# Introduction to
# MCAO 302: DATA SCIENCE USING PYTHON[4-1-0]
# About Course

- **Course Objectives:** The objective of this course is to analyze the data statistically and discover valuable insights from it. The course gives hands-on practice on predictive and descriptive modeling of the preprocessed data. In addition, the student also learns to apply mining association rules from the transactional data and mining text from the document will also be covered during the course.
- **Course Learning Outcomes** : At the end of the course, the student will be able to

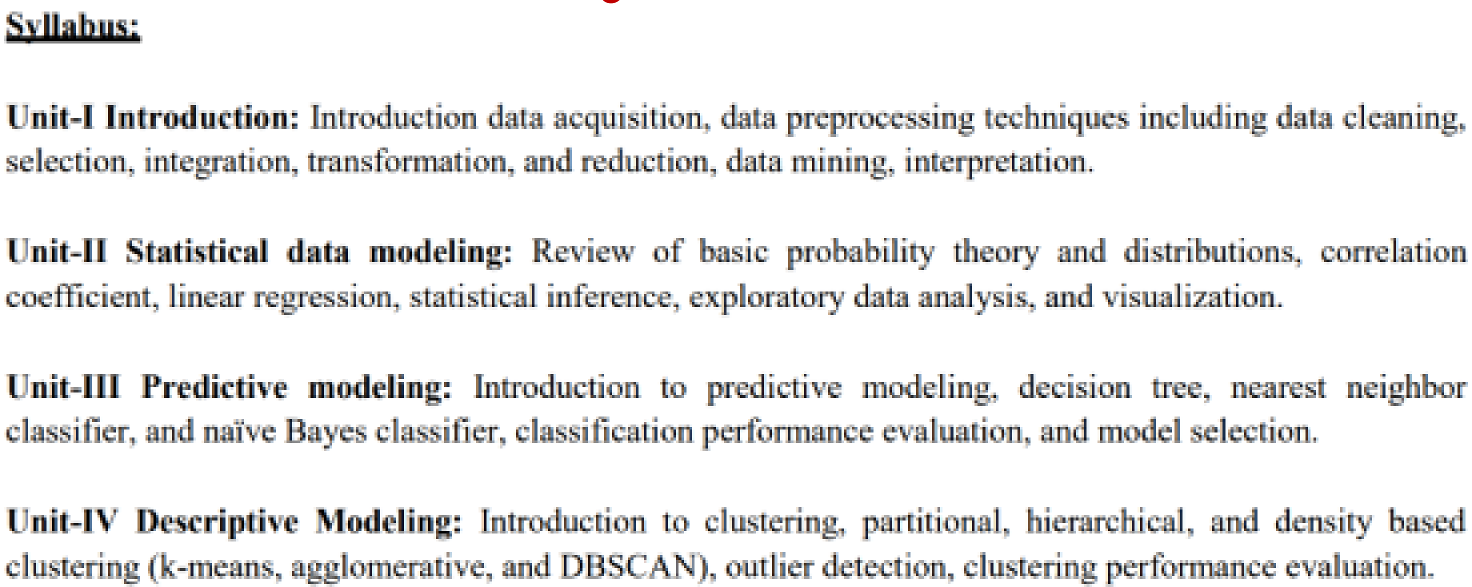
***CO1:*** *demonstrate proficiency with statistical analysis of data.*

***CO2:*** *develop the ability to build and assess data-based models.*

***CO3:*** *execute statistical analyses and interpret outcomes.*

***CO4:*** *apply data science concepts and methods to solve problems in real-world contexts and will communicate these solutions effectively.*

