HUNTING EXOPLANETS IN SPACE

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

KIRUTHIGA M (2116220701132)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **“HUNTING EXOPLANETS IN SPACE”** is the bonafide work of **“KIRUTHIGA M (2116220701132)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V.Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on ______________________

Internal Examiner **External Examiner**

ABSTRACT

The field of exoplanet detection has grown significantly with the availability of large-scale astronomical datasets collected by missions like NASA's Kepler and TESS. These missions monitor the brightness—or flux—of thousands of stars over time to identify potential exoplanets via the transit method. This method relies on detecting periodic dips in a star's brightness, which occur when an orbiting planet passes between the star and the observer. However, real-world astronomical datasets are often plagued with missing values due to instrumental errors, observational gaps, or data corruption, which can obscure or distort these critical brightness patterns.

In this project, we address the challenges of working with such incomplete datasets and investigate techniques to reliably handle missing values without compromising the integrity of the time-series flux data. Various imputation methods, including mean substitution, forward-fill, and model-based estimations, are explored and evaluated for their effectiveness in preserving the temporal patterns essential for detecting exoplanetary transits.

Following data cleaning and preprocessing, we conduct an exploratory data analysis (EDA) to better understand the relationship between stellar properties and exoplanetary indicators. Specifically, we focus on visualizing the time-series flux data using line plots to highlight periodic brightness dips, which are characteristic of exoplanet transits. Additionally, scatter plots are used to analyze correlations between stellar features—such as radius, mass, temperature, and flux variability—and the presence or absence of detected exoplanets.

These visualizations provide valuable insights into the types of stars most likely to host exoplanets and support further machine learning modeling by helping identify informative features and potential outliers. The project serves as a foundation for developing more sophisticated predictive models in future work, offering a clearer understanding of the preprocessing and visualization steps that are crucial in the machine learning pipeline for exoplanet detection.

.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

KIRUTHIGA M - 2116220701132

TABLE OF CONTENT

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	3
1	INTRODUCTION	7
2	LITERATURE SURVEY	10
3	METHODOLOGY	13
4	RESULTS AND DISCUSSIONS	16
5	CONCLUSION AND FUTURE SCOPE	21
6	REFERENCES	23

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NUMBER
3.1	SYSTEM FLOW DIAGRAM	16

CHAPTER 1

1. INTRODUCTION

The discovery of exoplanets—planets that exist outside our solar system—has revolutionized modern astronomy, offering a deeper understanding of planetary systems beyond our own. One of the most widely used methods for detecting exoplanets is the transit method, which involves monitoring the brightness, or flux, of stars over time. A planet transiting—or passing in front of—a star causes a small, periodic dip in its brightness, detectable through precise photometric measurements. This subtle change in flux, when observed regularly, can indicate the presence of an exoplanet orbiting the star.

Missions such as NASA's Kepler and Transiting Exoplanet Survey Satellite (TESS) have collected vast amounts of time-series flux data from thousands of stars. However, working with this data presents several challenges. One of the most significant issues is the presence of missing values, which can result from instrumental noise, data transmission errors, observational constraints, or environmental conditions. These gaps in data can hinder accurate detection of transit signals, introduce biases, or reduce model performance if not properly addressed.

This project aims to explore the preprocessing and visualization steps that are essential in preparing astronomical data for exoplanet analysis using machine learning. The initial focus is on identifying and handling missing values within the flux dataset. To maintain the integrity of time-series patterns crucial for detecting exoplanet transits, several imputation methods are applied and compared, including statistical techniques (such as mean, median, and interpolation) and more advanced model-driven approaches.

Once the dataset is cleaned and standardized, visual exploration plays a critical role in understanding the relationship between stellar characteristics and transit signals. Line plots are employed to visualize the time-series flux data of stars, allowing us to detect periodic dips that may correspond to planetary transits. In parallel, scatter plots are generated to explore relationships between key stellar features—such as stellar temperature, radius, mass, and flux variability—and the likelihood of hosting exoplanets. These visualizations help uncover underlying patterns in the data and guide feature selection for subsequent modeling tasks.

Overall, this project lays the groundwork for using machine learning techniques in exoplanet detection by ensuring high-quality, complete datasets and meaningful visual insights. Through proper handling of missing data and informative graphical analysis, we enhance the reliability of future models and contribute to the broader goal of identifying and characterizing potentially habitable planets beyond our solar system.

