NATIONAL STUDENTS' SPACE CHALLENGE 2023

DATA ANALYTICS

TEAM: NSCode

TEAM ID: T-0873032

OVERVIEW

TASK

This problem focuses on the analysis of exoplanet data provided by NASA. Exoplanets are celestial bodies that orbit stars outside our solar system, and their study provides insights into the diversity of planets in the universe. The dataset contains a wealth of information about these exoplanets, including their properties, host star characteristics, and detection methods. The primary task is to extract meaningful insights from this dataset through data exploration, preprocessing, feature engineering, and machine learning techniques.

APPROACH

The project begins with an in-depth exploratory data analysis and visualization to understand the dataset's structure, uncover trends, and visualize relationships between variables. Data visualization plays a vital role in interpreting and communicating the results effectively. Visualizations are created to analyze the relationships between exoplanet properties and host star characteristics. Insights are drawn regarding exoplanets, their properties, and their potential habitability, while also considering patterns and clusters within the data. Basic statistical measures are computed to provide an initial overview. Data preprocessing steps involve handling missing values, addressing outliers, and normalizing the data to ensure it is suitable for most of the machine learning models like Random Forest, XGBoost, etc. Feature engineering is employed to create new attributes that could provide valuable information for analysis.The core of the project revolves around model building. A classifier is developed to categorize exoplanets into three classes of habitability: uninhabitable, conservatively habitable, and optimistically habitable. This involves utilizing various machine learning algorithms, fine-tuning hyperparameters, and assessing model performance using K-Fold Cross Validation. Addressing class imbalance within the dataset is also a crucial aspect for this problem statement. Techniques such as resampling are employed to maintain uniformity in the number of labels for training and testing. The project leverages evaluation metrics like ROC curves and confusion matrices to quantify the performance of the classifier.

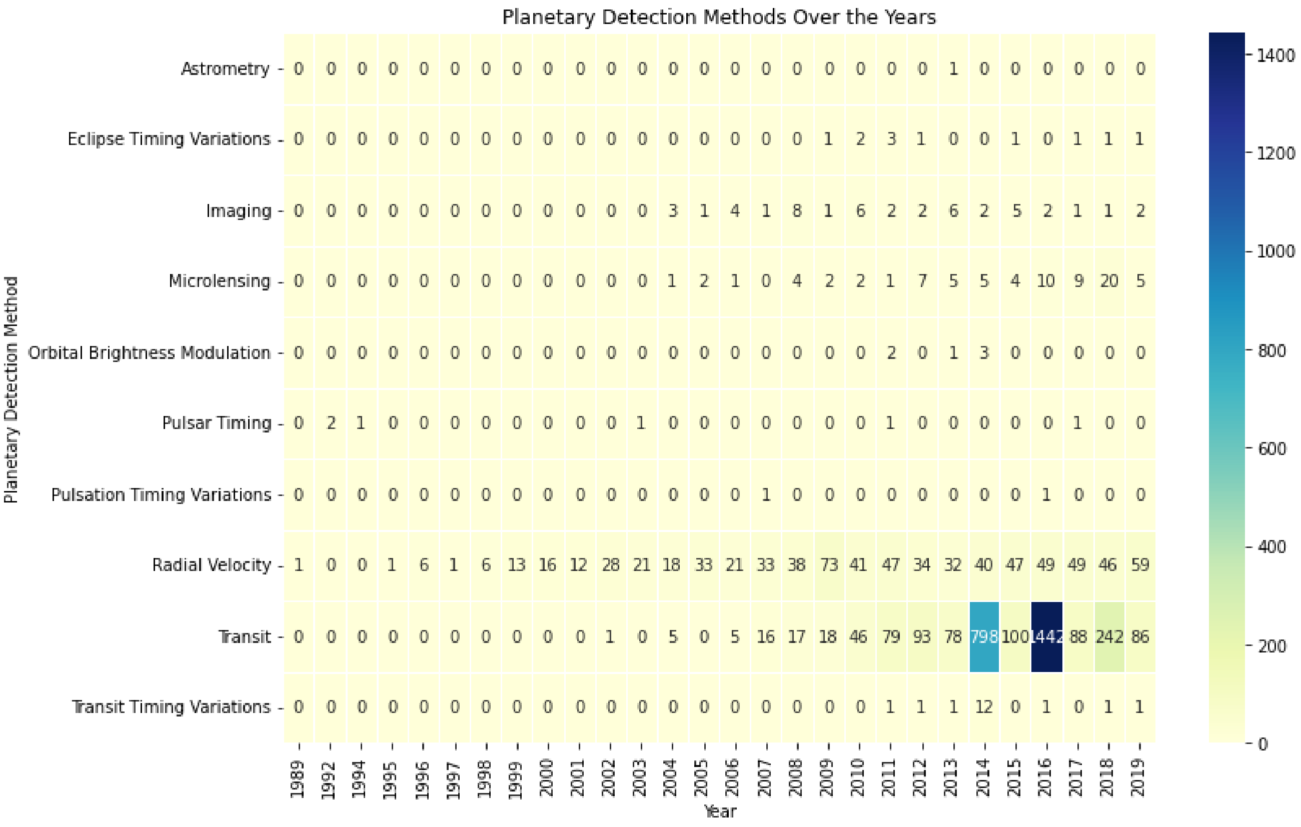
VISUALIZATION AND ANALYSIS OF DATASET

Data types in the dataset have not been explicitly provided, but it typically includes numerical data such as integers and floating-point numbers for measurements, and object data types for categorical information like names and labels. The following statistics have been evaluated.

Range : On the order of 10
Mean : 1.946e+33
Median : 0.879
Standard Deviation : 7.248e+35

It seems that the dataset likely requires normalization due to the presence of a wide range of numerical values and potential differences in scales among features. It exhibits a significant difference between the mean and median values, indicating a skewed distribution. The standard deviation is extremely high, which suggests a large degree of variability in the data. The range is also substantial, emphasizing the wide spread of values. These characteristics highlight the need for normalization to ensure that all features are on a similar scale for accurate analysis and modeling.

