LASSONDE SCHOOL OF ENGINEERING | YORK U

Final Project: Exploring and Predicting with a Dataset of Your Choice

Introduction to Artificial Intelligence and Logic Programming – EECS3401

Prepared by: Dr. Ruba Alomari[1]

Winter 2024

Contents

Instructions

This is a group project in groups of 3-4 students. Sign up under Final Project Groups on e-class.

The deadline to join a group is **February 15th**; students without a group will be randomly grouped by this deadline, and switching groups will not be allowed after this date. If randomly grouped, it is your responsibility to reach out to other group members and start communicating and discussing the dataset and workload distribution.

Students are expected to work in a group with other classmates. However, students who may have strong reasons to want to work individually should email the professor before the deadline shown above in order not to be randomly grouped.

This project is worth 25% of your final grade.

Check e-class for all due dates, sign-up sheet, and submission dropboxes.

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

LASSONDE SCHOOL OF ENGINEERING | YORK U

Project Overview

Objective: In this project, your group will select a recent dataset on a topic of interest to you, 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.

The objective of this project is to give you an opportunity to enhance your skills and expertise on a topic you feel passionate about, and is designed to encourage self-directed research.

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. Requesting to change the dataset after the deadline will result in a penalty of -2 marks of the final project grade.
3. Frame the problem and look at the big picture.
4. 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.
5. Prepare the data for modelling by performing any necessary cleaning, encoding, scaling, feature engineering, etc…. Data preprocessing must be done using a preprocessing pipeline.
6. Train and evaluate three machine learning algorithms on the prepared data. Use appropriate evaluation metrics to compare the performance of the models. Discuss and analyze findings and compare results of the three models.
7. For the best-performing algorithm, show 2+ graphs.

Deliverables:

Deliverable # 1: Final Project Report

Submit to eclass 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. Display the preprocessing pipeline in this section.

4. Training and evaluation of three machine learning algorithms. Discuss and analyze findings, and compare results of the three models. Include a performance comparison table.
5. 2+ graphs for the best performing algorithm.
6. Any limitations you have run into.
7. Next steps.
8. **Appendix 1**: Include 2 links in appendix 1
   - Dataset link: Link to your dataset.
   - Github link: Link to your executed Jupyter notebook on your 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. Failure to submit a working github link will result in -20 Marks of your total project mark.
   - Your jupyter notebook should stay accessible on your github until the final grades are posted and the terms ends.
9. **Appendix 2**: Include your source code with proper comments and attribution to any code you have reused.

Note that the report should be 5-6 pages using single-space 12-point font. The cover page and appendices do not count towards the page requirement.

Deliverable # 2: Jupyter Notebook

Submit to eclass your source code notebook **.ipynb** with the code already run and the output results included and saved in your notebook. Note that the dataset in your notebook should be loaded from a public repository. Do not save the dataset to or load it from your local drive.

Failure to submit a working notebook -50 Marks of your total project mark.

Deliverable # 3: Final Presentation

Record a 7-8-minutes video presentation of your project in an **.mp4** format.

- The videos will be available on eclass and will be played and discussed in class during the last 2 weeks.
- The presentation schedule will be posted on eclass, and group numbers will be picked by a random number generator.
- All group members must be present when their videos are being played and ready to answer questions about their project. Any member who is not present will receive a zero for the presentation component.
- You can’t do a live presentation in this class due to time constraints.

Rubric

#	Criteria	Mark
1	**Report –** Frame the problem and look at the big picture.	5
2	**Report –** A description of the dataset and 3+ graphs of EDA.	5
3	**Report –** Data cleaning and preprocessing.	10
4	**Report –** Train and evaluate three machine learning models.	20
5	**Report –** Discuss and analyze findings and compare the results of the three models.	20
6	**Report –** 2+ graphs for the best-performing algorithm.	5
7	**Report –** Overall organization, quality, and clarity.	10
8	**Code –** The code in Appendix 2 is clear, properly commented, and attributed when necessary.	5
9	**Jupyter Notebook –** Code covers all project requirements 1-6, results included, properly commented and attributed, no errors running the code.	50
10	**Github Link –** The Notebook is accessible, and results are included. No errors in the notebook.	20
11	**Presentation –** Quality, clarity, and coverage of the project requirements 1-6.	25
12	**Presentation –** Professionalism and overall quality of the video.	25
	Total	200