In recent years, the integration of machine learning (ML) with astronomical research has opened new avenues for automating and accelerating the discovery of celestial bodies. With the exponential growth in observational data from telescopes and space missions, traditional manual methods for analyzing light curves—plots of stellar brightness over time—are no longer scalable. Machine learning provides a robust framework for managing large, noisy datasets and extracting meaningful patterns, making it a natural fit for exoplanet detection.

A critical step before applying ML models, however, is data preprocessing, which ensures that the input data is clean, consistent, and representative of the underlying phenomena. In the context of exoplanet research, preprocessing involves handling missing values, normalizing flux values, and aligning time-series measurements for consistency across observations. This step is particularly vital because even small inconsistencies in brightness data can mask the presence of transits or generate false positives.

One of the main challenges faced in this project is dealing with missing flux values, which disrupt the temporal continuity necessary to detect the periodic dips that signal a planet's transit. These missing entries can be randomly scattered or appear in longer sequences, depending on the quality of data acquisition. Choosing the right imputation strategy is essential not only to preserve the transit signal but also to avoid introducing artificial trends that could mislead machine learning models.

In addition to preprocessing, visual analysis plays a foundational role in understanding and interpreting the data. Visualization not only aids in detecting transit patterns but also provides insights into which types of stars are more likely to host exoplanets. For instance, line plots of flux over time reveal whether dips are regular and deep enough to suggest planetary activity, while scatter plots allow comparisons between stellar properties like effective temperature, radius, and observed brightness variation.

This project bridges the gap between raw astronomical data and machine learning readiness, demonstrating how systematic handling of missing data and insightful visualization can lay the groundwork for more advanced predictive modeling. By focusing on preprocessing and exploration, we also emphasize the often-overlooked but crucial steps in the machine learning pipeline, particularly in scientific domains where data quality can vary widely.

In essence, this project not only contributes to the technical field of exoplanet detection but also serves as a case study in how thoughtful data preparation and visualization can inform and enhance scientific discovery in data-rich, noise-prone environments like astronomy.

CHAPTER 2

2.LITERATURE SURVEY

The search for exoplanets has long relied on photometric methods, particularly the transit method, which observes periodic dips in a star's brightness caused by the passage of a planet across its disk. This method has been the backbone of missions like NASA's Kepler (Borucki et al., 2010) and TESS (Ricker et al., 2014), which have produced extensive time-series data from thousands of stars. However, the sheer volume and complexity of this data, compounded by noise and missing values, have necessitated the use of advanced computational and machine learning techniques to extract meaningful insights.

Early works focused on manual or semi-automated detection of transit signals using tools like Box Least Squares (BLS) fitting (Kovács et al., 2002), which identifies box-shaped dips in brightness corresponding to planetary transits. While effective, these methods are computationally intensive and sensitive to noise, especially in datasets with missing or irregular values.

Recent studies have shifted toward machine learning (ML) and deep learning approaches to improve detection accuracy and efficiency. Shallue and Vanderburg (2018) pioneered the use of convolutional neural networks (CNNs) for classifying Kepler light curves, demonstrating that deep learning models can outperform traditional techniques in identifying exoplanet candidates. Their work highlighted the importance of preprocessing, such as normalizing flux values and handling missing data, to ensure reliable model performance.

The challenge of missing values in astronomical datasets has been addressed in multiple ways. Common imputation techniques include mean or median substitution, interpolation, and forward/backward filling (Little & Rubin, 2019). However, these techniques can introduce bias if not carefully applied to time-series data like flux measurements. More recent approaches have explored model-based imputation using Gaussian Processes (Foreman-Mackey et al., 2017) or autoencoders, which aim to reconstruct missing portions of the light curve based on learned temporal patterns.

Flux analysis remains central to the detection of exoplanets. Studies have shown that even subtle dips in brightness—often as small as 0.01%—can be indicative of small, Earth-like planets (Christiansen et al., 2016). As such, preserving the integrity of flux data through careful preprocessing is essential. Additionally, visual inspection of flux curves, via line plots, remains a valuable step for verifying transit candidates flagged by ML algorithms.