Let's explore more by plotting a heatmap to explore the various planetary detection methods used over the years:



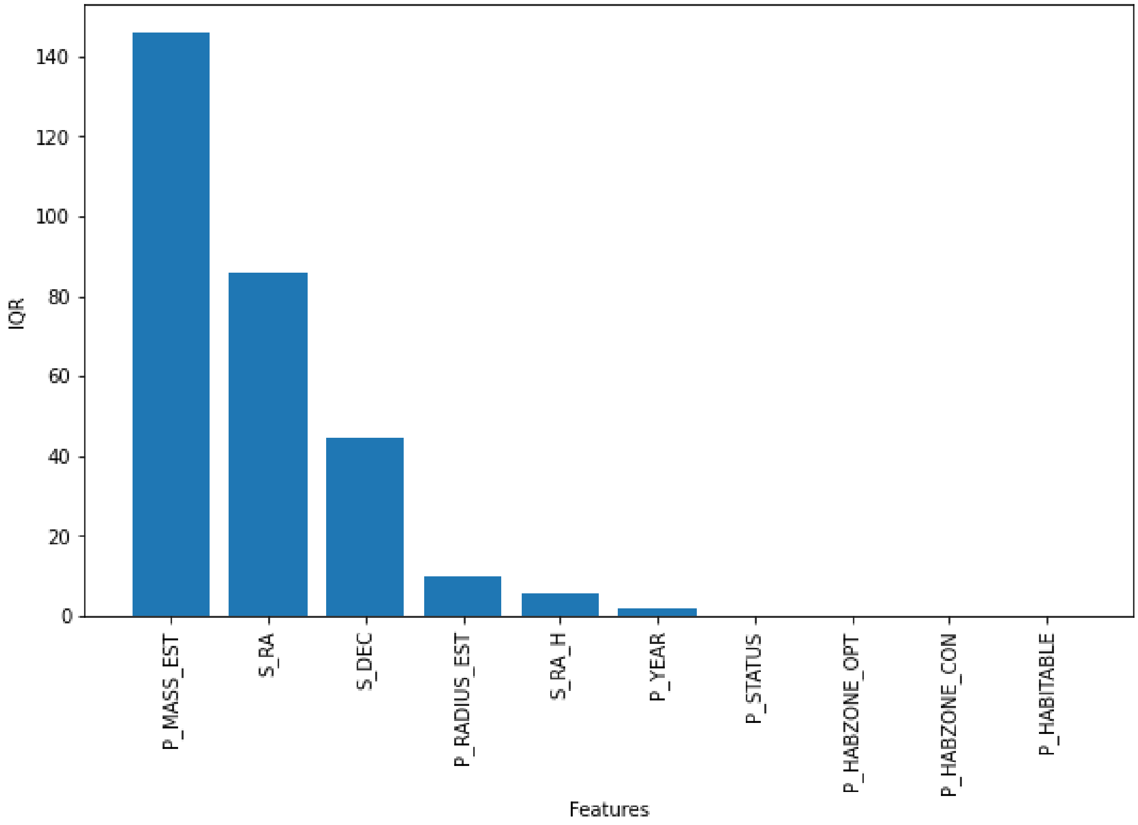
It is seen that the *Transit* method has an exceptionally large number of planetary detection compared to others over the years 2014-2018. The heatmap's concentration of detections suggests a combination of technological advancements, mission-specific data, and scientific interest that contributed to the high number of exoplanet detections during that time frame. NASA's *Kepler mission*, which was operational during the specified years, primarily used the Transit method. It contributed significantly to the discovery of exoplanets, potentially explaining the high number of detections in that period of 2014-2018.

Other Statistics:

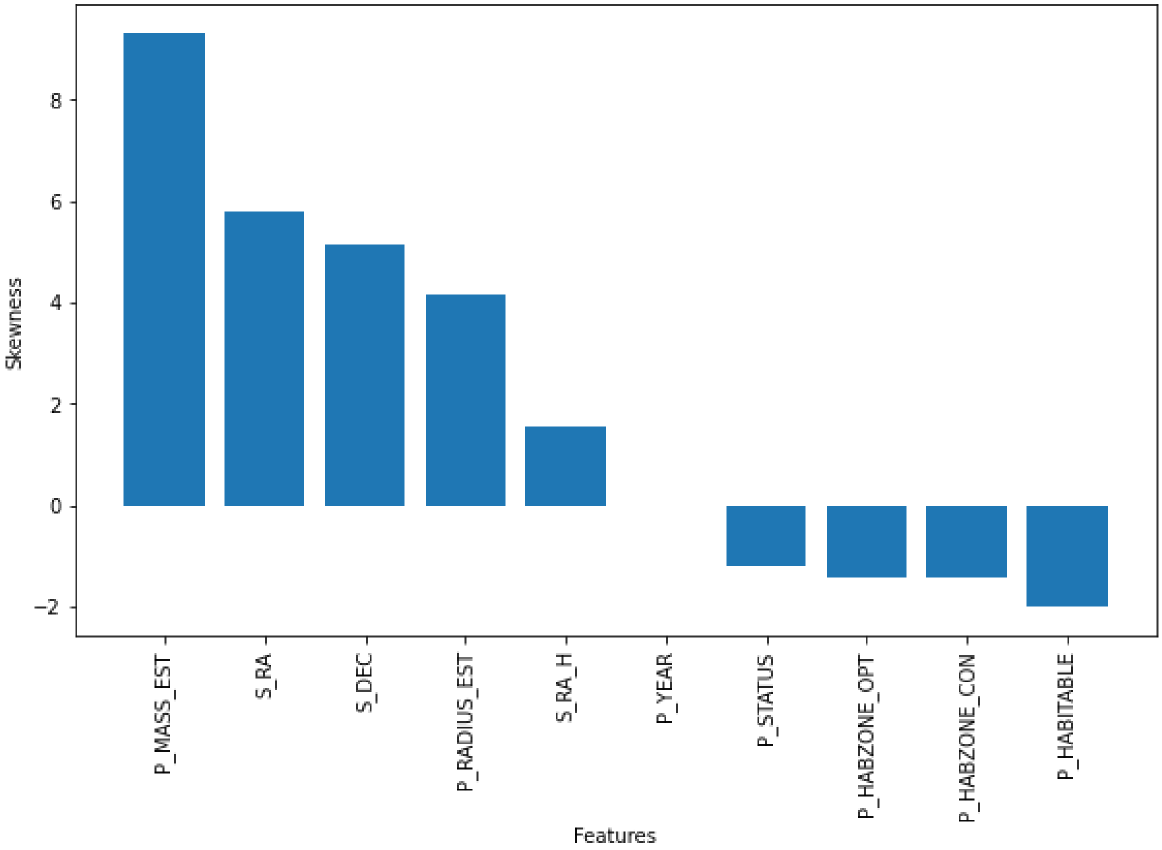
Planetary detection method that have identified the most

1) Uninhabitable planets (0) --→ **Transit**, which is *3076*
2) Conservatively habitable planets (1) --→ **Radial Velocity**, which is *12*
3) Optimistically habitable planets (2) --→ **Transit**, which is *29*

