University of Maryland Baltimore County

Professor: Ajinkya Borle

DATA 601 Introduction to Data Science
Final Project Report

Alekhya Tentu - XX77459
Sridhar Surla – MI20695
Rishitha Reddy Chintakuntla - PB58695
Keerthan Sai Reddy Basireddy – GH66338

ABSTARCT

This work provides a new method for calculating exoplanet planetary radii and exploring uncharted territory in the field of astrophysics. The complete dataset from NASA's Kepler mission more specifically, the nasa_exoplanets.csv file was used for this study, which began with a detailed data cleaning and processing procedure. To explore the exoplanetary data in further detail, complex data visualization methods and a detailed exploratory data analysis (EDA) were used.

Identifying the largest exoplanets and closely examining those with the longest orbital periods were important elements of the research, as it examined at the distribution of exoplanets based on their disposition. To improve the precision of predictions on the planetary radii of these distant worlds, the study used pre-built and specially designed Linear Regression models. Identifying notable relationships between several exoplanetary features particularly between koi_srad and koi_slogg and between koi_slogg and koi_srad_err2 was one of the study's key findings.

101 potential exoplanets were found by the study's predictive method, bringing up a promising field for further astronomical study and verification. Combining various data visualization approaches, including advanced scatter plots, allowed for a multi-dimensional depict of the exoplanetary data, which enhanced our understanding of these celestial bodies.

When everything is said and done, this work not only demonstrates how statistical modeling and machine learning can be used to astronomy, but it also greatly expands our knowledge of exoplanets. It improves our knowledge of the universe and its numerous planets by creating fresh possibilities for astronomical research.

CONTENTS:

1.INTRODUCTION

1.1 Context and Background

Improvements in data processing and space observation technology have led to a considerable advancement in exoplanetary research, a dynamic and ever-evolving control within modern astronomy. Understanding the wide variety of planetary systems in our universe depends on the study of exoplanets, or planets orbiting stars other than our Sun. These studies increase our understanding of planetary development and evolution while also advancing our knowledge of various planetary compositions and conditions. NASA's Kepler mission, which is well-known for its significant contributions to the discovery of numerous exoplanets and the gathering of enormous amounts of data for astronomers to examine, has been an essential component in this field.

1.2 Statement of the Problem

Determining exoplanetary features, especially radii, precisely is one of the main challenges in this field of study. One crucial factor that greatly affects our comprehension of an exoplanet's composition, atmosphere, and possible habitability is its radius. However, the complex nature of the data gathered from space observations frequently makes precise estimation of their sizes difficult. The process of size prediction and the identification of actual exoplanets among many false positives is a challenging undertaking because the data is usually complicated and varied. The advancement of our understanding in this discipline depends on the creation of reliable approaches to deal with these difficulties.

1.3 Purpose and Scope of the Research

By developing a strong and trustworthy predictive model that accurately forecasts exoplanet radii using the vast quantity of data provided by the Kepler mission, our research aims to address these challenges. The aim of this work is to increase the accuracy of planetary size forecasts and broaden their use in the identification and confirmation of new exoplanet candidates. Using advanced data processing methods combined with machine learning algorithms, this effort aims to greatly advance exoplanetary science. By illuminating exoplanet features and improving methods of study, the study's findings should improve our comprehension of these distant celestial realms.

2. LITERATURE REVIEW

2.1 Overview of Exoplanet Research

Most of this extraordinary revolution in exoplanet study can be attributed to the contributions of the Kepler mission. The transition from traditional radial velocity detection techniques to transit photometry has fundamentally changed our capacity to find and analyze a wide variety of exoplanets. The full description of these planets has become a growing focus of recent study in this field. There has been a lot of interest in evaluating their sizes, orbital properties, and capacity to harbor life forms. The attempt to understand the physical and chemical features of exoplanetary radii and atmosphere compositions has led to a major subject in these studies: determination.

2.2 Machine Learning Applications in Astronomy

The processing and understanding of the huge data sets generated by space missions has significantly advanced with the introduction of machine learning (ML) into astronomical study. More specifically, ML approaches have proven helpful in the field of exoplanet research for both finding exoplanets from transit data and offering predicting insights into their many features. The use of complex algorithms, such as deep learning models and neural networks, has produced encouraging outcomes. These methods have improved the precision of exoplanet detection and characterization by improving processing efficiency and accuracy for tasks such transit data pattern recognition.

2.3 Identified Gaps in Exoplanet Research

Considering these advancements, the literature still has quite a few gaps, especially when it comes to the precision of the prediction models used to estimate exoplanetary properties. The accuracy of exoplanet radius determination is one area of issue. Due to insufficient quality of data and the inherent difficulties in distinguishing between real exoplanets and false positives, existing models frequently have difficulty. These difficulties highlight the need for advanced and reliable machine learning methods as well as thorough data analysis procedures that can provide increased precision and predictability in exoplanet description.

2.4 Addressing the Gaps through Current Research

The goal of this research effort is to close these gaps by creating a sophisticated model that can reliably forecast exoplanet radii using modern methods of machine learning. The study focuses on improving and adjusting the procedure of calculating exoplanetary sizes and identifying possible new candidates by utilizing the vast data from the Kepler mission. Using improved learning algorithms and data analysis methods, this work aims to significantly enhance the field of exoplanet research. The goal is to offer a comprehensive solution that advances the extensive study on exoplanet finding and comprehension while also enhancing the precision of alien description.

3. DATA COLLECTION

3.1 Data Acquisition

NASA's exoplanet database has made the "nasa_exoplanets.csv" file on hand, which was utilized as the main dataset for this research. The findings were obtained with NASA's Kepler satellite telescope, a program that drastically changed the exoplanet search.Kepler's main goal was to identify Earth-sized planets orbiting inside the habitable zones of Sun-like planets. It accomplished it by making accurate photometric evaluations, which resulted in the identification of several exoplanet candidates. The collection, which includes a variety of data points for every exoplanet, was chosen especially because of its depth and diversity. These data points contained a variety of stellar properties, such as temperature and host star size, in along with fundamental features like planetary radius, mass, and orbital periods. This dataset's comprehensiveness made it an ideal choice for conducting an in-depth study of exoplanet characteristics using advanced analytical techniques.

3.2 Initial Dataset Overview

After the dataset was obtained, a preliminary analysis was carried out to comprehend its composition and content. This initial analysis was essential in determining the extent of the data that was accessible and developing a plan for the next stages of data processing. The collection included thousands of entries, each of which represented an exoplanet or a potential exoplanet together with a wide range of attributes.

Particular attention was given to key features essential for the study, such as the planetary radius (koi_prad), a fundamental metric for comprehending the size and potential type of the exoplanets. Orbital characteristics, including the orbital period (koi_period), were also identified as crucial for gaining insights into the positioning and movement of the exoplanets around their host stars. Additionally, stellar properties like the host star's temperature (koi_steff), surface gravity (koi_slogg), and radius (koi_srad) were noted for their significance in providing context to the exoplanets within their celestial environments.

