DATA MINING LAB

Course Code	23CS3551	Year	III	Semester	I
Course Category	PCC	Branch	CSE	Course Type	Practical
Credits	1.5	L-T-P	0-0-3	Prerequisites	Data Base Management Systems, Python Programming
Continuous Internal Evaluation :	30	Semester End Evaluation:	70	Total Marks:	100

Course Outcomes		
Upon successful completion of the course, the student will be able to:		
CO1	Apply various preprocessing techniques and Machine Learning methods on different datasets for a given problem.	L3
CO2	Implement various experiments in Jupyter Notebook Environment and Colab.	L3
CO3	Develop an effective report based on various learning methods implemented.	L3
CO4	Apply technical knowledge for a given scenario and express with an effective oral communication	L3
CO5	Analyze the outputs and visualizations generated for different datasets.	L4

Syllabus

Exp No.	CONTENTS	Mapped CO
1	Explore differentTools: Jupyter Notebook, PyTorch, TensorFlow, Google Colab, Kaggle. Explore the different datasets: Kaggle, UCI Machine Learning Repository.	**CO1,CO2,CO3 , CO4, CO5**
2	**Apply** essential data preprocessing techniques to clean and prepare a given dataset (handling missing values, normalization, encoding categorical variables), and **compute** descriptive statistics (mean, median, mode, standard deviation) to summarize the data distribution and identify potential outliers or biases.	**CO1,CO2,CO3 , CO4, CO5**
3	**Apply** the **K-Nearest Neighbors (KNN)** algorithm for both classification and regression problems. **Determine** the optimal number of neighbors (K) using **cross-validation** and **evaluate** using accuracy and **error metrics**.	**CO1,CO2,CO3 , CO4, CO5**
4	Implement the Decision Tree algorithm for both a classification problem. Perform parameter tuning (e.g., max depth, min samples split) and evaluate using accuracy, precision, recall, F1-score (for classification)	**CO1,CO2,CO3 , CO4, CO5**
5	**Apply** the **Naïve Bayes classification algorithm** on textual or categorical datasets and **analyze** its robustness using performance metrics like **confusion matrix, precision, recall**, and **accuracy**.	**CO1,CO2,CO3 , CO4, CO5**
6	**Implement** a **Simple Perceptron** and a **Multi-Layer Perceptron (MLP)** using the MNIST dataset, and **evaluate** their classification performance using **accuracy, precision, recall, and F1-score**.	**CO1,CO2,CO3 , CO4, CO5**

7	**Apply** the **Support Vector Machine (SVM)** algorithm for classification tasks on multiple datasets and **assess** performance using **confusion matrices**, **precision**, **recall**, **F1 score**, and **ROC-AUC curves**. Also, **visualize** the margin and support vectors where applicable.	**CO1,CO2,CO3, CO4, CO5**
8	a. **Implement** a **Simple Linear Regression** model to predict continuous output (e.g., house prices or CO_2 emissions). b. **Apply Logistic Regression** on classification datasets (eg: Breast Cancer, Titanic Survival, or Spam Detection)	**CO1,CO2,CO3, CO4, CO5**
9	**a. Implement the K-Means clustering algorithm on synthetic and real-world datasets (e.g., customer segmentation). Evaluate using Silhouette Score, Davies-Bouldin Index, and visual clustering results.** b. **Implement Hierarchical Agglomerative Clustering** and compare it with K-Means using **dendrograms**, **linkage criteria**, and **cluster validation indices**.	**CO1,CO2,CO3, CO4, CO5**
10	**Implement** the **Expectation-Maximization (EM)** algorithm for clustering on Gaussian Mixture Models. **Assess** clustering performance using **log-likelihood**, **BIC/AIC**, and visual interpretation of clusters on 2D datasets.	**CO1,CO2,CO3, CO4, CO5**
11	**Capstone Project**: **Design and implement** an end-to-end **Machine Learning pipeline** involving problem identification, dataset selection, preprocessing, algorithm selection (classification, regression, or clustering), model building, **parameter tuning**, and **performance evaluation** using appropriate metrics.	**CO1,CO2,CO3, CO4, CO5**

Learning Resources

Text Books

1. Data Mining concepts and Techniques, 3rd edition, Jiawei Han, Michel Kamber, Elsevier, 2011.
2. Machine Learning with Python for Everyone, Mark E.Fenner, First Edition, 2020,Pearson.
3. Machine Learning: A Probabilistic Perspective, Kevin P. Murphy, 2012, MIT Press

Reference Books

1. "Machine Learning:An Algorithmic Perspective", Second Edition,Stephen Marsland, CRC Press
2. "Machine Learning in Action",Peter Harrington, DreamTech
3. "Introduction to Data Mining", Pang-Ning Tan, Michel Stenbach, Vipin Kumar, 7th Edition, 2019.

E-Resources & other digital material

1. https://www.coursera.org/learn/machine-learning
2. https://github.com/atinesh-s/Coursera-Machine-Learning-Stanford