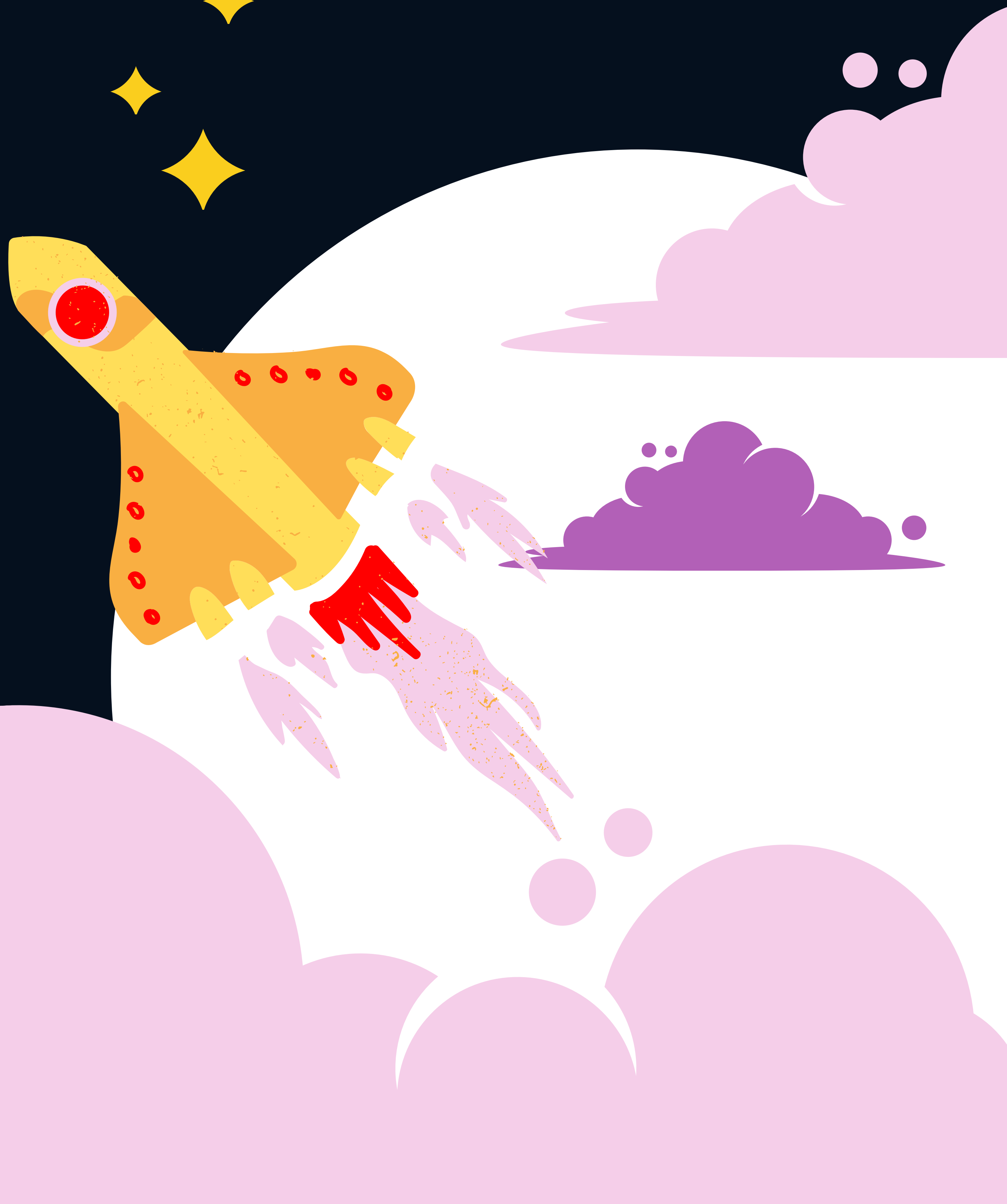
2023 Group 2 Project 4

Stars Classification

Welcome To Our Presentation

Cheila, Grace, Helen,
Kassem & Rahmi

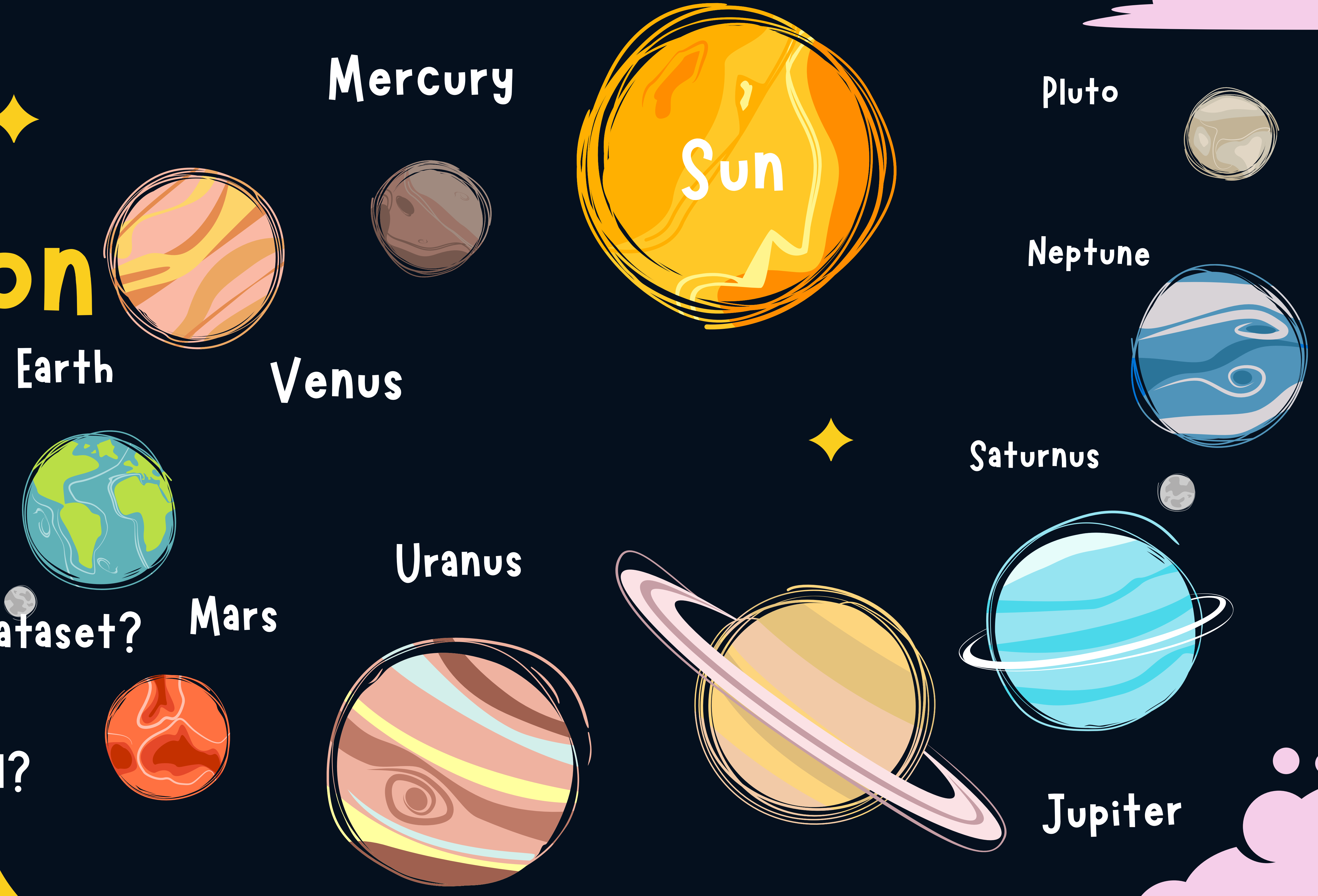


2023

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

2023

Introduction



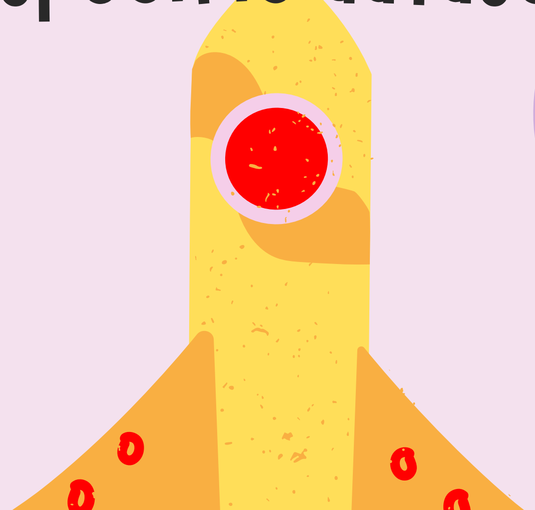
Why did we pick the stellar dataset?

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.



2023

Dataset & Pre-Processing

01

Dataset -
Kaggle, Raw File, Clean, Final Dataset

02

ETL -
Extract, Transform & Load

03

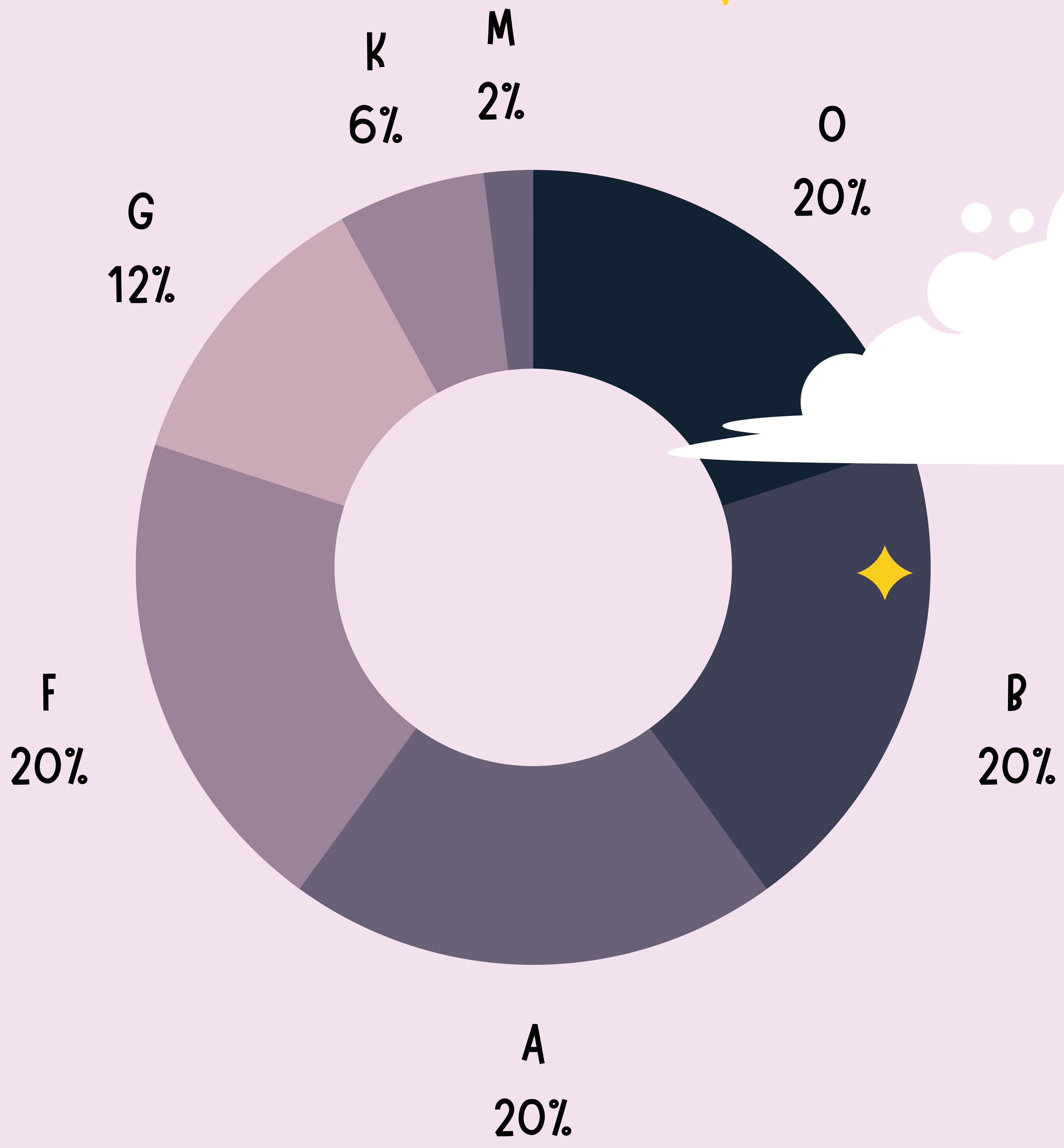
Pre-Processing

2023

How the dataset splits for each star classification

There 7 main types of classification:

OBAFGKM



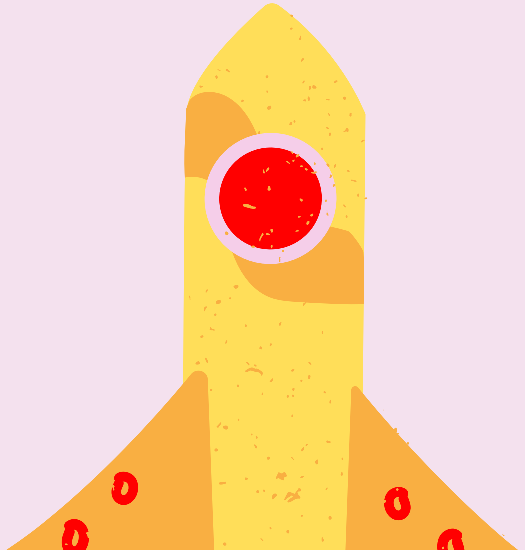
The Brightest Star...

... 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 and later with the data found from machine learning results



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

The Biggest (known) star...

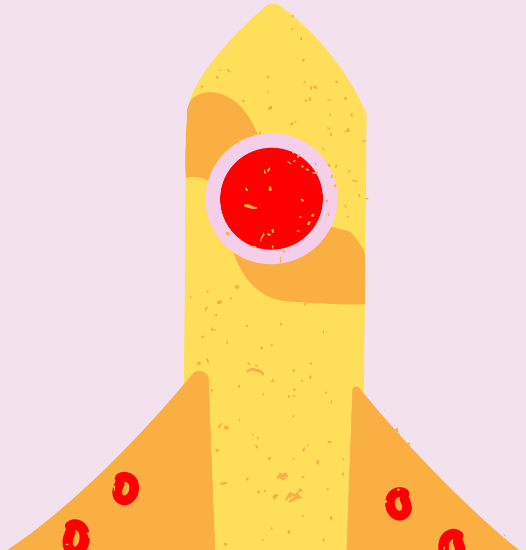
... in terms of size, is UY Scuti, a hypergiant star located in the Scutum constellation. Its size is estimated to be around 1,700 times that of the Sun. However, it is important to note that there may be other stars that are even larger but have not yet been discovered.

2023

ANALYSIS

Our analysis used machine learning...

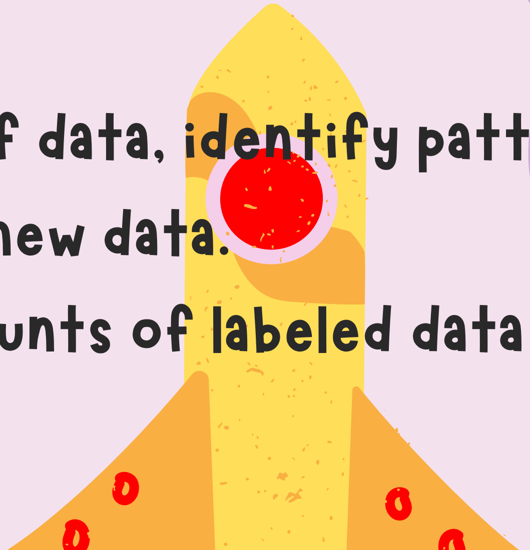
what did Cheila do?
what did Helen do?



2023

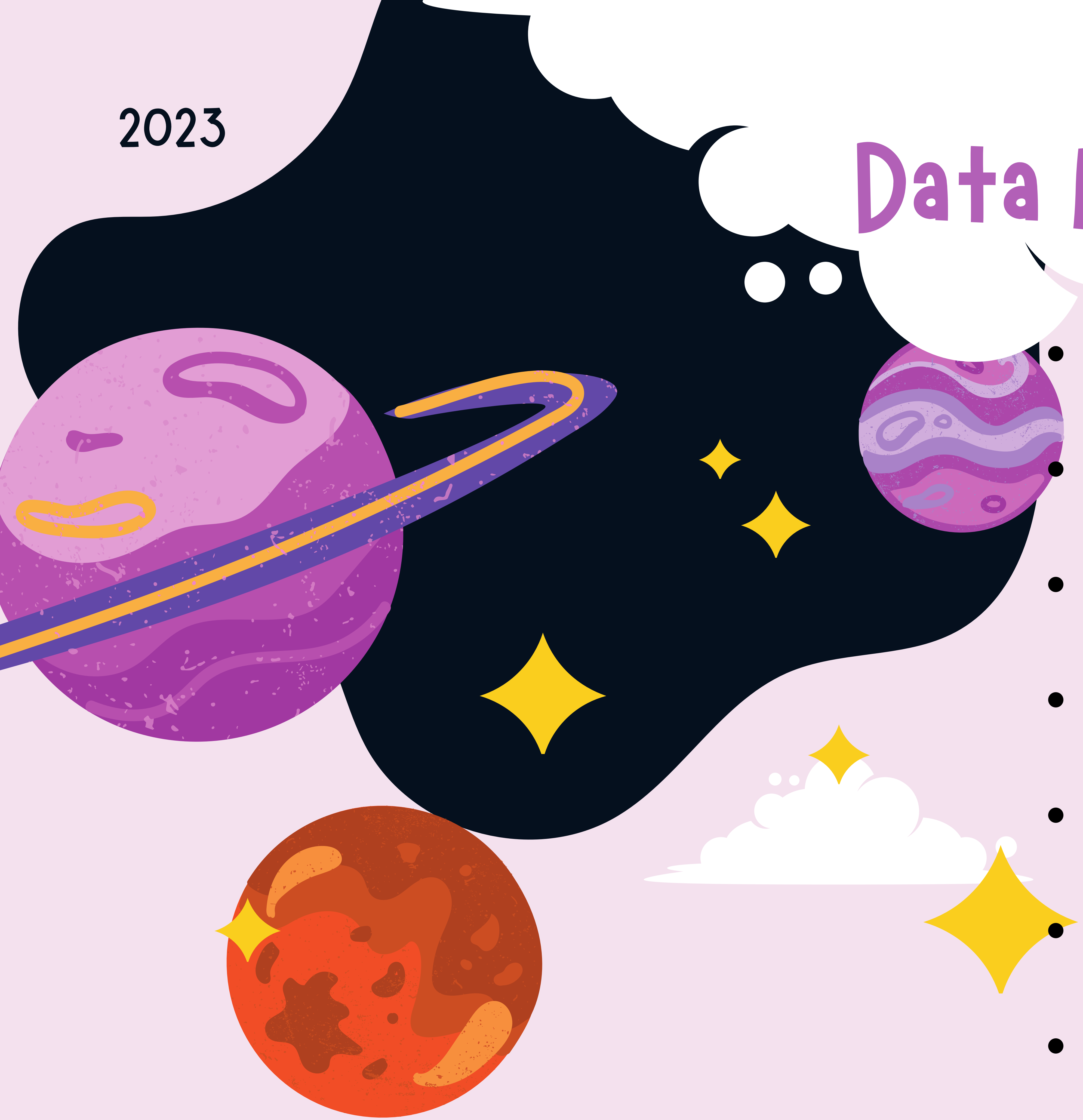
MACHINE LEARNING!

- 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

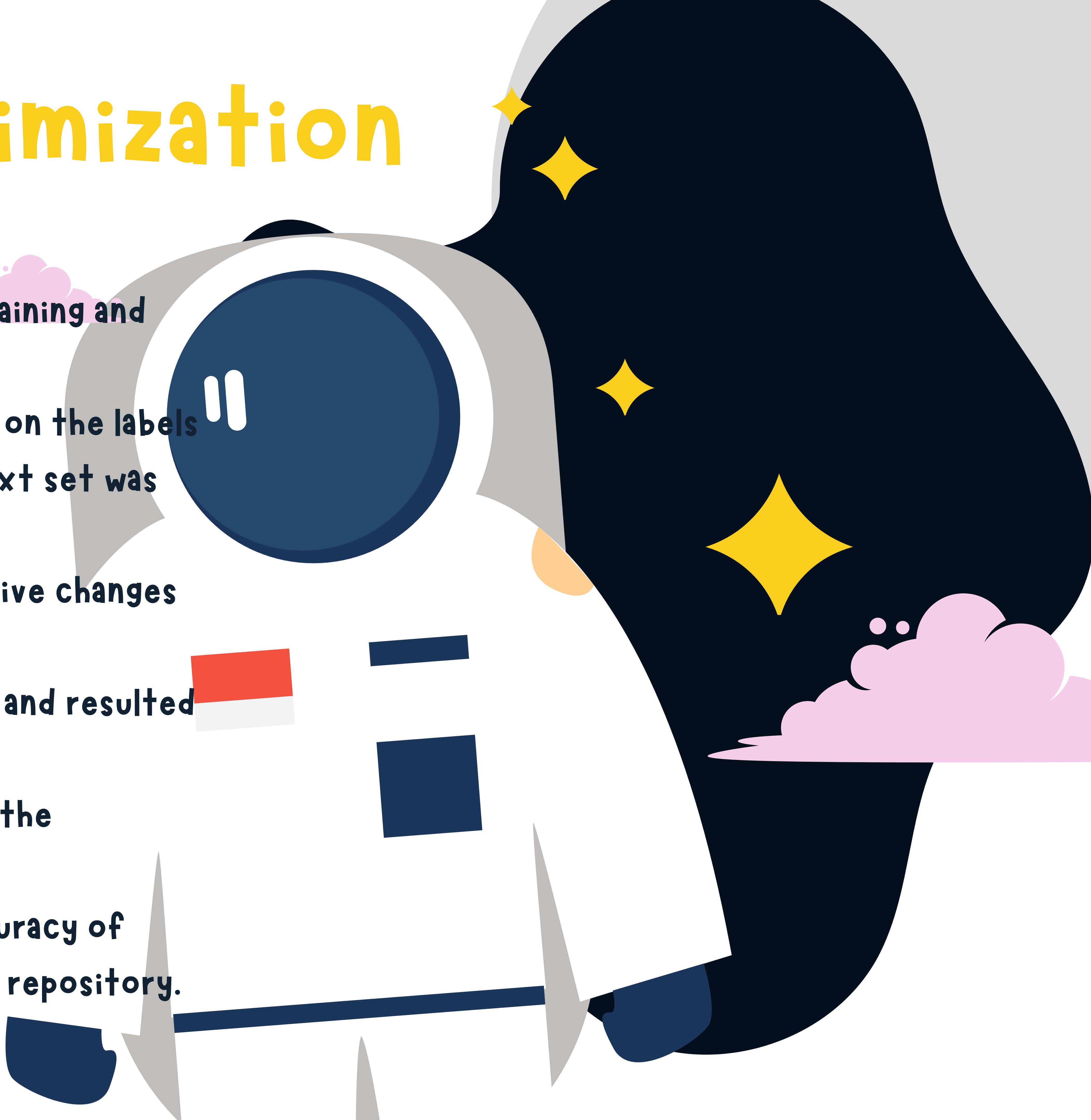
Data Model Implementation



- Created Python script to initialise, train, and evaluate the model on the "Final_Stars.csv" dataset
- Data cleaned, normalised, and standardised during pre-processing work
- Resource files show cleaning progress starting with "Star9999_raw.csv"
- "Clean_stars.csv" and "Final_Stars.csv" are fully formatted and cleaned datasets
- The model utilises data retrieved from SQL, visible in the Jupyter Notebook script
- A classification accuracy of 75.10% shows significant predictive power above the 75% threshold set in the rubric
- For loop used during pre-processing to clean target class, leading to higher classification accuracy

Data Model Optimization

- Data was separated into two sets during model training and testing to avoid compromising model accuracy
- Datasets were split using a loop through numbers on the labels
- One set of data was used for training, and the next set was used for testing after predictions were made
- Model optimisation was achieved by making iterative changes to the model
- Changes in model performance were documented and resulted in a slightly higher than 75.6% accuracy
- Overall, the model's performance is 75.6% after the optimisation process
- The final optimised model and analysis with an accuracy of 89.2% is saved in the "two-target" folder of our repository.



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

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 is printed or displayed at the end of the script as mentioned at 75.6% post the model optimisation process. Within the two-target folder is our final optimised model and analysis which contains an accuracy of 89.2%.

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!

Thank You

for your attention & any questions