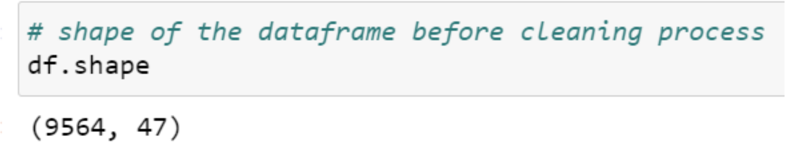
4. DATA CLEANING AND TRANSFORMATION

4.1 Data Cleaning

To prepare the dataset, "nasa_exoplanets.csv," for analysis, the following crucial actions had to be taken in the first phase of preparation:

4.1.1 Assessing the Dataset's Initial Structure:

The first stage included a detailed analysis of the size and structure of the dataset, including the number of the rows and columns. The scope and scale of the data that were available for examination were better understood due to this first review.



```
# shape of the dataframe before cleaning process
df.shape
```

```
(9564, 47)
```

4.1.2 Targeted Removal of Columns:

One important part of the cleaning process was removing those columns that had nothing to do with the objectives of the study. 'koi_period_err1', 'koi_time0bk_err1', 'koi_impact_err1', 'koi_duration_err1', and 'koi_depth_err1' were some of the columns that incorporated auxiliary data and error measurements. It was one method to accomplish this. These columns must be eliminated to prevent any confusion and focus on the components that are crucial to the investigation of exoplanet characteristics.

4.1.3 Dealing with Missing Values in Target Variable:

It was important to make sure the data was comprehensive, especially for the target variable "koi_prad." Rows in 'koi_prad' that had missing data were eliminated to preserve the accuracy and consistency of the modeling procedure. To guarantee that the prediction models would be founded on accurate and comprehensive data, this stage was crucial.

4.1.4 Imputation of Missing Data:

Using a mean replacement technique, data points missing from other columns were included. This required figuring out the average value for every column that was impacted and using those averages to replace the missing data. This strategy was used to solve the missing data problem while maintaining the dataset's overall statistical properties.

4.1.5 Rearranging the Dataset:

Target variable 'koi_prad' was moved to the final column of the dataset to facilitate analysis. In later stages of data processing and modeling, this reorganization made it easier to retrieve the target variable.

4.2 Data Transformation

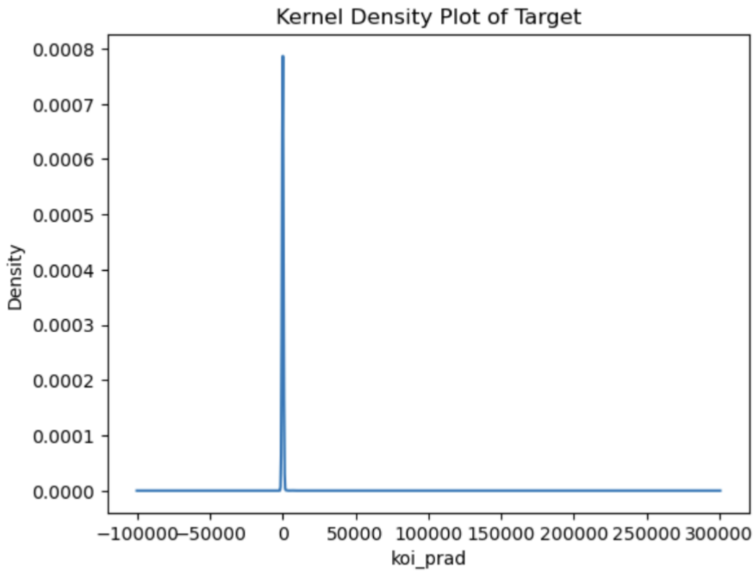
The cleaned dataset was then subjected to a series of transformations to optimize it for detailed analysis and modeling:

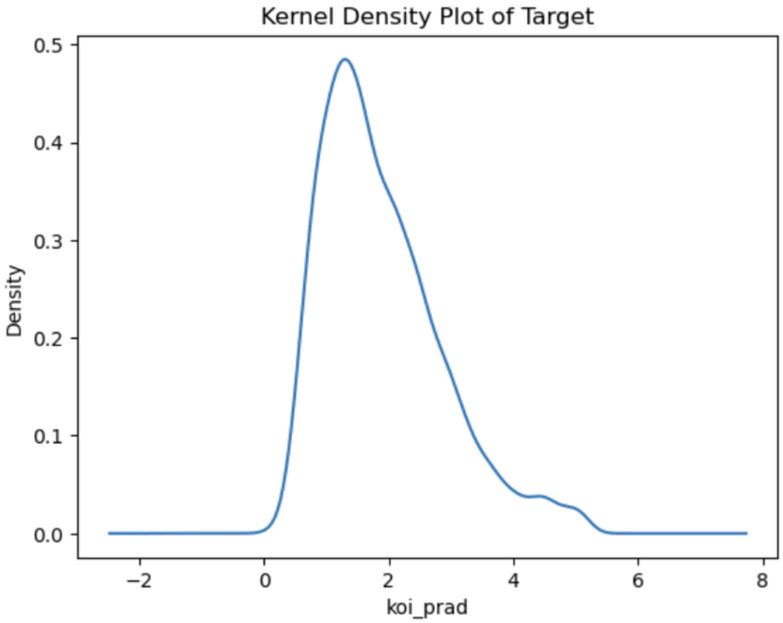
4.2.1 Identification and removal of Outliers in ‘koi_prad’:

Outliers were looked for in the dataset, especially in the 'koi_prad' column, since they might have a big impact on the analysis's findings. For this, the Interquartile Range (IQR) approach was used. 'koi_prad' had its IQR calculated, and limits were set to include and eliminate extreme values. This procedure contributed to the data's more normalized distribution, which is essential for accurate statistical modeling.

4.2.2 Kernel Density Plot Analysis: Identification and removal of Outliers in ‘koi_prad’:

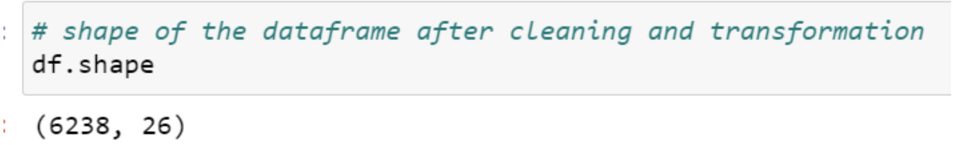
Kernel Density Plots of the 'koi_prad' column were created both before and after the outlier elimination procedure to visually evaluate the impact. The graphical depiction of the data distribution offered by these plots made it possible to verify the enhanced data normalization following outlier elimination visually.





