CMPS 460 Machine Learning Project – Spring 2024

Due by 11:59pm on Thursday, 2nd May 2024

The course project aim is to apply the end-to-end machine learning process and techniques to a real-world dataset. Then produce a comprehensive technical report that includes:

1. **Project Ideation and Dataset Selection**
    - Brainstorm project ideas and select a realistic dataset that aligns with the project objectives and provides sufficient complexity for ML model training. You may consider various sources such as Kaggle competitions, publicly available datasets, or private data that you have permission to access.
2. **Exploratory Data Analysis (EDA)**
    - Perform exploratory data analysis to understand the structure, distributions, and relationships within the dataset.
    - Produce insightful summary statistics and visualizations to draw relevant insights and identify potential patterns or anomalies.
3. **Data Cleaning and Preprocessing**
    - Address missing values, outliers, and duplicates in the dataset through appropriate techniques such as imputation, trimming, or removal.
    - Investigate relationships and correlations between attributes to inform feature engineering decisions.
    - Split the dataset into training, and testing sets to ensure robust evaluation of model performance.
    - Use techniques such as stratified sampling to maintain class balance in classification tasks.
4. **Model Selection and Training**
    - Design and train **at least three types of ML models**: one traditional ML model (e.g., Naïve Bayes, decision trees, KNN), one ensemble model (e.g., Random Forest, Gradient Boosting), and one deep learning model (e.g., neural networks).
    - Justify the choice of each model based on the dataset characteristics, complexity, and desired performance metrics.
5. **Model Evaluation and Comparison**
    - Evaluate the performance of trained models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, or MSE.
    - Compare the performance of different models to identify strengths, weaknesses, and areas for improvement.
6. **Model Optimization**
    - Enhance the performance of trained models using techniques such as feature ranking/selection, regularization, or hyperparameter tuning to optimize performance.
7. **Documentation and Reporting**
    - Organize and present all project work in a readable, tidy, and clean Jupyter notebook. Include:
        - A summary of the project objectives and motivations.
        - Clear and well-commented code, visualizations, and explanations to facilitate understanding and reproducibility.
        - Document all steps taken throughout the ML process, including data preprocessing, model training, and optimization.
        - Justify choices and provide explanations for decisions made during the project.
        - Highlight interesting findings, interpretation of results and conclusions.

Be meticulous and innovative in your approach! Consider novel techniques to address the project challenges.