Course 6C: DATA SCIENCE

(Skill Enhancement Course (Elective), Credits: 05)

I. Learning Outcomes: Students after successful completion of the course will be able to:

1. Develop relevant programming abilities.
2. Demonstrate proficiency with statistical analysis of data.
3. Develop the ability to build and assess data-based models.
4. Demonstrate skill in data management
5. Apply data science concepts and methods to solve problems in real-world contexts and will communicate these solutions effectively

II. **Syllabus:** (*(Total Hours: 90 including Teaching, Lab, Field training, Unit tests etc.)*

UNIT I (10 hours)

Introduction: The Ascendance of Data, What is Data Science? , Finding key Connectors, Data Scientists You May Know, Salaries and Experience, Paid Accounts, Topics of Interest, Onward.

Python: Getting Python, The Zen of Python, Whitespace Formatting, Modules, Arithmetic, Functions, Strings, Exceptions, Lists, Tuples, Dictionaries, Sets, Control Flow, Truthiness, Sorting, List Comprehensions, Generators and Iterators, Randomness, Object – Orienting Programming, Functional Tools, enumerate, zip and Argument Unpacking, args and kwargs, Welcome to Data Sciencester!

Visualizing Data: matplotlib, Bar charts, Line charts, Scatterplots.

Linear Algebra: Vectors, Matrices

UNIT II (10 hours)

Statistics: Describing a Single Set of Data, 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 Distribution, 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.

UNIT III (10 hours)

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.

Working with Data: Exploring Your Data, Exploring One-Dimensional Data, Two Dimensions Many Dimensions, Cleaning and Munging, Manipulating Data, Rescaling, Dimensionality Reduction.

Machine Learning: Modeling, What Is Machine Learning? Over fitting and under fitting, Correctness, The Bias-Variance Trade-off, Feature Extraction and Selection

UNIT IV (10 hours)

K-Nearest Neighbors: The Model, Example: Favorite 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, Interpreting the Model, Goodness of Fit.

UNIT V (10 hours)

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.

Neural Networks: Perceptron, Feed-Forward Neural Networks And Back propagation, Example: Defeating a CAPTCHA.

Clustering: The Idea, The Model, Example: Meetups , Choosing k, Example: Clustering Colors, Bottom-up Hierarchical Clustering.

III. References

1. Data Science from Scratch by Joel Grus O'Reilly Media
2. Wes McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython", O'Reilly, 2nd Edition, 2018.
3. Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data", O'Reilly, 2017.
4. Web resources:
    a. https://www.edx.org/course/analyzing-data-with-python
    b. http://math.ecnu.edu.cn/~lfzhou/seminar/[Joel_Grus]_Data_Science_from_Scratch_First_Princ.pdf
5. 9. Other web sources suggested by the teacher concerned and the college librarian including reading material.

IV. Co-Curricular Activities:

a) **Mandatory:** *(Training of students by teacher in field related skills: (lab:10 + field: 05):*

1. **For Teacher**: Field related training of students by the teacher in laboratory/field for not less than 15 hours on identifying, analyzing and presenting the data and then to predict the future instances.

2. **For Student**: Students shall (individually) search online and visit any of the agencies like Statistical cell, weather forecasting centers, pollution control boards, manufacturing industries, agriculture departments, etc. to observe the manual process going on to collect the data, maintain the data, present the data and to predict the data for future instances and submit a hand-written Fieldwork/Project work/Project work/Project work/Project work Report not exceeding 10 pages.

3. Max marks for Fieldwork/Project work/Project work/Project work/Project work/Project work Report: 05.
4. Suggested Format for Fieldwork/Project work/Project work/Project work/Project work: *Title page, student details, index page, details of place visited, observations, findings and acknowledgements*.
5. Unit tests (IE).

b) Suggested Co-Curricular Activities

1. Training of students by related industrial experts.
2. Assignments
3. Seminars, Group discussions, Quiz, Debates etc. (on related topics).
4. Presentation by students in related topics.