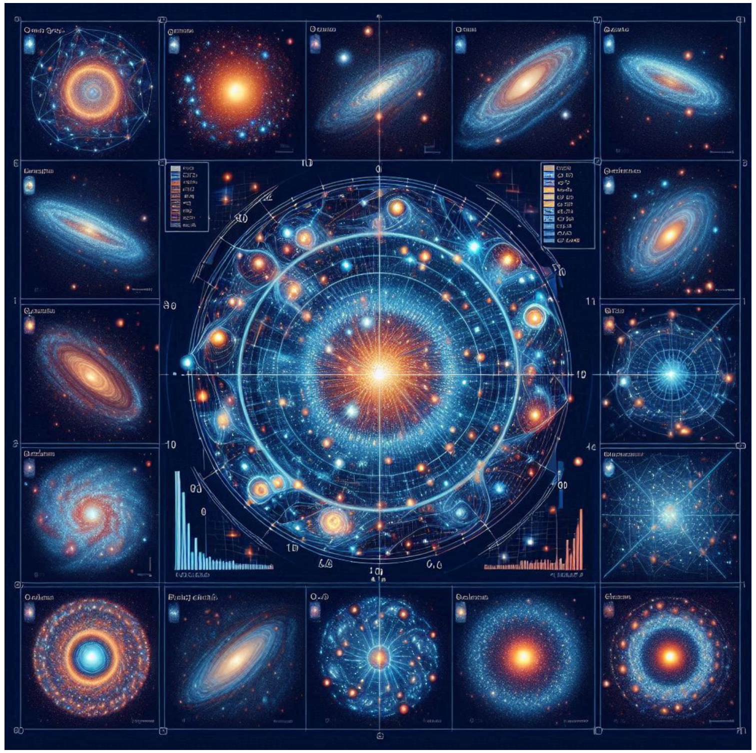
University of Palermo
Department of Physics and Chemistry



MACHINE LEARNING-BASED PHOTOMETRIC CLASSIFICATION OF GALAXIES, QUASARS AND STARS

Donato Cascio

ML4astro2

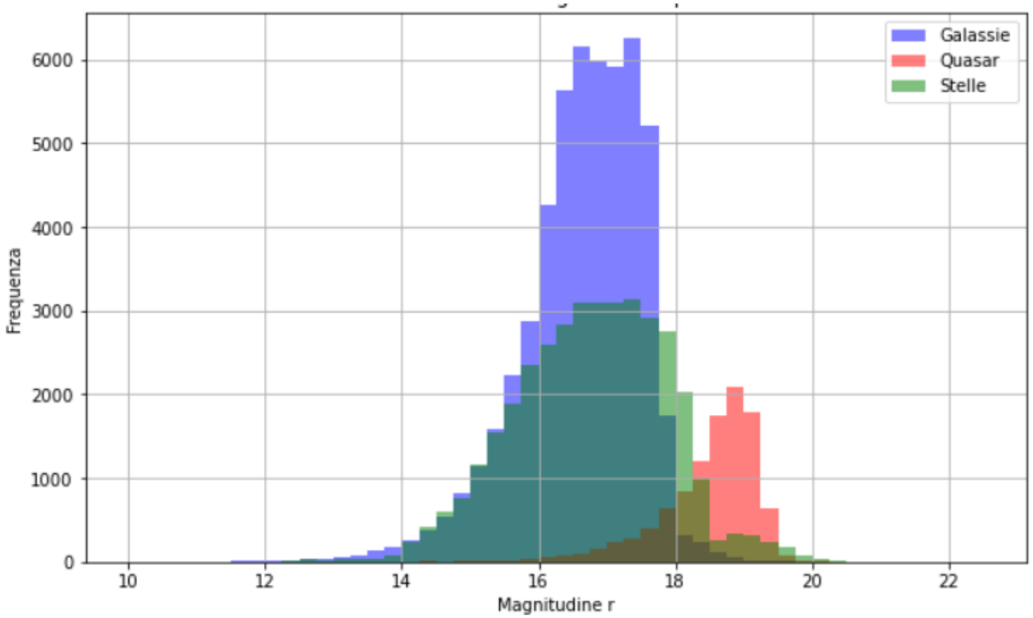
Catania 8-12 July 2024

Introduction to the data

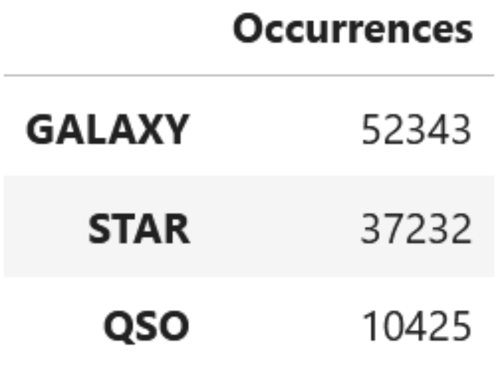
- **Data from Sloan Digital Sky Survey (SDSS)**

