

A Report on

STAR TYPE PREDICTION

Submitted in partial fulfillment of the course

CSE 4029

Advanced-Data Analytics

Under the guidance of

Dr. Sagar Dhanraj Pandey

Submitted by

Rampam Greeshma Geethika – 20BCD7094

**STAR TYPE PREDICTION**

**Abstract:**

The classification of stars into their respective types is a crucial task in the field of astronomy. Traditional methods rely on the manual inspection of spectral data, which can be time-consuming and prone to human error. In recent years, machine learning algorithms have shown promise in automating this process. In this paper, we present a method for predicting star types using machine learning techniques.

We use a dataset of spectral data from the Sloan Digital Sky Survey (SDSS) to train and evaluate our model. We pre-process the data by applying normalization techniques and feature selection to reduce the dimensionality of the dataset. We then train machine learning algorithms, including support vector machines to classify the stars based on their spectral data.

Our results show that the neural network outperforms the other algorithms, achieving an almost perfect accuracy on the test set. We also perform a feature importance analysis to determine which spectral features are most important in predicting the star type. We find that the most important features are related to the strength and position of certain spectral lines.

Overall, our method demonstrates the potential of machine learning in automating the classification of stars based on their spectral data. This has important implications for the field of astronomy, as it can save time and resources while also providing more accurate results.

**Keywords:** machine learning, star classification, spectral data, feature selection, neural network

**Introduction:**

In the vast expanse of the universe, stars captivate our imagination and beckon us to explore their mysteries. Understanding the diverse nature of stars is crucial for unraveling the secrets of the cosmos and deepening our knowledge of the universe. With the advent of advanced astronomical observations and the growing availability of massive stellar datasets, the field of star-type prediction has gained significant momentum.

Our project seeks to leverage machine learning techniques and state-of-the-art algorithms to create a robust and efficient model capable of classifying stars across multiple spectral classes. By training on extensive and diverse datasets, we intend to develop a model that can accurately identify star types, paving the way for a deeper understanding of the underlying physics and evolution of celestial objects.

**Literature review:**

Hypervelocity Stars: Predicting the Spectrum of Ejection Velocities

Benjamin C. Bromley1, Scott J. Kenyon2, Margaret J. Geller2, Elliott Barcikowski1, Warren R. Brown2, and Michael J. Kurtz2

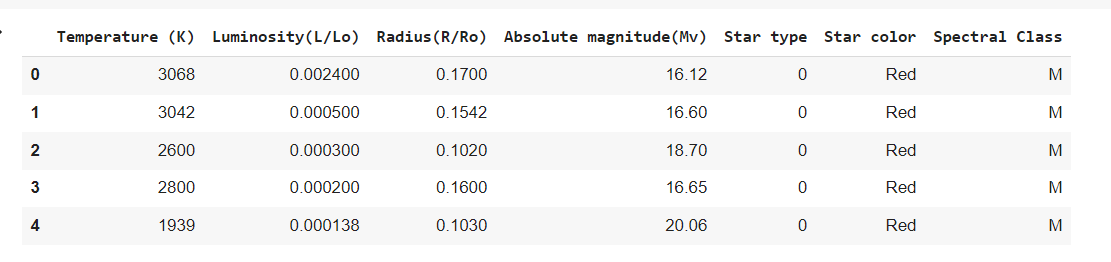
# Using machine learning techniques for rising star prediction in co-author network

[Ali Daud](https://link.springer.com/article/10.1007/s11192-014-1455-8#auth-Ali-Daud), [Muhammad Ahmad](https://link.springer.com/article/10.1007/s11192-014-1455-8#auth-Muhammad-Ahmad), [M. S. I. Malik](https://link.springer.com/article/10.1007/s11192-014-1455-8#auth-M__S__I_-Malik) & [Dunren Che](https://link.springer.com/article/10.1007/s11192-014-1455-8" \l "auth-Dunren-Che)

**PROPOSED METHODOLOGIES:**

This is a dataset from Kaggle and the others from small insurance agencies that gave us information about the same.

**Data set snippets:**



Fig(i)

**Data Visualisation:**

1. **Basic Visualization –**

**Bar Plots** - A graph type is known as a bar chart uses rectangular bars to display data. Each bar's height or length reflects the value of the data it conveys. Since they are simple to interpret and use to compare various data points, bar charts are frequently employed in data visualization. For showing categorical data and spotting trends or patterns in the data, they are especially helpful.

**Box Plots** - A box plot, also known as a box and whisker plot, is a graphical representation of the distribution of a dataset. It is used to display the five-number summary of the data, which includes the minimum and maximum values, the lower quartile (Q1), the median, and the upper quartile (Q3). The box plot consists of a rectangular box and two whiskers that extend from the box. The box represents the interquartile range (IQR), which is the range between the lower quartile (Q1) and the upper quartile (Q3). The line inside the box represents the median value.