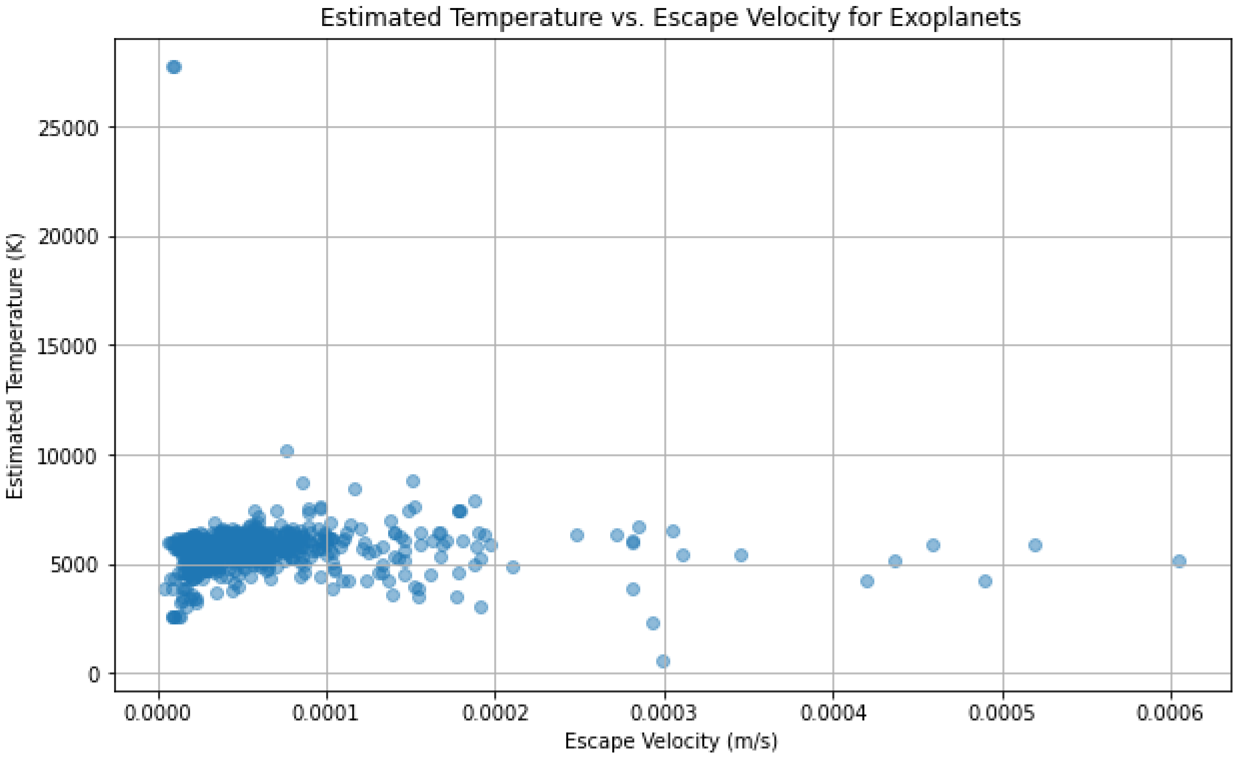
Top 10 features with the most Inter-Quartile Range :



Top 10 features with the most Skewness:

