CSE 445: Machine Learning

Assignment#1: Submit the program with explanations

Assignment#2: Viva based on submitted assignment

Summer 2025

Name: _____

Student ID: _____

Direction: You can consult any resources such as books, online references, and videos for this assignment, however, you have to properly cite and paraphrase your answers when it is necessary. There will be points for partial attempts. You can upload a typed or handwritten copy of your assignment to the canvas. However, you have to show the original copy of your assignment in case of arising questions about the authenticity. You should upload your assignment to canvas. There will be a few days of grace period for late submission, however, 20% of points will be deducted.

Submission guidelines: All our assignments will be based on 100 points so that we can assign weights to get your points for final grade. Whoever fails to submit within the time period assigned through NSU canvas but submit the assignment within the next two days will be punished by 20% deduction of points (e.g. start with -20 points within 100 points). The late submission will be open until semester through email to TA and CC to the instructor. After the grace period, the deduction will be 30-40%.

Note: Any keyword used in your assignment has to be elaborated first and then the abbreviation can be used. You have to use Interactive Python for solving your problem. Base ten numbers are listed as normal (e.g. 23), binary numbers are prefixed with 0b or format such as XX_2 and hexadecimal numbers are prefixed with 0x or format such as XX_{16}/XX_{HEX}

Problem (100 points): As a Machine Learning (ML) Engineer for a real estate firm in the Greater Boston area, you need to build a predictive model to estimate housing prices using the Boston Housing Dataset. Your goal is to apply data preprocessing, model training, evaluation metrics, and ensemble learning to determine the best-performing model. Use California Housing ML Model link for your reference:
https://colab.research.google.com/drive/1O1r4V8CtDv9phrsYfp8ct9sWR74aArvr?usp=sharing

Tasks:

1. **(15 points) Data Preprocessing and Early Data Analysis (EDA):**
    a. https://scikit-learn.org/1.0/modules/generated/sklearn.datasets.load_boston.html From the link load the dataset, handle missing values, and visualize key features.
    b. Apply feature scaling and one-hot encoding
    c. Comment and provide explanation of each method or function used in the code.
2. **(25 points) Train and Evaluate Regression Models:**
    a. Train Linear Regression, Decision Tree, Random Forest, and Support Vector Machine models.
    b. Compare performance using Mean Squared Error (MSE), Mean Absolute Error (MAE) and R^2 Score.
    c. Tune hyperparameters using GridSearchCV and RandomizedSearchCV
3. **(15 points) Convert Regression to Classification:**
    a. Convert the continuous price variable into categories (e.g., "Low", "Medium", "High" based on percentiles).
    b. Train Logistic Regression, Random Forest, and SVM classifiers.
4. **(15 points) Calculate Classification Metrics:**
    a. Compute Accuracy, Precision, Recall, F1-score for each model.
    b. Plot the Receiver-Operating Characteristics Curve (ROC) and calculate Area Under Curve (AUC) score.
5. **(15 points) Apply Ensemble Learning:**
    a. Implement Bagging, Boosting (e.g. Gradient Boosting, AdaBoost, LightGBM), and Stacking to improve classification accuracy.
    b. Identify the most important features using feature importance scores.
6. **(15 points) Insights & Conclusion:**
    a. Analyze and compare model performance for both regression and classification.
    b. Summarize key takeaways in each of the five steps, including importance and limitations.