Stellar Classification

Thomas Lomas

I. Motivation

My dataset has many features of a star and its stellar classification. With my research questions, I hope to see how much information is needed to determine a star's stellar classification. By looking at the feature importance of various models I hope to determine what information is important to the stellar classification and what might not be as important. I also hope to examine the correlations that are found between the features of the dataset and test the robustness of a random forest regressor on astronomical datasets. This could be helpful not only for scientists trying to classify stars but also for scientists gathering information. These models could be used by scientists to classify stars but also they could tell scientists what information to gather and what information they don't have to worry about as much.

II. Datasets

A. Dataset Description

The dataset consists of many features of a star such as temperature, luminosity, radius, absolute magnitude, star type, star color, and spectral class. The star type refers to where it would be placed on a Hertzsprung-Russell diagram and includes classifications such as brown type, red dwarf, white dwarf, main sequence, and supergiants. The dataset contains 240 rows and 7 columns.

B. Data Preparation

The dataset contains many rows with missing data which need to be removed before running the dataset through various machine learning models. These missing pieces of data are originally considered empty strings so I first had to replace empty strings with “NaN” and then I could use dropna to clear out the rows with missing values. The dataset also contained categorical data which I had to process. The spectral class column had data considered ordinal, so I used an ordinal encoder to transform the data into int data types. The star color had data considered nominal, so I used OneHotEncoder to transform the data into numerical data.

III. Research Questions and Related Work

A. Research Questions

I have five research questions that I look to answer using this dataset. My first research question is “What is the correlation between luminosity and temperature, and luminosity and radius of the dataset? Does the resulting correlation coefficient make sense given the formula for luminosity?”. This research question hopes to examine the correlations between some of the features of the dataset and see if it conforms to the formulas for luminosity. My second research question is “Can we build a model to predict the type of star given its characteristics such as temperature, luminosity, radius, absolute magnitude, star color, and spectral class?”. This research question is testing how well a model can be trained to predict the type of star given many features. Astronomers could then use models such as the ones tested to identify stars. My third research question is “How many features of the dataset can be removed before the accuracy of the decision tree model falls below 80%?”. This research question hopes to test how much information about a star is needed before a decision tree model is unable to accurately predict a star's spectral class. My fourth research question is “Can we build a model that predicts the absolute magnitude of a star given its characteristics such as temperature, luminosity, radius, star color, star type, and spectral class?”. This research question hopes to test how well a random forest regressor can perform on an astronomical dataset. My last research question is “How many features of the dataset can be removed before the accuracy of the random forest model falls below 80%?”. This research question hopes to test the robustness of the random forest model, specifically on an astronomical dataset.

B. Related Work

There is some other work done from this same dataset but most of those datasets mainly focus on visualizing the datasets rather than training models on it. Some of the work does look at training models but those models only look at classification models and not regression models as I do.

IV. Research Methods

I started this project by processing the data using the methods mentioned in 2B. After processing the data I then went into answering my research questions. I started by looking at how each of the features correlates to each other using a correlation heatmap. My main purpose for looking at the correlation heatmap was to look at the correlations found between luminosity and temperature, and luminosity and radius. After seeing the heatmap, I looked at the relationships between luminosity and temperature, and luminosity and radius through scatter plots to better visualize the correlation between the features. For my second research question, I used a decision tree classifier with a grid search cv technique for hyperparameter tuning to predict the spectral class of a star. I then looked at the confusion matrix, precision, recall, f1-score, and the feature importance of the model. For my third research question, I used the same model and technique for hyperparameter tuning as before but limited the number of features that the model had access to. I removed features by looking at the feature importance of the model and removing the feature with the highest feature importance. The first feature that I removed was temperature. After removing temperature, the feature with the highest feature importance was the star color “Red”. The next model ran without temperature and “Red” as a feature. After removing those features, the feature with the highest feature importance was “Blue-White”. The final model ran without temperature, “Red”, and “Blue-White”. I then started to answer my third research question by training a random forest regression model with a grid search cv technique for hyperparameter tuning. I looked at the MSE, RMSE, and R^2 of the training and testing data to ensure there was little to no overfitting/underfitting. I also looked at the feature importance which leads into my next research question. In my next research question, I did the same process as before with removing features until the accuracy of the model declined. In the case of a regression model, I looked at R^2. I first removed luminosity which revealed the next highest feature importance was radius. After removing luminosity and radius, the next highest feature importance was star type so it was removed from the model. This led to a very low R^2 and is where I stopped removing features from the model.

