Technical assignment

The objective of this homework is to put into practice the notions covered in the lectures on classification and neural networks. Your submissions should be uploaded on Moodle or shared with Ezekiel on GitHub no later than midnight (23:59) on 28th of October 2022. After this date, no submissions will be accepted.

Dataset

You can find the dataset file on Moodle (assignment section). The extensive dataset description is provided here. In summary, the dataset contains 10,000 observations of space taken by the SDSS. Every observation is described by 17 features and one class column that indicates if the object is a star, a galaxy or a quasar. These objects are defined as follows:

- A galaxy is a gravitationally bound system of stars, stellar remnants, interstellar gas, dust, and dark matter. Galaxies are categorised according to their visual morphology as elliptical, spiral, or irregular. Many galaxies are thought to have supermassive black holes at their active centres.
- A star is a type of astronomical object consisting of a luminous spheroid of plasma held together by its own gravity. The nearest star to Earth is the Sun.
- A quasar, also known as a quasi-stellar object, is an extremely luminous active galactic nucleus (AGN). The power radiated by quasars is enormous. The most powerful quasars have luminosities exceeding 1041 watts, thousands of times greater than an ordinary large galaxy such as the Milky Way.

Task

Your main task consists in building **two** classification models that categorize the observations into galaxies, stars, and quasars. The first model should be a random forest, whereas the second should be a neural network. Your program should contain the following steps:

1. Data preparation: Cleaning, transformation, and scaling.
2. Model creation: Two classes of models are required: a random forest and a neural network.
3. Cross-validation: Tune the hyper-parameters using at least two methods (*e.g.,* random search and grid search) and pick the best models.
4. Model evaluation: Evaluate the two models using the metrics covered in lectures.

5. Overfitting: Check if your model is overfitting and update it accordingly.

Your program should be written in Python and built upon the libraries Scikit-learn and TensorFlow.

Evaluation

Your submission will be evaluated based on the following criteria:

- **Completeness (25%):** The five tasks should be completed for the two models: the random forest and neural network. Besides, the program should be fault-free and readable, *i.e.,* there are comments that explain each step.
- **Performances (50%):** At each step, you will be evaluated based on the *correctness* and *outcomes* of your choices (*e.g.,* the choice of the model architecture and evaluation metrics).
- **Explainability (25%):** After the deadline, you will receive a short quiz to evaluate your understanding of the program and the rationales behind it. This quiz will be performed in the class.