The University of Texas at Dallas – Department of Computer Science
CS 4375 – Introduction to Machine Learning

Machine Learning Final Project

You are tasked with choosing one of the following project options to explore in-depth. Your goal is to apply machine learning techniques to distinct datasets, experiment with various methods, and compare the performance of different models. You should also analyze the results and present your findings in a report.

Option 1: Standard Supervised Learning on Structured Datasets

Description: Select one or more structured datasets (tabular data) to perform a classification task. You will experiment with multiple machine learning models (e.g., Decision Trees, Random Forests, KNN, Logistic Regression, etc.) and perform hyper-parameter tuning to optimize each model's performance.

Tasks:

- Choose structured datasets of interest.
    - Groups composed of less than 3 students will work with at least 4 databases.
    - Groups composed of 4 students will work with at least 5 databases.
    - Groups composed of 5 students will work with at least 6 databases.
- Exploratory Data Analysis
    - Evaluate the structure of dataset (variable types, missing values, distribution for variables, descriptive statistics, etc.)
- Potentially pre-process the data (e.g., scaling, missing imputation and one-hot encoding).
- Experiment with a minimum of four different machine learning models (use same models across all datasets).
- Apply techniques such as cross-validation and grid search for hyper-parameter tuning.
- Compare models' performances using appropriate metrics (e.g., accuracy, F1 score, AUC).
- Analyze feature importance or model interpretability where possible.

Dataset Ideas:

- **UCI Machine Learning Repository:** You can select datasets like the Titanic survival dataset, or Breast Cancer dataset. https://archive.ics.uci.edu/datasets
- **Kaggle Datasets:** Select any dataset of interest (e.g., Loan Default Prediction, or the Credit Card Fraud dataset). https://www.kaggle.com/datasets?tags=13302-Classification

Option 2: Text Classification

Description: You are provided with a text corpus, and your task is to build a text classification model (e.g., for sentiment analysis or topic classification). You are expected to experiment with various techniques such as *Bag of Words (BOW)* or *Term Frequency-Inverse Document Frequency (TFIDF)*, along with machine learning classifiers (e.g., Naive Bayes, Logistic Regression, Random Forests).

Tasks:

- Choose textual datasets of interest.
    - Groups composed of less than 3 students will work with at least 3 databases.
    - Groups composed of 4 or 5 students will work with at least 4 databases.
- Preprocess the text data: tokenization, stop-word removal, stemming, or lemmatization.
- Use BOW or TFIDF as feature extraction techniques.
- Build and compare models using different classifiers (e.g., Naive Bayes, SVM).
- **Bonus (30 points):** Experiment with deep learning models (e.g., RNN, LSTM, or other gated architecture).

Dataset Ideas:

- **IMDb Reviews:** Sentiment analysis on movie reviews (positive/negative classification). https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
- **Twitter Sentiment Analysis:** Classify reviews based on their sentiment (positive, neutral, negative). https://www.kaggle.com/datasets?search=sentiment+analysis&tags=13204-NLP
- **Spam Detection:** Use email or SMS datasets to classify messages as spam or not spam. **https://archive.ics.uci.edu/dataset/228/sms+spam+collection**
- **Text Classification (News):** Classify news messages according to their subject (e.g., science, religion, politics) https://www.kaggle.com/datasets/crawford/20-newsgroups/data

Option 3: Propose Your Own Project

Description: If you prefer to work on a project of your own choosing, you can propose a machine learning problem to be tackled. Your project must be approved by either the instructor or the teaching assistant before starting. You are free to explore any dataset, problem type (classification, regression, clustering, image processing), or advanced techniques (deep learning, reinforcement learning) based on your interests.

Requirements:

- Submit a project proposal until 04/05 outlining the dataset, problem statement, and the machine learning techniques you plan to use.
- You may also propose bonus points (up to a maximum of 30 points) for exploring extra topics beyond what was covered in class, subject to approval.
- Your project will be evaluated based on its complexity, relevance, and the rigor of the methodology applied.

Project Application: Each group must consist of **no more than 5 students**. By 04/05, **one student from each group** should submit a *kickoff document* via the e-learning platform. This document should include the names of all group members and indicate which project option has been chosen.

If your group selects Option 3 (your own project), the *kickoff document* must also contain a clear project proposal, detailing the problem statement, the dataset you plan to use, and the techniques you will apply. Ensure that your proposal reflects a project of comparable complexity to Options 1 and 2 to increase the chances of approval.

Deliverables for All Projects:

- **A written report that includes an introduction, methodology, brief description about the datasets used in the experiments, results, and analysis/conclusions.**
- **A short presentation (7 minutes max) of your findings.**
- **Your code (clean and well-documented) in a Jupyter notebook or script format.**

Evaluation Criteria:

- **Correctness and rigor of the methodology.**
- **Comparison and interpretation of model performance.**
- **Depth of analysis and understanding of the results.**
- **Quality of report and clarity of explanation.**