V. Findings/Results

For my first research question, I wanted to see the correlation coefficient of luminosity and radius, and luminosity and temperature because the formula for luminosity is $L = \sigma A T^4$ where L is luminosity, σ is the Stefan-Boltzmann constant, A is the surface area, and T is the temperature. A strong correlation between these features would be expected because of the formula but this is not the case. The correlation heatmap shows a correlation coefficient of 0.41 between luminosity and temperature and a correlation coefficient of 0.54 between luminosity and radius. Scatter plots of these features confirm the correlation coefficients.

Below is a table of the training accuracies, testing accuracies, accuracy gap, precision, recall, and f1-score of each of the models used for research questions 2 and 3. Best hyperparameter setups and feature importance can be found below the table.

	Training Accuracy	Testing Accuracy	Accuracy Gap	Precision	Recall	F1-Score
Original Model	93.148%	88.372%	4.776%	92.178%	88.372%	86.292%
Removing Temperature	92.407%	86.047%	6.367%	83.765%	86.047%	83.579%
Removing Temperature and Red	91.667%	86.047%	5.62%	83.278%	86.047%	84.245%
Removing Temperature, Red, and Blue-White	85.556%	79.070%	6.486%	68.162%	79.070%	72.655%

For my second research question, I used a decision tree classifier and grid search cv technique for hyperparameter tuning. My training/testing split for this data was 60% training data and 40% testing data. Through testing many splits, this showed the smallest gap between training and testing accuracy. For the hyperparameter tuning, I used gini and entropy for the criterion, a max depth of two through eight, max leaf nodes of five, ten, fifteen, and twenty, and max features of None, 0.2, 0.4, 0.6, 0.8. For the grid search, I used a scoring of accuracy and a cross-fold variation of 15. After running the model, the best hyperparameters ended up being entropy as the criterion, a max depth of 4, max features of 0.6, and max leaf nodes of 10. The feature importance of the model shows that temperature was the most relevant with a feature importance of 65.47% with absolute magnitude, star type, color red, color white, and color yellow-white not impacting the model at all.

For my third research question, I trained four separate models and only changed the hyperparameters for the last one. For the first model, I removed temperature from the features to see how the model handled it. The best hyperparameters for this model were an entropy criterion, a max depth of 6, max features of 0.8, and max-leaf nodes of 10. These have changed somewhat from the original model where its max depth increased by 2 and its max features went up by 0.2. The feature importance of this model showed that the colors red and blue were the most impactful to the model with red having a feature importance of 41.235% and blue having a feature importance of 31.718%. Absolute magnitude gained some importance with a feature importance of 6.941% with the star type, and color white and yellow-white still having no impact. For the next model, I removed temperature and red from the features. The best hyperparameter setup for this model was an entropy criterion, a max depth of 5, max features of None, and max leaf nodes of 15. The most impactful features for this model were the color blue-white and blue with blue-white having a feature importance of 30.206% and blue having a feature importance of 29.780%. All features had some impact on the model. For the final model, I removed the color blue-white from the features. The best hyperparameter setup for this model was an entropy criterion, a max depth of 6, max features of 0.4, and a max leaf nodes of 10. The most impactful feature for this model was the star type with a feature importance of 39.727%. All of the features had some impact on the model.

Below is a table containing the training and testing MSE, RMSE, and R^2 of the models used for research questions four and five. Below the table contains information about the best hyperparameter setups and the importance of the models' features.

