Predicting Exoplanets Using Big Data Techniques: A Classification Task
Summary and Conclusions

1. Summary

1.1 Introduction and Aim

The aim of this project was to explore the use of machine learning models for the classification of exoplanet candidates. Exoplanet classification is a critical task in the field of astronomy, with observational data being generated by space missions like NASA's Kepler. This research focused on utilising various machine learning algorithms based on their distinct characteristics and learning paradigms to classify these exoplanet candidates into confirmed exoplanets or false positives.

1.2 Methodology

The dataset used in the project consisted of various astronomical measurements related to exoplanet candidates, which were subjected to a preprocessing pipeline. The preprocessing involved subsetting data, checking for duplicates and outliers and handling missing data. Then, exploratory analysis techniques such as shown in Figure 1, and statistical tests were performed. In preparation for using the data to train machine learning models, the data was normalised to ensure all features contributed equally to the analysis. Then, the most relevant features that would contribute to the model's predictive power were chosen using domain expertise. Due to a class imbalance, the minority class, the confirmed disposition of the dataset had to be oversampled.

The dataset was split into training and testing subsets to evaluate the performance of the machine learning models. The hold-out validation method was employed as the evaluation protocol, as it provided reliable results without the need for more complex methods like K-fold cross-validation. Multiple models were implemented and trained on this dataset. The primary models evaluated were Naive Bayes, Logistic Regression, SVM, Random Forest, and MLP. These models were trained and evaluated using key metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC).

Figure 1: *EDA: Bivariate Analysis of the equilibrium temperature (koi_teq) against the disposition score (koi_score)*

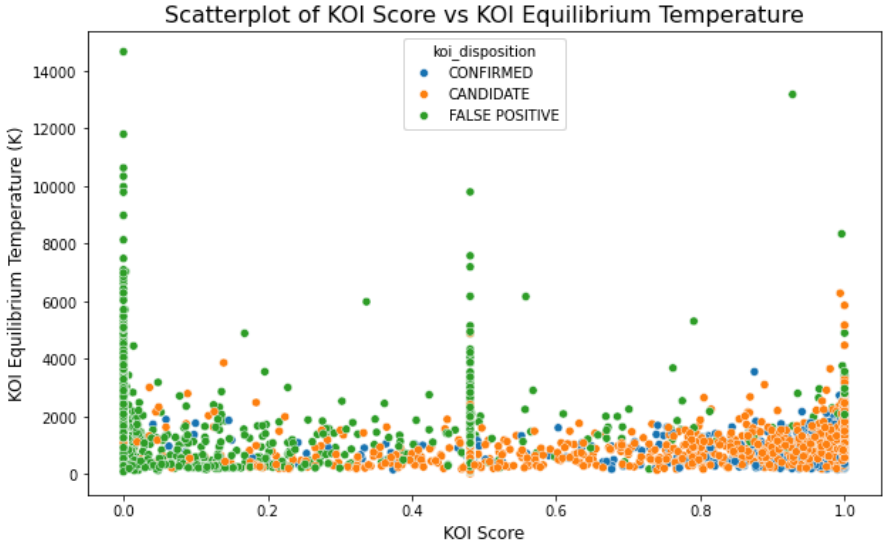
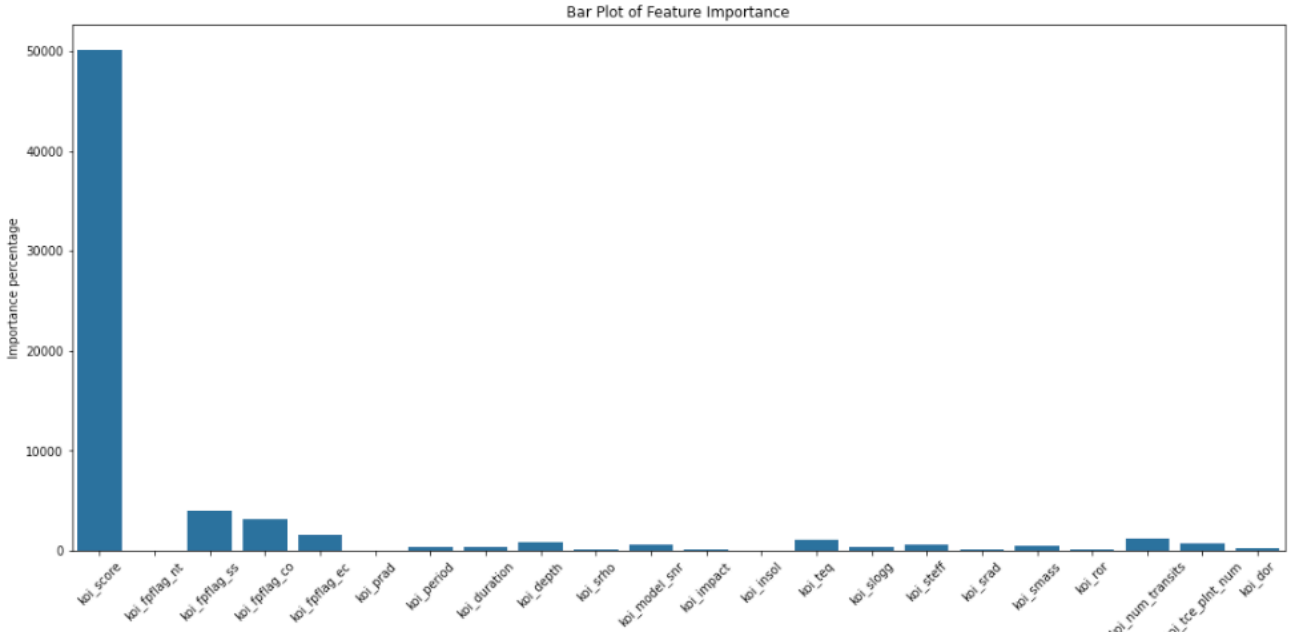


