2023 Group 2 Project 4: Demystifying Machine Learning

Mission Meteor

Welcome To Our Presentation

Cheila, Grace, Helen,
Kassem & Rahmi

2023

Let's get to know about stars with the help of machine learning.

2023

Introduction

Earth

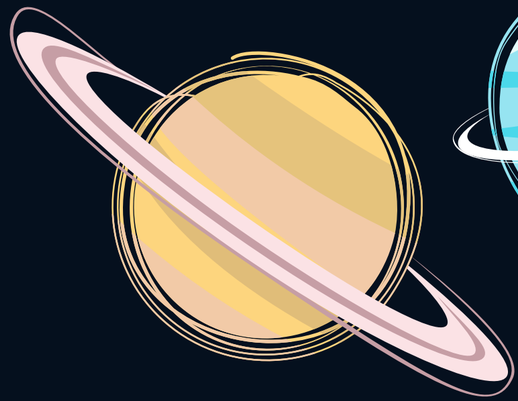
Mercury

Pluto

Neptune

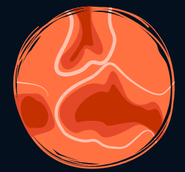
Saturnus

Uranus



Jupiter

Venus



Mars

Why did we pick the stellar dataset?

Why did the Sun go to school?
To get brighter!

2023

AIM!

The aim of this study is to utilise the Morgan-Keenan (MK) classification system, which incorporates the HR classification system, to categorise stars by their chromaticity and size using spectral data. The study will focus on categorising stars into the main Spectral Types using the Absolute Magnitude and B-V Color Index within a specific dataset.

Machine learning & Stars...

... in recent years there have been several studies using machine learning techniques to classify stars on spectra and only one study using a convolutional neural network achieved and accuracy of just over 90%

2023

Dataset & Pre-Processing

01

Dataset -
Kaggle, Raw File,
Clean, Final Dataset

02

ETL -
Extract, Transform
& Load

03

Pre-Processing

2023

Preparing the dataset for lift-off... I mean testing!

Part 1

Convert dtype: Float

Drop NaN

Reset index

Part 2

Calculating amag

$$M = m + 5(log_{10}p + 1)$$

Add amag column

Part 3

B-V colour index

8 categories (0-7)

with a For loop

AMag

2 types: Dwarf / Giant

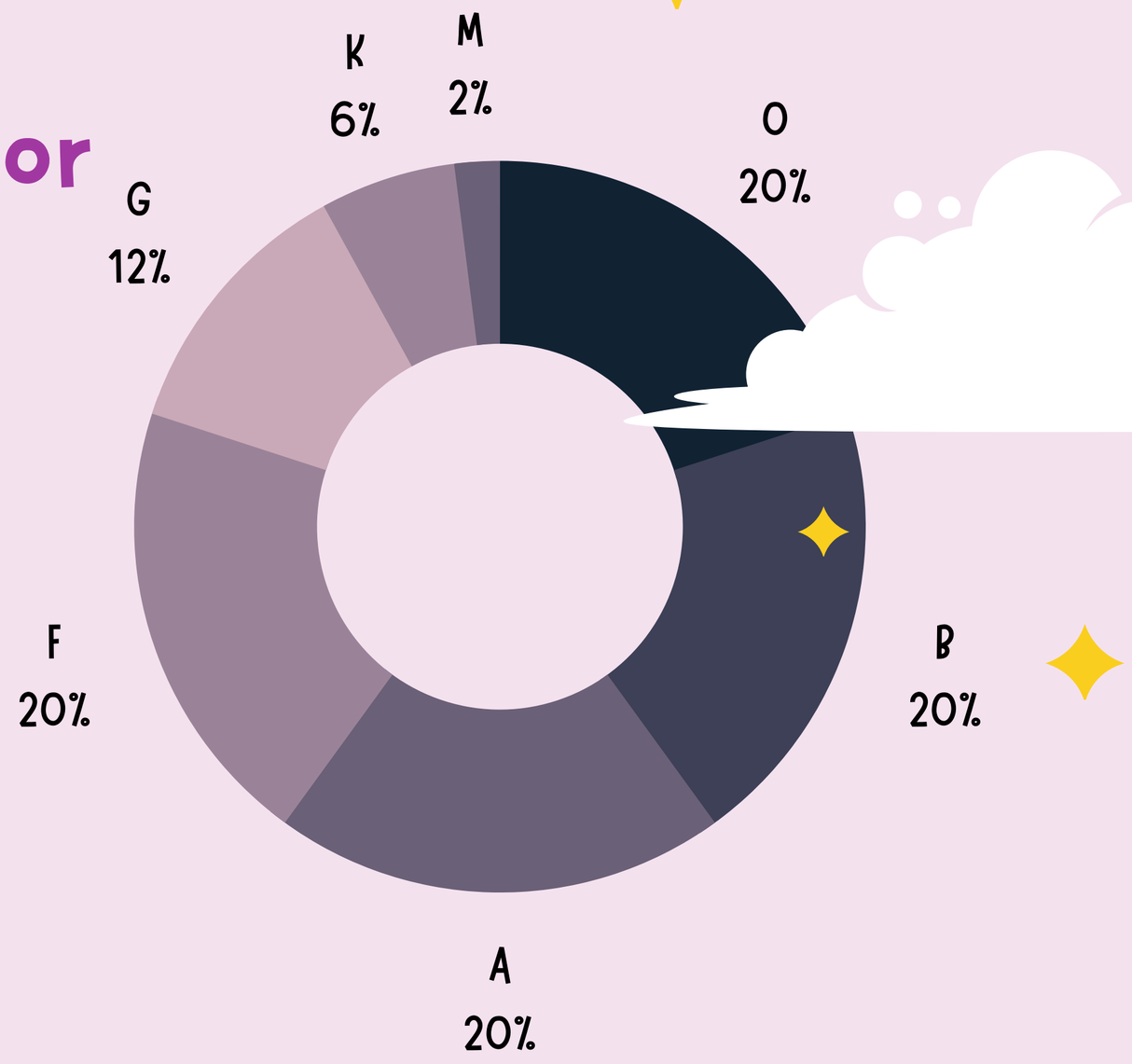
Intuitive For loop

2023

How the dataset splits for each star classification

There 7 main types of classification:

- O (Blue)
- B (Blue)
- A (Blue)
- F (Blue/White)
- G (White/Yellow)
- K (Orange/Red)
- M (Red)



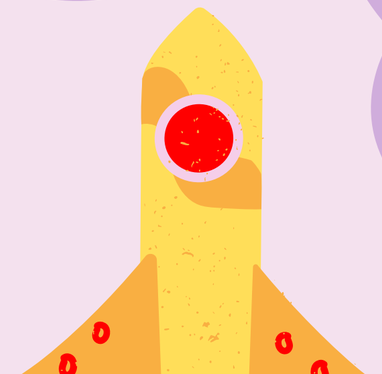
The Brightest Star...

... in the night sky is Sirius, also known as the Dog Star. It is located in the constellation Canis Major and has an apparent magnitude of -1.46, making it over 20 times more luminous than the sun.

2023

VISUALISATIONS

We created visualisations on the cleaned dataset



2023

Visualisations

Matplotlib

Statistical visualisations

Seaborn

Density plot visualisations

Tableau

Dashboard of multiple visualisations on stellar Spectral Types

2023

Tableau Dashboards

Rahmi's Stellar Classifications

Grace's Visual & Absolute Magnitude

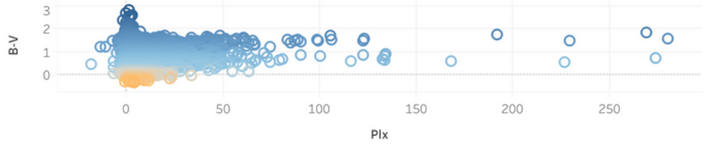
2023

Star Dataset - Summary

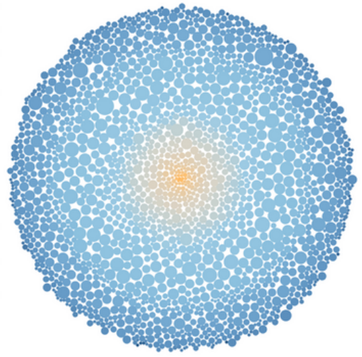
Total Number of Stars	Target Classes	Spectral Types	Stellar Parallax - Plx	
9,999	8	885	7.66	4.96
			avg	med

B-V
-0.359 2.835

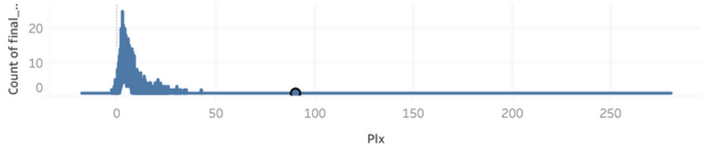
Plx vs B-V

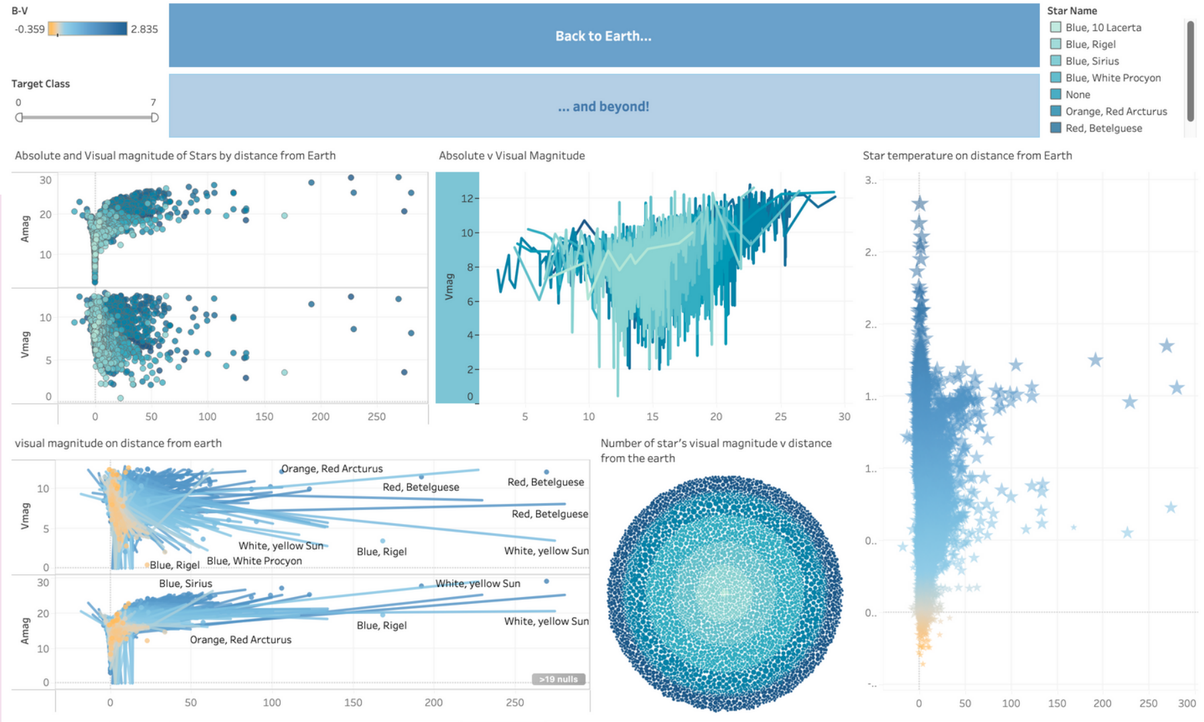


B-V of stars



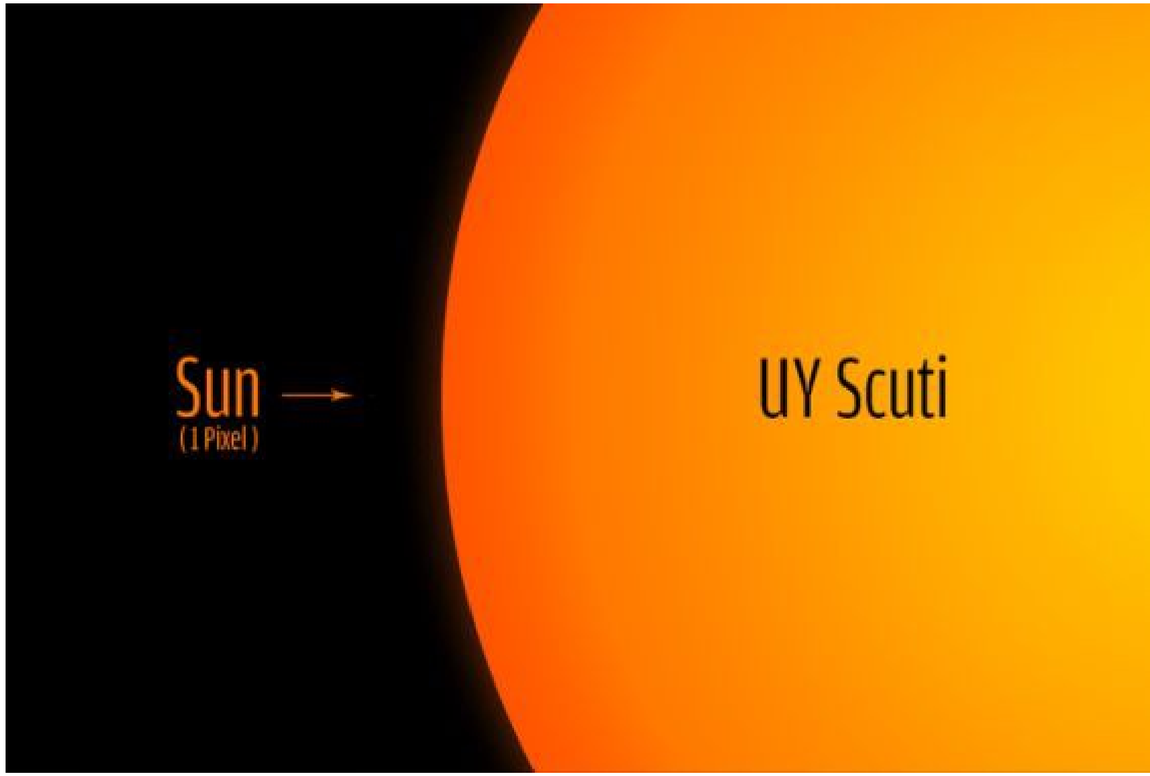
Plx of stars





The Biggest (known) star...

... in terms of size, is UY Scuti, a hypergiant star located in the Scutum constellation. Its size is estimated to be around 1,700 times that of the Sun. However, it is important to note that there may be other stars that are even larger but have not yet been discovered.



2023

MACHINE LEARNING!

- For a classification problem, we used supervised machine learning as it's typically more accurate.
- In supervised learning, labelled data is used, where each data point has a known outcome or "label" that the model is trained to predict based on input features.
- During training, the model learns to identify patterns and relationships between the input features and the corresponding labels. This knowledge is then used to predict the label for new, unseen data.
- We used Random Forest Classification, which is a collection of decision trees that work together to improve classification accuracy and prevent overfitting.
- We also used Support Vector Machines (SVMs), which finds the optimal hyperplane to separate the classes in a high-dimensional feature space.
- Supervised machine learning can handle large amounts of data, identify patterns and relationships that may not be immediately apparent to humans, and make accurate predictions on new data.
- Many common classification problems involve large amounts of labeled data that can be used to train a supervised model.

