Machine Learning Final Project for CPDS 2025

Project Overview

For this assignment, you will apply machine learning techniques to solve a real-world problem of your choice. The goal is to demonstrate your understanding of the ML pipeline from data preprocessing to model evaluation and interpretation. You should select a practical application domain such as stock selection, sentiment analysis, healthcare, marketing (customer segmentation, recommendation systems), trend prediction, or any other area that interests you. Your project should showcase the application of at least one of the techniques we have covered in class: neural networks, or clustering algorithms. You can also choose model such as decision trees, support vector machines, k-nearest neighbors or any other ML algorithm suitable to your problem. The emphasis should be on thoughtful application and analysis.

You are required to obtain your dataset from publicly available sources such as Kaggle, Yahoo Finance, UCI Machine Learning Repository, or or any public data. The dataset should be substantial enough to demonstrate meaningful ML analysis (minimum 1000 samples recommended) and should present a genuine challenge that requires careful consideration of feature selection, model choice, and evaluation metrics. You may use any Python libraries including scikit-learn, TensorFlow, pandas, numpy, and matplotlib for your implementation. Your analysis should include ML approache, discussion of why this model perform better than others for your specific problem, and interpretation of results in the context of the real-world application.

I have given some sample projects for you to pick from (next page). You are free to work on any other idea you have.

Deliverables

Submit a 1-2 page technical report (with proper references) that includes: (1) **Problem Introduction**: Clear description of the real-world problem, dataset characteristics, and why ML is appropriate for this task; (2) **Model Selection and Methodology**: Justification for chosen algorithms, preprocessing steps, feature engineering decisions, and evaluation metrics; (3) **Results and Analysis**: Discussion of strengths/limitations, and interpretation of findings in practical context; (4) **Conclusion**: Summary of key insights and potential real-world impact or applications. Additionally, submit your complete Python code (well-commented) and dataset (or clear instructions for accessing it). The project will be evaluated based on problem formulation (25%), technical implementation (30%), analytical depth (25%), and presentation quality (20%).

Deadline

You have complete and submit final project by **1**5 August, 2025 EOD.

Group

You can form a group of two or three for a larger project. For a smaller project you can work alone.

Some project suggestions

1. **Credit Card Fraud Detection Dataset:** Kaggle Credit Card Fraud Detection Dataset **Problem:** Identify fraudulent transactions from legitimate ones in highly imbalanced dataset. **ML Techniques:** Decision trees, SVM with class balancing, neural networks with anomaly detection approaches.

2. **Customer Churn Prediction for Telecom Dataset:** Kaggle Telecom Customer Churn Dataset **Problem:** Predict which customers are likely to cancel their service based on usage patterns and demographics. **ML Techniques:** Decision trees for interpretability, SVM for classification, clustering for customer segmentation.

3. **Medical Diagnosis Classification Dataset:** UCI Heart Disease Dataset, Breast Cancer Wisconsin Dataset **Problem:** Classify patients as having/not having a medical condition based on clinical measurements. **ML Techniques:** SVM for high-dimensional data, decision trees for clinical interpretability, neural networks for complex patterns.

4. **Movie Recommendation System Dataset:** MovieLens Dataset, TMDB Movie Dataset **Problem:** Recommend movies to users based on ratings history and movie features. **ML Techniques:** Clustering for user/item segmentation, neural networks for collaborative filtering, KNN for similarity-based recommendations.

5. **Real Estate Price Prediction Dataset:** Kaggle House Prices Dataset, Zillow Research Data **Problem:** Predict housing prices based on location, features, and market conditions. **ML Techniques:** Decision trees for feature importance, SVM for regression, neural networks for complex non-linear relationships.

6. **Social Media Sentiment Analysis Dataset:** Twitter API, Reddit Dataset, Amazon Product Reviews **Problem:** Classify sentiment of text data as positive, negative, or neutral for brand monitoring or market research. **ML Techniques:** SVM with text features, neural networks for sequence processing, clustering for topic analysis.

7. **Energy Consumption Forecasting Dataset:** UCI Individual Household Electric Power Consumption, Government Energy Data **Problem:** Predict electricity usage patterns for efficient grid management and cost optimization. **ML Techniques:** Neural networks for time series, clustering for usage pattern identification, decision trees for rule extraction.

8. **Employee Attrition Prediction Dataset:** Kaggle HR Analytics Dataset, IBM HR Dataset **Problem:** Predict which employees are likely to leave the company based on performance, satisfaction, and demographic factors. **ML Techniques:** Decision trees for HR interpretability, SVM for classification, clustering for employee segmentation.

9. **Network Intrusion Detection Dataset:** KDD Cup 1999 Dataset, NSL-KDD Dataset **Problem:** Detect malicious network activity and classify different types of cyber attacks. **ML Techniques:** SVM for anomaly detection, decision trees for rule-based classification, neural networks for complex attack patterns.

10. **Agricultural Crop Yield Prediction Dataset:** FAO Agricultural Data, Kaggle Crop Production Dataset **Problem:** Predict crop yields based on weather conditions, soil quality, and farming practices for food security planning. **ML Techniques:** Decision trees for agricultural rule extraction, neural networks for weather pattern analysis, clustering for regional analysis.

11. **E-commerce Product Category Classification Dataset:** Amazon Product Dataset, eBay Product Listings **Problem:** Automatically categorize products based on descriptions, images, and features for improved search and organization. **ML Techniques:** SVM for text classification, neural networks for multi-modal data, KNN for similarity-based categorization.