Suggested Teaching Guidelines for

Advanced Analytics using Statistics
PG-DBDA September 2022

Duration: 46 hours classroom and 44 hours Lab

Objective: To perform advanced analytics using Python & R skills and important mathematical concepts.

Prerequisites: Good Knowledge of Basic Mathematics

Evaluation method: Theory exam– 40%
Lab Exam - 40%
Internal exam- 20%

<u>List of Books / Other training materials</u>

Text Book:

1. Statics Using R by Sudha Purohit, Pub: Narosa
2. Practical Statistics For Data Scientists 2/Ed 50+ Essential Concepts Using R and Python by Peter Bruce (Shroff/O'Reilly Publisher)

Reference:

1. Beginning R – The Statistical Programming Languageby Dr. Mark Gardener PUB: WILEY
2. Art of Programming in R, by Norman Matloff
3. Statistics for Management by Levin
4. Business Analytics: Methods, Models, and Decisions by James R Evans
5. Introductory Statistics with R (Statistics and Computing) by Peter Dalgaard
6. R in a Nutshell by Joseph Adler (O'REILLY)
7. R Cookbook by Paul Teetor (O'REILLY)
8. The R Book, Second Edition
9. Statistics Using R, Shailaja Deshmukh, Sudha Purohit, Sharad Gore, Pub: Narosa
10. Statistical Inference via Data Science: A ModernDive into R and the Tidyverse by Chester Ismay (Chapman & Hall Publisher)
11. Statistics for Machine Learning: Implement Statistical methods used in Machine Learning using Python by Himanshu Singh (BPB Publications)
12. Statistics: Statistics for Beginners in Data Science: Theory and Applications of Essential Statistics Concepts using Python by AI Publishing

Note:

- **Each session mentioned is for theory and of 2 hours' duration. Lab assignments are indicatives; faculty needs to assign more assignments for better practice.**
- **Trainer has to teach the statistical and probability concepts involved here in detail**

Session 1, 2 and 3:

- Introduction to Analytics
- Data analytics Life Cycle:
- Discovery,
- Types of Data
- Pulling the data from CSVs
- Data preparation
- Model planning
- Model building implementation
- Quality assurance
- Documentation

- Management approval
- Installation
- Acceptance and operation
- Intelligent data analysis
- Calculating frequency counts of univariate and bivariate crosstabs and interpreting the normalize option of pandas.crosstab

Assignment –Lab:

1. Import csv file using R and perform ETL operation using dplyr package.
2. Import csv file using pandas options and calculate frequencies for categorical variables and mean, variance, standard deviation, coefficient of variation, skewness, kurtosis with numerical variables

Session 4 & 5:

- Random Variable
- Concepts of Correlation
- Covariance
- Outliers
- Producing graphs like bar chart, pie chart, histogram, boxplot, density plot, scatter plot with different options in pandas, matplotlib and seaborn libraries
- Detecting Outliers using Boxplot

Assignment –Lab:

1. Load any dataset and find out the relation between different variables graphically and also find outliers using boxplot
2. Load any dataset and find out the covariance between two fields and also find the correlation and determine how two fields are correlated. Also handle the outliers in the data.

Session 6 & 7:

- Sample Spaces and Events
- Concept of Probability: Addition, Multiplicative, Complement Rules
- Joint, Conditional and Marginal Probability
- Bayes' Theorem
- Usage of sklearn.BernoulliNB function to predict probabilities (predict_proba ( ) method)

Assignment –Lab: Load any dataset, apply Bayes' Theorem using sklearn.BernoulliNB function to demonstrate the output it gives.

Session 8 & 9:

- Probability Distribution
  - Discrete distribution – (Binomial, Poisson) Probability Mass Functions, Distribution Functions
  - Continuous distribution – (Normal) Probability Density Function, Distribution Function, Inverse of Distribution Function

Assignment –Lab: Calculating the probabilities in various scenarios for Binomial, Poisson and Normal Distributions

Session 10:

- Descriptive Statistical Measures
- Summary Statistics - Central Tendency & Dispersion (Mean, Median, Mode, Quartiles, Percentiles, Range, Interquartile Range, Standard Deviation, Variance, and Coefficient of Variation)

Suggested Teaching Guidelines for

Advanced Analytics using Statistics
PG-DBDA September 2022

Assignment –Lab: Load any dataset and find out the mean, median mode and other central tendencies of the dataset.

Session 11:

- Sample & population, Uni-variate and bi-variate sampling, re-sampling
- Sampling and Estimation: Sampling Distribution
- Concept of Confidence Interval
- Central Limit Theorem

Assignment –Lab: Load any dataset and Explore sampling techniques.

Session 12 & 13:

- Statistical Inference Terminology (types of errors, tails of test, confidence intervals etc.)
- Hypothesis Testing
- Parametric Tests: One sample t-test, paired t-test, 2 independent samples t-test, 1-Way ANOVA
- Non-parametric Tests- chi-Square, U-Test

Assignment –Lab: Load any dataset and Perform the hypothesis testing on correlated variables.

Session 14:

- Predictive Modelling (From Correlation to Supervised Segmentation):
    - Identifying Informative Attributes,
    - Segmenting Data by Progressive Attributive,
    - Models,
    - Induction and Prediction,
    - Supervised Segmentation,
    - Visualizing Segmentations,
    - Trees as Set of Rules,
    - Probability Estimation;

Assignment –Lab: Explore predictive modelling techniques.

Session 15 & 16:

- Simulation and Risk Analysis
- Monte Carlo Simulation Method
- Optimization, Linear
- Formulating any Linear Programming Problem (LPP) and solving it using Excel and Python options

Assignment –Lab:

1. Explore Monte Carlo simulation using Excel or Python
2. Explore different LP Problems for Linear Optimization options using Excel and Python (pulp or scipy)

Session 17:

- Decision Analytics:
    - Evaluating Classifiers,
    - Analytical Framework,
    - Evaluation,
    - Baseline,
    - Performance and Implications for Investments in Data;

Suggested Teaching Guidelines for

Advanced Analytics using Statistics
PG-DBDA September 2022

Session 18:

- Evidence and Probabilities:
    - Explicit Evidence Combination with Bayes Rule,
    - Probabilistic Reasoning;

Session 19:

- Business Strategy:
    - Achieving Competitive Advantages,
    - Sustaining Competitive Advantages

Session 20:

- Factor Analysis,
- Directional Data Analytics,

Assignment –Lab: Download dataset and perform factor analysis on it.

Session 21 & 22:

- Interactivity with ipwidgets in Jupyter Notebook
- Creating simple interactive graphics with ipwidgets
- Creating a simple What-if tool for predicting using ipwidgets

Assignment –Lab: Explore different datasets for generating interactive graphs and creating a simple what-if tool for predictions

Session 23:

- Generating Simple Interactive Applications in Shiny App with R
    - Interactive Boxplots, Histograms
    - Interactive scatter plots

Assignment –Lab: Creating Simple Interactive Shiny App with any dataset