ASSIGNMENT: Practical Machine Learning

Deadline:-28/04/2025 till 11:59 PM

Instructions:-

- **Read the Problem Statements Carefully**
  - Each problem describes a real-world scenario — make sure you fully understand the dataset, the goals, and the business context before starting.
- **Document Your Workflow**
  - For every step (Exploratory Data Analysis, Feature Engineering, Model Selection), write clear markdown or text-based explanations in your Report and make sure the document is in your own words.
- **Research and Use**
  - As this assignment is for your practice and learning too so If you lack the knowledge on any topic make sure to do research and then use it.
- **Focus on Data Quality**
  - Handle missing values, outliers, and inconsistent data carefully before moving to the modeling stage.
- **Use Visualizations Where Necessary**
  - Support your findings with charts and graphs — make your analysis interpretable and insightful.
- **Explain Your Model Choice**
  - For every task, justify why you chose a particular algorithm, preprocessing method, or evaluation metric.
- **Submit a Final Report**
  Include:
  - Problem understanding
  - Preprocessing steps
  - Model comparison and selection
  - Conclusion and recommendations
- **Final Submission Must Include:**
  - Python notebooks or scripts (.ipynb)
  - A clear README & PDF report

Plagiarism is strictly prohibited!

Submit your own original work. Any plagiarized content will result in a **zero**.

Problem 1: Healthcare Patient Readmission Risk Prediction

Scenario: You work for a healthcare provider that wants to reduce hospital readmissions by identifying high-risk patients who might need additional post-discharge care.

Dataset: The "Hospital Readmission" dataset contains 10,000 patient records with fields including:

Tasks:

1. Load the dataset and perform exploratory data analysis:
   - Display first and last 5 rows
   - Check correlation between features
   - Handle missing values appropriately
2. Apply appropriate scaling/normalization to the numeric features
3. Build classification models using:
   - Random Forest
   - Decision Trees
   - SVM
4. For each model, calculate and compare:
   - Accuracy
   - F1-score
   - ROC AUC
   - Log Loss
5. Use K-means clustering to segment patients into risk groups based on their features.
6. Apply PCA to reduce dimensionality and visualize the clusters.
7. Select the best performing model and explain why it's most appropriate for this healthcare scenario.

Dataset: "Diabetes 130-US hospitals for years 1999-2008" dataset”

Link:-
https://www.kaggle.com/datasets/brandao/diabetes?select=diabetic_data.csv

Problem 2: Retail Customer Segmentation and Sales Prediction

Scenario: A retail chain wants to improve its marketing strategy by segmenting customers and predicting future purchases. They need both a customer segmentation model and a sales prediction model that can inform targeted promotional campaigns.

Dataset: "Retail Customer Transactions" with 50,000 records containing:

Tasks:

1. Perform data preprocessing and exploratory analysis:
   - Check for data quality issues
   - Analyze correlations between variables
   - Create meaningful features from the raw data (feature engineering)
2. Customer Segmentation:
   - Apply both K-means and Hierarchical Clustering
   - Compare different distance measures (Euclidean, Manhattan,etc.)
   - Determine the optimal number of clusters using silhouette score or any other you know.
3. For each customer segment, build regression models to predict future purchase amounts:
   - Decision Trees
   - Random Forest
   - Gradient Boosting Machines
   - XGBoost
4. Evaluate regression models using:
   - MAE
   - RMSE
   - R^2 score
5. Apply PCA to reduce dimensionality and improve model performance.
6. Design a model selection and validation strategy that accounts for temporal aspects of customer behavior. [Optional]
7. Implement regularization techniques to prevent overfitting.
8. Create a final report that explains how the retail chain should use both the segmentation and prediction models together for maximizing marketing ROI.

Dataset: "Online Retail" dataset
Link:- https://archive.ics.uci.edu/dataset/352/online+retail

Problem 3: Financial Fraud Detection System

Scenario: A financial institution needs to build a robust fraud detection system that can identify fraudulent transactions while minimizing false positives. The system must handle imbalanced data, work in near real-time, and adapt to new fraud patterns.