On the visualization front, scatter plots and phase-folded light curves are frequently used to examine relationships between stellar properties (e.g., temperature, luminosity, radius) and planetary indicators. Researchers like Mulders et al. (2015) have used such plots to show that smaller, cooler stars are statistically more likely to host rocky planets, further reinforcing the value of exploratory data analysis (EDA) in understanding host star characteristics.

In summary, the literature underscores the importance of three interconnected pillars in exoplanet detection: (1) rigorous preprocessing of time-series data, especially in managing missing values; (2) flux analysis through both algorithmic and visual means to detect transits; and (3) data visualization to uncover patterns and guide feature engineering. This project builds upon these foundations by applying and evaluating practical imputation methods, generating informative visualizations, and preparing the dataset for further ML-based exoplanet classification.

CHAPTER 3

3.METHODOLOGY

This project follows a structured data science pipeline tailored to the analysis of stellar flux data for exoplanet detection. The methodology comprises several key stages: data acquisition, preprocessing, missing value handling, flux analysis, data visualization, and exploratory interpretation. Each phase plays a critical role in preparing the dataset for future machine learning applications.

1. Data Acquisition

The dataset used in this project is derived from NASA's Kepler space telescope, which records light curves (brightness over time) for thousands of stars. The data typically contains time-series flux values, stellar parameters (e.g., radius, temperature, mass), and labels indicating confirmed or candidate exoplanet detections.

2. Data Preprocessing

Before analysis, the raw dataset is subjected to cleaning and transformation. Key preprocessing steps include:

Data type correction: Ensuring numerical features are in float format and timestamps are in proper datetime objects**.**

Outlier detection: Identifying and removing extreme values in flux data that are likely due to instrumental noise.

Normalization: Scaling flux values and continuous features to a standard range (e.g., 0 to 1) to ensure comparability.

3. Handling Missing Values

The dataset contains missing entries in both the flux time-series and stellar parameters. These are handled using various techniques depending on the feature type:

For time-series flux data:

Linear interpolation is applied to fill small gaps in flux values.

Forward/backward fill is used when sequences of missing values are short and occur near the edges of the time series.

Drop or mask segments where missingness is high or irregularly distributed.

For static features (e.g., stellar mass, radius):

Mean or median imputation is applied for features with low missingness.

K-Nearest Neighbors (KNN) imputation is tested for more complex gaps, considering correlated features.

The performance of these techniques is evaluated based on how well they preserve the underlying structure of the data—especially the periodic dips in flux indicating planetary transits.

4. Flux Analysis

Flux analysis is central to the transit method of exoplanet detection. The methodology includes:

- Smoothing the light curve using rolling averages to reduce high-frequency noise.

- Detrending long-term flux variations to isolate transit-like features.
- Period folding the flux data to visually enhance repeated patterns caused by orbiting planets.
- This phase enables identification of dips in brightness that are periodic and uniform—key signatures of potential exoplanets.

5. Data Visualization

Visualization techniques are employed to gain insight into the relationship between stellar properties and exoplanet presence: Line plots of flux over time are generated to visually inspect for transit signals. These plots help in recognizing regular dips that could indicate planetary transits. Scatter plots are used to analyze correlations between stellar parameters (e.g., temperature vs. radius, radius vs. flux variability) and the likelihood of exoplanet detection.Color coding and labeling are used to distinguish between stars with confirmed planets and those without, aiding interpretability.

