Classification of Exoplanets

Machine Learning

Ishaan Lagwankar - PES1201700150
Richa Sharma - PES1201700662
Malavikka R - PES1201700794
Arpit Agarwal - PES1201701084

Problem Statement

Explore the efficacy of machine learning (ML) in characterizing exoplanets into different classes. Perform a detailed analysis of the structure of the data and propose methods that can be used to effectively categorize new exoplanet samples.

Dataset:
The source of the data used in this work is University of Puerto Rico's Planetary Habitability Laboratory Exoplanets Catalog (PHL-EC).

Preprocessing Data

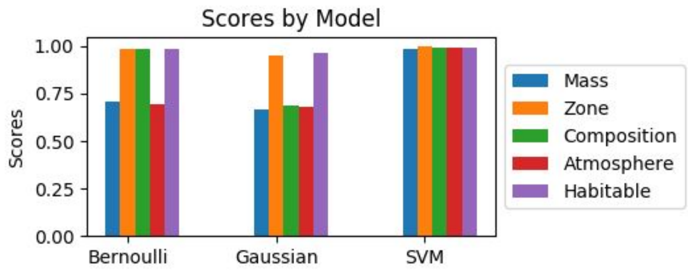
The dataset provided to perform analysis was a collection of observed explored exoplanets, with about 69 categorical and numerical attributes.

For the analysis, 5 output classes were considered.

1. Atmosphere
2. Zone
3. Habitable
4. Mass
5. Composition

Models Trained

1. K Nearest Neighbours
2. Decision Trees
3. Gaussian Naive Bayes
4. Bernoulli Naive Bayes
5. Support Vector Machines



Top 3 models

Accuracy Metrics

Model	Mass	Atmosphere	Composition	Zone	Habitable
K - Means	-	-	-	-	0.9849
Decision Trees	0.3170	0.6221	0.5270	0.0835	-
Bernoulli Naive Bayes	0.7817	0.6954	0.9845	0.9683	0.9832
Gaussian Naive Bayes	0.6645	0.6761	0.6864	0.9496	0.9651
Support Vector Machines	**0.9832**	**0.9935**	**0.9896**	**0.9942**	**0.9858**

Conclusion

- The best model observed was the Support Vector Machine, which gave accuracies of above 95% for each predictive class.
- But the model might be overfitting, as the dataset is inherently skewed, causing sampling to be nullified.
- Due to a high number of class ‘Inhabitable’ examples, the dataset was biased towards the value.
- But in the given conditions, the model performs with great accuracy.
- Please refer to the report for detailed explanations on the models.

Thank you!