Scientific Programming with Python - Final Project

You are tasked with analyzing a data set and presenting your insights from the data. Additionally, you are to achieve a classification model based on the data set you selected. It is recommended that you research your subject matter beyond the data itself to familiarize yourself with the topic. Your report should be divided into five sections as shown below. In each section (excluding intro/summary) add a reference to the relevant code. You may submit either a Powerpoint Presentation/PDF alongside your code, or a single Colab Notebook file with all your code and tasks (with output displayed). Make sure your code is clean, ordered by section and documented, and all output is relevant to the final report. Additionally, you must record a 6-8 minute video presentation explaining your work.

Intro

In this section, give an overview of the data set you received, including but not limited to:

- Subject matter
- Breakdown of features and their types
- Size of data set

Initial Data Analysis

In this section, perform data cleansing and basic data manipulation. Handle missing data, formatting, errors, and any blatant outliers. Explain your decisions. Include in your report:

- Feature statistical analysis
- Summary of data fixes

Include resulting data set in your report as a single CSV file representing the DataFrame after your manipulation.

Exploratory Data Analysis

In this section, analyze the cleaned-up data and present your observations. Include in your report:

- Feature correlation
- Analysis of each feature
- Visualization for relevant and interesting features
- Additional data cleansing performed based on deeper data exploration

Your visualization is dependent on your data, and it is your responsibility to choose the correct graphs and tables to display your data. Some examples:

- If your data has lon/lat coordinate values, you might want to map them
- If your data contains a time series, you may want to display different date-based plots (per month/year/week data)
- If there are clear linear correlations between features, you can use pivot tables to display aggregated data

Of course some visualization is data-independent, such as distribution plots and scatter plots.

Classification Model

You will show the results of 2 classification methods:

1. Gaussian Naïve Bayes: Select the 2 features that yield the best results for a GNB classifier and show a visualization of them as a 2-d plot.
2. Decision Tree: This will be your main result and is divided into 2 parts.
    a. Show a baseline decision tree classification report for your dataset. A baseline classification is one where all features were used, all rows containing NA values were dropped.
    b. Create a decision tree classifier based on your manipulated data set, selecting the most relevant features. Show your final decision tree classification report and model visualization (including tree representation, classification score and feature importance). Did you manage to improve the performance?
        i. Include the image file for the tree representation (even if you are submitting a notebook file).
        ii. Attempt at least one manual stop condition and show how this affects the classification (include tree visualization).

Summary

In this section, give a brief review of your results. Explain any issues you encountered and include insights from your analysis. Explain which classification performance measure best fits your data set.