Most recent: Data Release 18 (DR18)
https://www.sdss.org/dr18/

- Contains**100,000 patterns** described by **43 features** and related to three classes: **galaxies, stars, and quasars.**



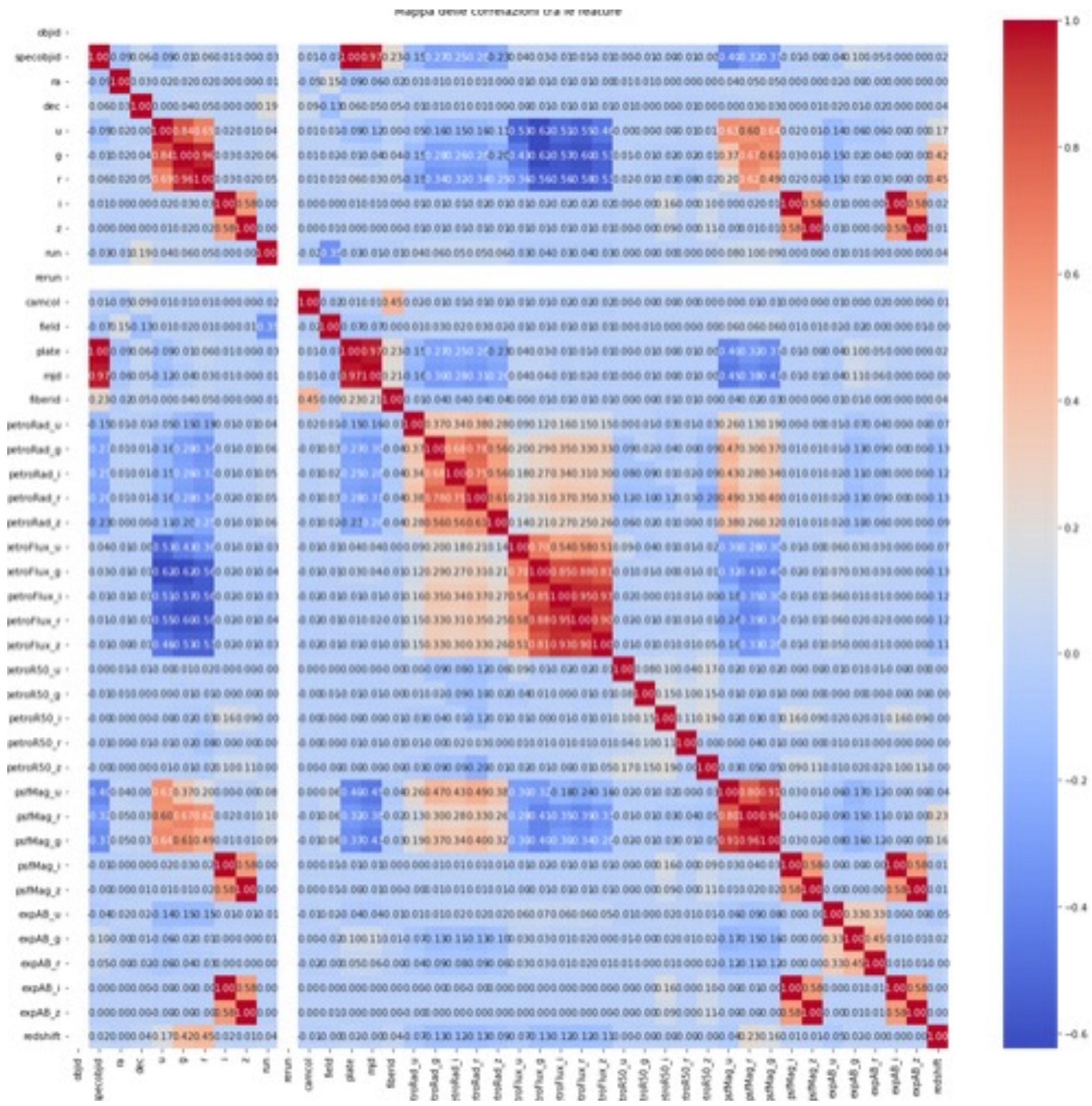
the distributions of the magnitude-r for the three classes