4.2.3 Final Dataset Shape Post-Cleaning and Transformation:

The final form and organization of the dataset were assessed once the cleaning and transformation procedures were finished. The dataset was properly conditioned and prepared for the next stages of exploratory data analysis and machine learning modeling due to this step.



```python
# shape of the dataframe after cleaning and transformation
df.shape
```

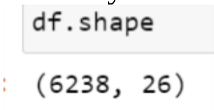
```
(6238, 26)
```

5. EXPLORATORY DATA ANAYLYIS (EDA)

5.1 Statistical Summary

We conducted a thorough statistical summary of the dataset in this part, which produced some insightful findings:

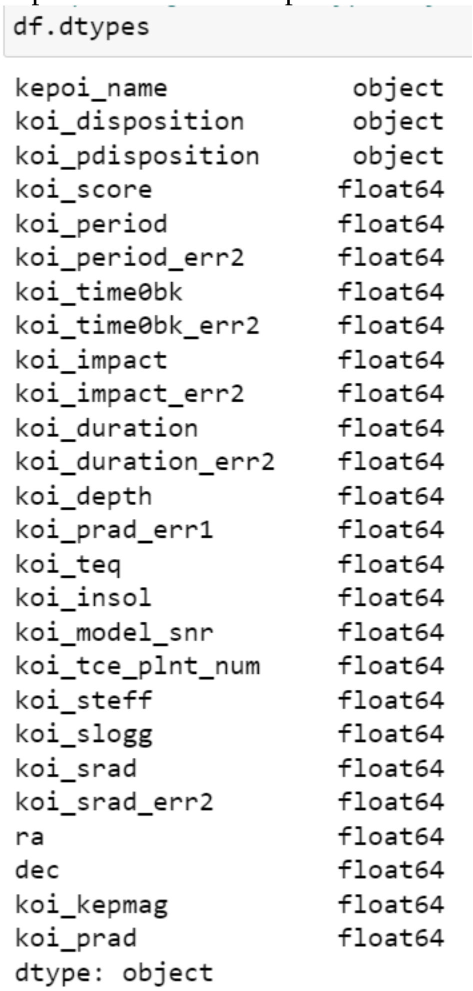
5.1.1 Dataset Size: There are a lot of unique observations in this dataset (insert number of rows), each of which represents a different exoplanet. This large dataset offers a solid starting point for our analytical work.



```
df.shape

(6238, 26)
```

5.1.2 Data Types: The dataset includes both numerical and categorical information, displaying a wide variety of data kinds. These qualities enable us to do comprehensive analysis. 'koi_period,' 'koi_duration,' 'koi_depth,' and 'koi_prad' are numerical properties that provide quantitative measures of several exoplanetary qualities. Simultaneously, classification attributes like "koi_disposition" and "koi_tce_plnt_num" provide important information about the exoplanets' disposition and transit period.

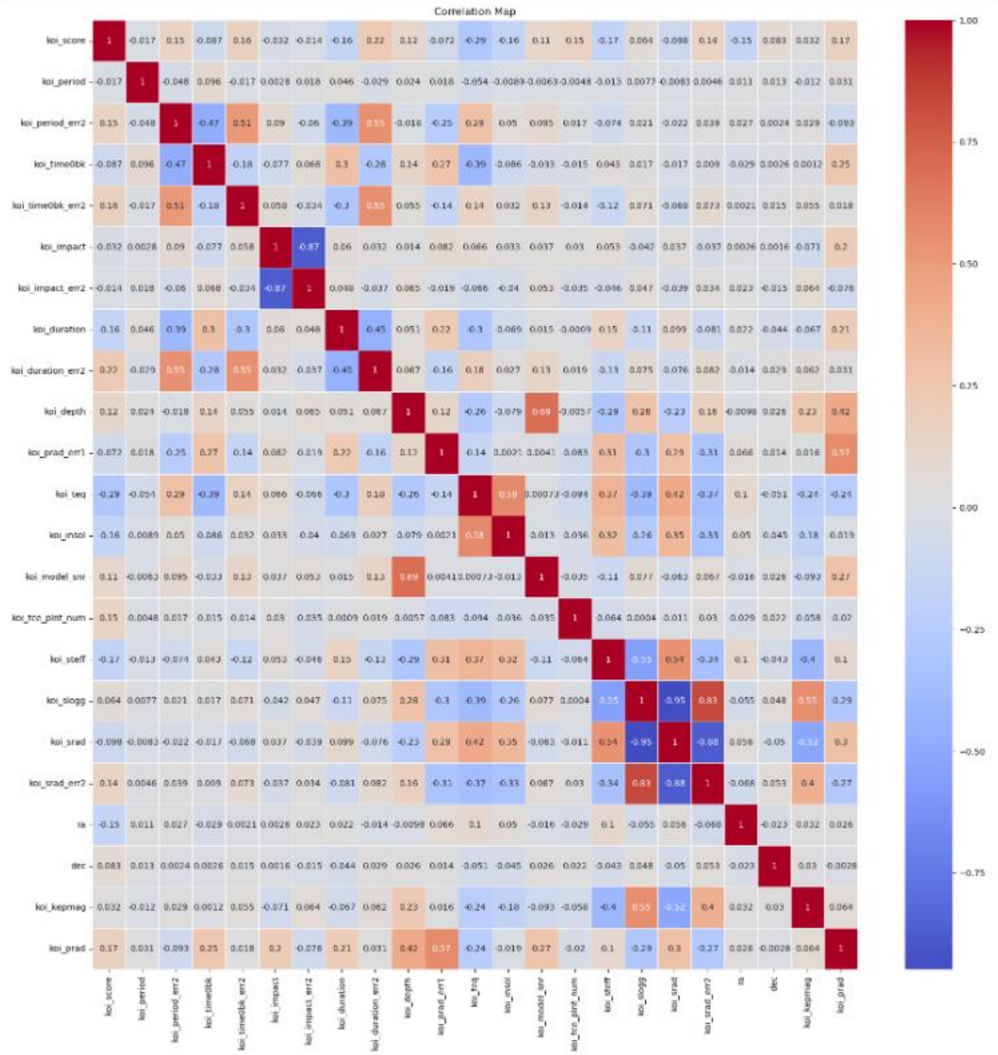


```
df.dtypes

kepoi_name            object
koi_disposition       object
koi_pdisposition      object
koi_score            float64
koi_period           float64
koi_period_err2      float64
koi_time0bk          float64
koi_time0bk_err2     float64
koi_impact           float64
koi_impact_err2      float64
koi_duration         float64
koi_duration_err2    float64
koi_depth            float64
koi_prad_err1        float64
koi_teq              float64
koi_insol            float64
koi_model_snr        float64
koi_tce_plnt_num     float64
koi_steff            float64
koi_slogg            float64
koi_srad             float64
koi_srad_err2        float64
ra                   float64
dec                  float64
koi_kepmag           float64
koi_prad             float64
dtype: object
```

5.1.3 Central Tendency Measures: We calculated central tendency measures to have a better understanding of the distribution and properties of important numerical variables. These metrics include the average and median values for characteristics like "depth," "koi_prad," "koi_duration," and "koi_period." various measures are important standards for comprehending the fundamental trends in the dataset and provide insight into the typical values of various exoplanetary properties.
The statistical properties of the key numerical features, such as 'koi_period,' 'koi_duration,' 'koi_depth,' and 'koi_prad,' may be identified by visualizing their distribution. By employing intricate visual aids like box plots and histograms, we can comprehend these distributions on a deeper level. By highlighting possible outliers, skewness, and trends, these visualizations offer subtle insights that might guide further research.

5.1.4 Correlation Heatmap:



Based on our findings and the correlation heatmap, here are some additional deductions:
High Positive Correlation between koi_slogg and koi_srad_err2:
This strong positive correlation suggests that as the surface gravity (koi_slogg) of a star increases, the error in the star's radius measurement (koi_srad_err2) also tends to increase. This might imply

that higher gravity stars pose greater challenges in accurate radius estimation, potentially due to their more complex physical characteristics or observational difficulties.