Dataset: "Financial Transactions" dataset containing 1 million records with:

Tasks:

1. Data preparation:
    - Handle the severe class imbalance (explore SMOTE, class weights, etc.)
    - Create temporal features to capture behavioral patterns
    - Scale features appropriately
    - Split data maintaining temporal order (time-based validation)
2. Feature engineering and selection:
    - Apply PCA for dimensionality reduction
    - Use Random Projections to handle high-dimensional sparse features
    - Evaluate feature importance
3. Build an ensemble detection system using:
    - Random Forest
    - Gradient Boosting
    - XGBoost
4. Evaluation with emphasis on:
    - Precision-recall tradeoff (optimize for business cost)
    - F1-score and ROC AUC
    - False positive rate at various thresholds
    - Model explainability (for regulatory compliance)
5. Implement a structural risk minimization approach to balance model complexity and generalization.
6. Design a monitoring system to detect concept drift and model degradation over time.

Dataset: "Credit Card Fraud Detection" dataset

Link: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Alternative Dataset: "IEEE-CIS Fraud Detection" dataset

Link: https://www.kaggle.com/competitions/ieee-fraud-detection/data

Problem 4: Healthcare Appointment No-Show Prediction

Scenario: A healthcare clinic is struggling with patient no-shows for scheduled appointments, which disrupts their operations and reduces efficiency. They want to develop a predictive model to identify patients likely to miss appointments so they can implement targeted reminders and interventions.

Tasks:

1. Data exploration and preparation:
   - Analyze the relationship between patient characteristics and no-show rates
   - Handle categorical variables appropriately (encoding)
   - Create relevant features from appointment date/time information
   - Visualize key patterns related to appointment adherence
2. Feature engineering:
   - Calculate the time interval between scheduling and appointment dates
   - Create patient history features (prior no-shows, appointment frequency)
   - Generate day-of-week and time-of-day features
   - Explore interactions between patient demographics and appointment characteristics[Optional]
3. Classification model development:
   - Implement and compare three classifiers:
     - Random Forest
     - Logistic Regression
     - Gradient Boosting
   - Address the class imbalance issue (if present)
   - Generate classification reports and compare F1-scores for all models
4. Patient segmentation:
   - Apply K-means clustering to identify distinct patient attendance patterns
   - Profile each cluster according to demographic and behavioral characteristics

   - Recommend targeted intervention strategies for each segment
5. Model evaluation and improvement:

   - Analyze confusion matrices to understand model errors
   - Apply regularization techniques to prevent overfitting
6. Feature importance analysis:

   - Apply PCA to reduce dimensionality of the feature space
   - Identify the most important predictors of appointment no-shows
   - Create visualizations of feature importance
   - Provide actionable insights for the clinic based on the findings

Dataset: "Medical Appointment No Shows" dataset

Link: https://www.kaggle.com/datasets/joniarroba/noshowappointments

Problem 5: Household Energy Consumption Analysis and Forecasting

Scenario: EnergySmart, a utility company, wants to help residential customers reduce their electricity bills while optimizing grid load balancing. They need to analyze consumption patterns and develop a forecasting system that can predict daily and weekly energy usage.

Tasks:

1. Data preparation and time series analysis:
   - Resample the minute-level data to hourly and daily aggregates
   - Handle missing values appropriately for time series data
   - Create temporal features (hour, day, week, month, season)
   - Visualize and analyze consumption patterns across different time scales
2. Energy usage pattern discovery:
   - Apply hierarchical clustering to identify distinct usage patterns
   - Compare different distance measures (Euclidean, DTW)
   - Determine the optimal number of clusters
   - Characterize each cluster according to time of day/week and sub-metering values
3. Forecasting model development:
   - Create lagged features and rolling statistics for prediction
   - Implement and compare three regression models:
     - Random Forest
     - XGBoost
     - Gradient Boosting Machine
   - Evaluate models using MAE, RMSE, and R^2 metrics
4. Feature engineering and selection:
   - Generate derived features that capture seasonal patterns
   - Apply PCA to reduce dimensionality of the feature space
   - Identify the most predictive variables for energy consumption
5. Anomaly detection:
   - Identify unusual consumption patterns using statistical methods

    - Propose a simple threshold-based approach for detecting anomalies

6. Insights and visualization:

    - Create basic visualizations showing daily and weekly consumption patterns
    - Provide three key recommendations for optimizing energy usage based on the analysis

Dataset: "Household Electric Power Consumption" dataset

Link:
https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption