Final Project Initialization

DS - 440: Data Mining

Submitted by: Antonio Cascio & Krystian Confeiteiro
Submitted to: **Dr. Jack**
Embry-Riddle Aeronautical University
Daytona Beach, FL, 32114

Due: **November 03, 2022**

November 03, 2022

1 Our Chosen Dataset

1.1 Initial Thoughts

Given the Astronomy and astrophysics background of our group, we have decided to go with a dataset[1] that focuses on *astrophysical* data. Given that data science is such a large field that spans every field in industry *and* non-industry fields, we believe this dataset is more than appropriate for our final project.

1.2 Preliminary Ideas

For our chosen dataset, we have come up with the following ideas:

(a) Predict the types of stars (stellar classification) given all of the data–which will be used as the *features* for the machine learning model(s)

(b) Predict stellar distances given that our dataset has an adequate amount of surrounding data to calculate said distances as well as training a matching learning model to predict distances, as well

(c) Potentially use this data set in tandem with past research from Antonio and Krystian to make predictions for:

- Automated vetting of light curves (and with neural networks)
- Work on the actual Gaia DR2 data itself where we test the integrity of the data itself which could potentially be published if our findings are valuable

from these ideas we can combine and add to them to finalize our final project. Additionally, we could

(a) Predict the chemical composition of their atmospheres based on observed spectra

(b) Predict stellar ages using metallicity

These ideas work towards answering the questions that can be asked.

2 Data Visualizations

2.1 Galactic Positions on Celestial Sphere

The positions of the stars will be given in galactic coordinates, we could plot their positions over a 3D representation of the celestial sphere, which would be appropriate to display the cluster of stars in the data set.

2.2 Distributions

To visualize the distributions of the dataset, we can use `seaborn`[2] or `matplotlib`[3] and use histograms, empirical cumulative distributions, bi-variate distributions and more. We can use these types of visualizations in tandem with appropriate visualization of machine learning models to accurately visualize the data set *and* the accompanying machine learning models.

2.3 Machine Learning

While we are still deciding which machine learning model to apply, we are researching models like: nearest neighbor models, gaussian process models, Naive-Bayes models, decision tree models, and more. Using visualization techniques demonstrated in our midterm 1 project, we can use libraries like: `TensorBoard`, `Matplotlib`, `Seaborn`, `Plotly`, `Bokeh`, `Yellowbrick`, `SHAP`, `LIME`, and `ALVIS`.

[1] https://www.kaggle.com/datasets/solorzano/783k-gaia-dr2-stars/data
[2] https://seaborn.pydata.org/index.html
[3] https://matplotlib.org/stable/