Project 1 Machine Learning Techniques

Star Type Classification Dataset from Kaggle

Author: Jerrin C. Redmon

Kaggle Link:https://www.kaggle.com/datasets/brsdincer/star-type-classification

Abstract:

This project asks us to compare four machine learning classification techniques. We find a dataset of our choice that will fit appropriately with the four techniques. The four we are comparing our logistic regression, SVC, Decision Trees, and Random Forest. With this four techniques we may choose to run them through a randomized search or voting classifier. For our case the randomized search cross validation technique will be used to determine the best model for this dataset.

Background:

The dataset used for this project is the Star Type Classification dataset from Kaggle. This dataset goes over the different kinds of stars found in our universe and how each are different compared to size, luminosity, and temperature. I choose this dataset since I have an interest in space and I always wondered how different types of stars compare to each other.

Exploratory Analysis:

-The stars dataset contains 240 sample types and 7 columns with no found null variables.

Variable	Data Type
Temperature	Integer
L	Float
R	Float
A_M	Float
Color	Object
Spectral_Class	Object
Type	Integer

Machine Learning Models and Methods:

-Methods

First I imported necessary libraries and looked into the dataset. Everything was in order and no columns needed to be removed. So for the first step of the data cleaning process replaced the categorical variables with dummy variables to keep everything as a number so the machine learning techniques will perform well.

With the data cleaned I needed to prepare the data for the machine learning models. First I set and x and y based on the Type variable which represents each kind of star. After this I split the data which the split train test and proceeded with scaling and encoding the data.

Finally with the data prepared and cleaned, I begin the machine learning techniques. I imported the necessary libraries, prepared a loop of each model with their parameters. With all this done I run the models through a randomized search cross validation to find which score is the best;, then took the scores and placed them into a data frame.

Results:

Model	Best_Score	Best_Params
SVC	0.989609	{'kernel': 'linear', 'C': 10}
Logistic Regression	0.984211	{'solver': 'lbfgs', 'penalty': 'l2'}
Random_Forest	1.0	{'max_depth': 5, 'criterion': 'gini'}
Decision_Tree	0.994872	{'max_depth': 4, 'criterion': 'entropy'}

Conclusion:

-We find that overall that all techniques performed very well. However, Random_Forest performed the best with a score of 1.0