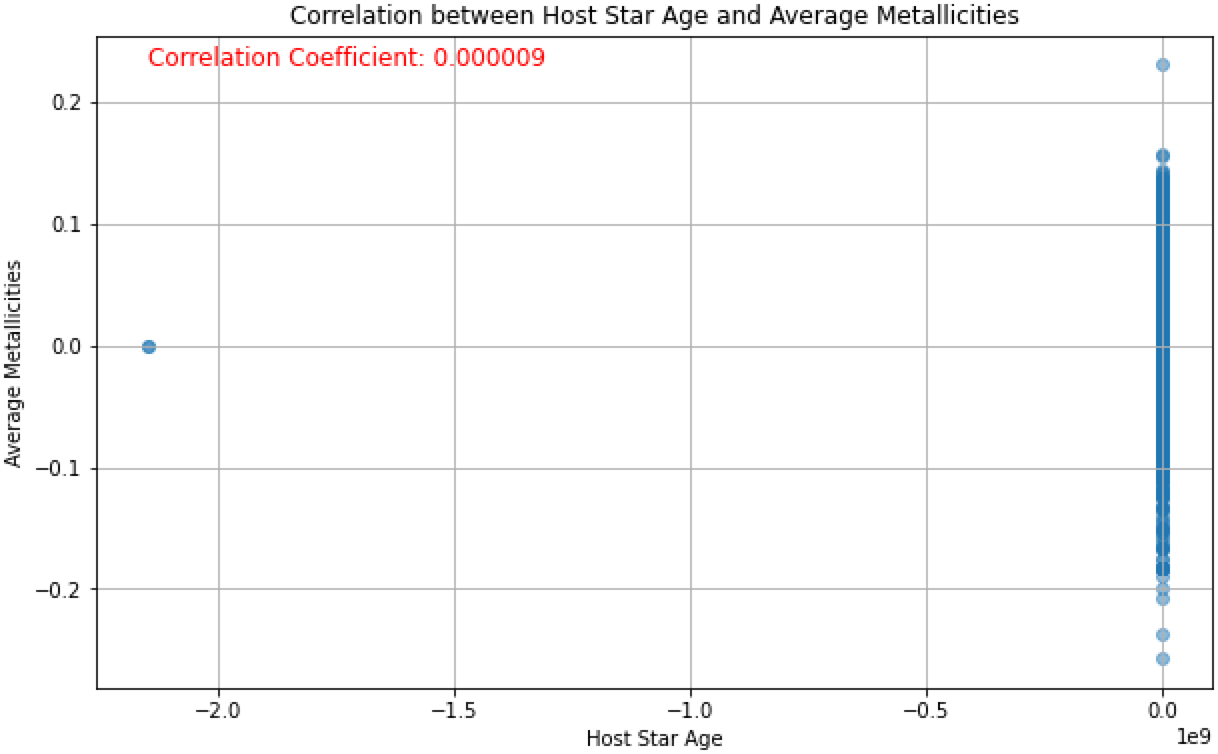


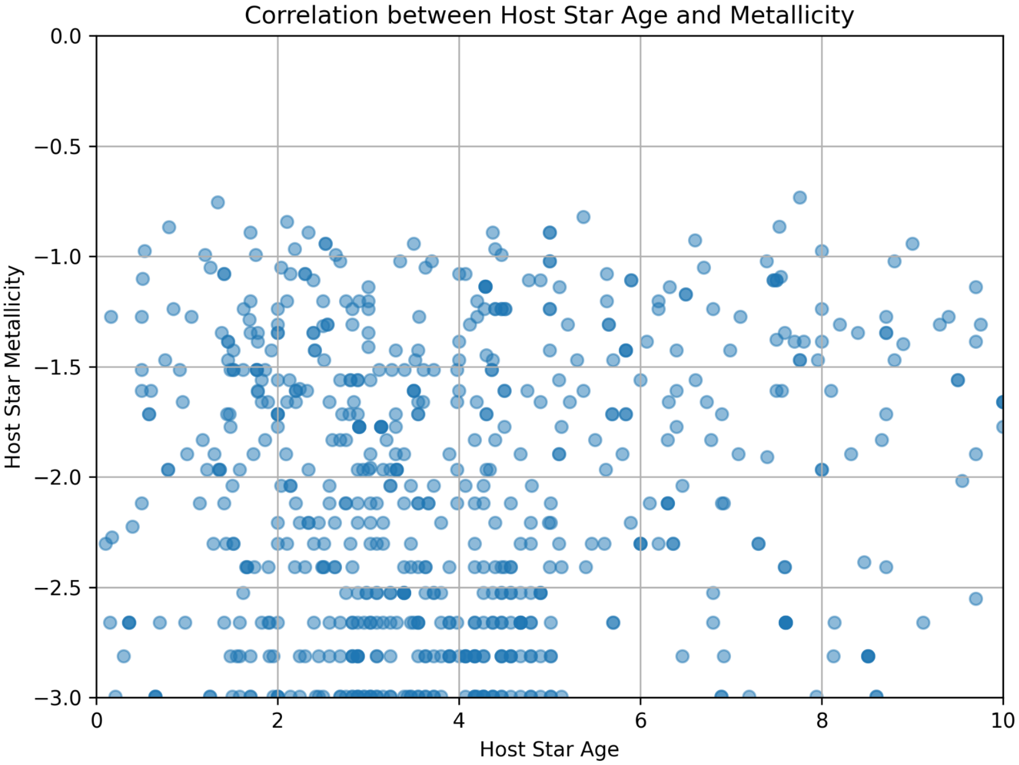
One of the problems we have faced is the class imbalance. To address the classification bias (class imbalance) in the dataset, various techniques can be applied. These include resampling methods such as oversampling the minority class with synthetic data or undersampling the majority class, employing algorithms robust to imbalanced data like ensemble methods, utilizing cost-sensitive learning to assign different misclassification costs, using evaluation metrics tailored for class imbalance, stratified sampling to maintain the class distribution in training and testing sets, ensemble techniques, data augmentation if additional data is available, treating the problem as an anomaly detection task when the imbalance is severe, and adjusting classification thresholds to favor the minority class.



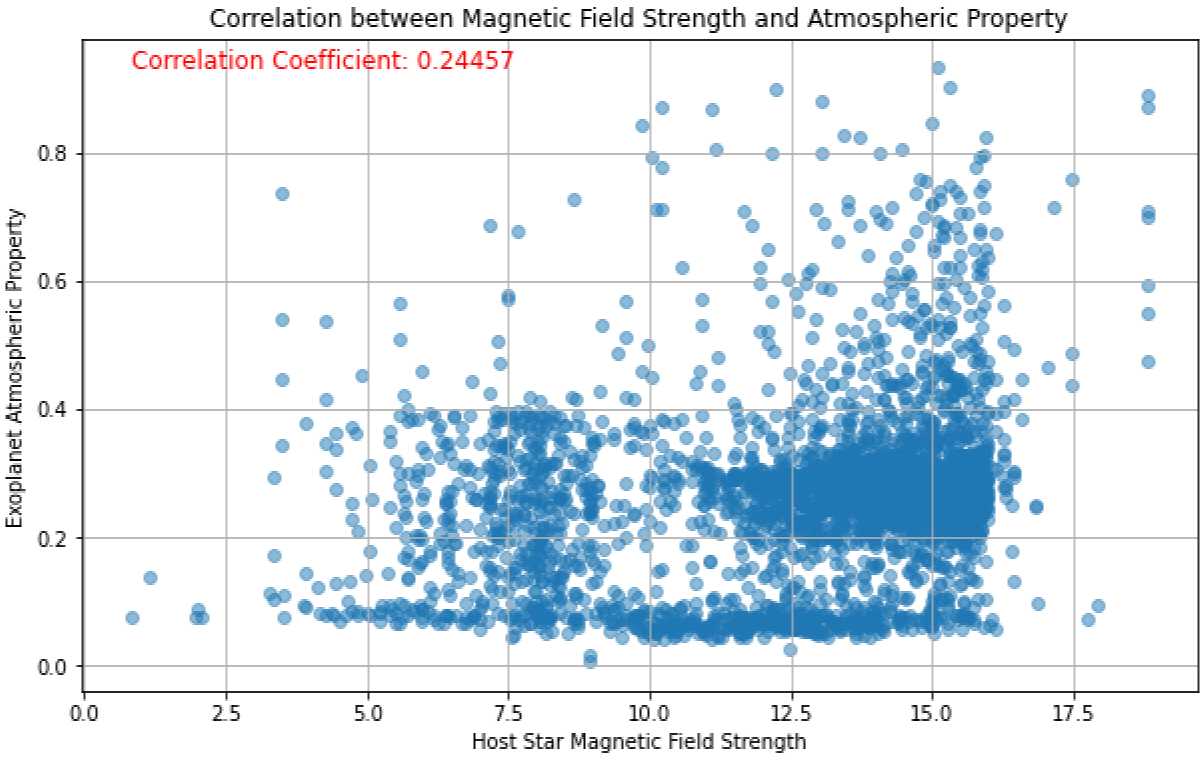
2.1 Above is the graph for Escape Velocity vs Estimated Temperature. The escape velocity plot shows a positive correlation between estimated temperature and escape velocity. As temperature increases, escape velocity also tends to increase. The physical explanation is that planets with higher mass and smaller radius, hence higher density, will have stronger surface gravity and higher escape velocities. Temperature correlates positively with escape velocity because more irradiated planets are heated to higher temperatures, expanding their atmosphere and reducing gravity and escape velocity. The scatter plot shows most planets have escape velocities above 10 km/s, sufficiently high to retain substantial atmospheres. However, planets closer to their host star are more susceptible to atmospheric escape through non-thermal processes like stellar wind stripping despite their high escape velocities. Overall, the escape velocity plot provides insights into how irradiation, temperature, planet mass and atmospheric retention are connected.

2.1.1 The estimated escape velocities for most exoplanets seem high enough, above 10 km/s, to retain atmospheres against thermal escape processes like Jeans escape. However, non-thermal escape mechanisms driven by stellar radiation like ion pickup and hydrodynamic escape can still erode atmospheres over time. Planets orbiting very close to their host stars receive tremendous irradiation, heating their upper atmospheres and increasing thermal escape rates orders of magnitude above Jeans escape. Additionally, the stellar wind from active host stars can directly strip atmospheres through momentum transfer and sputtering. Magnetospheric shielding against stellar wind can mitigate these non-thermal losses. The wide range of exoplanets discovered provides opportunities to study escape and atmospheric retention mechanisms in diverse environments.