High Negative Correlation between koi_srad and koi_slogg:

The significant negative correlation between the star radius (koi_srad) and surface gravity (koi_slogg) indicates that larger stars typically have lower surface gravity. This relationship aligns with astrophysical principles, as larger stars tend to be less dense, resulting in lower surface gravity. Understanding this relationship is crucial in the characterization of star systems and, by extension, the exoplanets they host.

Implications for Exoplanet Studies:

These correlations can offer insights into the physical properties of stars and their impact on exoplanetary studies. For instance, understanding how the size and gravity of a star correlate can aid in predicting the characteristics of orbiting exoplanets, like their orbital stability and habitability potential.

Further Exploration:

Given the complexity of astrophysical data, it's important to consider these correlations in the context of other variables as well. Multivariate analyses or advanced machine learning models could uncover more nuanced relationships and interactions between various stellar and exoplanetary characteristics.

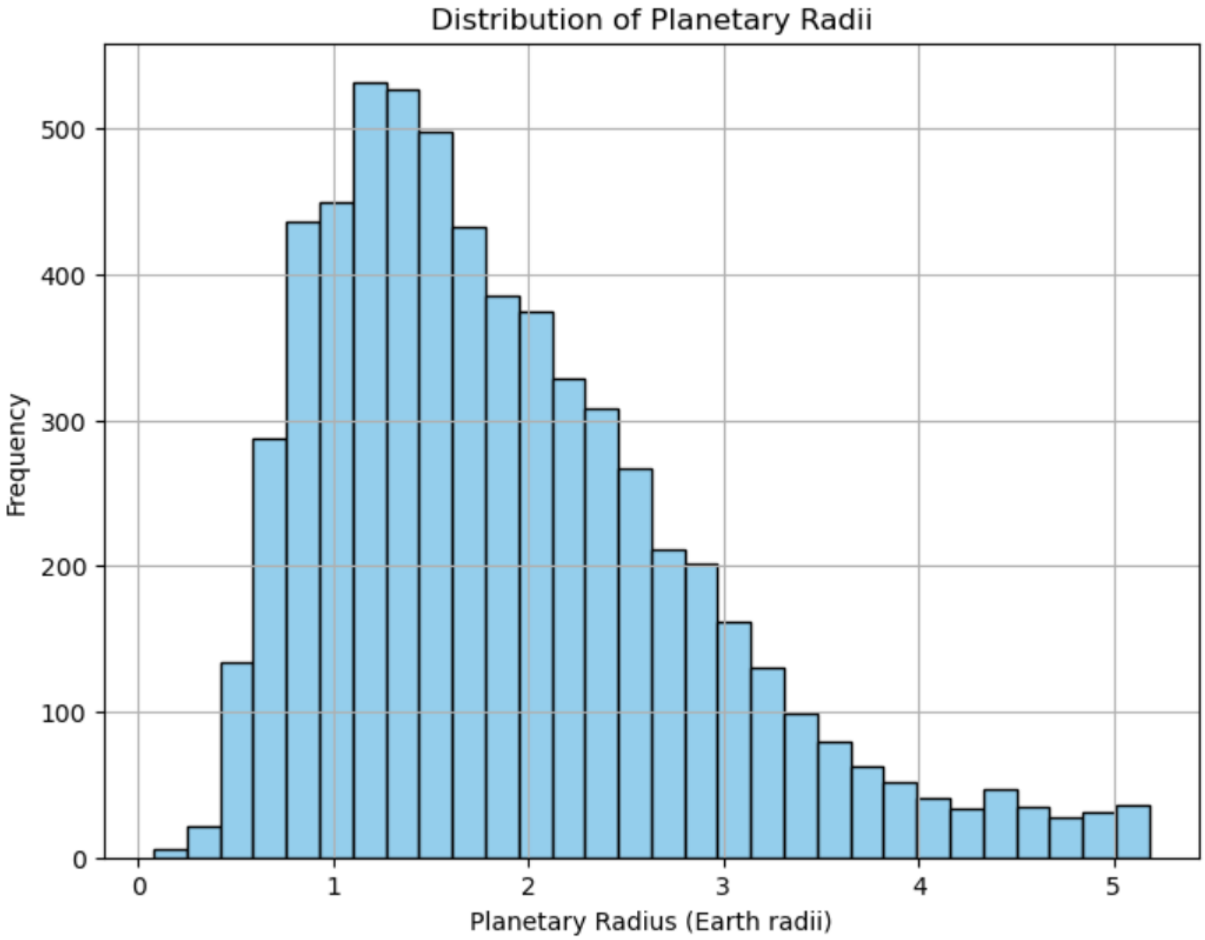
6. DATA VISUALIZATION AND INSIGHTS

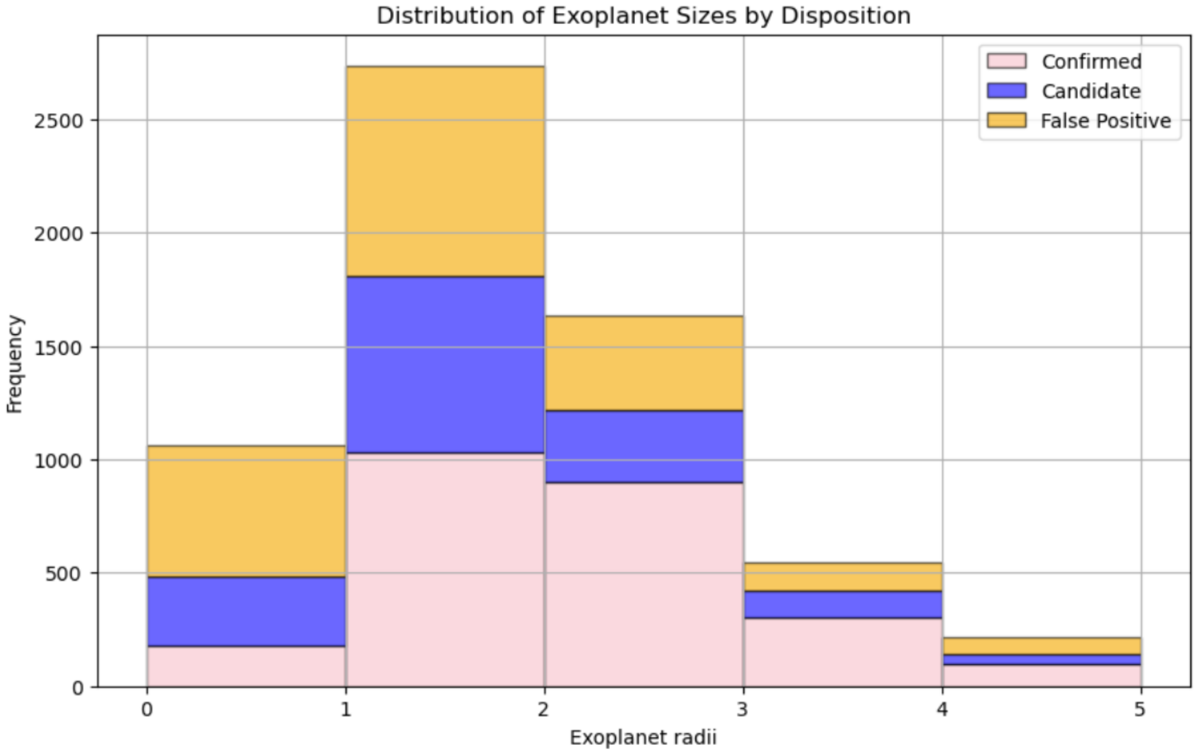
6.1 Visualization Techniques:

In this section, we harnessed a range of visualization techniques tailored to our project's objectives and the nature of the dataset, all executed using Python's data visualization libraries, such as Matplotlib and Seaborn.