2023

Data Model Journey: Using B-V

- Began by experimenting with MLP and unsupervised machine learning. After more research we quickly realised this was wrong.
- Initial tests with RFC showed us that we needed more datapoints to train the model, so we introduced feature engineering.
- Realised we hadn't stratified our test and train set, had to fix this! (We thought we cracked it when we got 99% acc, we had not).

```
RFC Training Data : 1.0
RFC Testing Data: 0.9979338842975206
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.72	0.50	0.59	112
2	0.78	0.83	0.80	257
3	0.83	0.86	0.85	491
4	0.74	0.67	0.70	436
5	0.78	0.86	0.82	547
6	0.61	0.51	0.55	87
7	0.00	0.00	0.00	4
cy			0.77	1936
vg	0.56	0.53	0.54	1936
vg	0.77	0.77	0.77	1936

- We tested mulitple different models with different hyperparameters. (MLP, GNB, LR, SVM, RFC, KNN).
- Began to run in depth GridSearchCV to find ideal hyperparameters. (This took a very long time).
- Realised two targetclass groups did not have enough data to pull from so they weren't being predicted at all.
- More problems to solve!

Data Model Journey:

Using B-V Part 2

- We attempted to solve this problem with more feature engineering, this was not successful as we were simply increasing the datapoints proportionally.
- Secondly we attempted to balance the class weights, however this was still returning non-existent F1 scores for 2 Target Classes.
- After more research we used the resampler SMOTETomek from the Imblearn library to try to fix this, whilst it was a partial fix, it lowered our overall accuracy.
- Finally, we decided we needed to upsample, this was deliberated over for a long time as we didn't want to introduce any unnecessary bias, however to create a robust model we were willing to take the risk.
- In our final model we used RFC, GridSearchCV and upsampling was used to achieve higher F1 scores accross the board and also an accuracy of 91%.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.54	0.63	0.58	112
2	0.80	0.74	0.77	257
3	0.86	0.82	0.84	491
4	0.73	0.69	0.71	436
5	0.82	0.75	0.78	547
6	0.50	0.78	0.61	87
7	0.09	0.75	0.16	4
:y			0.74	1936
/g	0.54	0.64	0.56	1936
/g	0.77	0.74	0.75	1936

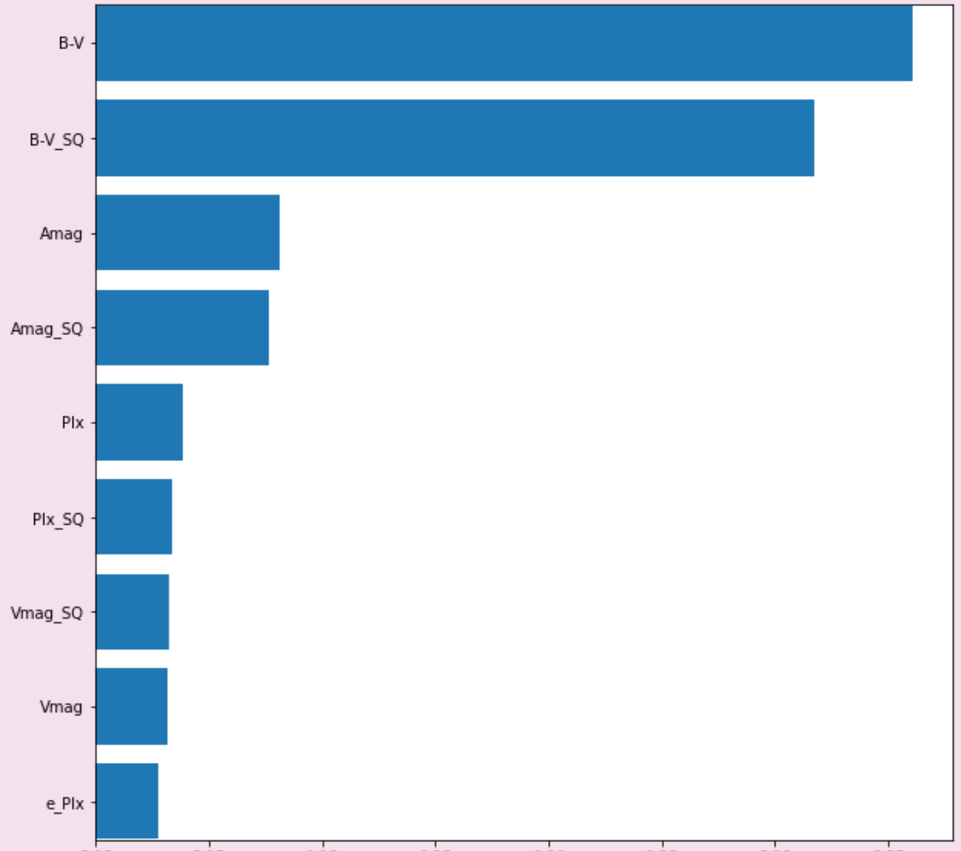
Final Results for B-V Model

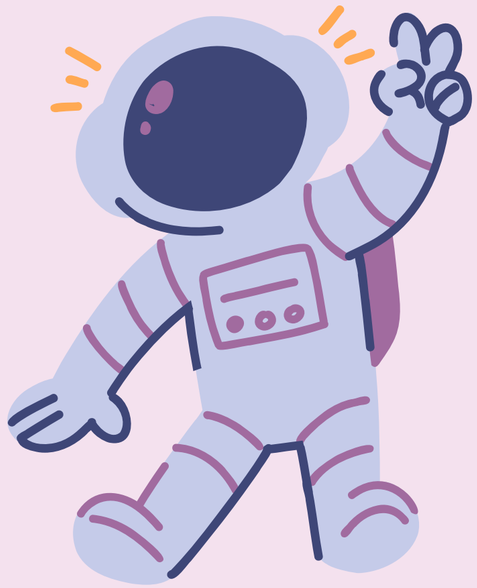
Before

RFC Training Data : 0.874

RFC Testing Data: 0.778