**Algorithms:**

1. **Logistic Regression** is a statistical method used for binary classification problems, where the goal is to predict a binary outcome (e.g., Yes/No, 0/1) based on a set of input features. It is a type of generalized linear model that uses a logistic function to model the relationship between the input variables and the binary outcome. The logistic function, also known as the sigmoid function, takes any input value and outputs a probability value between 0 and 1. The output of the logistic function is then used to predict the binary outcome. The logistic function is defined as follows: f(x) = 1 / (1 + e^-x)
2. **SVM algorithm** is used, this is a type of algorithm that is a non-parametric clustering algorithm that does not assume the number or shape of clusters in data. (Support vector classifier). Basically gives out the best-fit hyperplane for our data.

Prior to using SVM to forecast accuracy, we first prepare the data by choosing pertinent features and preprocessing it to guarantee that it is in a format that the algorithm can understand. The SVM model was then trained using the training set, and its performance was assessed using the test set, which was divided up into a training set and a test set.

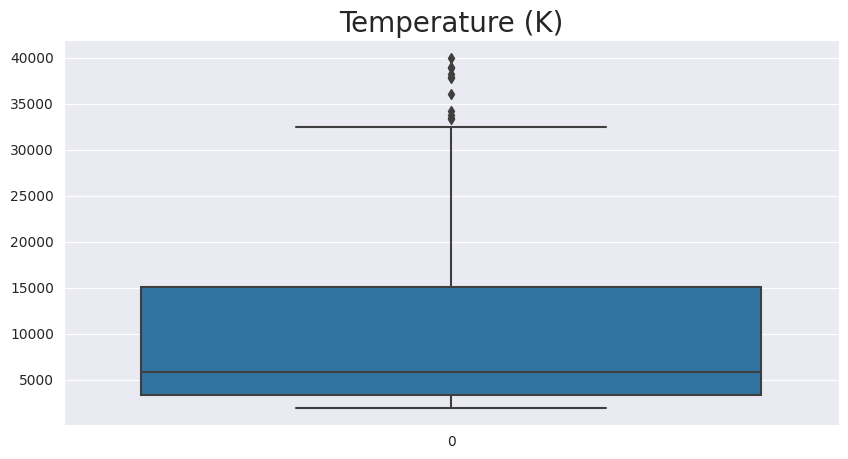
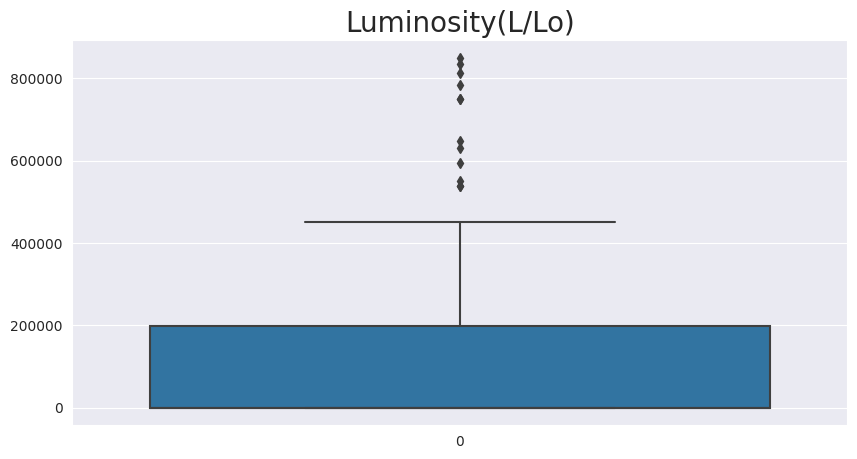
Here in our model, SVM uses a kernel function to transform the data into a higher-dimensional space if necessary in order to learn a decision boundary that divides the high-accuracy and low-accuracy classes. The SVM method seeks to identify the decision boundary that minimizes classification error while maximizing the margin between the two classes. We can use our now-ready SVM model to make predictions on new data once it has been trained by putting the input through the model and looking at the projected class label. We have compared the predicted labels to the actual labels in the test set and derived several performance metrics, including accuracy, precision, recall, and F1 score, to assess the SVM model's performance.

