, Inserting & Deleting Columns & Rows, Cell, Address of a cell, Components of a cell – Format, value, formula, Use of paste and paste special

**Functionality Using Ranges**: Ranges, Selecting Ranges, Entering Information Into a Range, Using AutoFill .

**15 Hours**

**Unit 2**

**Creating Formulas:** Using Formulas, Formula Functions – Sum, Average, if, Count, max, min, Proper, Upper, Lower, Using AutoSum,

**Advance Formulas**: Concatenate, VLOOKUP, Hlookup, Match, Countif, Text, Trim

**Spreadsheet Charts**: Creating Charts, Different types of charts, Formatting Chart Objects, Changing the Chart Type, Showing and Hiding the Legend, Showing and Hiding the Data Table.

**15 Hours**

**Unit 3**

**Data Analysis:** Sorting, Filter, Text to Column, Data Validation

**PivotTables:** Creating PivotTables, Manipulating a PivotTable, Using the PivotTable Toolbar, Changing Data Field, Properties, Displaying a PivotChart, Setting PivotTable Options, . Adding Subtotals to PivotTables

**Spreadsheet Tools:** Moving between Spreadsheets, Selecting Multiple Spreadsheets, Inserting and Deleting Spreadsheets Renaming Spreadsheets, Splitting the Screen, Freezing

Panes, Copying and Pasting Data between Spreadsheets, Hiding , Protecting worksheets.
**09 Hours**

**Course Outcomes**

At the end of this course students will be able to

1. Apply basic excel functionalities
2. Write basic and advanced formulae in Excel
3. Draw charts using the data in Excel
4. Use Excel for data analysis

**Text Books**

1. Microsoft Excel 2019: Data Analysis & Business Model, L Winston Wayne, PHI Learning Pvt. Ltd. (11 October 2019)

**********

| **ANALYSING, VISUALIZING AND APPLYING DATA SCIENCE WITH PYTHON** | | | |
|---|---|---|---|
| Course Code | IS2601-1 | CIE Marks | 50 |
| L:T:P | 1:0:2 | SEE Marks | 50 |
| Number of Teaching Hours | 13+26 | Credits | 02 |

Course Objective:

1. To learn how to use python for data science.
2. To understand and use all the tools and libraries of python for data science.

**Unit 1**

Data Analysis libraries: will learn to use Pandas DataFrames, NumPy multi-dimensionalarrays, and SciPy libraries to work with a various dataset.

Pandas, an open-source library, and we will use it to load, manipulate, analyze, and visualize various datasets.

15 Hours

**Unit 2**

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.

Descriptive Statistics, Basic of Grouping, ANOVA, Correlation,

15 Hours

**Unit 3**

Polynomial Regression and Pipelines, R-squared and MSE for In-Sample Evaluation, Prediction and Decision Making Grid Search, Model Refinement, Binning, Indicator variables

9 Hours

Lab Work:

1. Demonstrate knowledge of Data Science and Machine Learning.
2. Apply Data Science process to a real life scenario.
3. Explore New York City - 311 Complaints and Housing datasets.

4. Analyze and Visualize data using Python.
5. Perform feature engineering exercise using Python.
6. Build and validate predictive machine learning model using Python.
7. Create and share Actionable Insights to real life data problems.

**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.

**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, Wiley Publication.
3. Python for Data Science and Visualization -Beginners to Pro, Udemy.


**E-resources**

1. https://www.simplilearn.com/top-python-libraries-for-data-science-article
2. https://www.geeksforgeeks.org/libraries-in-python/
3. https://www.w3schools.com

********

| INTRODUCTION TO MACHINE LEARNING | | | |
|---|---|---|---|
| Course Code | IS2003-1 | CIE Marks | 50 |
| L:T:P | 3:0:2 | SEE Marks | 50 |
| Number of Teaching Hours | 39+26 | Credits | 04 |

**Course Objective:**

1. To understand basics of machine learning in data science.
2. To understand various basic machine learning algorithm that can be used withvarious type of data.

