Star type detection using Machine Learning

Narendra Bhatkar
UID : 229126
Roll No : 17

Department of Physics, St. Xavier's College

Abstract

Astronomical classification of stars is a crucial and essential endeavour that sheds light on the behaviour of celestial objects. The goal of this project is to use the logistic regression approach in machine learning to categorise stars from the data based on their various characteristics, such as colour, luminosity, and so on. We want to create a predictive model that can divide stars into different classifications like main sequence, red giant, white dwarf, and others by utilising logistic regression.

Introduction

Machine Learning is a subfield of artificial intelligence that involves the development of algorithms and the creation of models and data-driven predictions.

Machine learning can be classified into 4 types,

1. **Supervised Learning :** Here the model is first trained with a given set of data and then make predictions based on traning.
2. **Unsupervised Learning :** Here the model identifies patterns in data without any prior labelled outcomes.
3. **Semi - Supervised learning :** It is basically a combination of both Supervied and Unsupervised learning.
4. **Reinforcement learning :** Here the model learns to make decision by interacting with the environments and is recieves feedback in the form of a reward.

For this project we will be looking at the Supervised Learning part of Machine learning which consists of two types

- **Linear regression :** Linear regression is a significant scientific approach used to model the association between a dependent variable and one or more independent variables. The primary goal of linear regression is to identify the optimal straight line (or hyperplane in higher dimensions) that minimises the discrepancy between the predicted values and the observed values of the dependent variable. The coefficients and intercept that characterise the slope of this line are estimated.
- **Logistic regression :** Logistic regression is a statistical methodology employed to construct a model that estimates the likelihood of a binary outcome, usually represented as either 0 or 1, by considering one or more predictor variables. In contrast to its nomenclature, logistic regression is

frequently utilised for classification tasks rather than regression analysis. The utilisation of this approach is particularly advantageous in cases when the dependent variable is categorical and the primary purpose is to estimate the likelihood of an observation being assigned to a specific class.

The logistic function, commonly known as the sigmoid function, transfers any real-valued input to a value between 0 and 1, and is used in logistic regression to transform the linear combination of predictor variables. This transformed output is the predicted probability that an observation belongs a specific class (positive class).

Mathematically, the logistic regression model can be represented as:

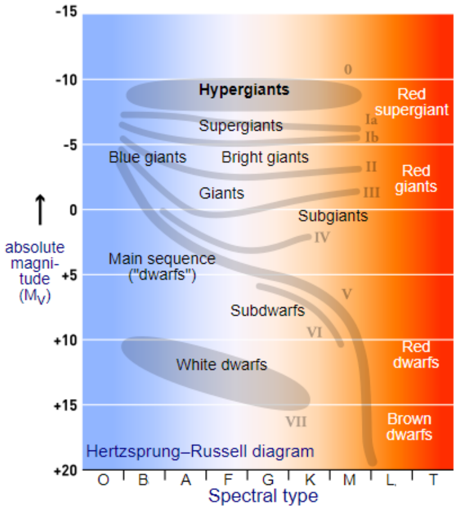
$$\rho = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + ..)}}$$

where e is the euler's contant and β is the total number of classes.

Logistic regression consists of 2 classification binary and multiclass.When the target variable has two classes (binary classification), binary logistic regression is utilised. The goal is to predict the probability of an observation belonging to one of the two classes.Whereas when there are more than 2 variable classes then the regression used is called multiclass logistic regression.We will be using the multiclass logistic regression.

HR Diagram

The Hertzsprung-Russell diagram, also known as the H-R diagram, is used in astronomy to depict the absolute magnitudes (intrinsic luminosity) of stars against their spectral types (temperatures).It was derived from charts begun in 1911 by the Danish astronomer Ejnar Hertzsprung and the American astronomer Henry Norris Russell, both of whom worked independently.The stars are arranged from bottom to top by decreasing magnitude (increasing luminosity) and from right to left by increasing temperature (spectral class) on the diagram.Within the diagram, the Sun's stars within that specific spiral arm tend to gather together in clearly defined zones where they stand out particularly distinctly. The group of stars commonly referred to as main sequence spans from the hot and radiant stars positioned on upper left corner of this particular image, finally shifting across to cooler and fainter representations that come into frame down at lower right turn.You will find large yet bright specimens, coolly referring to themselves as giants and super-giants, which take form on the scene and have a strong visual impact in the upper right quadrant, whereas amusingly, tiny yet luminous hot stars- white dwarfs position themselves in the lower left corner.The Sun is located about halfway along the main sequence, where stars spend the majority of their lifespan. As stars consume hydrogen in their cores converting it into helium, they grow brighter and at the same time cooler (owing to expansion), thus transitioning away from the middle region of main sequence into sectors above towards supergiants and giants.



Procedure

The dataset used for the project is taken from kaggle which consists 240 stars with different features based on which we need to classify the star types. Information on the target classes is given below,

0 - Brown Dwarf

1 - Red Dwarf

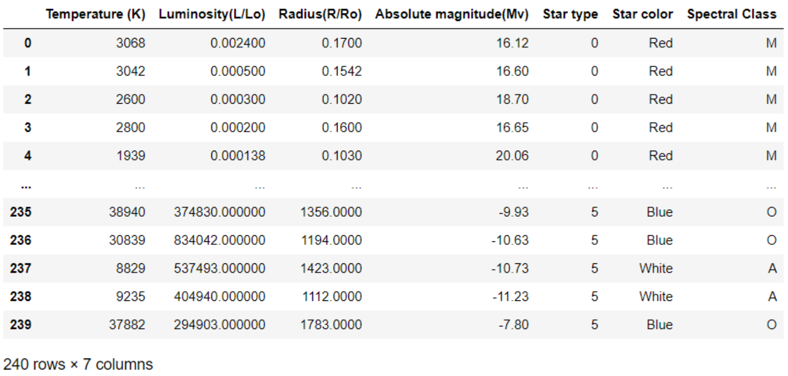
2 - White Dwarf

3 - Main Sequence

4 - Supergiants

5 - Hypergiants

We start with importing the necessary libraries and reading the csv file. The data set consists of several features of stars like Absolute Tempreature, Relative Luminosity, Relative Radius, Absolute Magnitude, Star Type, Star Colour.

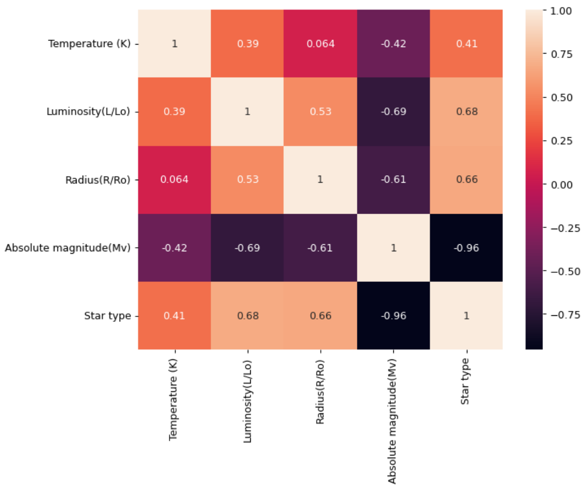


	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute magnitude(Mv)	Star type	Star color	Spectral Class
0	3068	0.002400	0.1700	16.12	0	Red	M
1	3042	0.000500	0.1542	16.60	0	Red	M
2	2600	0.000300	0.1020	18.70	0	Red	M
3	2800	0.000200	0.1600	16.65	0	Red	M
4	1939	0.000138	0.1030	20.06	0	Red	M
...	...	...	...	...	...	...	...
235	38940	374830.000000	1356.0000	-9.93	5	Blue	O
236	30839	834042.000000	1194.0000	-10.63	5	Blue	O
237	8829	537493.000000	1423.0000	-10.73	5	White	A
238	9235	404940.000000	1112.0000	-11.23	5	White	A
239	37882	294903.000000	1783.0000	-7.80	5	Blue	O

240 rows × 7 columns

Data File

Then we analyse the data and look for any missing columns in the data. Here there are no null sets present in the data. Next we check for any repeated data present and if present then we clean the data accordingly. Now we look for any correlation in the features present in our data. If any two or more features present in the data provide similar information to our model then we can eliminate one of the features as it won't help improve in the learning of our model. However, in this case we can see that none of the features from our dataset are corelated.