# Syllabus



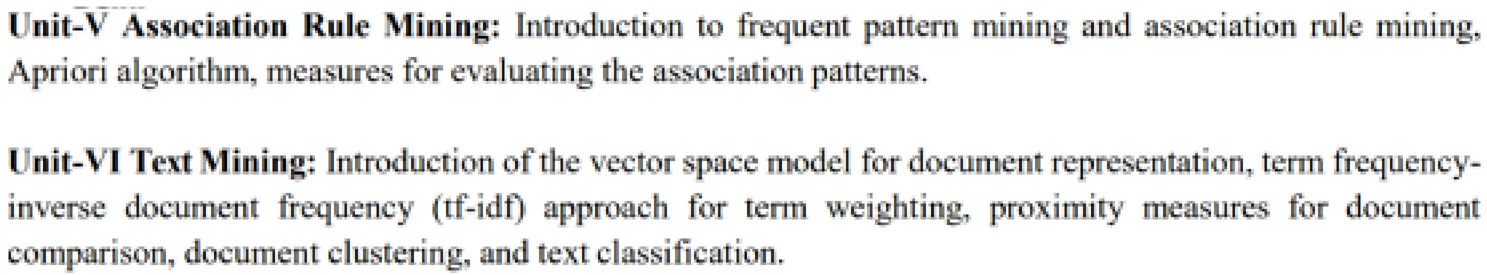
**Syllabus:**

**Unit-I Introduction:** Introduction data acquisition, data preprocessing techniques including data cleaning, selection, integration, transformation, and reduction, data mining, interpretation.

**Unit-II Statistical data modeling:** Review of basic probability theory and distributions, correlation coefficient, linear regression, statistical inference, exploratory data analysis, and visualization.

**Unit-III Predictive modeling:** Introduction to predictive modeling, decision tree, nearest neighbor classifier, and naïve Bayes classifier, classification performance evaluation, and model selection.

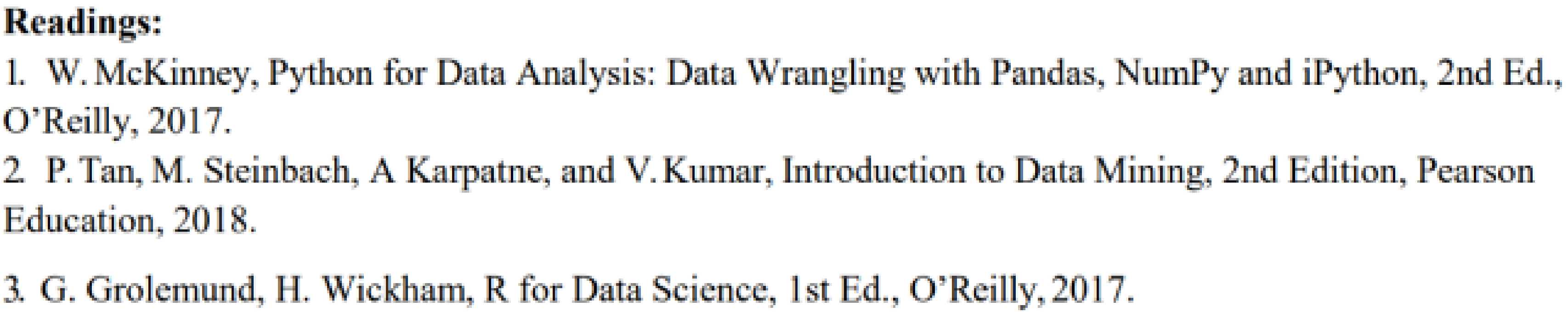
**Unit-IV Descriptive Modeling:** Introduction to clustering, partitional, hierarchical, and density based clustering (k-means, agglomerative, and DBSCAN), outlier detection, clustering performance evaluation.



**Unit-V Association Rule Mining:** Introduction to frequent pattern mining and association rule mining, Apriori algorithm, measures for evaluating the association patterns.

**Unit-VI Text Mining:** Introduction of the vector space model for document representation, term frequency-inverse document frequency (tf-idf) approach for term weighting, proximity measures for document comparison, document clustering, and text classification.

# Readings



**Readings:**

1. W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy and iPython, 2nd Ed., O'Reilly, 2017.
2. P. Tan, M. Steinbach, A Karpatne, and V. Kumar, Introduction to Data Mining, 2nd Edition, Pearson Education, 2018.
3. G. Grolemund, H. Wickham, R for Data Science, 1st Ed., O'Reilly, 2017.

# Grading Scheme

- Max. Marks: 100

- End-term Examination: 70 marks
- Internal Assessment: 30 marks
  - Surprise Tests and Scheduled Minor: 15-20 marks
    - **One** Minor (10-12)
    - Minimum **TWO** Surprise Tests (5-8)
  - Assignment: 10-15 marks
    - Mini-projects and Presentations

# Grading Scheme

- Max. Marks: 100
  - End-term Examination: 70 marks

- Internal Assessment: 30 marks
    - Surprise Tests and Scheduled Minor: 15-20 marks
        - **One** Minor (10-12)
        - Minimum **TWO** Surprise Tests (5-8)
    - Assignment: 10-15 marks
        - Mini-projects and Presentations

“Teachers can **open** the **door**, but you must *enter* it yourself.”

—Chinese proverb

Good Luck!