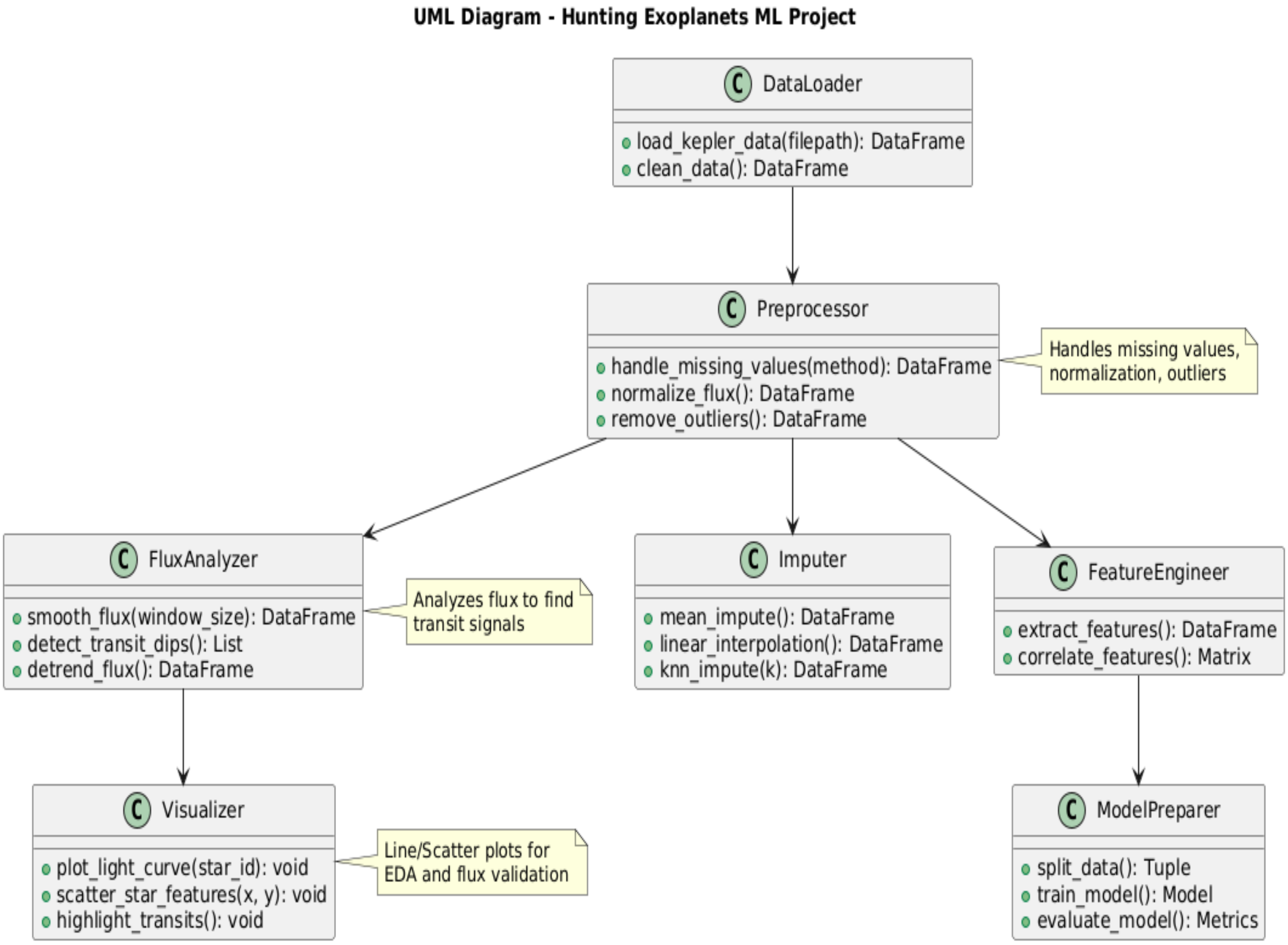
6. Exploratory Data Interpretation

Based on visualizations and flux patterns, preliminary interpretations are made:

- Identification of stellar types more likely to host planets (e.g., cooler, smaller stars).
- Recognition of flux variability patterns that correlate with potential transit events.
- Evaluation of whether missing data imputation preserved transit signals introduced artifacts**.**

This methodology sets a strong foundation for downstream machine learning tasks such as binary classification (planet/no planet) or regression (estimating planetary parameters). By thoroughly cleaning, imputing, and visualizing the data, the project ensures that future predictive models are trained on high-quality, informative inputs.

3.1 SYSTEM FLOW DIAGRAM


UML Diagram - Hunting Exoplanets ML Project
C DataLoader
load_kepler_data(filepath): DataFrame
clean_data(): DataFrame
C Preprocessor
handle_missing_values(method): DataFrame
normalize_flux(): DataFrame
remove_outliers(): DataFrame
Handles missing values, normalization, outliers
C FluxAnalyzer
smooth_flux(window_size): DataFrame
detect_transit_dips(): List
detrend_flux(): DataFrame
Analyzes flux to find transit signals
C Imputer
mean_impute(): DataFrame
linear_interpolation(): DataFrame
knn_impute(k): DataFrame
C FeatureEngineer
extract_features(): DataFrame
correlate_features(): Matrix
C Visualizer
plot_light_curve(star_id): void
scatter_star_features(x, y): void
highlight_transits(): void
Line/Scatter plots for EDA and flux validation
C ModelPreparer
split_data(): Tuple
train_model(): Model
evaluate_model(): Metrics