```
predictions:   [3,2,2,2,4,2,5,3,5,2]
actual labels: [3,3,2,1,5,1,5,3,5,2]
```



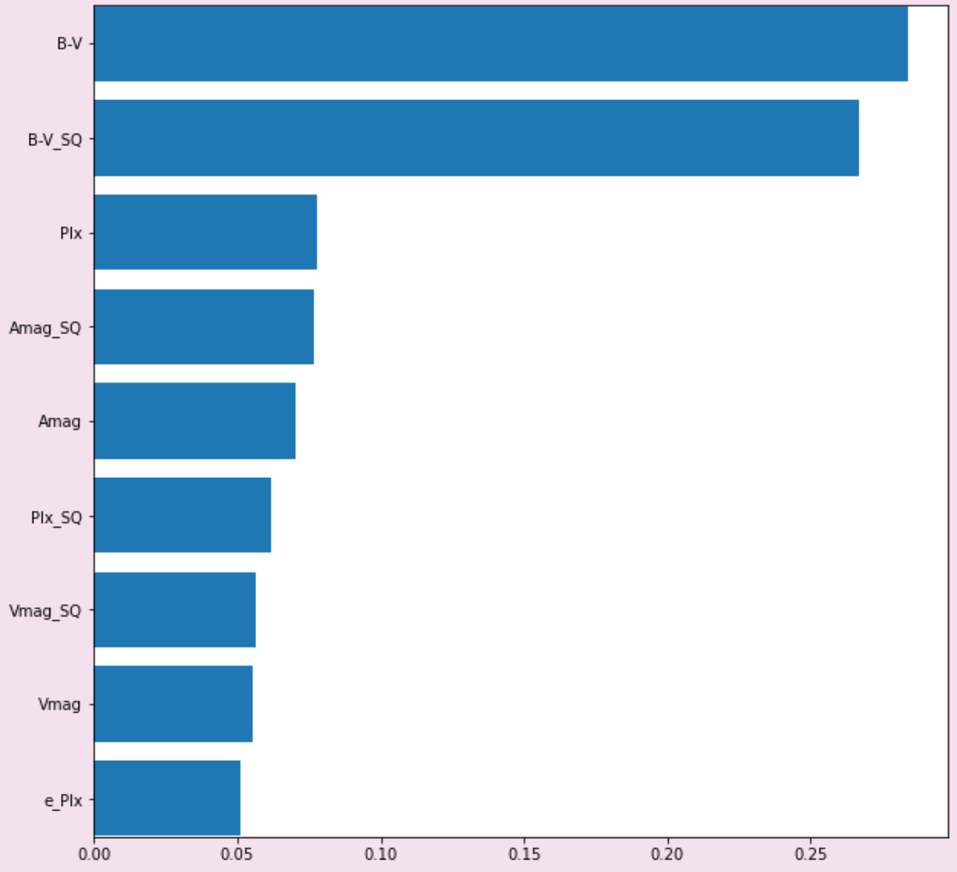


After

RFC Training Data : 0.933

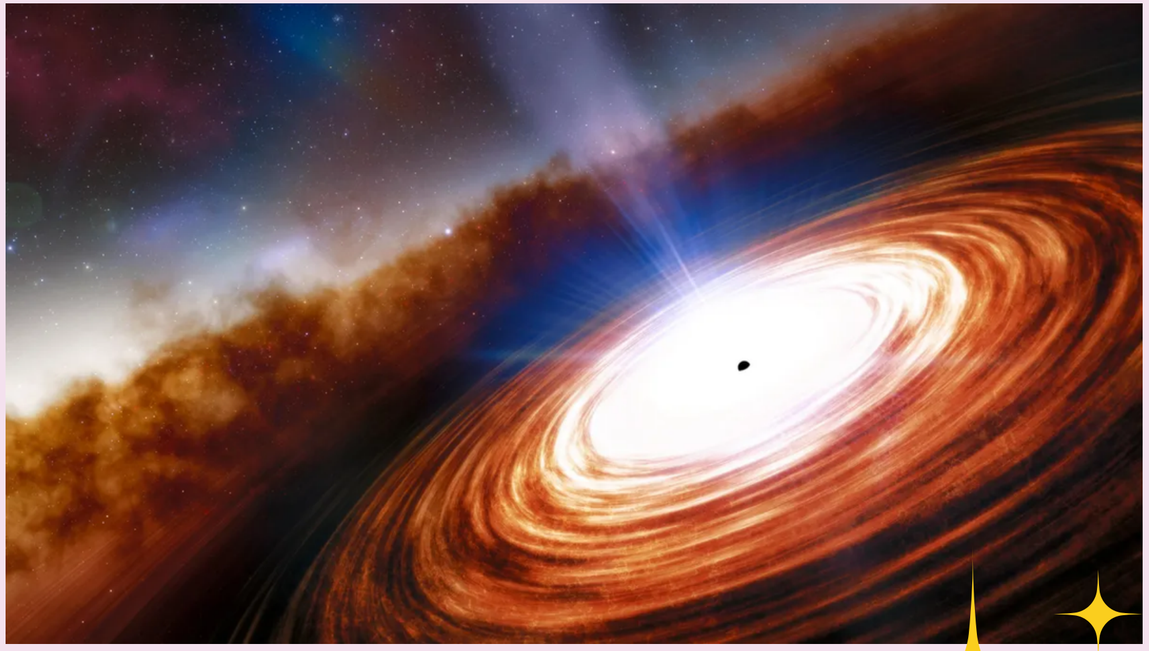
RFC Testing Data: 0.912