We now move towards our machine learning model, we split our data into two categories train data and test data.

Trainig Data : This specific portion of the dataset is employed for the purpose of instructing or training the machine learning model. The dataset comprises input samples (features) paired with their corresponding known output or goal values. During the training phase, the model acquires knowledge of the patterns, correlations, and dependencies that are observable in the training data. The model adjusts its internal parameters in order to decrease the disparity between its predictions and the actual target values of the training data.

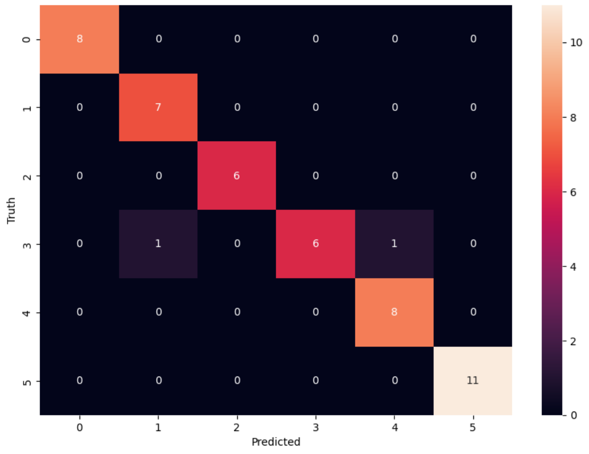
Testing Data : Once the model has been trained, it is essential to evaluate its performance on new, unseen data to determine how well it generalises to new data. The test data is a distinct subset of the dataset that was not exposed to the model during training. It is used to evaluate the efficacy of the model and its ability to accurately predict new data. The test data aid in estimating the model's performance in real-world scenarios.

The reasoning behind utilizing distinct training and test datasets is to decrease the issue of overfitting, which arises when a model gets excessively specific to the training data and exhibits poor results when applied to newer data. Through the process of assessing the model's performance on a separate and distinct dataset, (test data) we can acquire a better understanding of its ability to generalize beyond the training data.

In our case we have split the data in 80 : 20 ratio where 80 percent is the traning data and 20 percent is the testing data and accordingly the result is obtained.

Results and Conclusion

Our model was abel to predict the star data with an accuracy score of approximately 96 percent which is considered a very good accuracy score. However there is an error of 4 percent in our data which can be found out by plotting the confusion matrix.



Here from the test data we can see that the model has incorrectly predicted 2 stars one which belongs to the main sequence but is predicted as supergiant and secondly which again belongs to main sequence but is predicted as red dwarf. The test data consisted of 20 percent of the original data that is from 240 star dataset 48 datasets were used as test data out of which 2 were wrongly predicted which matches with the accuracy of 96 percent determined by the training data set.

However when it come to a huge amount of data our model may not work that accurately since the data may be imbalanced and because of which our model may be baised to the features having more number of data. Also our model is bason on the assumption that the features are independent of each other but in most cases it is not necessary that the features may be independent.

References

1. HR Diagram : https://www.britannica.com/science/Hertzsprung-Russell-diagram.

2. dataset : https://www.kaggle.com/datasets/deepu1109/star-dataset

3. Logistic regression in data analysis: An overview, July 2011, International Journal of Data Analysis Techniques and Strategies 3(3):281-299.

4. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis by Frank E. Harrell

Codelink

stellar project

ORIGINALITY REPORT

12%	9%	2%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.britannica.com Internet Source	3%
2	medium.com Internet Source	2%
3	Submitted to University of North Texas Student Paper	2%
4	www.ijnrd.org Internet Source	1%
5	Submitted to Jacobs University, Bremen Student Paper	1%
6	Jesus Tamez Villarreal, Sophia Barton. "Stellar Classification based on Various Star Characteristics using Machine Learning Algorithms", Journal of Student Research, 2023 Publication	1%
7	Submitted to University of Greenwich Student Paper	1%
8	speedypaper.x10.mx Internet Source	

		1%
9	teachersinstitute.yale.edu Internet Source	1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		