Correlation between Host Star Age and Average Metallicities
Correlation Coefficient: 0.000009
Average Metallicities
0.2
0.1
0.0
−0.1
−0.2
−2.0
−1.5
−1.0
−0.5
0.0
1e9
Host Star Age

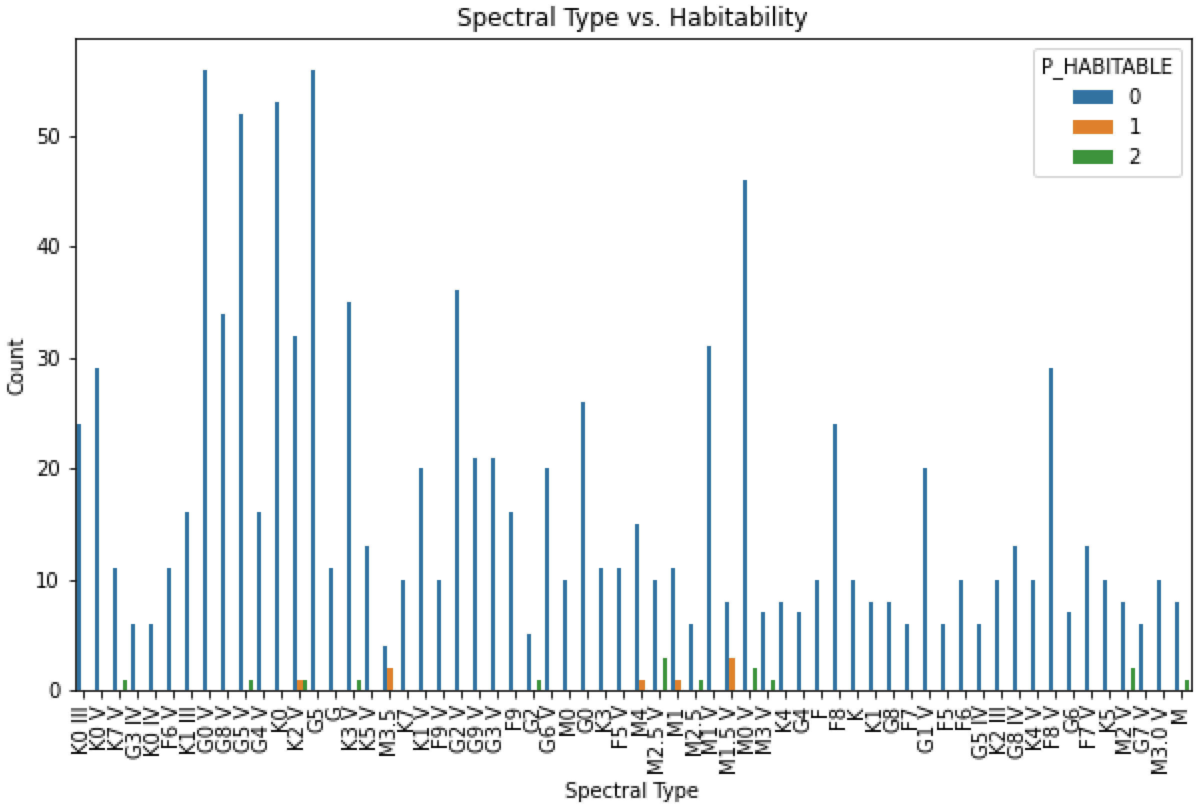

Correlation between Host Star Age and Metallicity
Host Star Metallicity
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
0
2
4
6
8
10
Host Star Age

2.2 , **2.2.1** The correlation coefficient between host star age and metallicity is very close to zero, indicating essentially no correlation in the dataset. This is contrary to the expected pattern of stellar metallicity decreasing over time from nuclear fusion processes. Older stars burn more hydrogen and helium, reducing their atmospheric metal content. Possible explanations include selection biases in the discovered exoplanets not representing the true population, errors in age and metallicity estimates, or complex star formation histories. The lack of age-metallicity correlation warrants further investigation through stellar atmosphere modeling and planet formation simulations. Additional exoplanet discoveries and spectral characterization of their host stars can build better understanding of how stellar evolution influences composition.



2.3 , **2.3.1** There is a moderate positive correlation of 0.24457 between host star magnetic field strength and exoplanet atmospheric properties like estimated ESI. Stronger stellar magnetic fields can help shield their planets' atmospheres from erosion by charged particles in the stellar wind through deflection and magnetic reconnection. The interaction between the two magnetospheres, modulated by the stellar wind, can also influence atmospheric composition over time through ion escape processes. Highly irradiated planets around active stars are subject to the most atmospheric loss, but a protective magnetic field helps preserve the atmosphere. Detailed observations of exoplanet atmospheres combined with models of magnetospheric shielding and escape can further elucidate these complex dynamics. The magnetic fields of host stars have clear implications for the habitability of their planets.

2.4 Spectral types categorize stars according to their surface temperature and spectral features. The major spectral types in order of decreasing temperature are O, B, A, F, G, K, M. This sequence is also associated with the star's mass and luminosity.