**Unit 1**

Foundations of Machine Learning: What is machine learning? Applications of Machine learning, Understand Data, Types of machine learning: Supervised, Unsupervised, Reinforcement Learning, Theory of learning: feasibility of learning, error and noise, training versus testing, theory of generalization, bias and variance, learning curve Measures of Similarity and Dissimilarity: Transformation, similarity and dissimilarity between simple attributes, Euclidean distance, Minkowski distance, Similarity measures for binary data: Simple matching coefficient, Jaccard coefficient, Cosine similarity, Correlation. Classification: Preliminaries; General approach to solving a classification problem; Confusion Matrix, Decision tree induction, How decision tree works?, Hunt's algorithm, Design issues, Methods for expressing attribute test conditions, Measures for selecting best fit, Algorithm for decision tree

induction; Rule-based classifier: How rule based classifier works, Rule ordering schemes, Sequential covering algorithm, Nearest-neighbor classifier: Selecting K value, KNN algorithm **15 Hours**

**Unit 2**

Association Analysis–1: Problem definition, Frequent item set generation, Apriori principle, Candidate generation and pruning, Rule Generation in Apriori algorithm. Association Analysis – 2: FP-Growth algorithm, Evaluation of association patterns, Effect of skewed support distribution, Sequential patterns.

Cluster Analysis: Different types of clustering: Hierarchical vs partitional, Exclusive vs overlapping, Fuzzy clustering, Complete vs partial. Types of clusters: Well separated, Prototype based clusters, Graph based clusters, Density based clusters, Conceptual clusters, K-means clustering algorithm, centroids and objective functions, Choosing initial centroids, time space complexity of K-means, Kmeans additional issues, Strengths and weakness of k-means, Agglomerative hierarchical clustering, Key issues in hierarchical clustering, Strengths and weaknesses, DBSCAN algorithm, Cluster Evaluation: Overview, Unsupervised cluster evaluation using cohesion and separation **15 Hours**

**Unit 3**

Parametric Methods: Introduction Maximum Likelihood Estimation, Bernoulli Density, Multinomial Density, Gaussian (Normal) Density, Evaluating an Estimator: Bias and Variance, The Bayes' Estimator Parametric Classification Regression Tuning Model Complexity: Bias/Variance Dilemma, Model Selection Procedures.

Nonparametric Methods: Introduction, Nonparametric Density Estimation, Histogram Estimator, Kernel Estimator, k-Nearest Neighbor Estimator, Generalization to Multivariate Data, Nonparametric Classification, Condensed Nearest Neighbor, **09 Hours**

Lab Work:

1. Demonstrate importing a dataset, identifying, and handling missing values, encoding categorical data and feature scaling using machine learning libraries.
2. Implement the FIND-S algorithm for finding the most specific hypothesis based on a given set of training data samples. Read the training data from a .CSV file.
3. Demonstrate the working of Candidate-Elimination algorithm to output a description of the set of all hypotheses consistent with the training examples.
4. Construct a decision tree based on ID3 algorithm. Use an appropriate dataset for building the decision tree and apply this knowledge to classify a new sample
5. Demonstrate application of Linear regression to predict the stock market prices of any organization.
6. Demonstrate the use of Support Vector Machine algorithm for a regression problem on any preferred dataset and evaluate the performance of the model
7. Write a program to implement k-Nearest Neighbor classification algorithm on the iris flower dataset and visualize the results.
8. Demonstrate image segmentation using K-means clustering algorithm and visualize the results

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

1. Develop an appreciation for what is involved in learning models from supervised learning.
2. To use various type of Machine learning model
3. To implement various ML algorithms on data models

**Text Books/References:**


1. C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
2. Ethem Alpaydin, Introduction to Machine Learning, Second Edition, 20043. Practical Data Mining" by Monte F. Hancock, Auerbach Publication.
3. T. M. Mitchell, "Machine Learning", McGraw Hill, 1997.
4. R. O. Duda, P. E. Hart and D. G. Stork Pattern Classification, Wiley Publications, 2001.
5. T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2e, 2008.
6. P. Flach, "Machine Learning: The art and science of algorithms that make sense of data", Cambridge University Press, 2012.
7. K. P. Murphy, "Machine Learning: A probabilistic perspective", MIT Press, 2012. 6 M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of Machine Learning", MIT Press, 2012


*****

| **COMPUTATIONAL DATA ANALYTICS** | | | |
|---|---|---|---|
| Course Code | IS2004-1 | CIE Marks | 50 |
| L:T:P | 3:0:2 | SEE Marks | 50 |
| Number of Teaching Hours | 39+26 | Credits | 04 |

**Course Objective:**

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

**Unit 1**

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.

