Proposal: Stellar Classification

Philipp Fößl

April 2025

1 Motivation

I'm intrigued by the cosmos, spending my free time absorbing videos, documentaries, and more about it. So, I've chosen the "Stellar Classification Dataset - SDSS17"[fed] from Kaggle for a project. This dataset contains 100,000 observations with 18 attributes, allowing for a comprehensive analysis.

My main goal is to accurately differentiate between galaxies, quasars, and stars using machine learning techniques. This is crucial in astronomy, as precise classification lays the groundwork for various astronomical studies. By automating this process, I aim to improve the efficiency and reliability of astronomical research, freeing up astronomers to delve deeper into data interpretation and scientific inquiry.

Stellar classification involves utilizing different metrics to distinguish between celestial objects, making it a suitable task for this project.

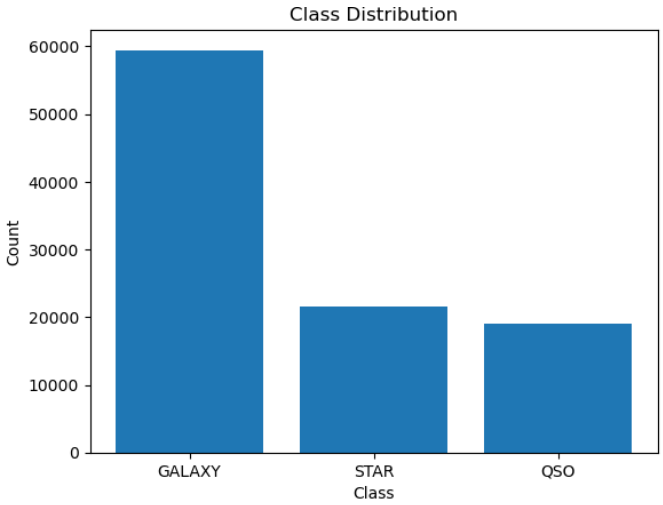

Figure 1: Distribution of stellar classes

2 Learning task

2.1 Training Experience

The data utilized for this project originates from the "Stellar Classification Dataset - SDSS17" on Kaggle. It's derived from the "DR17" dataset from the Sloan Digital Sky Survey (SDSS), which was last updated in January 2021. Since the "DRx" releases are cumulative, this dataset encompasses prior release data as well. Notably, within the astronomy category, this dataset enjoys significant popularity, indicated by its high usability score of 10 on Kaggle and one of the most votes in this category. Numerous notebooks on the site showcase classification accuracies exceeding 90%.

	obj_ID	alpha	delta	u	g	r	i	z	run_ID	rerun_ID	cam_col	field_ID	spec_obj_ID	class	redshift	plate	MJD	fiber_ID
0	1.237661e+18	135.689107	32.494632	23.87882	22.27530	20.39501	19.16573	18.79371	3606	301	2	79	6.543777e+18	GALAXY	0.634794	5812	56354	171
1	1.237665e+18	144.826101	31.274185	24.77759	22.83188	22.58444	21.16812	21.61427	4518	301	5	119	1.176014e+19	GALAXY	0.779136	10445	58158	427
2	1.237661e+18	142.188790	35.582444	25.26307	22.66389	20.60976	19.34857	18.94827	3606	301	2	120	5.152200e+18	GALAXY	0.644195	4576	55592	299
3	1.237663e+18	338.741038	-0.402828	22.13682	23.77656	21.61162	20.50454	19.25010	4192	301	3	214	1.030107e+19	GALAXY	0.932346	9149	58039	775
4	1.237680e+18	345.282593	21.183866	19.43718	17.58028	16.49747	15.97711	15.54461	8102	301	3	137	6.891865e+18	GALAXY	0.116123	6121	56187	842

Figure 2: Samples of the dataset

2.2 Learning task

This project uses the K-Nearest Neighbours (KNN) algorithm to classify stars based on their features. KNN is a simple, effective algorithm for classifying data. I want to use the KNN algorithm to create a model that can accurately categorise stellar objects based on their similarities to other data points.

2.3 Performance measure

In terms of performance evaluation, my target is to attain a precision rate exceeding 90% for the model. Precision, in this context, signifies the accuracy of positive predictions made by the classifier. Achieving such high precision is imperative as it ensures that the model's classifications of stellar objects are overwhelmingly accurate.

3 Related work

On Kaggle.com there are many different notebooks that use that dataset. The most upvoted of them is one that uses a Support Vector Machine Classifier and a Random Forest Classifier.

Stellar Classification - 98.4% Acc 100% AUC[BEY]
This notebook starts by looking at the data, then removes outliers and handles data that is not evenly distributed. It then divides the data into two sets for training and testing. Finally, it uses two different classifiers to analyse the data. The two classifiers used are a Support Vector Machine Classifier and a Random Forest Classifier.

4 Plan

- Data Preparation: Gather and clean the stellar dataset, ensuring it's ready for analysis. Split the data into training and testing sets.
- Feature Selection: Identify key features for classification, focusing on those most indicative of star types.
- Algorithm Implementation: Develop the KNN algorithm, considering distance metrics and neighbor weighting options.
- Evaluation: Assess the performance of the KNN model using standard metrics like precision, recall, and accuracy.
- Fine-tuning: Fine-tune the model by adjusting parameters such as the number of neighbors and distance metrics, optimizing for the best performance.
- Analysis and Refinement: Interpret the model's results, identifying areas for improvement and potential insights into star classification. Iterate on the model design based on these findings to enhance its accuracy and generalization ability.

5 Risk management

Even though I do not suspect that something bad happens that prevents me from finishing my goal with this dataset. If something doesnt work out with this dataset I would try another dataset from the astronomy category called "Star Type Classification / NASA"[Bar]. This dataset is significantly smaller in size but has also many votes and a usability of 10.

References

[Bar]Baris Dincer. Star type classification / nasa. Zugriff am: 03.04.2025.

[BEY] BEYZA NUR NAKKAS¸. Stellar classification - 98.4 Zugriff am: 03.04.2025.

[fed]fedesoriano. Stellar classification dataset - sdss17. Zugriff am: 03.04.2025.