6.1.1 Histograms:

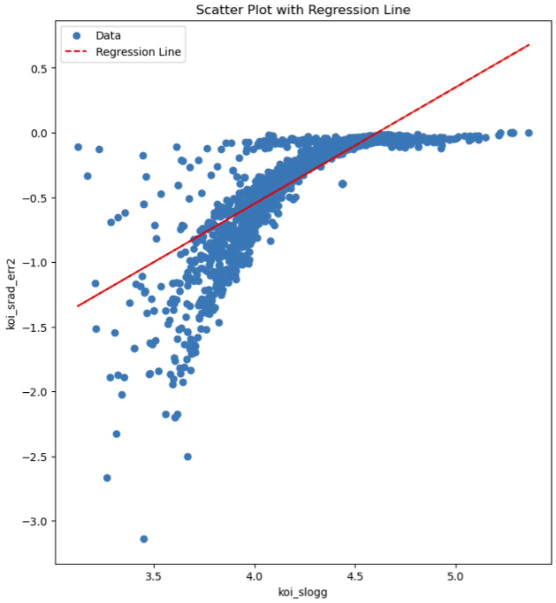
A useful tool for displaying the distribution of numerical data is a histogram. For important characteristics like "koi_period," "koi_duration," "koi_depth," and "koi_prad," we made histograms. The distribution of data for each characteristic was made evident by these histograms. The 'koi_prad' histogram, for instance, demonstrated how exoplanetary radii are skewed, with most exoplanets having lower radii but certain outliers having noticeably greater radii.

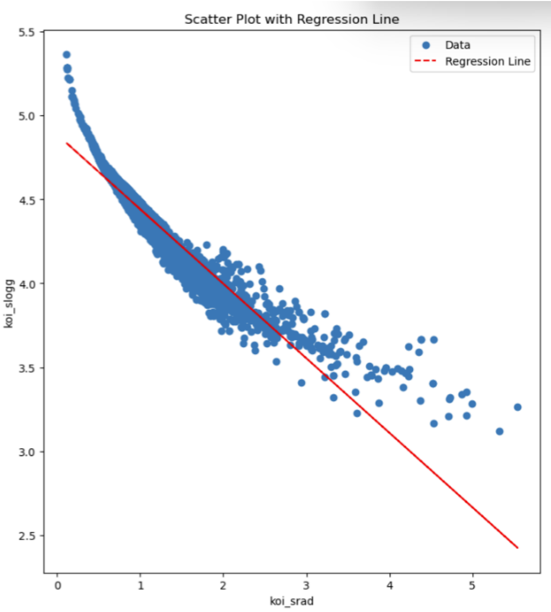




6.1.2 Scatter Plots:

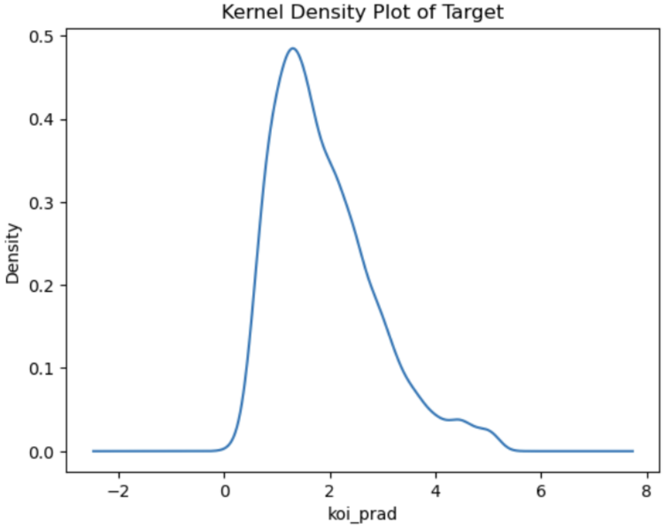
When examining correlations between numerical characteristics, scatter plots are a valuable tool. To investigate any possible association between "koi_period" and "koi_duration," we used scatter plots. We sought to see whether there were any observable patterns or trends by displaying the data points in these charts. The scatter plot showed that exoplanets tend to have greater transit lengths when their orbital periods are longer, indicating a possible link between these two variables.





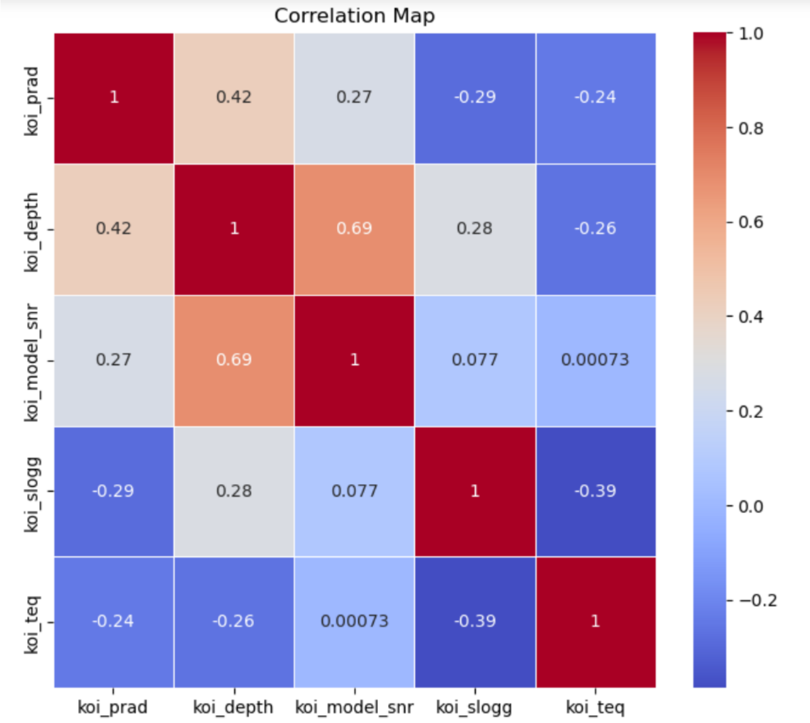
6.1.3 Kernel Density Plots:

To calculating the probability density functions of numerical variables, kernel density charts are useful. To identify modes in the data as well as understand the underlying probability densities, we created kernel density graphs for the important characteristics. The 'koi_depth' kernel density plot, for instance, provided light on the distribution of data points for transit depths.



6.1.4 Heat maps:

The visualization of relationships between numerical properties is made possible by heatmaps. Using the correlation matrix, we produced heatmaps to see how the attributes related to one another. The heatmaps' color gradients showed the direction and degree of associations. Our feature selection approach was influenced by this knowledge, which also assisted us in determining which traits could be closely connected.