Data visualization, Data import and visualization, Introduction to various plots **15 Hours**

**Unit 2**

Frequentist Hypothesis Testing, Z-Tests, Power Analysis

Linear regression, diagnostics, visualization, Likelihoods Inference, Fitting a line with Likelihood, Model Selection with one predictor **15 Hours**

**Unit 3**

Bayesian Inference, Fitting a line with Bayesian techniques, Multiple Regression andInteraction Effects, Information Theoretic Approaches **09 Hours**

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.

**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.

**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.
3. Hadley Wickham, Garret Grolemund, *R for Data Science*, Shroff Publisher/O'Reilly 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.


**********

| **WEB DATA MINING** | | | |
|---|---|---|---|
| Course Code | IS1104-1 | CIE Marks | 50 |
| L:T:P | 3:0:0 | SEE Marks | 50 |
| Number of Teaching Hours | 39 | Credits | 03 |

**Course Objective:**

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

**Unit 1**

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
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. **15 Hours**

**Unit 2**

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, Stopwords Removal, Stemming, Web Page Preprocessing, DuplicateDetection, Inverted Index and Its Compression, Inverted Index, Search using Inverted Index, Index Construction, Index Compression, Latent Semantic Indexing, Singular ValueDecomposition, Query and Retrieval, Web Search, Meta Search, Web Spamming.
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. **15 Hours**

**Unit 3**

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. **09 Hours**

**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.

**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.

*****

| INTRODUCTION TO BIG DATA | | | |
|---|---|---|---|
| Course Code | IS1105-1 | CIE Marks | 50 |
| L:T:P | 3:0:0 | SEE Marks | 50 |
| Number of Teaching Hours | 39 | Credits | 03 |

**Course Objective:**

1. To understand how to use Big data frameworks and APIs.
2. To conceptualize data analysis and to learn about various data processing and pipelining strategies.
3. To understand and visualize map-reduce computing paradigm.

**Unit 1**

Classification of Digital Data, Structured and Unstructured Data, Introduction to Big Data: Characteristics Evolution – Definition, Data Warehouse, Hadoop ecosystem in Brief, Map Reduce: Mapper – Reducer – Combiner – Partitioner – Searching – Sorting – Compression - Terminologies used in Big Data Environments, Functional Programming in Scala: Basic Syntax type inference- Parameters-Recursive arbitrary collections, ConsList-Arrays-Tail recursion Higher order functions. MapReduce Template-Pattern Matching syntax, objects in Scala. Apache Spark: -Resilient Distributed Datasets -Creating RDDs, Lineage and Fault tolerance, DAGs, Immutability, task division and partitions, transformations and actions, lazy evolutions, and optimization - Formatting and housing data from spark RDDs--Persistence.

**15 Hours**

**Unit 2**

HADOOP: Data format, analyzing data with Hadoop, Hadoop streaming, Hadoop pipes – design of Hadoop 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. 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. **15 Hours**

**Unit 3**

Data frames, datasets, Apache Spark SQL, Setting up a standalone Spark cluster-: spark-shell, Curriculum for B.Tech. Artificial Intelligence and Data Science: 2022-26 190 basic API, Modules-Core, Key/Value pairs and other RDD features, MLlib-examples for bi-class SVM and logistic regression. MongoDB: Why Mongo DB - Terms used in RDBMS and Mongo DB - Data Types - MongoDB Query Language. Stream and Graph Processing on Spark. **09 Hours**

**Course Outcomes:**

At the end of the course student will be able to

1. Solve problems through a map-reduce approach.
2. Implement data analytics solutions using general data pipelining.

3. Apply scaling up machine learning techniques and associated computing techniques and technologies.
4. Identify the characteristics of datasets and compare the trivial data and big data for various applications.
5. Use Hadoop-related tools such as HBase, Cassandra, Pig, and Hive for big data analytics

**Text Books/References:**


1. Learning Spark: Lightning-Fast Big Data Analysis', Holden Karau , Andy Konwinski, Patrick Wendell and MateiZaharia, O'Reilly; 1st edition , 2015.
2. Eric Sammer, "Hadoop Operations", O'Reilley, 2012
3. Michael Minelli, Michelle Chambers, and AmbigaDhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.
4. 'High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark', Holden Karau, Rachel Warren, O'Reilly; 1st edition, 2017.
5. 'Programming in Scala: A Comprehensive Step-by-Step Guide', Martin Odersky,Lex Spoon and Bill Venners, Artima Inc; Version ed. edition , 2008.
6. "MongoDB: The Definitive Guide", Shannon Bradshaw, Eoin Brazil, Kristina Chodorow, O'Reilly; 3rd edition,2019.