	Train MSE	**Train RMSE**	**Train R^2**	**Test MSE**	**Test RMSE**	**Test R^2**
Original Model	1.248	1.117	99.738%	1.735	1.317	98.387%
Removing Luminosity	1.276	1.13	99.594%	1.861	1.364	98.27%
Removing Luminosity and Radius	1.63	1.277	99.056%	1.789	1.338	98.337%
Removing Luminosity, Radius, and Star Type	55.788	7.469	73.617%	80.013	8.945	25.613%

For my fourth research question, I used a random forest regression model with a grid search cv technique for hyperparameter tuning. The best training/testing split I could find for the model was 75% training and 25% testing. For the hyperparameter tuning, my setup was 25-50 estimators, max features of square root, log2, and none, max depth of three, six, and nine, and max leaf nodes of none, ten, twenty, thirty, and forty. For the grid search cv, I used a scoring method of negative mean squared error and a cross-fold variation of 5. For this model, the best hyperparameter setup was a max depth of 9, max features of none, max leaf nodes of 30, and 25 estimators. The most impactful feature for this model was luminosity with a feature importance of 67.097%. The only feature that did not impact the model was the color white.

For my fifth research question, I trained three separate models and again only changed the hyperparameters for the last model. For the first model, I removed luminosity from the features. The best hyperparameter setup for this model was a max depth of 9, max features of none, max leaf nodes of 20, and 50 estimators. The most impactful feature for this model was radius with a feature importance of 98.783%. All of the features for this model impacted it in some way. For the next model, I removed luminosity and radius. The best hyperparameter setup for this model was a max depth of 6, max features of none, max leaf nodes of 10, and 50 estimators. The most impactful feature for this model was the star type and had a feature importance of 97%. There were only two features that didn't impact the model at all which were the colors blue and red. For the final model, I removed luminosity, radius, and star type. I also increased the number of estimators to 400 in an attempt to increase R^2. The best hyperparameter setup for this model was a max depth of 6, max features of square root, max leaf nodes of 30, and 200 estimators. The most impactful features for this model were temperature and spectral class with temperature having a feature importance of 63.744% and spectral class having a feature importance of 20.208%. All features impacted the model in some way

VI. Conclusions

In conclusion, this dataset does not conform to the expectations that come from the Stefan-Boltzmann law. In theory, there should be a larger correlation between luminosity and temperature, and luminosity and radius. The deviations from the theory could be due to many factors. In this case, the confirmation using scatter plots rules out an issue with correlation coefficients showing linear relationships. This means that there are most likely processes going on within the data that affect the correlation coefficients. There is most likely a limited range of data as this only gives characteristics of 215 stars but there could also be hidden factors to luminosity that are not explored within this dataset. Through various tests, I can also conclude that decision tree classifiers do a good job of predicting the spectral class of a star given many characteristics. If you want to push the model to its limits, you could still achieve above 80% accuracy while only giving the model a couple of characteristics. In the real world, a decision tree classifier could be used if given many characteristics but if you don't have as much data, it is better to use a more advanced algorithm to identify stars. In terms of random forest regression models on astronomical datasets, we can see that these models are quite robust and can make predictions with very good accuracy. Even with little characteristics, the model is still able to gain accuracy way over 90%.

VII. Future Work

In the future, I would like to look at more advanced algorithms, specifically for classification. I would like to test spectral class classification with more advanced algorithms as a decision tree classifier did not have as good of accuracy as I was hoping for. I would also like to work with a more advanced dataset. This dataset was decent but I don't feel there were enough stars to use these models in the real world. There are hundreds of billions of trillions of stars in the universe so it is hard to gather meaningful real-world information when you are only looking at 215 of them.

VIII. Acknowledgment and Reflection on the Use of LLMs

I did use LLMs throughout my project. Most of the code used was gathered from in-class examples and not from LLMs or outside sources. I did use LLMs to help with some debugging that I had to do especially with the regression models. I could not figure out why my R^2 for the training data was so low even though it was doing so many passes until I asked ChatGPT. I realized that I was scaling the data even though random forest models do not need scaled data and it was confusing the model. I also used ChatGPT to help explain some of the forms of accuracy of a model. For example, I asked ChatGPT what the MSE of a model actually meant. All in all, I did use ChatGPT to actually help me with some of the coding, most of the time I used it to better understand some of the ideas going on.

IX. References

Raja Ahmed Ali Khan. 2024. Astronomical Data. Retrieved November 25th, 2024 from https://www.kaggle.com/datasets/datascientist97/astronomical-data?resource=download

Devra AI. 2024. Exploring Astronomical Data with Machine Learning, Retrieved December 6th, 2024 from

https://www.kaggle.com/code/devraai/exploring-astronomical-data-with-machine-learning