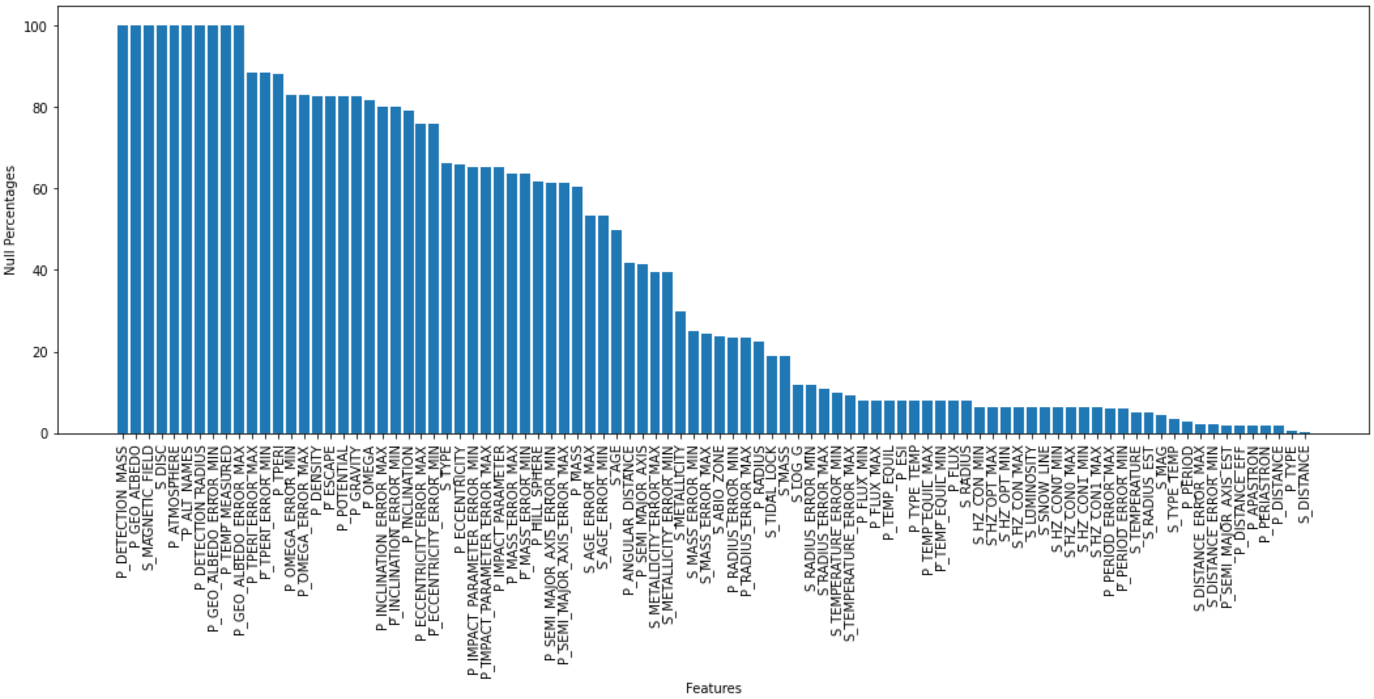


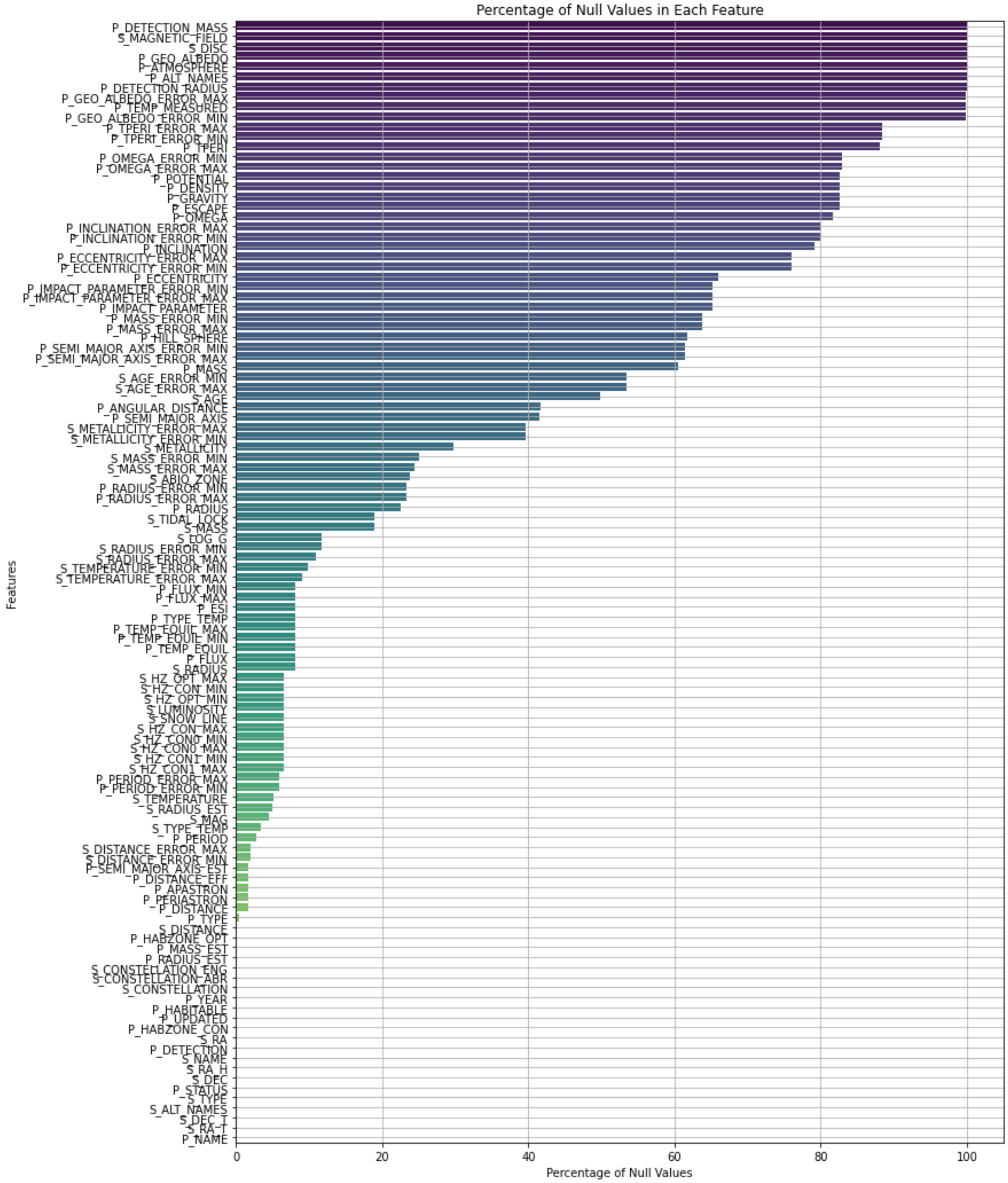
2.4.1 The categorical plot above shows hotter spectral types like O, B and A have very few or no habitable planets, while cooler K and M type stars host more habitable exoplanets. This aligns with expectations based on stellar characteristics. Hot stars have short lifetimes incompatible with biological evolution. Their intense ultraviolet radiation is also damaging for atmospheres and surface life. In contrast, K and M dwarfs have stable lifetimes of trillions of years and lower luminosity UV emission, making their planets more geologically and chemically suitable for life.

2.4.2 The distribution of habitable exoplanets clearly matches our knowledge of how different spectral types impact habitability. The harsh conditions around hot stars severely inhibit habitability. But small, cool K and M stars allow more exoplanets to remain within stable, non-desiccating orbits for billions of years with less sterilizing radiation. Their prevalence and longevity contribute to more confirmed habitable planets compared to unstable, short-lived hot stars. Stellar spectral type is a key factor determining exoplanet habitability.

2.4.3 There is a weak negative correlation of approx -0.18 between exoplanet size/density and host star temperature and spectral type. Hotter stars tend to have larger, lower density planets on average. Possible explanations include greater irradiation from hot stars influencing inflated planetary radii and lower densities. Planet formation and composition may also be affected by the stellar radiation environment and luminosity. However, the correlation is quite weak, indicating other factors also determine planet size and density. More data on confirmed exoplanets and their host stars can provide greater insights into these relationships.

