Project 1

Deadline: November 10, 2024, 11:59pm Eastern Time

Description:

In Project 0, you have brainstormed ideas and identified a problem, dataset, and a probable solution. In this project, you are expected to explore and analyze the dataset, and extract insights following a thought process that will allow you to acquire good understanding of the data and the problem. It will also enable you to come up with an effective solution. This project is focused on the exploration and analysis of a dataset and problem, while you will also create a ML model at the end. Provide supporting visualizations with their analysis wherever needed.

1. Show overall descriptive statistics of your dataset; number of data points, number of descriptive features, type of features, your target feature and its type, descriptive features for different target feature values. (10 points)
2. Determine if any features have missing data and what should be done with the missing data. Explain why the decision was made for each feature. If there is no missing data, explain how you would handle missing data and why. Provide supporting visualizations with their analysis. (10 points)
3. Explore your features further in their distributions and plot their bar and box plots. How are individual features distributed? Show outliers for each feature. Do you think any of the outliers may impact your analysis? Why? Provide supporting visualizations with their analysis. (20 points)
4. What data pre-processing techniques do you apply? E.g., encoding features, missing values, scaling, etc. Explain each process and why you use it. (10 points)
5. Analyze distribution of your target variable. Is it balanced or imbalanced? Do you think any of these may cause a problem and why? Provide supporting visualizations with their analysis. (10 points)
6. What kind of ML approaches and algorithms do you choose to use and why? E.g., supervised, regression, classification, binary, multi-class, split rate of data, logistic regression, SVM, decision trees etc. Provide supporting visualizations with their analysis. (10 points)
7. What evaluation metrics you used to evaluate the performance of your model. Discuss the results of your model as to which model performs better and why this would be the case. How would your model perform based on the results? What shortcomings your model has and possible implications? What would you do to improve the results? (15 points)
8. Reflect on your thought process, steps and explain what kind of stages and processes you have gone through to make decisions in each step. For instance, what led you to choose the evaluation metric you use? what motivated your selection of ML algorithms for prediction? why did you choose the preprocessing techniques you used? Provide supporting visualizations with their analysis. (15 points)

You need to submit to iCollege:

- Codebase as (in .py not Jupyter notebook). –comment your code as needed.
- Dataset
- Report (up to three pages). You don’t have to repeat the same information from project 0. You can just refer to project 0 as needed, e.g., (see Project 0).