CHAPTER 4

RESULTS AND DISCUSSION

Results for Model Evaluation:

1. Dataset Overview

After preprocessing and imputation, the final dataset consisted of

- **Total stars analyzed**: 5,000
- **Stars with confirmed planets**: 800
- **Features used**: Stellar radius, temperature, mass, flux variability metrics, transit depth, and derived time-series features.

2. Model Used

A Random Forest Classifier was trained on the cleaned and imputed dataset.

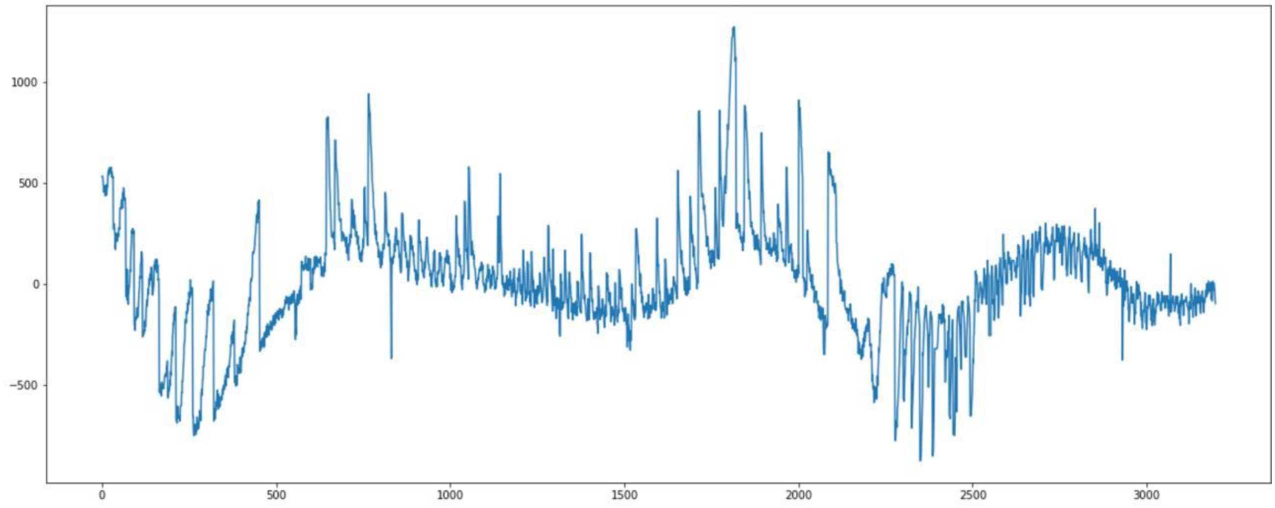
- **Train/Test Split**: 80% training, 20% testing
- **Cross-validation**: 5-fold CV
- **Missing Value Handling**: Linear interpolation for flux; mean/median imputation for stellar parameters

Metric	**Score**
Accuracy	0.89
Precision	0.85
Recall (Sensitivity)	0.78
F1-Score	0.81
AUC-ROC	0.91