FEATURE ENGINEERING




Percentage of Null Values in Each Feature
Features
P_DETECTION_MASS
S_MAGNETIC_FIELD
S_DISC
P_GEO_ALBEDO
P_ATMOSPHERE
P_ALT_NAMES
P_DETECTION_RADIUS
P_GEO_ALBEDO_ERROR_MAX
P_TEMP_MEASURED
P_GEO_ALBEDO_ERROR_MIN
P_TPERI_ERROR_MAX
P_TPERI_ERROR_MIN
P_TPERI
P_OMEGA_ERROR_MIN
P_OMEGA_ERROR_MAX
P_POTENTIAL
P_DENSITY
P_GRAVITY
P_ESCAPE
P_OMEGA
P_INCLINATION_ERROR_MAX
P_INCLINATION_ERROR_MIN
P_INCLINATION
P_ECCENTRICITY_ERROR_MAX
P_ECCENTRICITY_ERROR_MIN
P_ECCENTRICITY
P_IMPACT_PARAMETER_ERROR_MIN
P_IMPACT_PARAMETER_ERROR_MAX
P_IMPACT_PARAMETER
P_MASS_ERROR_MIN
P_MASS_ERROR_MAX
P_HILL_SPHERE
P_SEMI_MAJOR_AXIS_ERROR_MIN
P_SEMI_MAJOR_AXIS_ERROR_MAX
P_MASS
S_AGE_ERROR_MIN
S_AGE_ERROR_MAX
S_AGE
P_ANGULAR_DISTANCE
P_SEMI_MAJOR_AXIS
S_METALLICITY_ERROR_MAX
S_METALLICITY_ERROR_MIN
S_METALLICITY
S_MASS_ERROR_MIN
S_MASS_ERROR_MAX
S_ABIO_ZONE
P_RADIUS_ERROR_MIN
P_RADIUS_ERROR_MAX
P_RADIUS
S_TIDAL_LOCK
S_MASS
S_LOG_G
S_RADIUS_ERROR_MIN
S_RADIUS_ERROR_MAX
S_TEMPERATURE_ERROR_MIN
S_TEMPERATURE_ERROR_MAX
P_FLUX_MIN
P_FLUX_MAX
P_ESI
P_TYPE_TEMP
P_TEMP_EQUIL_MAX
P_TEMP_EQUIL_MIN
P_TEMP_EQUIL
P_FLUX
S_RADIUS
S_HZ_OPT_MAX
S_HZ_CON_MIN
S_HZ_OPT_MIN
S_LUMINOSITY
S_SNOW_LINE
S_HZ_CON_MAX
S_HZ_CON0_MIN
S_HZ_CON0_MAX
S_HZ_CON1_MIN
S_HZ_CON1_MAX
P_PERIOD_ERROR_MAX
P_PERIOD_ERROR_MIN
S_TEMPERATURE
S_RADIUS_EST
S_MAG
S_TYPE_TEMP
P_PERIOD
S_DISTANCE_ERROR_MAX
S_DISTANCE_ERROR_MIN
P_SEMI_MAJOR_AXIS_EST
P_DISTANCE_EFF
P_APASTRON
P_PERIASTRON
P_DISTANCE
P_TYPE
S_DISTANCE
P_HABZONE_OPT
P_MASS_EST
P_RADIUS_EST
S_CONSTELLATION_ENG
S_CONSTELLATION_ABR
S_CONSTELLATION
P_YEAR
P_HABITABLE
P_UPDATED
P_HABZONE_CON
S_RA
P_DETECTION
S_NAME
S_RA_H
S_DEC
P_STATUS
S_TYPE
S_ALT_NAMES
S_DEC_T
S_RA_T
P_NAME
0
20
40
60
80
100
Percentage of Null Values

As seen in the graphs above, the **top 10** features with highest null percentages are - P_DETECTION_MASS, P_GEO_ALBEDO, S_MAGNETIC_FIELD, S_DISC, P_ATMOSPHERE, P_ALT_NAMES, P_DETECTION_RADIUS, P_GEO_ALBEDO_ERROR_MIN, P_TEMP _MEASURED, P_GEO_ALBEDO_ERROR_MAX with a null percentage of 100 % and P_TPERT_ERROR_MIN with a percentage of approx 83%. Rest of the features have decreasing null percentage values.

FEATURE REDUCTION

Highly correlated features can be identified by calculating the correlation matrix between the numeric features and filtering values above a certain threshold like 0.6. Some highly correlated feature pairs identified are:

- P_MASS and P_MASS_EST (correlation = 1.0)
- P_RADIUS and P_RADIUS_EST (correlation = 1.0)
- P_MASS_ERROR_MIN and P_MASS_EST (correlation = -0.586)
- P_MASS_ERROR_MAX and P_MASS_EST (correlation = 0.513)

An appropriate feature reduction method for this dataset would be Principal Component Analysis (PCA). PCA is a linear dimensionality reduction technique that transforms the data into a lower dimensional space while retaining most of the variance. It is suitable here because:

- The dataset contains highly correlated variables which can be reduced using PCA.
- Visualizing high dimensional data is easier when reduced to fewer dimensions.
- PCA is a commonly used technique for astronomical data.

The scatter plots before PCA show positive correlation between P_MASS, P_RADIUS and their estimated values. The distribution seems concentrated along the diagonal. After applying PCA, the dimensions are reduced and the scatter plot will show most variance along the first few principal components. The distribution may appear more spherical and less correlated.

At first we imputed NaN values with 0. Then we imputed them with the mean value of all the rows. But we didn't get good accuracy in that. Finally we used KNN imputation to impute the NaN values. With the KNN Imputer we got the best accuracy among all.
So, a suitable imputation technique for the missing values in this dataset is K-Nearest Neighbors imputation. The reasons for choosing KNN imputation are:

- It can handle both numeric and categorical features unlike mean/median imputation.
- It uses nearby similar samples to impute missing values while preserving data distribution.
- It performs better than removal or basic imputation when the dataset has a mix of feature types.
- The dataset contains many features so finding similar samples is feasible.

PCA

Principal Component Analysis (PCA) is a vital technique in machine learning used to tackle high-dimensional data by reducing its complexity while preserving essential information. It is particularly valuable for dimensionality reduction and feature extraction. PCA begins with data standardization to ensure consistent scaling, followed by the computation of the covariance matrix to assess feature relationships. Eigenvalues and eigenvectors are derived from this matrix, with eigenvalues representing explained variances and eigenvectors defining the principal components. The most informative principal components are selected based on their contribution to the total variance, effectively reducing the data's dimensionality. By projecting the original data onto these selected components, PCA simplifies complex datasets, mitigates computational challenges, and enhances machine learning model performance, making it an indispensable tool in data preprocessing and analysis.

PCA was applied on the features and correlations among different features were checked and highly correlated features were identified. Applying PCA on the best performing model, i.e. XGBoost Classifier with **80** components with **7 fold cross validation** gave a cross validation score of **99.8664%** (**+/- 0.1204%**) and there was no observed change in resulting accuracy compared to when only XGBoost was used.

