Data Science Fall 2025

Assignment Bayes Classifier
Due date: Oct. 10th 11:59PM
Please start early working on this assignment.

Learning objectives:

- You can work in teams of two or individually, if in teams of two, indicate by Thursday Oct. 2nd. by the end of the class, indicate in your submission what part each one of you did. If you decide to work in teams of 2, keep in mind to coordinate how to commit/push since you will be modifying the same code. Don't wait until the last minute, otherwise you won't be able to submit.
- Understand how a Bayes Classifier works and why it works in the way it does
- As stated in class, it is important to understand how features (variables) are related to each other
- Understand how to evaluate a classifier with multiple labels

Deliverables
Remember that in Jupyter Notebooks, there are two types of cells, code and markdown. Please use the markdown syntax to format your cells that correspond to your analysis and insights.
https://www.datacamp.com/cheat-sheet/markdown-cheat-sheet-23

0. Create a virtual environment in python for this assignment, but if you create such venv in the same directory of your assignment, make sure that it is added into a .gitignore file on the root of your project, you don't want to include the entire venv as part of your submission, only the code and the dataset. Look into git's documentation to understand how to use the .gitignore.

1. You can download the Jupyter Notebook (JN) indicated in the resources section below and modify to suit what is being asked. If you find some existing code on the Internet, cite it and explain how you modified and adapted it.
2. For this assignment, impute missing values with mean values (use existing methods from scikit-learn or other package for imputing values, don't do it manually)
3. Remove the feature class from the dataset (last column) you will not use it
4. The new label to predict is the feature type (which can take the following values Compact, Large, Midsize, Small, Sporty, Van)
5. Before working with the data, randomly choose 6 observations one for each type, remove them from the dataset and use them to predict their value and evaluate your model to predict observations never seen when training/testing your model. Since the true values are known, evaluate how close they were predicted by the model. Depending on what the results are, discuss what could explain the difference between the predicted and true value.
6. Conduct a preliminary data exploratory analysis on the data, use graphs from libraries such as matplotlib and seaborn (most of this work was already done in your previous assignment). Since you already did a data exploratory analysis, use a summary with a few graphs that provide a summary of the dataset, don't redo the entire data exploratory analysis
7. Build a Bayes Classifier model to label what type of car it is (Compact, Large, Midsize, Small, Sporty, Van), given the other feature values

8. Explain what the Bayes Classifier model learn
9. Evaluate your model using 10-fold Cross validation
10. Prepare a confusion matrix to explain the results, including the metrics associated with its evaluation (accuracy, precision, recall, specificity).
11. Explain in your own words what is the meaning of evaluating your model and the results you obtained in step 10.

Dataset:
The dataset provides specifications for 93 new car models for the 1993 year, among the features are given to evaluate price, mpg ratings, engine size, body size, and features.
https://github.com/computingcelts/f25-ds-examples/blob/main/datasets/cars.csv

How to submit:

- Create a Jupyter Notebook named "bayes_classifier.ipynb" into your hw-2 directory, commit and push to your repo (you can do multiple commits and push) in fact is strongly recommended to do it multiple times and make sure to check on GitHub on the web to ensure your submission went through.
- Make at least 3 push of your code (mandatory) to ensure you are progressing to complete the assignment and also practices good git version control practices

Resources:

And here is the GitHub repo with examples for tons of models, this can help you to organize your JN.
https://github.com/scikit-learn/scikit-learn/tree/main/examples

A good example of Bayes Classifier with sci-kit learn
https://www.datacamp.com/tutorial/naive-bayes-scikit-learn