Stellar Classification

Creation of Logistic Regression and K-Means machine learning classification models, *Alessandro Sciorilli*

The following report presents the methodology and results of applying of Logistic Regression and K-Means machine learning classification models to a Stellar Classification Dataset obtained from Kaggle. These models are designed to correctly predict the class (labeled as 'y') of various celestial objects, specifically galaxies, quasars (QSOs) and stars. The prediction is based on a range of features: alpha, delta, u, g, r, i, z, and redshift. These machine learning classification models have the potential to assist scientists in precisely categorizing celestial objects based on images captured by telescopes.

After importing the required libraries and reading the CSV file into a Pandas dataframe, I conduct data exploration and visualization. This reveals that **'GALAXY'** is the most frequently detected celestial object (with almost **60** thousand observations), followed by **'STAR'** and **'QSO'** (both around **20** thousand observations). Plotting the correlation matrix shows a strong positive correlation between **redshift** and the parameters **'i'** and **'r'** (both at **+49%** and **+43%,** respectively).

After separating the **features (X)** from the label **'class' (y)** and splitting the dataset into a **70/30** ratio for training and testing, I **rescale** the dataset by standardizing the values. Then, I fit the **Logistic Regression model** to the training dataset. During **5-Fold Cross Validation**, I achieve a very solid average **accuracy** of **0.95**. Moving forward, I use the regression model to make predictions on the test dataset. I create a dataframe displaying the real and predicted classes, along with the prediction probabilities.

To evaluate the model's effectiveness, I generate and plot a **confusion matrix**, resulting in a high count of true-positive and true-negative results. The machine learning model's effectiveness is also confirmed by the **classification report**. **Precision** for the **'GALAXY'** class is **0.96**, indicating **96%** of predicted **'GALAXY'** objects were indeed **'GALAXY'**. Similarly, for the **'QSO'** and **'STAR'** classes, precision is **0.95**, meaning **95%** of predicted **'QSO'** objects were **'QSO'** and **95%** of predicted **'STAR'** objects were **'STAR'**. Regarding **F1-scores**, values are **0.96** for **'GALAXY'**, **0.91** for **'QSO'**, and **0.97** for **'STAR'**. The model's **accuracy** across the entire test dataset is **0.96**, meaning **96%** of predictions were correct. Overall, the classification report underlines the model's strong performance, characterized by elevated precision, recall, and F1-scores across most classes.

I cluster my data using the **K-Means** algorithm. Since K-Means is an unsupervised method, I don't need to split my dataset into training and test portions anymore. I use the **Elbow method** to find the best number of clusters for my data, the number (k) where SSE decreases the fastest. I plot the Distortion Score Elbow, and I see that the best number of clusters is **4**.

After that, I run the KMeans clustering and create a plot that shows the clusters and their central points, called **centroids**. As the final step, I calculate **purity scores** using different distance measures. The results show that, no matter which distance measure I use, the clusters produced

by the algorithm are about **59.44%** pure. This means that around **59.44%** of the data points in each cluster belong to the same true class or category. These results is not very robust: even if they might indicate some level of clustering consistency, further analysis is needed to get a better understanding of the results.

In summary, for this particular dataset, opting for the Logistic Regression algorithm is a better choice when predicting the correct class of an observed celestial object. This is due to significantly higher chances of accurate classification compared to applying the k-means clustering algorithm.

Dataset link: https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17