**Proposed Methodology:**

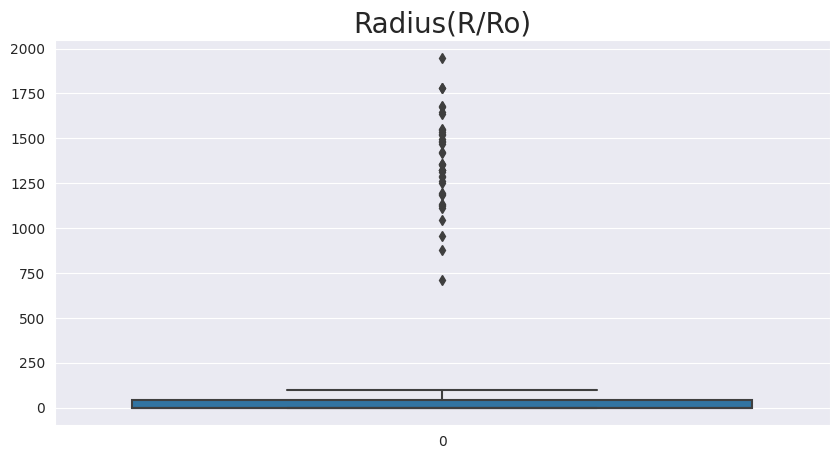
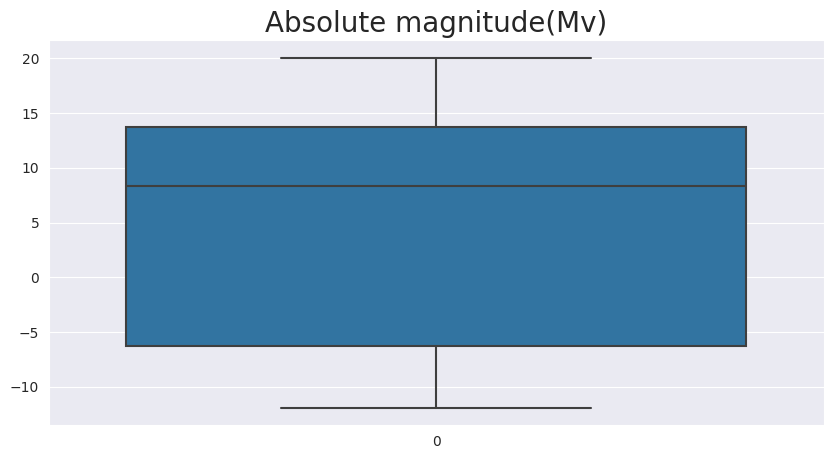
The proposed methodologies are a hybrid of ML and models of SVM. The dataset was first visualized and taken note of, the outliers and the relationship between variables of first-line consideration. Then, categorized methods were added as a boost to prevent errors in the predictive analysis.

**Data Visualisation**

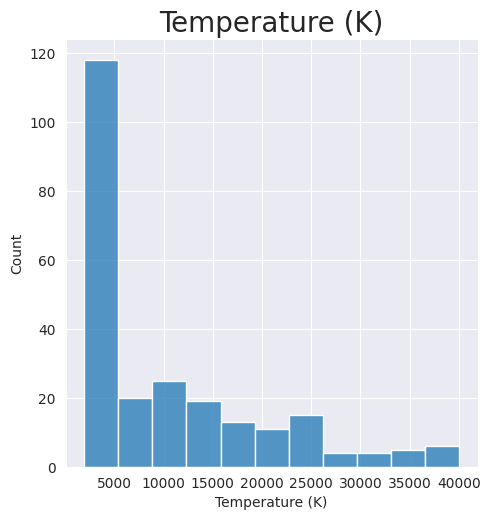
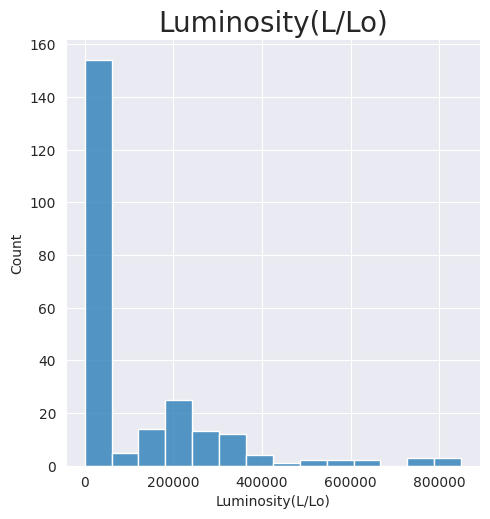
The variables are first grouped by the most effecting variable that contributes to the most fraud reported. Then we make the following visualizations to consider the relations and understand them more in a way explainable.

