Final Project Deliverables
Instructions

Overview of the Final Project

The Final project is an individual assignment in which you are expected to do the following:

1. Select a dataset:
    a. This should be a publicly available dataset that is suitable for supervised machine learning.
    b. The data volume of the dataset should be consistent with cloud machine learning capabilities – approximately 30,000 instances (rows).

2. Formulate a research question:
    a. The question must be answerable by having the model predict either continuous values (regression) or classes (classification).
    b. Questions that are answerable using descriptive statistics do not qualify.

3. Identify relevant published research papers:
    a. Papers should help you identify appropriate supervised machine learning algorithms for use in your model.
    b. Papers should help you identify metrics for deciding if your model produces significant results.

4. Using the cloud resources provided, develop a supervised machine learning model using the techniques presented in the course including:
    a. Data preprocessing
    b. Pipeline building using Jupyter Notebooks
    c. Feature engineering
    d. Model training, testing, and evaluation
    e. Dimensionality reduction and model tuning

Overview of the Final Project Deliverables

The Final Project Deliverables have two parts:

1. A Final Report document
2. A public GitHub repository that contains all code and data for the project

These two items are interrelated in that there is a section in the Final Report in which the GitHub repository is identified and the URL for the repository is provided.

Tools

The Final Project Proposal is a word processing document. You may use any word processor. Please convert your document to PDF before submitting.

Starter Files

While there is no template file provided for the Final Report, we recommend that you begin with your Final Project Proposal document. Some sections of these documents overlap. Please remember to revise any content that appeared in your proposal to make it match the reality of how your project was conducted.

Assignment Details

Revise your initial proposal based on your experimental results. Provide details about the experiments that you conducted. What have you learned from the analysis? What decisions did you make during the entire process? What evidence did you base these decisions on? What phase of the framework did you spend most of your time while building machine learning pipelines? What challenges or obstacles did you encounter while working on your research project? Are there anything you would have done differently if you had time to conduct? Reflect on what you have learned so far. Your final report should include the following sections:

- Introduction
- Literature Review
- Data
- Methodology
- Results
- Discussion & Conclusion
- GitHub Repository
- References

In the *GitHub Repository* section, you must include the link to your public GitHub repository containing the data set and all the Jupyter notebooks used to conduct your experiments. Note that you should also provide a main Jupyter notebook that explains the steps that are required to reproduce your work, e.g., which notebook runs first, which comes next, etc. Make sure to describe how to reproduce the same results you obtained throughout the entire process.

Deliverables

Please submit 1 PDF file to the Canvas submission activity using the file naming conventions described below.

Naming the Submission File

Name your file using the following scheme:

```
surname_givenname_final_project_report.pdf
```

If this were my own submission, I would name the file as follows:

```
trainor_kevin_final_project_report.pdf
```

Due By

Please submit this assignment by the date and time shown in the Weekly Schedule.

Last Revised
2024-09-30