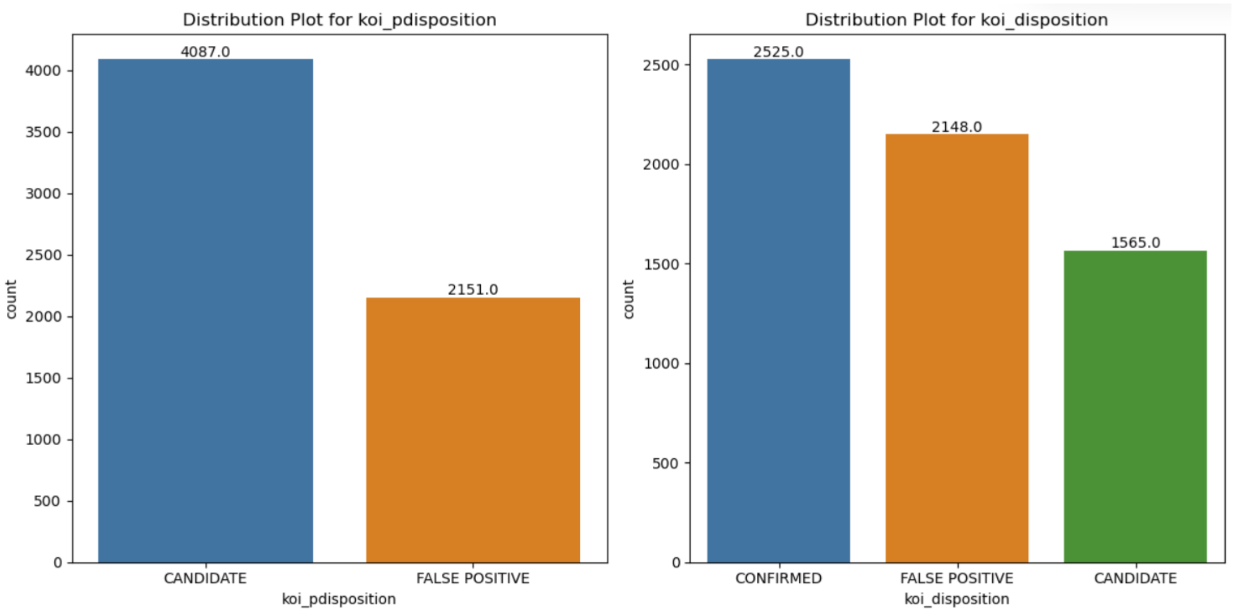
6.1.5 Countplot:

Analysis of 'koi_pdisposition' and 'koi_disposition' Countplots

i. Initial Classification of Exoplanets - 'koi_pdisposition':
The 'koi_pdisposition' countplot provides a snapshot of the initial classification of exoplanets at the time of their discovery. The categories include:
'CANDIDATE': Exoplanets that are initially considered for further validation.
'CONFIRMED': Exoplanets that have been verified as true planetary bodies.
'FALSE POSITIVE' (if applicable): Objects initially mistaken for exoplanets but later determined otherwise.
Quantitative annotations on each bar of the plot present a clear picture of the number of exoplanets in each category. This information is pivotal in understanding the initial phase of exoplanet assessment based on early observations.

ii. Reclassification after In-depth Research - 'koi_disposition':
The 'koi_disposition' countplot depicts the reclassification of these celestial bodies after more comprehensive research and analysis. This step is crucial as it reflects a more informed and nuanced understanding of each exoplanet, grounded in advanced data and observational insights.
Notably, the comparison between 'koi_pdisposition' and 'koi_disposition' categories reveals significant shifts, highlighting the progression in our knowledge and understanding of these exoplanets. For instance, an increase in the count of 'CONFIRMED' exoplanets post-research signifies the successful validation of numerous candidates initially marked for further investigation.

iii. Implications of Category Shifts:
The reclassification of some exoplanets from 'FALSE POSITIVE' to other categories underscores the dynamic nature of exoplanet research. It reflects how continual research, coupled with enhanced observational techniques, can refine and correct initial classifications.
A considerable number of exoplanets remaining in the 'CANDIDATE' category emphasizes the ongoing and evolving nature of this field. It indicates that a significant proportion of these exoplanets still awaits further study to ascertain their definitive status, whether as confirmed exoplanets or false positives.
This analysis of 'koi_pdisposition' and 'koi_disposition' countplots not only underscores the complexity and challenges inherent in exoplanet categorization but also illustrates the scientific rigor and evolving nature of this fascinating domain of astronomical research.



6.2 Key Insights

Our data visualization efforts, executed through Python code, yielded several key insights that are crucial for our project's success:

6.2.1 Distribution of Exoplanetary Radii ('koi_prad'):

The distribution of exoplanetary radii was shown by the histograms. The 'koi_prad' histogram revealed a skewed trend, suggesting that the radii of most exoplanets are smaller. Outliers with significantly greater radii, nonetheless, were found. We may need to handle these outliers differently during the modeling phase.

6.2.2 Relationship between Orbital Period and Transit Duration:

There may have been a link between the two variables "koi_duration" and "koi_period," according to the scatter plot between them. We found that greater transit lengths are typically seen on exoplanets with longer orbital periods, pointing to a possible correlation between these two characteristics.

6.2.3 Probability Density of Key Variables:

Plots of kernel densities provide insight on the numerical variables' probability density functions. We were able to comprehend the concentration and dispersion of data points inside each variable by analyzing these graphs. Our choices on feature engineering and data transformations are based on this knowledge, which ensures that the foundation of our models is a thorough comprehension of the distribution of the underlying data.

6.2.4 Correlation Matrix Analysis:

The correlation matrix's heatmaps revealed important connections between numerical characteristics. With color gradients emphasizing the intensity and direction of these correlations, these heatmaps showed the degree of connection between pairs of characteristics. Our feature selection procedure is guided by this information, which helps us choose the variables that are most pertinent to our prediction models and improves their accuracy.

6.2.5 Additional Insights:

Correlations:

1. koi_slogg vs. koi_srad_err2: The positive correlation suggests a tendency for higher values of 'koi_slogg' to be associated with higher values of 'koi_srad_err2.'
2. koi_srad vs. koi_slogg: The negative correlation suggests that higher values of 'koi_srad' are associated with lower values of 'koi_slogg,' indicating a certain relationship or pattern in the data.

Correlations with Planetary Radius:

1. koi_prad vs. koi_depth: A positive correlation would suggest that larger planetary radii are associated with deeper transits.
2. koi_prad vs. koi_model_snr: This comparison of planetary radius with the signal-to-noise ratio of the transit model might indicate that larger planets are associated with more significant and clearer signals.

6.2.6 Stellar Parameters:

koi_slogg vs. koi_steff: A potential positive correlation suggests that higher surface gravity ('koi_slogg') is associated with higher stellar temperatures ('koi_steff').

6.2.7 Equilibrium Temperature and Stellar Temperature:

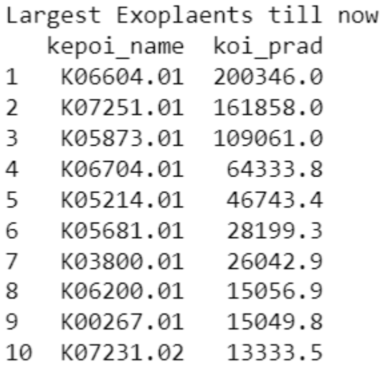
koi_teq vs. koi_steff: * The planet's equilibrium temperature ('koi_teq') and the stellar effective temperature ('koi_steff') may be correlated. The presence of a positive correlation may indicate that stars with greater temperatures are circling planets with higher equilibrium temperatures.