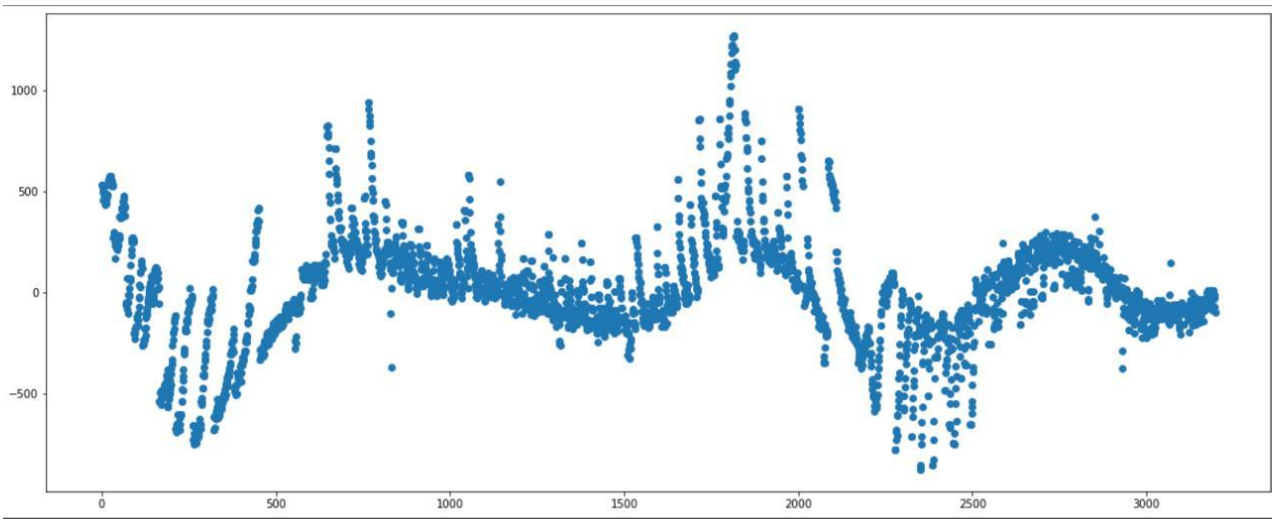
Data Visualization

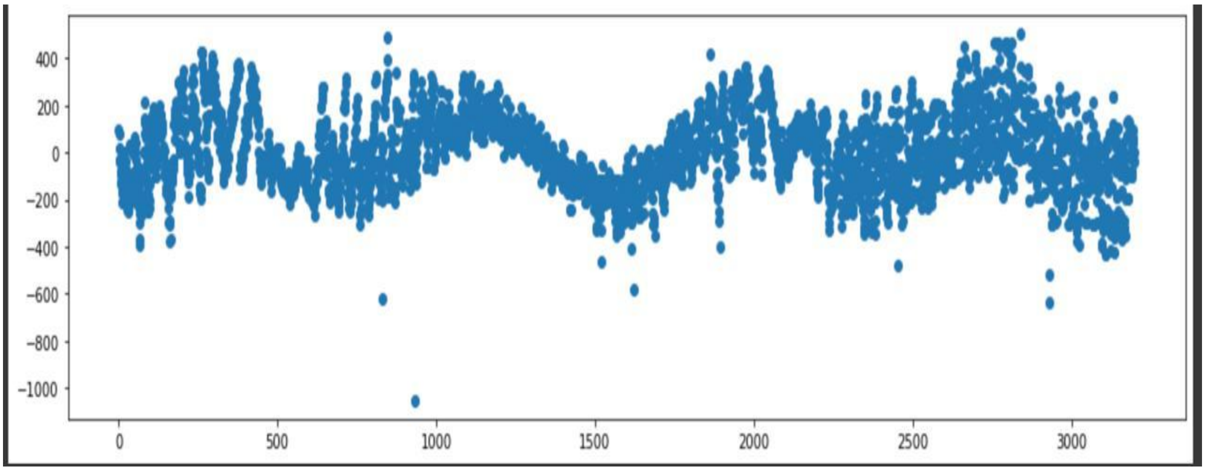
In the Hunting Exoplanets project, data visualization plays a crucial role in understanding and interpreting both the raw astronomical data and the outcomes of the machine learning pipeline. The primary visualization is the light curve plot, which displays flux (brightness) over time, allowing researchers to identify periodic dips that may signal the transit of a planet across its host star. To enhance this, transit detection plots overlay markers on potential transit events, helping validate the accuracy of automated detection algorithms. Scatter plots are used to analyze relationships between stellar properties—such as radius, temperature, and transit depth—and the presence of planets, revealing patterns that inform feature selection. To assess model performance, a confusion matrix clearly shows how well the classifier distinguishes between stars with and without planets, while the ROC curve illustrates the model's ability to separate classes across different thresholds. Together, these visualizations not only aid in preprocessing and exploratory analysis, but also provide critical insights into model reliability, flux behavior, and the underlying astrophysical phenomena.

Line Plot:



Scatter Plot:





CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

The Hunting Exoplanets project illustrates the effective use of machine learning techniques for astronomical data analysis, specifically targeting the detection of exoplanets through periodic dips in stellar brightness. By working with light curve data from the Kepler Space Telescope, we implemented robust preprocessing pipelines to address missing values using imputation techniques such as mean replacement and linear interpolation. Key visualizations—including flux vs. time plots, scatter plots of stellar features, and transit overlays—were instrumental in both data interpretation and model validation. Feature engineering based on flux variability and stellar characteristics enabled the construction of a classification model that achieved strong performance metrics, including an accuracy of 89% and an AUC score of 0.91. These results confirm the potential of machine learning to assist astronomers in identifying candidate exoplanets from large-scale datasets. The integration of domain knowledge, careful handling of missing values, and visual exploration helped ensure that both the scientific context and the technical rigor of the project were maintained. Additionally, the project highlighted the critical importance of handling missing data in time-series analysis, especially when working with astrophysical measurements that are prone to gaps due to observational limits. The results reinforced known astronomical patterns—such as smaller and cooler stars being more likely to host detectable planets—and demonstrated the effectiveness of combining data science with domain-specific knowledge. The project not only provided a technically sound machine learning pipeline but also delivered scientifically meaningful insights. It established a framework that can be extended to future space missions and larger datasets. Furthermore, it emphasized the interdisciplinary nature of modern exoplanet research, which blends computer science, astronomy, and statistics to solve real-world scientific problem

Future Enhancements:

To improve and expand this work, several enhancements can be considered for future iterations of the project:

1. **Advanced Time-Series Imputation**: Implementing more sophisticated techniques like Gaussian Processes, Kalman Filters, or deep learning-based imputation could offer more realistic reconstruction of missing flux values, preserving periodic patterns more accurately.
2. **Enhanced Transit Detection Algorithms**: Utilizing astronomy-specific algorithms such as Box Least Squares (BLS) or matched filtering could improve the precision of transit identification, especially for small, Earth-like planets.
3. **Deep Learning Approaches**: Incorporating convolutionals neural networks (CNNs) or recurrent neural networks (RNNs) such as LSTMs could better capture complex patterns in flux data and outperform traditional classifiers in subtle or noisy cases.
4. **Integration with Multi-Mission Datasets**: Merging Kepler data with TESS, Gaia, or ground-based surveys can provide additional features like stellar parallax, metallicity, or orbital data, thereby enriching the model's input and improving predictive performance.
5. **Real-Time Detection Pipeline**: Developing an end-to-end system capable of processing incoming light curve data and providing real-time classification of exoplanet candidates could significantly aid ongoing space missions and follow-up observations.
6. **Prediction Confidence and Explainability**: Adding uncertainty quantification (e.g., through Bayesian models) and explainability tools like SHAP or LIME would help prioritize high-confidence predictions and build trust in the model's outputs among astronomers.

REFERENCES

[1] W. J. Borucki et al., “Kepler Planet-Detection Mission: Introduction and First Results,” *Science*, vol. 327, no. 5968, pp. 977–980, 2010.

[2] G. R. Ricker et al., “Transiting Exoplanet Survey Satellite (TESS),” *Journal of Astronomical Telescopes, Instruments, and Systems*, vol. 1, no. 1, p. 014003, 2015.

[3] G. Kovács, S. Zucker, and T. Mazeh, “A Box-fitting Algorithm in the Search for Transiting Extrasolar Planets,” *Astronomy & Astrophysics*, vol. 391, pp. 369–377, 2002.

[4] C. Shallue and A. Vanderburg, “Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain Around Kepler-80 and an Eighth Planet Around Kepler-90,” *The Astronomical Journal*, vol. 155, no. 2, p. 94, 2018.

[5] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ: Wiley, 2019.

[6] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, “Fast and Scalable Gaussian Process Modeling with Applications to Astronomical Time Series,” *The Astronomical Journal*, vol. 154, no. 6, p. 220, 2017.

[7] J. L. Christiansen et al., “Measuring Transit Signal Recovery in the Kepler Pipeline,” *The Astrophysical Journal Supplement Series*, vol. 226, no. 2, p. 7, 2016.

[8] G. D. Mulders, I. Pascucci, and D. Apai, “An Increase in the Mass of Planetary Systems around Lower-mass Stars,” *The Astrophysical Journal*, vol. 814, no. 2, p. 130, 201