MANGALORE UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE

CSS 454: DATA SCIENCE		
Hours/Week: 4 **Credits : 4**		**I.A. Marks: 30** **Exams. Marks: 70**

Course Outcomes:

CO1: Students will develop relevant programming abilities.
CO2: Students will demonstrate proficiency with statistical analysis of data.
CO3: Students will develop the ability to build and assess data-based models.
CO4: Students will execute statistical analyses with professional statistical software.
CO5: Students will demonstrate skill in data management.

	UNIT-I	**12 Hrs.**

Introduction: The Ascendance of Data, What Is Data Science?, Motivating Hypothetical:, Finding Key Connectors, Data Scientists You May Know, Salaries and Experience, Paid Accounts, Topics of Interest, Onward. **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, Enumerate, zip and Argument Unpacking, args and kwargs. **Visualizing Data:** matplotlib, Bar Charts, Line Charts, Scatterplots. **Linear Algebra:** Vectors, Matrices.

	UNIT-II	**12 Hrs.**

Statistics: Describing a Single Set of Data: Central Tendencies, Dispersion. Correlation, Simpson's Paradox, Some Other Correlation Caveats, Correlation and Causation. **Probability:** Dependence and Independence, Conditional Probability, Bayes's Theorem, Random Variables, Continuous Distributions, The Normal Distributions, The Central Limit Theorem. **Hypothesis and Inference:** Statistical Hypothesis Testing, Example: Flipping a Coin, Confidence Intervals, P-hacking, Example: Running an A/B Test, Bayesian Inference. **Gradient Descent:** The Idea Behind Gradient Descent, Estimating the Gradient, Using the Gradient, Choosing the Right Step Size, Putting It All Together, Stochastic Gradient descent. **Getting Data:** stdin and stdout, Reading Files: The Basics of text Files, Delimited Files. Scraping the Web: HTML and the Parsing Thereof, Example: O'Reilly Books About Data. Using APIs: JSON(and XML) Using an Unauthenticated API, Finding APIs Example: Using the Twitter APIs, Getting Credentials. **Working with Data:** Exploring Your Data: Exploring One-Dimensional Data, Two Dimensions, Many Dimensions; Cleaning and Munging, Manipulating Data, Rescaling, Dimensionality Reduction.

	UNIT-III	12 Hrs.

Machine Learning: Modelling, What Is Machine Learning? Over fitting and Under fitting, Correctness, The Bias-Variance Trade-off, Feature Extraction and Selection. **K-NearestNeighbours:** The Model, Example: Favourite Languages, The Curse of Dimensionality. **Naive Bayes:** A Really Dumb Spam Filter, A More Sophisticated Spam Filter, Implementation, Testing Our Model. **Simple Linear Regression:** The Model, Using Gradient Descent, Maximum Likelihood Estimation. **Multiple Regression:** The Model, Further Assumptions of the Least Squares Model, Fitting the Model, Goodness of Fit, Digression: The Bootstrap, Standard Errors f Regression Coefficients, Regularization. **Logistic Regression:** The Problem, The Logistic Function, Applying the Model, Goodness of Fit, Support Vector Machines. **Decision Trees:** What Is a Decision Tree? Entropy, The Entropy of a Partition, Creating a Decision Tree, Putting It All Together, Random Forests.

	UNIT-IV	12 Hrs.

Neural Networks: Perceptrons, Feed-Forward Neural Networks, Backpropagation, Example: Defeating a CAPTCHA. **Clustering:** The Idea, The Model, Example: Meetups, Choosing k, Example: Clustering Colors, Bottom-up Hierarchical Clustering. **Natural Language Processing:** Word Clouds, n-gram Models, Grammars, An Aside: Gibbs Sampling, Topic Modeling. **Network Analysis:** Betweenness Centrality,Eigenvector Centrality:Matrix Multiplication,Centrality;Directed Graphs and PageRank. **Recommender Systems:** Manual Curation, Recommending What's Popular, User-Based Collaborative Filtering, Item-Based Collaborative Filtering. **Database and SQL:** CREATE TABLE and INSERT, UPDATE, DELETE, SELECT, GROUP BY,ORDER BY, JOIN, Subqueries, Indexes, Query Optimization, NoSQL, **MapReduce:** Example: Word Count, Why MapReduce? MapReduce More Generally, Example: Analyzing Status Updates, Example: Matrix Multiplication, An Aside: Combiners.

REFERENCE BOOK:

1. Joel Grus,Data Science from Scratch: First Principles with Python, 1st Edition, ORIELLY Publications, 2015.

2. Rachel Schutt, Cathy O'Neil Doing Data Science: Straight Talk from the Frontline, 3rd Edition, O'Reilly Publication, 2014