Project: Exploring and Predicting with a Dataset of Your Choice

Introduction to Artificial Intelligence and Logic Programming – EECS3401

Prepared by: Dr. Ruba Alomari[1]

Instructions

This is a group project in groups of 2-3 students. Sign up under Final Project Groups on e-class.

The deadline to join a group is September 30th; students without a group will be randomly grouped by this deadline, and switching groups will not be allowed after this date.

Students are encouraged to work in a group with other classmates. However, students who wish to work individually should email the professor by the deadline shown above in order not to be randomly grouped.

This project is worth 20% of your final grade.

Check e-class for all due dates.

Project Overview

Objective: The objective of this project is to give students the opportunity to select a recent dataset, perform exploratory data analysis (EDA), prepare the data for modelling, train and evaluate a variety of machine learning models, and present the results of their analysis.

Tasks:

1. Select a recent (within the last 3 years) dataset from a reputable source (such as Kaggle, UCI Machine Learning Repository, or a government agency). The dataset should be large enough to support meaningful analysis and predictions, and should be accompanied by a clear description of the data and its source.
   You can use https://datasetsearch.research.google.com/ for dataset search.
   Your dataset must be approved by the professor. The dataset sign-up sheet is available on e-class. No two groups can use the same dataset.
2. Frame the problem and look at the big picture.
3. Perform EDA on the dataset to understand the distribution of the data and identify any trends or patterns. Use appropriate visualization techniques to explore the data and communicate your findings. Show 3 graphs of EDA.
4. Prepare the data for modelling by performing any necessary cleaning, encoding, scaling, feature engineering, etc…

[1] This project references steps adapted from Geron's book: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.

5. Train and evaluate three machine learning algorithms on the prepared data. Use appropriate evaluation metrics to compare the performance of the models. Discuss the performance and summarize the results in a performance comparison table.
6. For the best-performing algorithm, show 3 graphs.

Deliverables:

- A **report** (in PDF format) that includes:
  1. Framing the problem and looking at the big picture.
  2. A description of the dataset and 3 graphs of EDA.
  3. Data cleaning and preprocessing.
  4. Training and evaluation of three machine learning algorithms, analyze findings, and compare results.
  5. 3 graphs for the best performing algorithm.
  6. Any limitations you have run into.
  7. Appendix 1: Source code with proper comments and attribution to any code you have reused.
  8. Appendix 2:
     - Link to your dataset.
     - Link to your executed Jupyter notebook on github. The notebook should contain the code for your machine learning models and should show the results. Your code should be properly commented, and you must attribute any code you are using from someone else.

  Note that the report should be no less than 5 pages and no more than 6 pages using single-space 12-point font. The cover page and appendices do not count towards the page requirement.

- **Presentation**: More on the presentation format and duration will be posted when the number of groups is finalized on Sept. 30. The presentation schedule will be published at a later date. The order of the groups presenting will be chosen randomly.

Rubric

Criteria	Mark
Frame the problem and look at the big picture.	5
A description of the dataset and 3 graphs of EDA.	5
Data cleaning and preprocessing.	10
Train and evaluate three machine learning models and compare the performance of the models.	30
3 graphs for the best-performing algorithm.	10
The code in the appendix is clear, properly commented, and attributed when necessary. Results are included in the Jupyter Notebook uploaded to github.	20
Presentation: Clarity, organization, and overall quality of the presentation.	20