```
[1,3,3,1,6,4,5,4,6,4,6,6,7,5,0,7,6,4,6,1,4,0,0,5,4 3,0,6,7
[1,3,3,1,6,4,5,4,6,4,6,6,7,5,0,7,6,4,6,1,4,0,0,5,4,3,0,6,7
```



Did you know...

Most black holes form from the remnants of a large star that dies in a supernova explosion. (Smaller stars become dense neutron stars, which are not massive enough to trap light.)



But they are pretty.

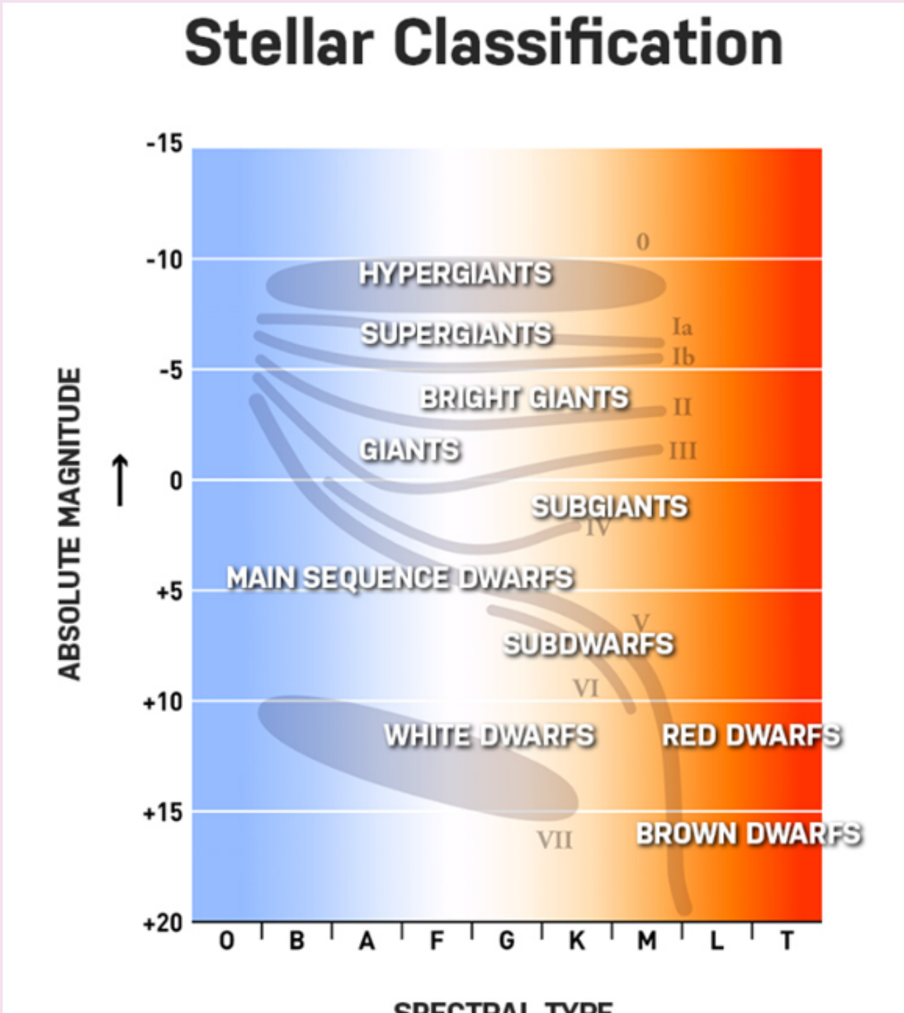
BLACK HOLES WILL SPAGHETTIFY YOU AND EVERYTHING ELSE.

SVM Linear and Non-Linear Classification (Giants and Dwarfs)

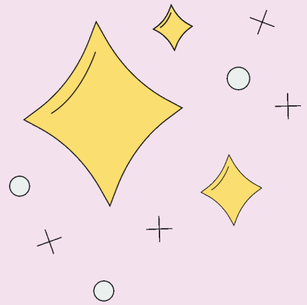
Linear and Non-linear in SVM refer to the types of decision boundaries that can be created by the algorithm, depending on whether the classes are linearly separable or not in the input feature space.

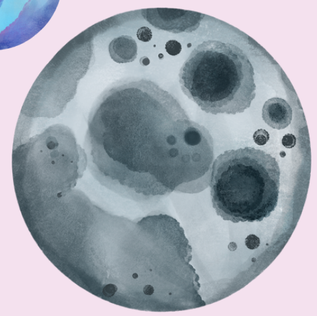
- The target column has been revised for 2 classes:
- •Giants (1): I, II, III, VII
- •Dwarfs (0): IV, V, VI
- •Others (9): Others (dropped)

Stellar Classification Chart (Hertzsprung–Russell diagram)



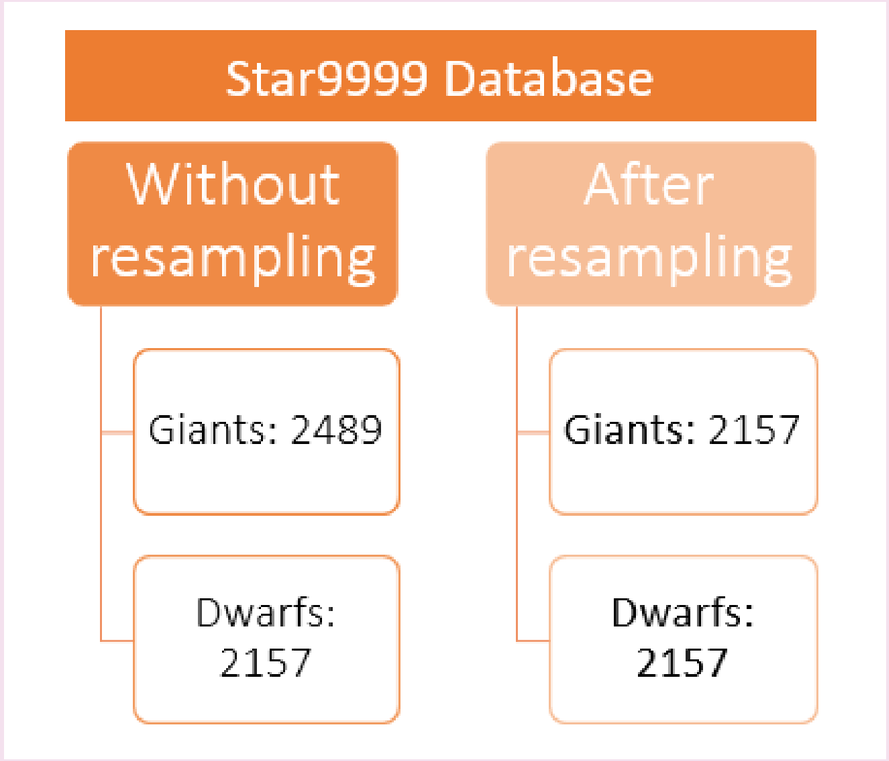
Giants and Dwarfs Database

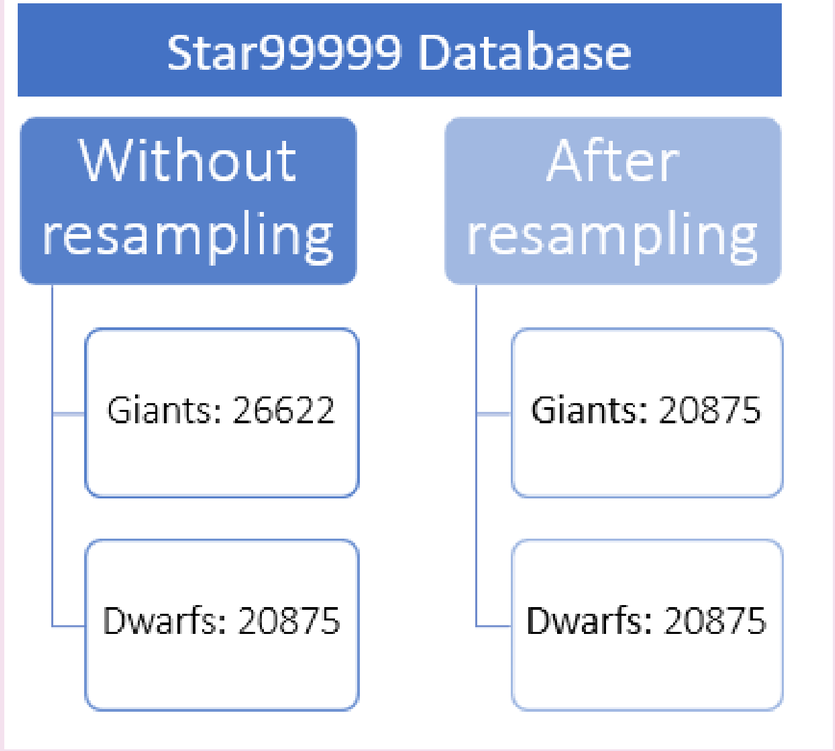




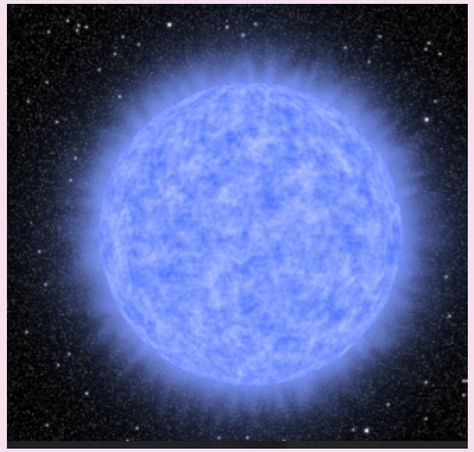
