Author Names: Joshua Nibbe, Ethan Hawk, Ashley Darnell

DATA-151-A

Prof. Goebert

Due: 10/21/21 by class time

DATA-151 Semester Project Documentation for Part 2

- ## Data collection
    - For our data collection, we downloaded the file from the UCI Machine Learning Repository and went through the Student Performance. There were two files for Student Performance which were for Mathematics and Portuguese.
- ## Exploratory data analysis
    - For our exploratory data analysis, the data was read in the file from our computer(s) to the Jupyter Notebook. The main problem when the data was read in was that the columns were separated by semicolons and quotes in the original excel file.
    - The solution for that problem was creating a new excel file where each variable has its own column in the sheet.
    - Correlations
        - Scatterplots
        - Heatplots
        - Histograms

- ## Demonstrated progress

    - For our exploratory data analysis, the data was read in the file from our computer(s) to the Jupyter Notebook. The main problem when the data was read in was that the columns were separated by semicolons and quotes in the original excel file.
    - The solution for that problem was creating a new excel file where each variable has its own column in the sheet.
    - The demonstrated progress consisted of reading in the file to the Jupyter Notebook, calculating correlations, displaying those correlations on scatter plots, heatplots, and histograms.
    - For the scatter plots, the group looked at specific relationships between variables in the heatplot and analyzed what variables most correlated with each other either negatively or positively.
        - From what the group found, there were many interesting correlations that had r-values that were closer to zero than what the group expected.
            - At the bottom right of the heatplot, the final grades for each trimester were closely correlated in the positive direction which means their correlations were closer to one than zero.
            - That makes sense because the final grades for each trimester will impact Portuguese students' performance.
        - For Box-Whisker Plots, our group mainly analyzed the relationships between higher education and final grade, payment and final grade for the third trimester, failures and absences, failures and daily alcohol

consumption, failures and weekly alcohol consumption, and ultimately failures and final grade for the third trimester.

- For scatter plots, our group analyzed the relationship between absences and final grade for the third trimester.

Choices

- We chose the Portuguese class file because it contained more data. Therefore, the file seems to be worth the work.
- We chose to keep our same schedule because it seems to flow somewhat smoothly with each others' daily schedules.

Data reduction methods (if any)

- None

Schedule: The project schedule will go as follows:

- Layer 1: Understanding the Problem
  - Sept 23th - 27th
    - Complete a project proposal/outline. - All team members
- Layer 2: Data Collection, Understanding, and Preparation
  - Sept 28th - Oct 8th
    - Begin data cleaning - All team members
    - Exploratory Analysis - Josh
    - Fit and refine best models - Ashley
    - Enter models on Python - Ethan

        - Oct 8th-21st
            - Create presentation and report on exploratory analysis - All team members
    - Layer 3: Modelling & Evaluation
        - Oct 22nd - Oct 25th
            - Research related papers - All team members
        - Oct 25th-Nov 8th
            - Begin research paper
            - Intro- Ethan
            - Body - Ashley
            - Conclusion - Ethan
        - Nov 8th - Nov 16th
            - Finalize paper & prepare for demonstration - All team members
    - Layer 4: Finalize Model and Deploy
        - Nov 17th- Nov 30th
            - Evaluate feedback of paper and revise - All team members
        - Nov 30th - Dec 16th
            - Work on final model - All team members