RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

Artificial Intelligence and Data Science, IV-Semester

AD404: DATA SCIENCE

Rationale:

- The purpose of this subject is to cover the underlying concepts and techniques used in Data Science. Some of these techniques can be used in Data Analysis & in prediction.
- To understand modern way to get insights from the data.

Prerequisite: - The students should have knowledge of Probability and Statistics.

Course Outcomes: After completing the course student should be able to:

- To expose students to various perspectives and concepts in Data Science
- To Understand the Concept of Advance Excel
- Data visualization techniques and ability to implement data visualization techniques
- Student should be able to get insights from the data.

Unit-I: Introduction to Data Science-Types of Data: structured and unstructured data, Data Science Road Map: Frame the Problem; Understand the Data, Data Wrangling, Exploratory Analysis, Extract Features, Model and Deploy Code. Graphical Summaries of Data: Pie Chart, Bar Graph, Pareto Chart, Histogram. Measures of central tendency of Quantitative Data: Mean, Median, Mode. Measures of Variability of Quantitative Data: Range, Standard Deviation and Variance. Probability: Introduction to Probability, Conditional Probability.

Unit II: Unstructured Data Analytics- Importance of Unstructured Data, Unstructured Data Analytics: Descriptive, diagnostic, predictive and prescriptive data Analytics based on Case study. Data Visualization: boxplots, histograms, scatterplots, features map visualization, t-SNE **. Overview of Advance Excel-** Introduction, Data validation, Introduction to charts, pivot table, Scenario manager, Protecting data, Excel minor, Introduction to macros.

Unit III: Statistical & Probabilistic analysis of Data, Multiple hypothesis testing, Parameter Estimation methods, Confidence intervals, Correlation & Regression analysis, logistic regression, Shrinkage Methods, Lasso Regression, Bayesian statistics. L1 and L2 regularizations. **POWERFUL DATA ANALYSIS**—SUMIFS, SUMPRODUCT, VLOOKUP | XLOOKUP, INDEX + MATCH, Handling Formula Errors, Dynamic Array Formulas, Circular References, Formula Auditing, Pivoting.

Unit IV: Data Manipulation With Pandas- Introduction to Pandas, understanding DataFrame, Missing Values, Data operation, String Manipulation, Regular Expressions and Data learning, Outlier and Error. Visualization tool in Python: Representation of Pie Chart, Bar Chart, Histogram, Scatterplots using Python. Data Analysis, performance metrics, ROC curve, types of errors, Overfitting & Under fitting, evaluating performance of learning model: Holdout, Random sampling, cross validation and Bootstrap method. Bagging & boosting, Gradient Boosting, Random Forests, Committee Machines.

Unit V: Introduction to Business Intelligence- Introduction, Types of Business Intelligence, Modern Business Intelligence Tools, Modern Business Intelligence. **Data Science and Ethical Issues-** Unfair discrimination, Reinforcing human biases, Lack of transparency. Discussions on privacy, security, ethics, Role of Next-generation data scientists.

Reference Books

1. The Data Science Workshop, Anthony So, Thomas V. Joseph, Robert Thas John, Andrew Worsley, and Dr. Samuel Asare, Packt Publication
2. Python Data Science Handbook, Jake VanderPlas, OREILLY
3. The Data Science HandBook, Wiley Publication.
4. Principles of Data Science, Packt Publication.
5. Microsoft Excel 2019: Data Analysis & Business Modelling, L. Winston Wayne, PHI
6. Data Collection: Methods, Ethical Issues & Future Directions (Government Procedures and Operations: Ethical Issues in the 21st Century), by Susan Elswick, Nova Science Publishers Inc