2 Databases used for this analysis:

- Star99999_raw.csv
- Star9999_raw.csv
- To balance the dataset, a resampling technique was used .





SVM – 2 Classes (Giants and Dwarfs)



The Support Vector Machine classifier was set up accordingly for the following scenarios:

- Linear classifier: linear kernel, and random_state=42.
- Non-Linear classifier: RBF kernel, and the additional parameters C=1.0 and gamma = 1.0.

We can observe a slight accuracy improvement for non-linear classifier (RBF kernel) specifically for the 9999 database.

	Test Accuracy - Database 9999		Test Accuracy - Database 99999	
	RBF Kernel	Linear Kernel	RBF Kernel	Linear Kernel
Before Balancing	**0.851**	0.840	0.830	0.817
After Balancing	**0.851**	0.840	0.830	0.819

The oldest (known) star...

... in the universe is SM0313, which is estimated to be around 13.6 billion years old. It is a metal-poor star located in the Milky Way galaxy, and its age was determined by analysing its spectrum to measure its chemical composition and other properties. There may be other stars in the universe that are even older, but SM0313 is currently considered the oldest known star.

2023

Why did the big star refuse to go out with the small star?

Because they knew they would never make a good match – they were simply not on the same scale!

2023

ANALYSIS

What did we learn from our dataset?

2023

Analysis

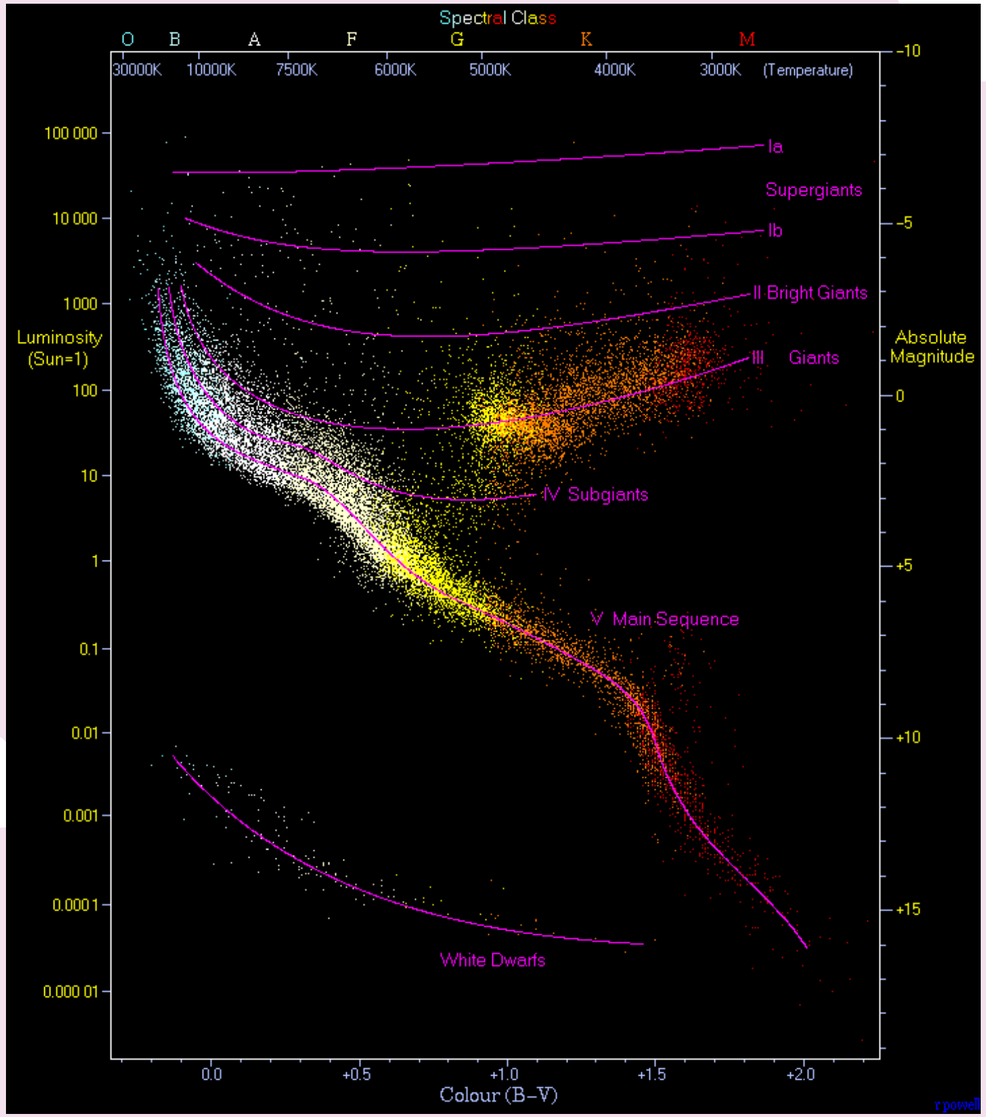
We ran the model optimisation and evaluation process by making iterative changes to the model and the resulting changes in model performance

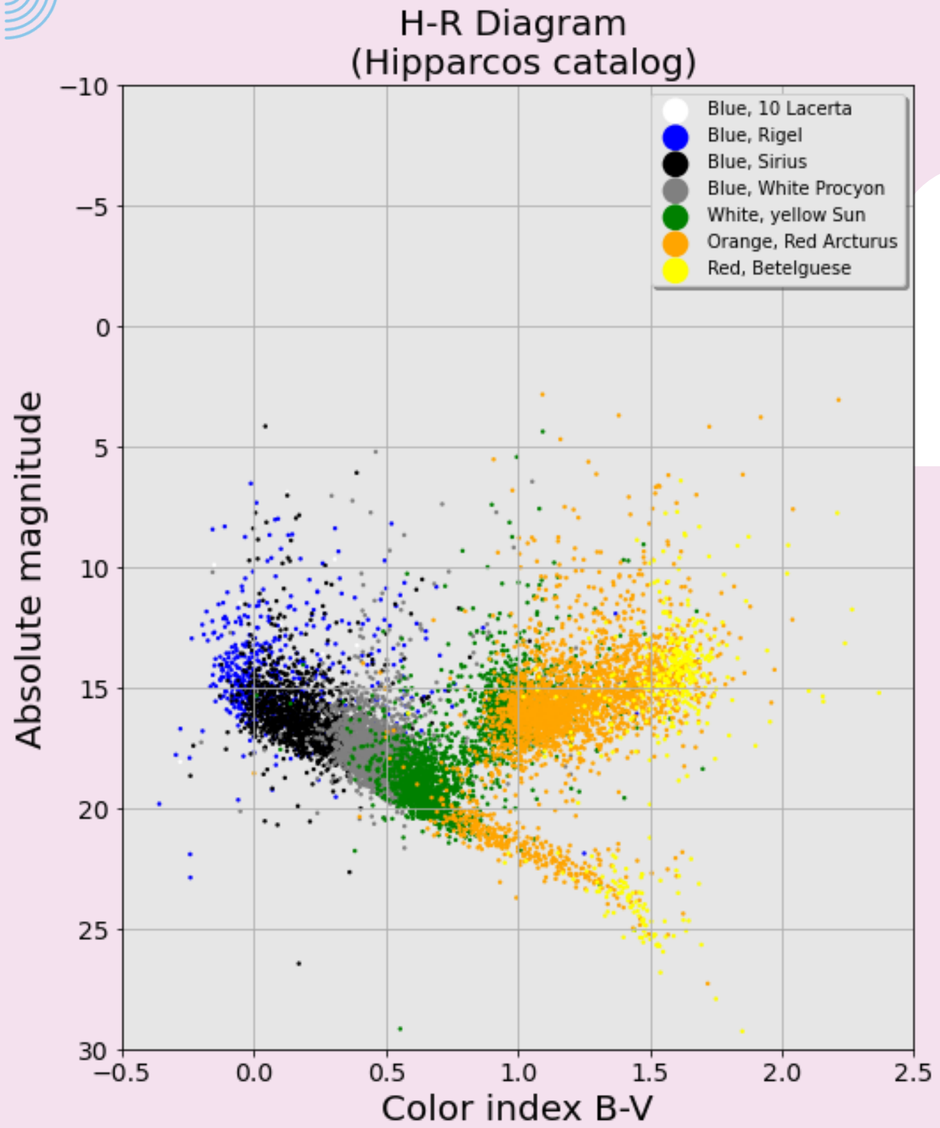
Machine Learning

Overall model performance for B-V is printed or displayed at the end of the script as mentioned began at 76% and after optimisation ended at 91%.

Overall model performance for AMAG began at 77% and after optimisation ended at 85%.

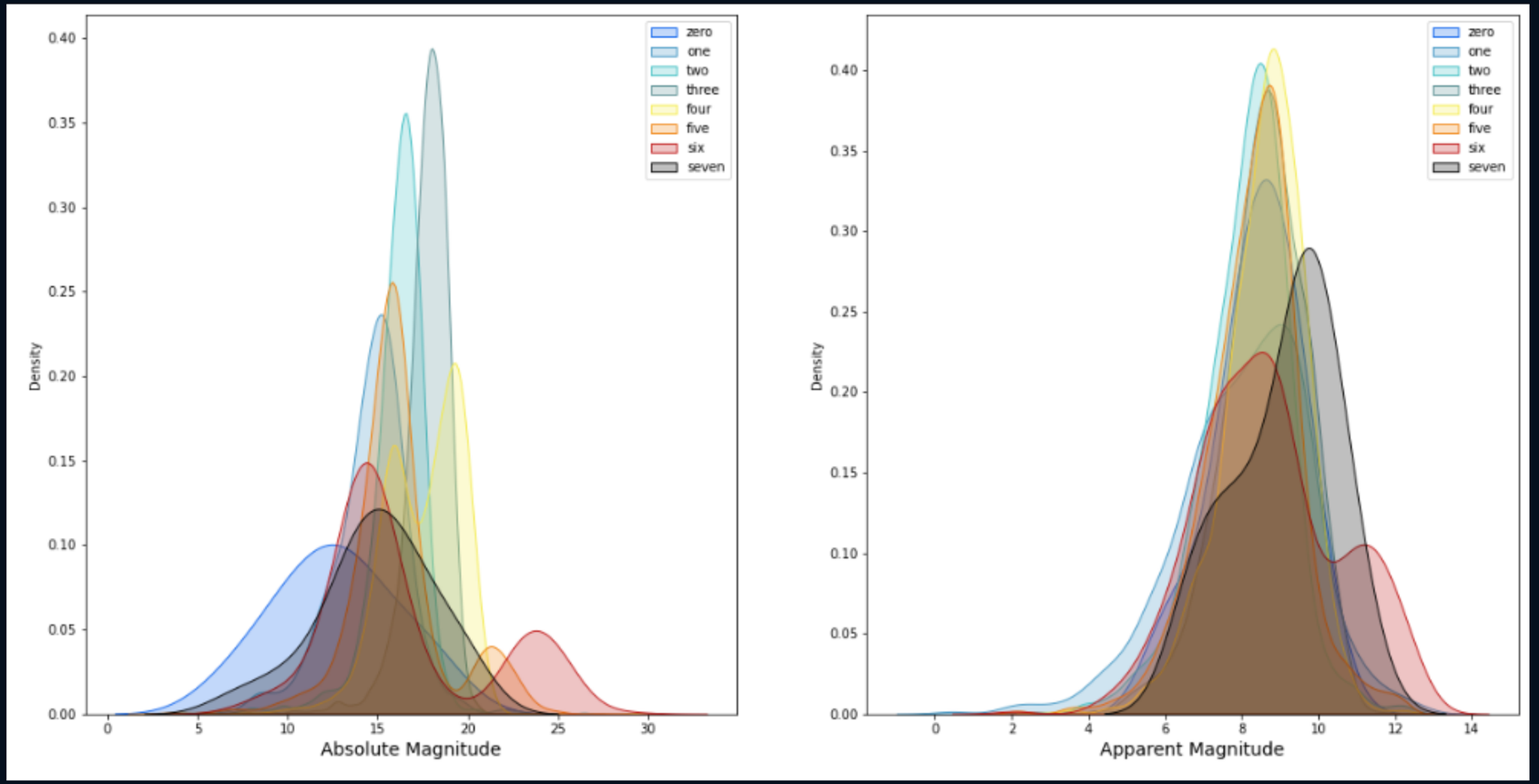
Does our data fit the catalog?





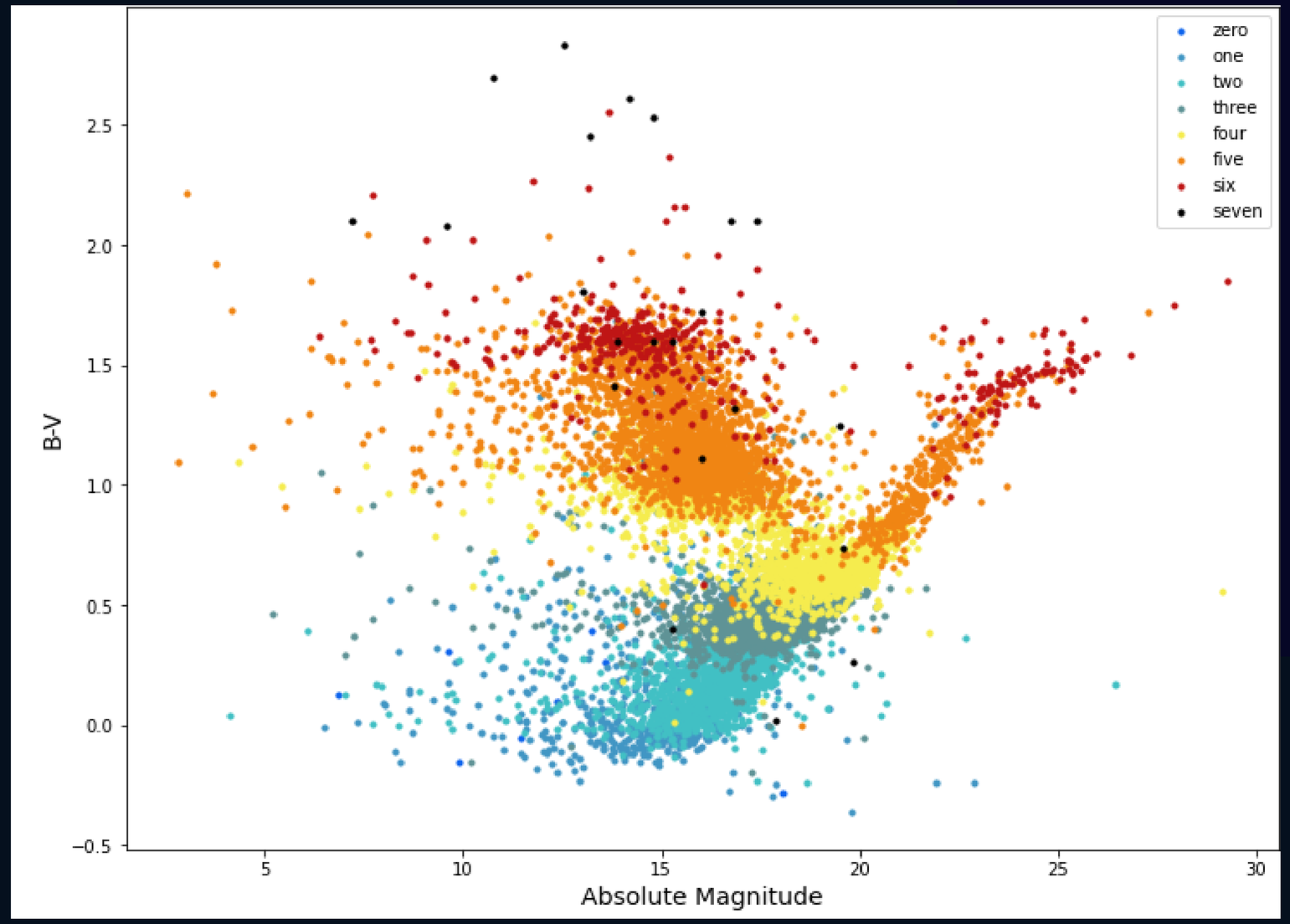
2023

Data has normal distribution



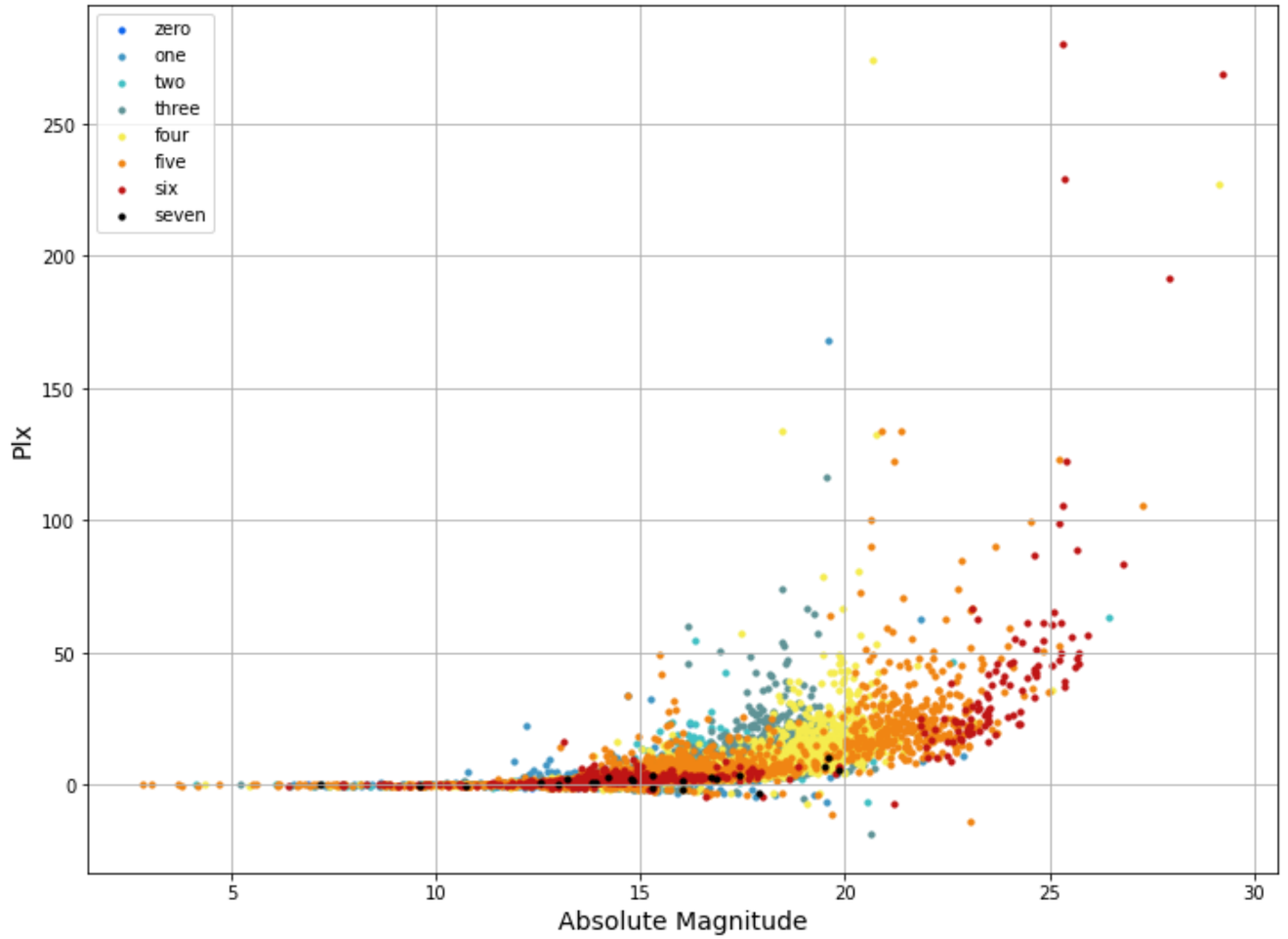
2023

Brightness vs Temperature



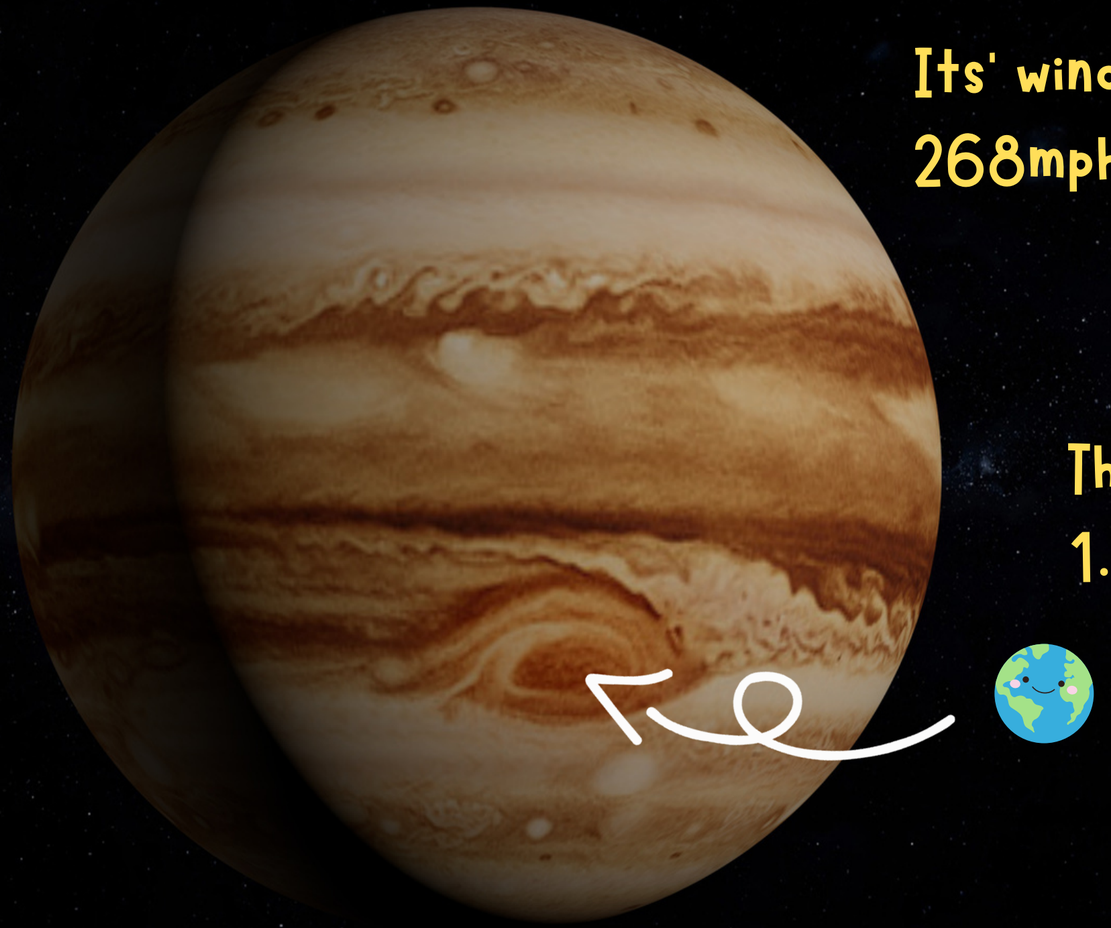
2023

In a galaxy far, far away..



Did you know?

Jupiters' largest storm – the 'Great Red Spot' was first observed in 1665.



Its' wind speeds are up to 268mph (432kmh).

The Earth can fit inside it 1.3 times..

Data has confirmed it is at least 310km deep but it is believed it could be as much as 500km.

Conclusions

- Resampling is incredibly useful with building robust models with imbalanced data, as accuracy is not the end all. A robust classification model should have strong F1 scores accross the board.
- Supervised Machine Learning gave us the highest accuracy, specifically classification models.
- Our two datasets had 10,000 rows and 100,000 rows. We did not experience any difference in our result by using the larger dataset.
- RandomForestClassifier is a data-hungry model, meaning feature engineering became invaluable to us.
- Non Linear Support Vector Machines was more accurate than Linear for this classification model, this is supported by our dataset, as you can see there aren't really any linear relationships going on.
- It's difficult to obtain very accurate data from very far away stars, meaning there is overlap and outliers in the data, making exact classifications incredibly difficult.

Thank You

for your attention!

Any questions?