CSH202: DATA SCIENCE WITH PYTHON

Hours/Week: 4 I.A. Marks: 30

Credits: 4 Exam. Marks: 70

Course Learning Objectives: Students will try to learn,

1. The probability distributions and density estimations to perform analysis of various kinds of data
2. The statistical analysis techniques using Python and R programming languages.
3. Expand the knowledge in R and Python to use it for further research.
4. The students will be able to carry out data analysis/statistical analysis effectively visualize the data.

Course Outcomes: After completing the course, the students will be able to,

CO1: Understand the fundamentals of data analytics and study the basic concepts of Excel spreadsheet Functions.

CO2: Realize the importance of filtering functions, charts and tables.

CO3: Identify the importance and usage of R package and its features

CO4: Learn the fundamentals of python programming

C05: Understand the various search methods and visualization techniques.

C06: Learn to use various techniques for mining data stream and applications using Map Reduce Concepts.

C07: Introduce programming tools PIG & HIVE in Hadoop echo system.

UNIT-I 12Hrs.

Introduction To Core Concepts And Technologies: Introduction, Terminology, data science process, data science toolkit, Types of data, Example applications. Data collection and management: Introduction, Sources of data, Data collection and APIs, Exploring and fixing data, Data storage and management, using multiple data sources. Data analysis: Introduction, Terminology and concepts, Introduction to statistics, Central tendencies and distributions, Variance, Distribution properties and arithmetic, Samples/CLT. Python: The Basics Getting Python, The Zen of Python, Whitespace Formatting, Modules, Arithmetic, Functions, Strings, Exceptions, Lists, Tuples, Dictionaries, Sets, Control Flow, Truthiness, The Not-So-Basics, Sorting, List Comprehensions, Generators and Iterators, Randomness, Regular Expressions, Object Oriented Programming, Functional Tools.

UNIT-II 12Hrs.

Mathematical Preliminaries And Statistical data modeling: Review of basic probability theory and distributions, correlation coefficient, linear regression, statistical inference, exploratory data analysis and visualization. Scores and Rankings: The Body Mass Index (BMI), Z-scores and Normalization, Advanced Ranking Techniques, Clyde's Revenge, Arrow's Impossibility Theorem. Statistical Analysis: Statistical Distributions, Sampling from Distributions, Statistical Significance, Permutation Tests and P-values, Bayesian Reasoning, data acquisition, data preprocessing techniques including data cleaning, selection, integration, transformation and reduction, and interpretation. Visualizing Data: Exploratory Data Analysis, Developing a Visualization Aesthetic, Chart Types, Great Visualizations, Reading Graphs, Interactive Visualization. Mathematical Models: Philosophies of Modeling, A Taxonomy of Models, Baseline Models, Evaluating Models, Evaluation Environments, Simulation Models. Linear Algebra: The Power of Linear Algebra, Visualizing Matrix Operations, Factoring Matrices, Eigen values and Eigen vectors, Eigen value Decomposition.

UNIT-III **12Hrs.**

Machine Learning: Modeling, Over fitting and Under fitting, Correctness, The Bias-Variance Trade-off, Feature Extraction and Selection. Degrees of Supervision, Supervised Learning , Unsupervised Learning , Semi-supervised Learning , Feature Engineering Linear and Logistic Regression: Linear Regression, Better Regression Models, Regression as Parameter Fitting, Simplifying Models through Regularization, Classification and Logistic Regression, Issues in Logistic Classification. Classification: Measuring Distances, Nearest Neighbor Classification, Graphs, Networks, and Distances, Naive Bayes, Apriori algorithm Decision Tree Classifiers, Boosting and Ensemble Learning, Support Vector Machines, Decision Trees and Random Forests, Random Forest Regression, Principal Component Analysis, Manifold Learning. Clustering: Introduction to clustering, partition, hierarchical, and density based clustering (k-means, agglomerative, and DBSCAN), outlier detection, clustering performance evaluation. k-Means Clustering, Gaussian Mixture Models, Kernel Density Estimation, Quality & Validity of clustering methods Cluster analysis software.

UNIT-IV **12Hrs.**

Advanced Prediction and Neural Networks: Introduction to predictive modeling , decision tree, nearest neighbor classifier and naïve Baye's classifier, classification performance evaluation and model selection. ARIMA model and SARIMA Model. Neural Networks: Supervised Learning Neural Networks, Perceptrons, Adaline, Back propagation Multilayer Perceptrons, Radial Basis Function Networks, Unsupervised Learning Neural Networks, Competitive Learning Networks, Hebbian Learning. Fuzzy Set Theory: Introduction to Neuro, Fuzzy and Soft Computing, Fuzzy Sets, Basic Definition and Terminology, Set-theoretic Operations, Member Function Formulation and Parameterization, Fuzzy Rules, Introduction to Fuzzy Reasoning, Extension Principle and Fuzzy Relations. Genetic Algorithm: Difference between Traditional Algorithms and GA, The basic operators, Schema theorem, convergence analysis, stochastic models, applications in search and optimization. Encoding, Fitness Function, Reproduction, Cross Over, Mutation, Application of Genetic Algorithm. Neuro Fuzzy Modeling: Adaptive Neuro-Fuzzy Inference Systems, Architecture, Hybrid Learning Algorithm, Learning Methods that Cross-fertilize ANFIS and RBFN, Coactive Neuro Fuzzy Modeling, Framework Neuron Functions for Adaptive Networks, Neuro Fuzzy Spectrum. Applications of Data Science, Technologies for visualization, Bokeh (Python), recent trends in various data collection and analysis techniques, various visualization techniques, application development methods of used in data science.

REFERENCE BOOKS:

1. Skiena, Steven S. The Data Science Design Manual. Springer, 2017.
2. VanderPlas, Jake. Python Data Science Handbook: Essential tools for working with data. O'Reilly Media, Inc., 2016.
3. Joel Grus, Data Science from Scratch: First Principles with Python, 1st Edition, ORIELLY Publications, 2015. ,
4. Rachel Schutt, Cathy O'Neil Doing Data Science: Straight Talk from the Frontline, 3 rd Edition, O'Reilly Publication, 2014
5. W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy and iPython, 2 nd Ed., O'Reilly, 2017.
6. Cathy O'Neil, Rachel Schutt, Doing Data Science, Straight Talk from The Frontline. O'Reilly, 2013.
7. M. Mitchell, An Introduction to Genetic Algorithms, Prentice-Hall, 1998.