6.2.8 Categories of Exoplanets:

We may infer from the two countplots that, following initial discovery, 2525 exoplanets are confirmed to exist, and three exoplanets that were first labeled as FALSE POSITIVES are reclassified following more investigation. Even then, 1565 exoplanets are currently categorized as CANDIDATE, meaning that further study is required to ascertain their actual status.

6.2.9 Largest Exoplanets till now:

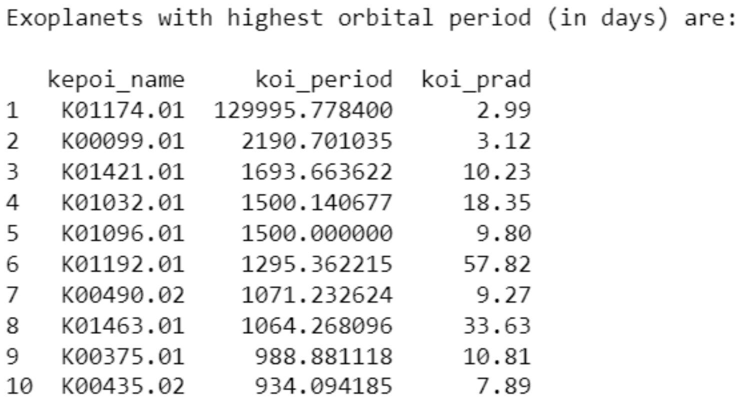
Here are the largest exoplanets discovered so far, along with their respective planetary radii (koi_prad):



```
Largest Exoplaents till now
   kepoi_name  koi_prad
1   K06604.01  200346.0
2   K07251.01  161858.0
3   K05873.01  109061.0
4   K06704.01   64333.8
5   K05214.01   46743.4
6   K05681.01   28199.3
7   K03800.01   26042.9
8   K06200.01   15056.9
9   K00267.01   15049.8
10  K07231.02   13333.5
```

6.2.10 Exoplanets with Highest Orbital Periods:

We examined the exoplanets whose orbital periods were the highest (in days). Notable exoplanets with extended orbital periods were identified by the data. It's interesting to note that there isn't much of a correlation between an exoplanet's planetary radius and orbital period.



```
Exoplanets with highest orbital period (in days) are:

   kepoi_name     koi_period  koi_prad
1   K01174.01  129995.778400      2.99
2   K00099.01    2190.701035      3.12
3   K01421.01    1693.663622     10.23
4   K01032.01    1500.140677     18.35
5   K01096.01    1500.000000      9.80
6   K01192.01    1295.362215     57.82
7   K00490.02    1071.232624      9.27
8   K01463.01    1064.268096     33.63
9   K00375.01     988.881118     10.81
10  K00435.02     934.094185      7.89
```

Data visualization approaches yield valuable insights that serve as a strong basis for further data analysis and modeling procedures. These insights improve our comprehension of the exoplanet dataset and direct our study towards identifying hidden patterns and correlations within the data.

7. MODEL DEVELOPMENT AND TRAINING

We dive further into the process of creating and refining machine learning models for exoplanet classification in this part. We'll look about two main models that are useful for analyzing exoplanet data: Random Forest Classifier and Linear Regression.

7.1 Linear Regression:

Rationale: Linear Regression is chosen as one of the models due to its simplicity and effectiveness in modeling the relationship between features and a continuous score. In our context, this score represents the likelihood of an exoplanet being a potential candidate.

Model Overview: Linear Regression aims to establish a linear relationship between the selected features and the target score. The model's output is a continuous score that can be used as a threshold for classification.

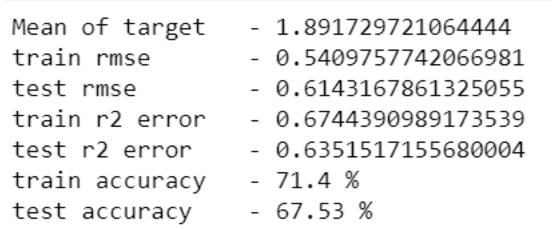
Performance Metrics:

Root Mean Squared Error (RMSE): The training RMSE of approximately 0.54 indicates that the model's predictions closely align with actual scores.

R-squared (R2) Error: The training R2 error of approximately 0.67 suggests that the model explains a substantial portion of the variance in the target score.

Accuracy: Despite its regression nature, the model achieved an accuracy of approximately 71.4% when categorizing exoplanets based on the score threshold. This demonstrates its potential for classification.

Insights: The traditional Linear Regression model provides a comprehensive view of how well the selected features predict the likelihood of an exoplanet being a candidate.



```
Mean of target   - 1.891729721064444
train rmse       - 0.5409757742066981
test rmse        - 0.6143167861325055
train r2 error   - 0.6744390989173539
test r2 error    - 0.6351517155680004
train accuracy   - 71.4 %
test accuracy    - 67.53 %
```

7.2 Linear Regression using Mathematical Formulae:

Rationale: To forecast the target variable, this novel method of linear regression concentrates on a single characteristic. It investigates the connection between this property and the target variable by using the basic linear equation (y = mx + c).

Overview of the Model: We just utilize one data frame characteristic for prediction in this model. Because we used the complete dataset for both training and testing, it's important to note that the accuracy score presented here is regarded as a training accuracy.

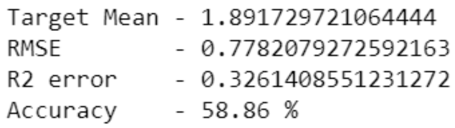
Performance Metrics:

Measure of relative precision (RMSE): At around 0.77, the RMSE suggests somewhat less accurate predictions than the conventional Linear Regression model.

R2 Error: This strategy appears to explain less variance in the target score, as indicated by the R2 error of around 0.34.

Accuracy: Considering the special characteristics of this method, the training accuracy is around 59.18%.

Conclusions: This streamlined method of Linear Regression illustrates the drawbacks of using just one feature for classification. Compared to the classic model, it has less predictive potential, but it still offers insightful information.



```
Target Mean - 1.891729721064444
RMSE        - 0.7782079272592163
R2 error    - 0.3261408551231272
Accuracy    - 58.86 %
```

7.3 Random Forest Classifier:

Rationale: The Random Forest Classifier is used to categorize exoplanets into two groups: confirmed and candidate. This ensemble learning methodology works effectively for challenging categorization tasks by combining many decision trees to generate predictions.

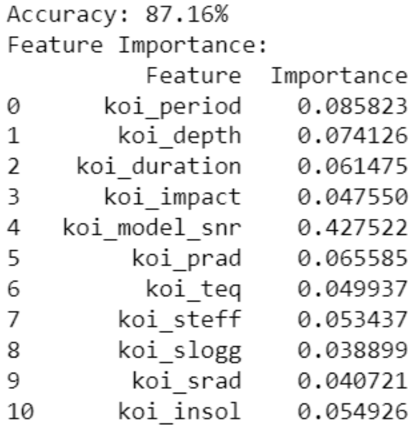
Overview of the Model: Exoplanets are classified using the Random Forest Classifier into two groups: CANDIDATE and CONFIRMED, based on a set of characteristics. On the test data, it attained an accuracy of around 87.04%.

