LEPL1109 - Statistics and Data Sciences
HACKATHON 2 - Classification: Stars, Galaxies and Quasars

Deadline: December 4, 2023

Lastname	**Firstname**	**Noma**
<Lastname 1>	<Firstname 1>	12345678
<Lastname 2>	<Firstname 2>	12345678
<Lastname 3>	<Firstname 3>	12345678
<Lastname 4>	<Firstname 4>	12345678

Please, read carefully the following guidelines:

- Answer in English or French.
- **Do not modify the layout of the document.** Tour answers will be imported into gradescope for correction. Each answer must be contained in the "zone" predefined by the template.
- Clearly cite every source of information (even for pictures!);
- Whenever possible, use the `.pdf` format when you export your images: this usually makes your report look prettier[1];
- Do not forget to also submit your code on Moodle.

Image source [1].

The dataset [2] consists of 100,000 observations of space taken by the SDSS (Sloan Digital Sky Survey). Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.

The project aims at building a ternary classifier for the following 3 classes: star, galaxy or quasar.

[1] This is because `.pdf` is a vector format, meaning that it keeps a perfect description of your image, while `.png` and other standard formats use compression. In other words, this means you can zoom as much as you want on your figure without decreasing image resolution. For simple plots, vector formats can also save a lot of memory space. On the other hand, we recommend using `.png` when you are plotting many data points: large scatter plots, heatmap, etc.

Question 1.1

Describe, briefly, your dataset (size, variables type, missing values, etc.).

Answer to 1.1

Your answer to question 1.1 in this box.

Question 1.2

Based on a study of the features distribution (variance, number of unique values, number of missing values, etc.), can you identify some features that do not provide useful information for the classification task? Explain your analysis and remove those features from the dataset.

Answer to 1.2

Your answer to question 1.2 in this box

Question 1.3

What are the drawbacks (if any) of choosing a small test set (in proportion)? On the contrary, what are the consequences (if any) of a relatively large testing set (in proportion)?

Answer to 1.3

Your answer to question 1.3 in this box

Question 2.1

Are the ternary classes balanced? What are the proportions of data in each class? Briefly, justify your answer and add a visualization.

Answer to 2.1

Your answer to question 2.1 in this box.

Question 2.2

What would be the expected performance of a random classifier on this dataset?

Answer to 2.2

Your answer to question 2.2 in this box

Question 2.3

Compute the correlation matrix of the dataset and plot it. Do you want to discard features based on this observation? Write clearly you decision rule.

Answer to 2.3

Your answer to question 2.3 in this box

Question 2.4

Why do we scale data? Justify properly, whether it is necessary or not for your feature set (X) and which scaler did you use.

Answer to 2.4

Your answer to question 2.3 in this box

Question 3.1

Explain the idea of K-fold cross-validation and why it is useful. How the choice of K (in the cross-validation) impacts the bias and the variance of the scores obtained on the different folds? Choose and justify the number of folds you consider in this project.

Answer to 3.1

Your answer to question 2.1 in this box.

Question 3.2

Explain your methodology of model evaluation. More precisely, explain which hyperparameters you tune and the values you test for each of them. Next, provide the best hyperparameters configuration for each of the three models as well as their CV F1 score.

Answer to 3.2

Your answer to question 3.2 in this box.

Question 3.3

Based on your answers to previous questions, select a final model that you will keep as classifier. Justify.

Answer to 3.3

Your answer to question 3.3 in this box.

Question 3.4

Plot the precision-recall curve for the three methods, one figure for each class. What happens to the precision and recall when the threshold tends to 0? And when it tends to 1? Explain and, if possible, establish a link with Question 2.1. For each class, for each method: what threshold would you use?

Answer to 3.4

Your answer to question 3.4 in this box.

Question 4.1

Use the test set to estimate the precision, recall and F1 score of your final model and validate its performance on unseen data. Observe if the scores are similar to the ones estimated with your cross-validation. Are you satisfied by the performance of your classifier, in view of the task for which it will be used?

Answer to 4.1

Your answer to question 4.1 in this box.

References

[1] Stellar classification dataset - sdss17. https://www.kaggle.com/code/satoru90/stellar-classification-dataset-sdss17. Accessed: 2023-11-24.

[2] Nuno Ramos Carvalho. Sdss galaxy subset, March 2022.