VIGNAN'S Foundation for Science, Technology & Research (Deemed to be UNIVERSITY) -Estd. u/s 3 of UGC Act 1956

Department of Computer Science and Engineering

Year & Section: III & **Academic Year:** 2025-26

MACHINE LEARNING LAB MANUAL

LAB MANUAL – MODULE 1: Supervised Learning & Regression

EXPERIMENT 1: Data Preprocessing and Feature Engineering

Problem Statement: Given the Titanic Dataset, clean and preprocess the data to make it suitable for classification.

Tasks:

- Load the dataset using Pandas.
- Identify and handle missing values using appropriate strategies.
- Encode categorical variables using Label Encoding and One-Hot Encoding.
- Split the dataset into train and test sets (80:20).
- Apply feature scaling (StandardScaler or MinMaxScaler).

Expected Outcome:

- Cleaned dataset ready for classification
- Report shapes of original and processed datasets
- Visualizations of missing data before and after cleaning

EXPERIMENT 2: Regression Modeling for House Price Prediction

Dataset: House Prices - Advanced Regression Techniques

Problem Statement: Predict the house sale price based on features like square footage, number of rooms, location, etc.

Models to Implement:

- Linear Regression (Univariate and Multivariate)
- Polynomial Regression
- Ridge and LASSO Regression

Expected Outcome:

- Print model coefficients and intercepts
- Plot actual vs predicted values for all models
- Report R^2, MAE, and RMSE for each model
- Comment on overfitting/underfitting observations

EXPERIMENT 3: Heart Disease Classification Using Logistic Regression

Dataset: Heart Disease UCI Dataset

Problem Statement: Predict whether a patient is likely to have heart disease.

Models to Implement:

- Logistic Regression

Tasks:

- Train the model using 4 validation strategies:
  - Simple hold-out validation
  - K-fold cross validation
  - Stratified K-fold cross validation
  - Leave-One-Out (LOO) validation
- Evaluate performance with Accuracy, Precision, Recall, F1 Score
- Plot the confusion matrix

Expected Outcome:

- Tabulate and compare validation scores
- Graph performance metrics and confusion matrix

EXPERIMENT 4: Feature Selection on a Breast Cancer Dataset

Dataset: Breast Cancer Wisconsin Dataset

Problem Statement: Select the most informative features to predict cancer diagnosis.

Tasks:

- Apply Filter Method: Chi-Square test
- Apply Wrapper Method: Forward and Backward Selection
- Apply Embedded Method: Elastic Net Regularization
- Evaluate model performance with and without feature selection using Logistic Regression

Expected Outcome:

- List selected features in each method
- Tabulate performance comparison with selected vs full feature sets

---

EXPERIMENT 5: Dimensionality Reduction and Impact Analysis

Dataset: Wine Quality Dataset

Problem Statement: Evaluate the effect of dimensionality reduction on classification accuracy.

Tasks:

- Apply PCA and LDA for reducing dimensionality
- Train logistic regression and decision tree on:
    - Original dataset
    - PCA-reduced dataset
    - LDA-reduced dataset

Expected Outcome:

- Compare and tabulate accuracy, F1 Score across three datasets
- Visualize decision boundaries (2D PCA/LDA plots)

LAB MANUAL – MODULE 2: Classification, Clustering & Ensembles

---

EXPERIMENT 6: Classifiers Comparison

Dataset: Iris Dataset

Problem Statement: Classify iris flower species using various supervised classifiers.

Models to Implement:

- Decision Tree (CART)
- k-Nearest Neighbors (k-NN)
- Fuzzy k-NN
- Multi-layer Perceptron

Expected Outcome:

- Report accuracy, precision, recall, F1 score for all models
- Visualize decision boundaries (use PCA for 2D projection)

EXPERIMENT 7: Ensemble Models for Binary Classification

Dataset: Bank Marketing Dataset

Problem Statement: Predict if a client will subscribe to a term deposit based on campaign data.

Models to Implement:

- Random Forest (Bagging)

- AdaBoost and Gradient Boosting
- XGBoost
- Stacking Classifier

Expected Outcome:

- Compare performance using ROC-AUC
- Print feature importances
- Show how ensemble models outperform base classifiers

EXPERIMENT 8: Clustering and Evaluation

Dataset: Mall Customers Dataset

Problem Statement: Cluster customers based on spending score and income.

Tasks:

- K-Means Clustering (find k using Elbow method)
- Fuzzy C-Means Clustering
- Spectral Clustering
- Self-Organizing Maps (SOM)

Expected Outcome:

- Visualize clusters
- Evaluate clustering using Silhouette Score and Davies–Bouldin Index

---

EXPERIMENT 9: Handling Imbalanced Data

Dataset: Credit Card Fraud Detection Dataset

Problem Statement: Improve classification performance on a highly imbalanced fraud dataset.

Tasks:

- Visualize class imbalance
- Apply Random Over/Under Sampling
- Apply SMOTE and ADASYN
- Train Decision Tree and Logistic Regression on original and balanced data

Expected Outcome:

- Compare F1 score, precision, recall on imbalanced vs balanced data
- Visualize confusion matrices

EXPERIMENT 10: Content-Based Recommendation System

Dataset: MovieLens 100K

Problem Statement: Build a movie recommender based on user ratings.

Tasks:

- Use cosine similarity for content-based filtering
- Generate top-5 recommendations for a given user
- Visualize rating distributions and similarity heatmaps

Expected Outcome:

- Recommend movies based on user profile
- Show similarity matrix and explain recommendation logic

SUBMISSION GUIDELINES:

- Use Jupyter notebooks with markdown cells for explanation
- Include all graphs, metric tables, and observations
- Maintain separate notebook per experiment
- Submit GitHub or ZIP folder link for evaluation