	Occurrences
GALAXY	52343
STAR	37232
QSO	10425

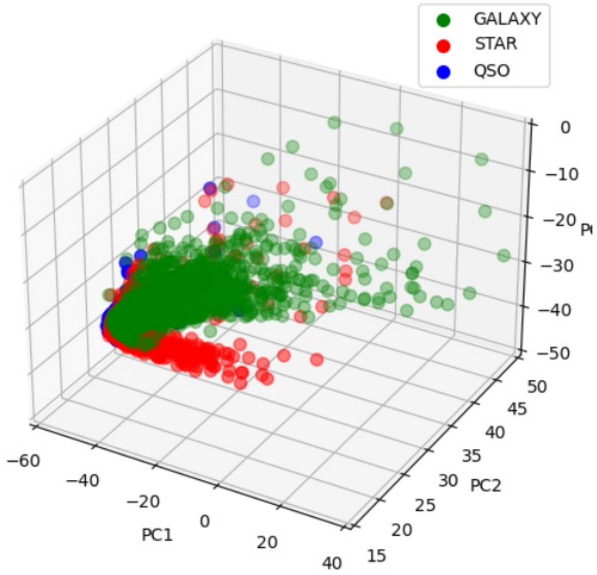
not balanced

Correlation Matrix



- □ We analyzed the data using the correlation matrix, it was observed that one feature, 'rerun', is constant.
- □ Additionally, we have seen that several features are highly correlated (correlation coefficient > 0.95). This allowed a reduction in dimensionality (36 features from the initial 43)

Modeling Workflow



□ To get an idea of the complexity of the separability problem between classes, a plot 3D with Principal Component Analysis (PCA) is shown.

n_components=3

overlap of the data

□ **The implemented model** is therefore composed of the following steps:

1. Data Standardization

2. Correlation Matrix and Elimination of Features with Correlation > 0.95

3. Feature Selection through a Feature Importance Procedure: led to a further dimensionality reduction, resulting in seventeen final features

4. Classification

Classifier Analysis and Parameter Optimization

- □ **Various classifiers were analyzed:**
- Random Forest (RF);
- XGBoost (XGB);
- Support Vector Machine (SVM);
- k-Nearest Neighbors (KNN)

For each of them, the relevant parameters were evaluated using grid-search procedures within a cross-validation framework

The SVM classifier, being a binary classifier, was evaluated using both the one-vs-one and one-vs-all strategies.

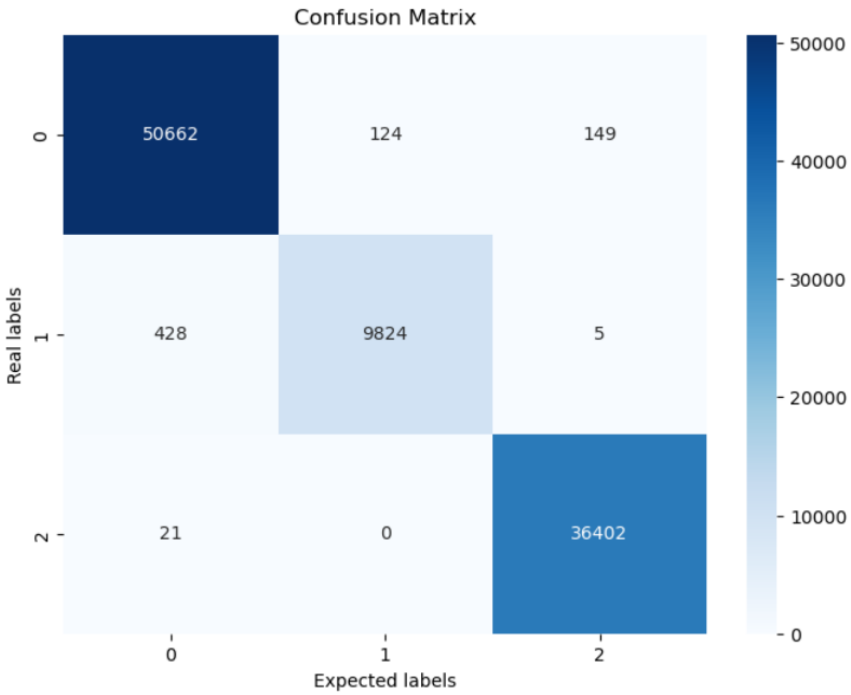
Results

□ The best result was obtained with **Random Forest** classifier

- Accuracy: 0.9926

- Mean Class Accuracy: 0.9779

- F1 score: 0.9925



□ Mean Class Accuracy is affected by the sensitivity on the class (with fewer examples) quasars

□ the parameter configuration that allowed it is as follows

- class_weights ={'STAR': 1.3, 'GALAXY': 1.0, 'QSO': 2.5}

- n_estimators=301, max_samples=0.75

Conclusions and Future Perspectives

- The dimensionality reduction, besides allowing a more computationally intensive analysis, led to an increase the accuracy of about **8**%.
- improving overall performance means enhancing the model's performance for the quasars class
- We plan to use Ensemble methods for further performance improvement.

- **Thank you for your attention**