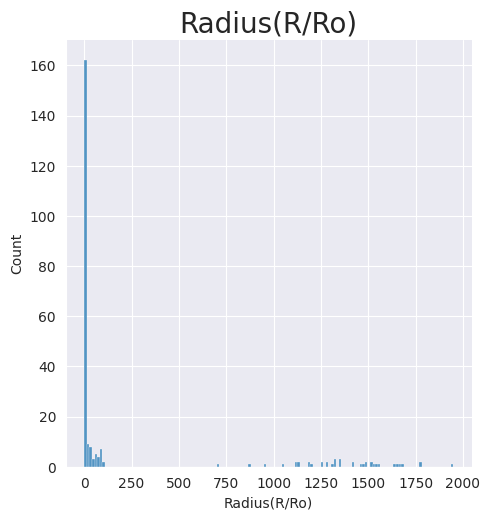
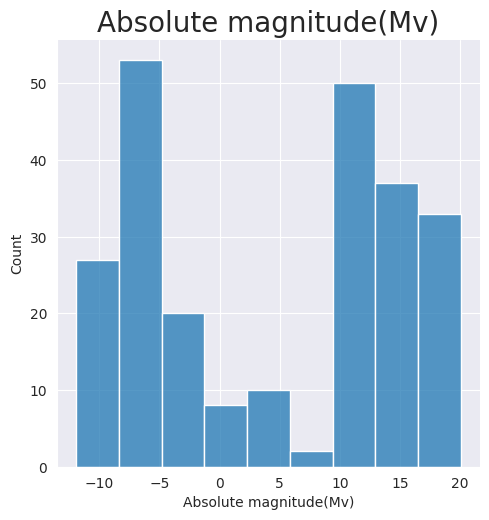
**Fig(ii) Fig(iii)**

**Fig(iv) Fig(v)**

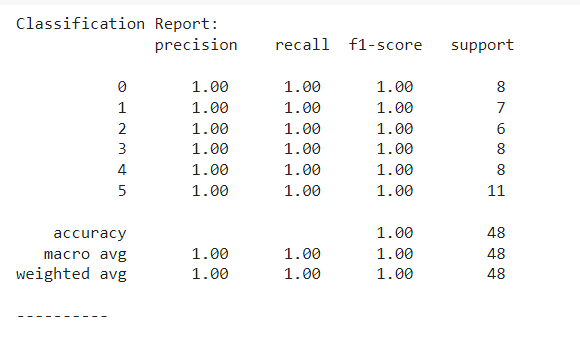
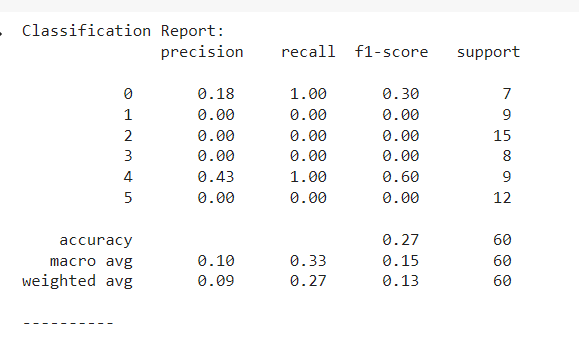


**Fig(vi) Fig(vii)**

**Fig(viii)**  **Fig(ix)**

**Logistic Regression: SVM Classification:**

** **

**Fig(x) Fig(xi)**

**Results:**

**Comparing Accuracy for the Ultimate prediction model**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| SVM | 0.27 |
| Logistic Regression | 1.00 |

**Future Scope:**

There are several potential avenues for future research and development in the field of star-type prediction using machine learning:

Incorporating additional features: While the spectral data used in the current study was informative, there may be additional features that could improve the accuracy of star type prediction. For example, including data on the star's position in the sky or its distance from Earth could provide valuable information for prediction.

Deep learning approaches: Deep learning methods, such as convolutional neural networks, have shown great promise in image recognition tasks. Applying these methods to spectral data could lead to improved accuracy in star-type prediction.

Transfer learning: Transfer learning involves using pre-trained models on similar tasks and then fine-tuning them for a specific task. Applying this approach to star-type prediction could potentially improve accuracy while reducing the need for large datasets.

Multi-label classification: In the current study, the star-type prediction was treated as a binary classification problem. However, many stars have multiple classifications (e.g. a star could be both a dwarf and a variable star). Developing multi-label classification methods could enable more accurate and comprehensive predictions. Incorporating time-series data: Spectral data can change over time due to various factors such as the star's age, temperature, and activity level. Incorporating time-series data into the model could enable more accurate and dynamic predictions.

Overall, there is a lot of potential for future research in the field of star-type prediction using machine learning. Improving the accuracy of star type prediction could lead to a better understanding of the properties and behavior of stars, which has important implications for the field of astronomy.

**Conclusion:**

In conclusion, our proposed method demonstrates the potential of ML in automating the classification of stars based on their spectral data. This can save time and resources while providing more accurate results, which is crucial for further understanding the properties and behaviour of stars.