Figure 2: *Feature Selection: Feature importance bar plot*



1.3 Results

The Naive Bayes model served as the benchmark model, in which more complex models are compared to. The performance metrics of this model revealed an accuracy of 80.4%, a precision of 83.3%, a recall of 80.4%, and F1 score of 80%, and AUC of 0.91. This means that the model was performing reasonably well. The Logistic Regression model was a significantly better model with an accuracy, precision, recall, and F1 score of approximately 98.4%, and an AUC of 0.99, reflecting its strong predictive power. The Random Forest model achieved even better performance metrics, achieving an accuracy, precision, recall, and F1 score of approximately 98.8%. It also achieved an AUC of 1.0, indicating perfect model performance. The Support Vector Machine also performed well but not as well as the Logistic Regression or Random Forest models, with an accuracy, precision, recall, and F1 score of approximately 97.7%, and an AUC of 0.99. Finally, the Multilayer Perceptron model also performed exceptionally well having an accuracy, precision, recall and an F1 score of approximately 98.5% and an AUC of 1.0. These results which are outlines in Table 1, indicate that all the models are suitable for accurately classifying this dataset, with the random forest model emerging as the best model.

Table 1: *Performance Metrics of Machine Learning Models*

Model	Accuracy	Precision	Recall	F1 Score	AUC
Naive Bayes	80.4	83.3	80.4	80.9	0.91
Logistic Regression	98.4	98.4	98.4	98.4	1.0
Random Forest	98.8	98.8	98.8	98.8	1.0
Support Vector Machine	97.7	97.7	97.7	97.7	0.99
Multilayer Perceptron	98.5	98.5	98.5	98.5	1.0

The final step involved applying the Random Forest model to a dataset of real-world exoplanet candidates. Predictions made by the model indicated that out of 1,928 candidates, 1,151 were classified as confirmed exoplanets, while 831 were identified as false positives. These results highlight the utility of machine learning in the field of exoplanet discovery, as such predictions can help astronomers prioritise candidates for further investigation.

Table 2: *Features added to the subset candidate dataframe after prediction*

features	probability	predicted_label
(-0.48324562968804785, 0.0, 0.0, 0.0, 5.297507...	[0.8605400009895087, 0.1394599990104914]	FALSE POSITIVE
(0.5207744708144647, 0.0, 0.0, 0.0, 2.96258664...	[0.012505626829781583, 0.9874943731702184]	CONFIRMED
[0.5147443200606958, 0.0, 0.0, 0.0, 0.05243742...	[0.012505626829781583, 0.9874943731702184]	CONFIRMED
[0.3921312547340627, 0.0, 0.0, 0.0, -0.2198082...	[0.0195241394531426, 0.9804758605468574]	CONFIRMED
(0.5217794959400929, 0.0, 0.0, 0.0, 2.69968471...	[0.012505626829781583, 0.9874943731702184]	CONFIRMED

1.4 Discussion

The research successfully addressed the primary aim, which was to use machine learning for exoplanet classification. The models were thoroughly tested, and the Random Forest model stood out due to its high accuracy and ability to make predictions with real-world data. Thus, the project answered the research questions of whether machine learning could be used to classify exoplanet candidates effectively. The project's methodology was robust, employing domain knowledge for feature selection and implementing thorough evaluation metrics allowed for a comprehensive assessment of each model's performance in classifying the exoplanet candidates. This ensured that the selected Random Forest model was based on solid, data-driven decisions.

One of the project's major contributions is its demonstration of how machine learning can be used to deal with large astronomical datasets effectively. Given the increasing amount of exoplanet data from missions such as Kepler, TESS, and others, the need for fast, automated classification systems becomes more urgent. The performance of many of the models not only proved them to be suited for identifying confirmed exoplanets but also highlighted the ability of machine learning to distinguish between true exoplanet discoveries and false positives, which can significantly aid in refining the focus of astronomers.

2. Conclusion

In conclusion, this project highlights the immense potential of machine learning in astronomy, especially in the discovery and classification of exoplanets. The application of machine learning in this domain will likely continue to grow, driven by the increasing availability of astronomical data and the need for faster, more accurate analytical tools. By improving both the speed and precision of exoplanet discovery, machine learning can significantly enhance our understanding of the universe and aid in identifying new planets outside our solar system.

Future work in exoplanet classification using big data analysis and machine learning can focus on improving data quality and feature selection by utilising automated techniques and refining datasets. Addressing class imbalance and exploring advanced models, such as XGBoost or deep learning, could enhance performance for large datasets. Incorporating time-series data and integrating astrophysical models could further improve classification accuracy. Validation through real-world observations remains essential. Transfer learning could adapt models to new datasets, and collaboration with citizen science initiatives might enhance model robustness. Lastly, real-time prediction systems could allow for immediate analysis of space mission data, streamlining the discovery process.

Word count: 977 (without Table 1)