FF No.: 654

COURSE CODE: MD2201 **COURSE NAME: DATA SCIENCE**

Course Prerequisites:

1. Linear Algebra Basics
2. Central Tendency & Measures of Dispersion – Mean, Mode, Median
3. Probability
4. Some exposure to programming environment – C programming; Python

Course Objectives:

1. Understand data processing pipeline
2. Perform dimensionality reduction operations
3. Optimize the performance of functions
4. Apply descriptive statistics tools
5. Deduce meaningful statistical inferences
6. Use unsupervised classification algorithms
7. Use supervised classification algorithms
8. Utilize the data science principles for an entire project life cycle as a case study

Credits: 5 **Teaching Scheme Theory: 3** Hours/Week
Tut: 1 Hours/Week
Lab: **2** Hours/Week

Course Relevance:

The course is offered in S.Y. B.Tech. to all branches of Engineering
Data Science is a multidisciplinary field. It uses scientific approaches, procedures, algorithms and frameworks to extract knowledge and insight from a huge amount of data.
Data Science uses concepts and methods which belong to fields like information technology, Mathematics, Statistics, Computer Science etc.
Data Science influences the growth and improvements of the product by providing a lot of intelligence about customers and operations, by using methods such as data mining and data analysis.
The course is relevant to all branches of Engineering and beyond, since data is generated as an obvious outcome of many processes.

SECTION-1

- Introduction to Data Science
  Role of data scientist, introduction to R, R studio; introduction to univariate and multivariate systems, understanding databases, Data Processing - Data collection; Data preparation; Data visualization techniques and inferences - scatter plot, scatter matrix, histogram, box plot. **(6 Hours)**
- Normal distribution, evaluating normal distribution, Binomial distribution, confidence Intervals, central limit Theorem, hypothesis testing, inference for numerical data – t-distribution, paired data, ANOVA **(8 Hours)**
- Vector norms, distances & projections, discriminants, Principal Component Analysis, Optimization: constrained and unconstrained, Gradient Descent **(6 Hours)**

SECTION-2

- Supervised Learning – line fitting, residuals, correlation; line fitting by least squares regression; outliers in linear regression; Inference for linear regression; Multiple regression; Model selection; Logistic regression, Nearest Neighbor Classification – Knn; Naïve Bayes Classification – Bayesian methods, Bayes algorithm; Classification using decision trees and learners **(9 Hours)**
- Unsupervised Clustering - K-means clustering; Evaluation of model performance – Confusion matrices, sensitivity, specificity, kappa statistics, precision, recall, F-measure, ROC curve etc.; Methods of cross-validation, Bootstrapping; Meta-learning through ensemble approach – Bagging, boosting, Random Forests strategies. **(7 Hours)**
- Classifier performance measurement metrics – Training & Testing strategies – Resubstitution, Hold-out, Cross validation, Bootstrap ; Confusion matrix, Performance measures – Accuracy, Error rate, Sensitivity, Specificity, Precision, Recall, F-Measure, Receiver Operating Characteristics curves **(4 Hours)**

List of Tutorials:

1. Data Visualization
2. Distances and Projections
3. Singular Value Decomposition
4. Principal Component Analysis
5. Optimization
6. Normal & Binomial Distribution
7. Hypothesis Testing
8. ANOVA test
9. Linear Regression
10. Logistic Regression
11. Nearest Neighbor Classification

12. Decision Trees based classification
13. Naive Bayes classification
14. Clustering
15. Evaluation of model performance
16. Bagging & Boosting approaches

List of Practicals: (Any Six)

1. Data visualization
2. Unconstrained Optimization
3. Hypothesis Testing
4. Linear regression
5. Logistic Regression
6. Nearest Neighbor classification
7. Naive Bayes classification
8. Clustering
9. Classifier performance using Confusion matrix and other attributes
10. Cross Validation methods

List of Course Projects:

1. Movie recommendation system
2. Customer Segmentation using Machine Learning
3. Sentiment analysis
4. Uber Data analysis
5. Loan prediction
6. HVAC needs forecasting
7. Customer relationship management
8. Clinical decision support systems
9. Development of machine learning solutions using available data sets (multiple projects)
10. Fraud detection

