PUCIT

INTRODUCTION TO DATA SCIENCE

Subject Code: STDS101

Total Hours: 30 Credits: 3

Course Learning Objectives (CLO)

The objective of this course is to make students learn how to use tools for acquiring, cleaning, analyzing, exploring, and visualizing data; making data-driven inferences and decisions; and effectively communicating results. These will be accomplished through course activities on the following data science topics:

UNIT 1: Maths and Programming behind Data Science [06 hours]

Introduction to Pandas and Matplotlib, Probability

Descriptive statistics, Distributions, Functions and random variables,

Hypothesis testing, P-value, statistical inference.

UNIT 2: Data Preprocessing and Exploratory Analysis [06 hours]

Crisp DM, Numpy, Scipy, Matrix Manipulation and data representation as matrices,

Data Transformation: Categorical to dummy, One hot encoding, Data Scaling,

Data Visualization, Dimensionality reduction, including principal component analysis

UNIT 3: Introduction to Machine Learning [06 hours]

Regression: Linear and Logistic

Naive Bayes Classifiers, Introduction to (Decision Tree, Kmeans and expectation maximization for clustering, Market Basket Analysis (association rules))

UNIT 4: Evaluation of Models and Model selection [04 hours]

Bias-Variance tradeoff

Overfitting and Underfitting

Cross-validation

MSE, RMSE, MAE, R2, Type 1, Type 2 error

Performance Metrics (Confusion Matrix, Precision, Recall, F1 Score, Specificity, ROC-AUC)

UNIT 5: Model deployment and Introduction to Advanced Topic [08 hours]

Data scraping and data acquisition via APIs and open sources, Model deployment

Introduction to NLP, Text Analysis, Recommendation Systems, Time series analysis, Speech and Imager Data Analysis, Ethics of big data.

Course Outcomes: On completion of this course, students are able to:

- Transform and clean messy datasets.
- Apply exploratory tools such as clustering and visualization tools to analyze data.
- Perform linear regression analysis to real-life problems.
- Use methods such as logistic regression, nearest neighbors, decision trees, support vector machines, and neural networks to build a classifier.
- Acquire data through web-scraping and data APIs.
- Apply dimensionality reduction tools such as principal component analysis.
- Evaluate outcomes and make decisions based on data.
- Effectively communicate results.

SKILL BASED EXERCISE (SBE):

Note: - These Projects/activities are only indicative; the faculty member can innovate

Assignments/ Mini Projects on: -

- Linear Regression Implementation (Programming Exercise 1)
- Logistic Regression Implementation (Programming Exercise 2)
- Model Evaluation and Selection Coursework
- Bayesian Analysis Coursework
- Conduct hypothesis testing to determine if the average income of males and females in a given population is significantly different. Use the income dataset from Kaggle or any other available source.

- Build a probability model to predict the likelihood of a customer defaulting on their credit card payment. Use the credit card default dataset from UCI Machine Learning Repository or any other available source.
- Implement principal component analysis (PCA) to reduce the dimensions of the iris dataset. Visualize the data using scatter plots.
- Medical Insurance (We have already have it ask Divya)
- Apply data transformation techniques (categorical to dummy, one hot encoding, data scaling) to the titanic dataset. Build a logistic regression model to predict the survival of passengers.
- Build a linear regression model to predict the house prices using the Boston Housing dataset from Scikit-learn.
- Implement the k-means clustering algorithm to group the customers based on their purchase behavior. Use the retail dataset from UCI Machine Learning Repository or any other available source. Visualize the results using scatter plots.
- Evaluate the performance of different regression models (Linear Regression, Lasso Regression, Ridge Regression) on the housing prices dataset. Use metrics such as mean absolute percentage error and root mean square error.

Textbooks and References:

There is no required textbook for the class. However, students may find it useful to consult the following textbooks for reference.

1. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython

   2nd edition, Wes McKinney, O'Reilly Media (2017)

2. **Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**

   Aurélien Géron, O'Reilly Media (2017)

Other Relevant Books

3. **Data Science from Scratch: First Principles with Python**, Joel Grus, O'Reilly Media (2015)

   Code

4. **Doing Data Science: Straight Talk from the Frontline**

Cathy O'Neil, Rachel Schutt, O'Reilly Media (2013)

5. **Learning the Pandas Library: Python Tools for Data Munging, Analysis, and Visualization**

   Matt Harrison, CreateSpace Independent Publishing Platform (2016)

6. **Mining of Massive Datasets**

   Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman, 2nd ed., Cambridge University Press (2014)

   Download from book website

7. **Data Mining: Concepts and Techniques**

   Jiawei Han, Micheline Kamber, and Jian Pei 3rd ed., Morgan Kaufmann (2011)

8. **An Introduction to Statistical Learning**

   Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani Springer Texts in Statistics (2015)