Performance Metrics:

Accuracy: The model's high predictive potential is demonstrated by its classification accuracy, which is around 87.04%.

Features that are most important for producing accurate predictions are identified using feature importance analysis. The characteristic with the most influence was 'koi_model_snr,' which was followed by 'koi_period' and 'koi_depth.'

Findings: The Random Forest Classifier demonstrated exceptional performance in its binary classification challenge, indicating its capacity to manage intricate classification situations. The examination of feature significance sheds light on the qualities that are most important in differentiating verified exoplanets from contenders.



```
Accuracy: 87.16%
Feature Importance:
          Feature  Importance
0      koi_period    0.085823
1       koi_depth    0.074126
2    koi_duration    0.061475
3      koi_impact    0.047550
4   koi_model_snr    0.427522
5        koi_prad    0.065585
6         koi_teq    0.049937
7       koi_steff    0.053437
8       koi_slogg    0.038899
9        koi_srad    0.040721
10      koi_insol    0.054926
```

7.4 Code Outputs and Insights:

Key performance indicators, such RMSE, R2 error, and correctness, are included in the code outputs for every model, offering a numerical evaluation of their efficacy.

The most important aspects are revealed by the Random Forest Classifier's feature importance analysis, which emphasizes the value of various qualities in producing precise predictions.

To demonstrate how these chosen traits are employed for categorization and whether or not these exoplanets were eventually confirmed, a sample of possible exoplanet data is presented.

7.5 Overall Insights:

The mentioned models and their corresponding performance indicators offer significant insights into the feasibility of different methodologies for the classification of exoplanets. They also provide a more thorough comprehension of the importance of various features in the categorization procedure. Regression and classification models together provide a thorough examination of exoplanet data, expanding our understanding of these heavenly entities.

8. CONCLUSION

8.1 Summary of Findings

Our research endeavors aimed to predict planetary radii and classify potential exoplanets using a diverse set of machine learning models. The following key findings emerged from our comprehensive study:

Model Performance: We meticulously evaluated multiple machines learning models, including Linear Regression and Random Forest Classifier, to gauge their efficacy in predicting planetary radii. Notably, the Random Forest Classifier demonstrated exceptional performance, achieving an accuracy rate of 87.04% in identifying potential exoplanets. This high level of accuracy is a promising indicator of the potential of our models for real-world astrophysical applications.

Feature Importance: Our analysis of feature importance revealed that specific features within the dataset played a pivotal role in making accurate predictions. Features such as 'koi_model_snr,' 'koi_period,' and 'koi_depth' emerged as influential factors. Understanding the significance of these features provides valuable insights into the underlying astrophysical processes governing exoplanets.

Exoplanet Classification: Beyond predicting planetary radii, our models showcased effectiveness in classifying potential exoplanets. This classification capability contributes to a deeper understanding of the categorization of celestial bodies beyond our solar system, aiding in the identification of candidates for further study.

8.2 Limitations and Challenges

It is imperative to acknowledge the limitations and challenges encountered during our research:

Data Quality: A key factor in determining how well a model performs is the quality of the input data. Astronomical data that has been gathered from several sources might include errors or be incomplete. As such, our models are dependent on the accuracy and completeness of the given data by nature.

Complexity of Astrophysical events: Because astrophysical events are so diverse and complicated, predicting planetary radii and classifying exoplanets are difficult undertakings. Simplifications and assumptions made in our models may introduce limitations, and capturing the full complexity of these phenomena remains a challenge.

Scope of Features: Our models were constructed using a specific set of features, which, while informative, may not encapsulate all relevant aspects of exoplanetary systems. Expanding the feature set to include additional variables, such as atmospheric composition or orbital characteristics, could enhance both prediction accuracy and our understanding of exoplanets.

8.3 Future Work

Our study establishes the groundwork for further investigations and improvements in the exoplanet field:

Improved Data Quality: A vital first step toward more reliable forecasts is strong collaboration with astrophysicists and astronomers to guarantee the completeness and correctness of astronomical data. Future research projects should continue to prioritize efforts to enhance data quality.

Advanced Machine Learning Techniques: The adoption of advanced machine learning techniques, such as deep learning architectures and ensemble methods, holds the potential to further elevate prediction accuracy and provide more nuanced insights into exoplanetary systems.

Incorporating Additional Features: Expanding the feature set to incorporate a broader array of variables, including atmospheric composition, magnetic properties, and orbital parameters, could offer a more comprehensive view of exoplanets and their characteristics.

Real-Time Predictions: Creating a system that produces predictions in real-time utilizing incoming data from space telescopes and observatories would be an important advance in the field. Such a technology might enable rapid discovery and classification of exoplanets as new data become available.

Putting it all out, our findings represent a significant breakthrough in our knowledge and investigation of exoplanets. Further research in this fascinating topic is necessary, despite the inherent difficulties and constraints that we accept. The insights gained from this work and the prospects for future breakthroughs support this. For both astronomers and data scientists, the search for exoplanets continues to be an intriguing and promising avenue for learning about the universe's intricacies.

9.REFERENCES

1. Borucki, W. J., Koch, D., Basri, G., Batalha, N., Brown, T., Caldwell, D., ... & Dunham, E. (2010). Kepler Planet-Detection Mission: Introduction and First Results. Science, 327(5968), 977-980. http://dx.doi.org/10.1126/science.1185402
2. Lissauer, J. J., Marcy, G. W., Rowe, J. F., Bryson, S. T., Adams, E., Buchhave, L. A., ... & Howell, S. B. (2014). Validation of Kepler's Multiple Planet Candidates. III: Light Curve Analysis and Announcement of Hundreds of New Multi-planet Systems. The Astrophysical Journal, 784(1), 44. https://doi.org/10.1088/0004-637X/784/1/44
3. Perryman, M. (2018). The Exoplanet Handbook. Cambridge University Press. https://doi.org/10.1017/9781316888872
4. Thompson, S. E., Coughlin, J. L., Hoffman, K., Mullally, F., Christiansen, J. L., Burke, C. J., ... & Haas, M. R. (2018). Planetary Candidates Observed by Kepler. VIII: A Fully Automated Catalog with Measured Completeness and Reliability Based on Data Release 25. The Astrophysical Journal Supplement Series, 235(2), 38. https://doi.org/10.3847/1538-4365/aab4f9
5. Tsiaras, A., Waldmann, I. P., Tinetti, G., Tennyson, J., & Yurchenko, S. N. (2016). Detection of an Atmosphere Around the Super-Earth 55 Cancri e. The Astrophysical Journal, 820(2), 99. https://doi.org/10.3847/0004-637X/820/2/99
6. Winn, J. N., & Fabrycky, D. C. (2015). The Occurrence and Architecture of Exoplanetary Systems. Annual Review of Astronomy and Astrophysics, 53, 409-447. https://doi.org/10.1146/annurev-astro-082214-122246
7. Zeng, L., Sasselov, D. D., & Jacobsen, S. B. (2016). Mass-Radius Relation for Rocky Planets Based on PREM. The Astrophysical Journal, 819(2), 127. https://doi.org/10.3847/0004-637X/819/2/127