List of Course Seminar Topics:

1. Data wrangling
2. Predictive modeling
3. Data analytics in life science (multiple topics)
4. Ensemble modeling techniques
5. Text pre-processing
6. Feature scaling for machine learning
7. Multivariate normal distribution applications
8. Distance metrics and their applications
9. Visualization techniques such as Chernoff's faces
10. Tree based algorithms
11. Ridge regression
12. LASSO

List of Course Group Discussion Topics:

1. PCA and ICA
2. Hierarchical and nonhierarchical systems
3. Linear - Non linear regression
4. Parametric-non parametric estimation
5. Overfitting and underfitting in the context of classification
6. Linear and Quadratic discriminant analysis
7. Regression v/s classification
8. Classifier performance measures
9. Supervised and unsupervised learning
10. Various clustering approaches
11. Classifiers and classifier combinations
12. Balancing errors in hypothesis testing
13. Standard sampling practices for a successful survey for reliable sample data

List of Home Assignments:

Case Study: A very large number of resources are available for data generated out of case study. Unique Home assignments will be set up for all groups

Surveys: Principles of surveying will be implemented by groups to demonstrate use of data science principles in home assignments

Assessment Scheme:

Mid Semester Examination - 10 Marks
Presentation - 15 Marks
Laboratory - 10 Marks
Course Project - 10 Marks
Home Assignment - 10 Marks

Group Discussion - 15 Marks
End Semester Examination - 10 Marks
Comprehensive Viva Voce - 20 Marks

Text Books: ***(As per IEEE format)***

1. ‘A Beginner’s Guide to R’ – Zuur, Leno, Meesters; Springer, 2009
2. ‘Introduction to Data Science’ – Igual, Segui; Springer, 2017
3. ‘Mathematics for Machine Learning’ – Diesenroth, Faisal, Ong; Cambridge University Press, 2017
4. ‘Machine Learning with R’ – Lantz, Packt Publishing, 2018

Reference Books: ***(As per IEEE format)***

1. ‘Elements of Statistical Learning’ - Hastie, Tibshirani, Friedman; Springer; 2011
2. ‘Data Science from Scratch’ - Grus; Google Books; 2015
3. ‘The art of Data Science’ - Matsui, Peng; 2016
4. ‘Machine Learning for absolute beginners’ - Theobald; Google Books; 2017

Moocs Links and additional reading material: www.nptelvideos.in

1. https://www.edx.org/course/machine-learning-fundamentals-2
2. https://www.edx.org/course/foundations-of-data-analysis-part-1-statistics-usi
3. https://www.coursera.org/learn/statistical-inference/home/welcome
4. https://www.coursera.org/learn/data-scientists-tools/home/welcome

Course Outcomes:

Upon completion of the course, student will be able to –

1. Apply Data processing & data visualization techniques - 3
2. Implement dimensionality reduction & optimization techniques for enhancing data suitability - 5
3. Perform Descriptive and Inferential statistical analysis for building reliable predictions - 4
4. Implement Supervised algorithms for classification and prediction - 4
5. Implement Unsupervised classification algorithms - 3
6. Evaluate the performance metrics of supervised and unsupervised algorithms - 2
7. Demonstrate complete Data Science life cycle with case studies - 4

Future Courses Mapping:

1. Deep Learning
2. Reinforcement Learning

3. DBMS
4. Big Data
5. Data Mining
6. Information Retrieval
7. Recommendation Systems
8. Cloud Computing – AWS
9. IOT
10. Artificial Intelligence
11. Pattern Recognition
12. Natural Language Processing
13. Computer Vision
14. Machine Vision
15. Fault Diagnosis
16. Optimization
17. Bioinformatics
18. Computational Biology
19. Econometrics
20. Supply Chain
21. Ergonomics
22. Operations Research
23. Nano-informatics

Job Mapping:

Job opportunities that one can get after learning this course

1. Data Scientist
2. Data Analyst
3. AI Engineer
4. Data Architect.
5. Data Engineer.
6. Statistician.
7. Database Administrator.
8. Business Analyst
9. Business Intelligence Developer
10. Infrastructure Architect
11. Enterprise Architect
12. Machine Learning Engineering
13. Machine Learning Scientist