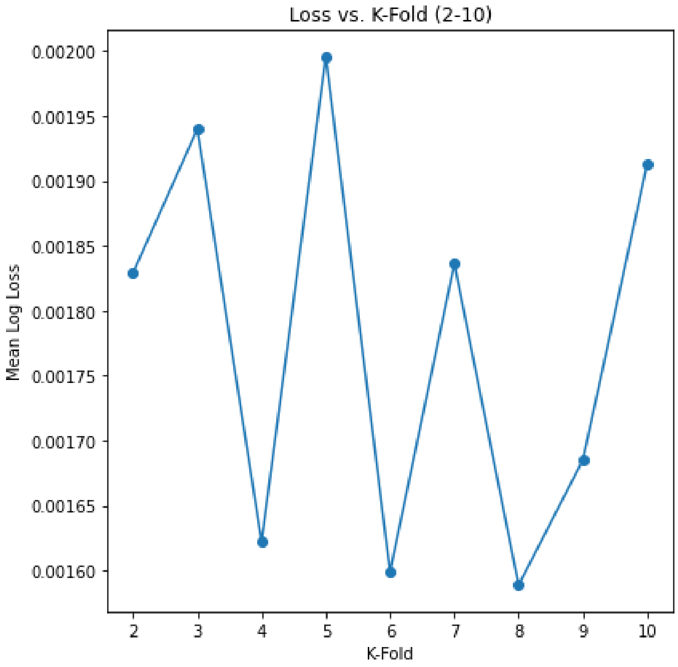
HABITABILITY CLASSIFICATION

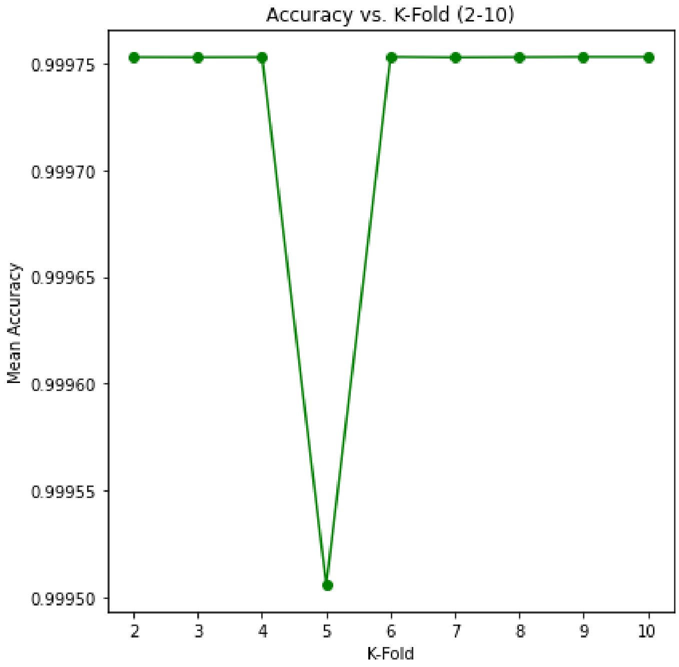
K-FOLD IMPLEMENTATION

K-Fold Cross-Validation is a crucial technique in machine learning that systematically assesses and refines predictive models. It involves dividing the dataset into k subsets, where the model is trained and evaluated k times, with each fold taking turns as the validation set while the others serve as training data. This process provides a more reliable estimate of a model's performance, reducing the risk of overfitting or underfitting. Cross-validation helps in hyperparameter tuning, feature selection, and overall model improvement, ultimately leading to more robust and accurate machine learning models, making it an essential tool for model assessment and refinement in data science projects.

XGBoost Classifier

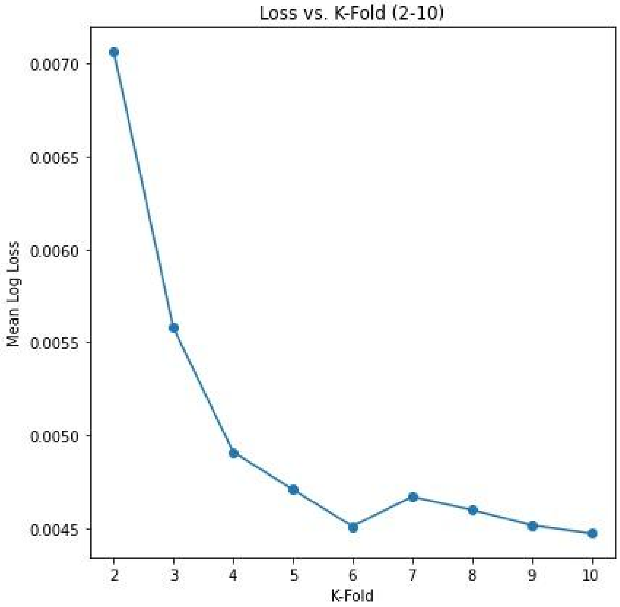
The XGBoost classifier, short for "Extreme Gradient Boosting," operates by iteratively training an ensemble of decision trees to make predictions. It begins with an initial simple model, typically a shallow decision tree, and then sequentially builds a series of decision trees, each aiming to correct the errors of its predecessor. During each iteration, XGBoost assigns higher weights to misclassified samples, emphasizing the importance of getting these instances right. Additionally, it employs a regularization term in its objective function to control the complexity of the trees, preventing overfitting. XGBoost optimizes this objective function using gradient descent, finding the best possible set of weights for each tree and their contribution to the final prediction. By combining the predictions of these trees, XGBoost creates a highly accurate and robust classifier that excels in various machine learning tasks, delivering state-of-the-art results in many real-world applications.

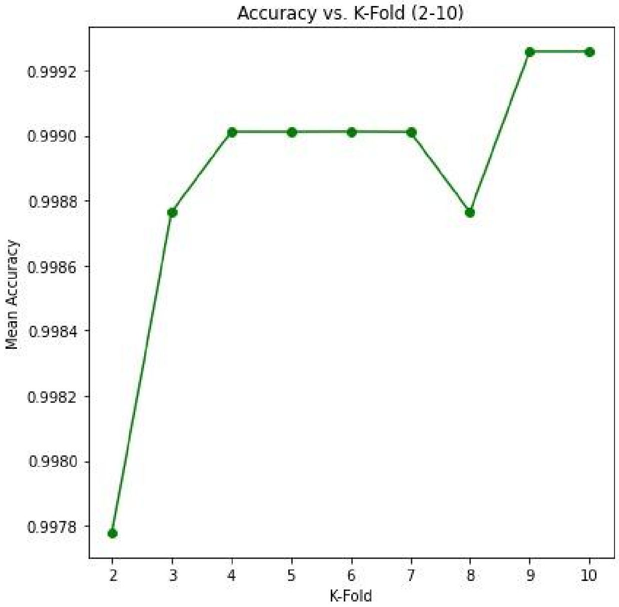




Random Forest Classifier

The Random Forest Classifier is an ensemble learning method that operates by constructing a multitude of decision trees during the training phase. Each tree in the forest is built independently and is based on a random subset of the training data, known as bootstrapping. Additionally, at each split point in a tree, a random subset of features is considered for determining the best split, which enhances diversity among the trees. During classification, each tree in the forest provides a prediction, and the final prediction is determined by majority voting (for classification tasks) or averaging (for regression tasks) across all trees. This ensemble approach reduces overfitting by combining the predictive power of multiple trees, providing robust and accurate results while mitigating the risk of individual tree biases. Random Forests are particularly effective for handling high-dimensional data, noisy datasets, and complex classification problems, making them a popular choice in various machine learning applications.

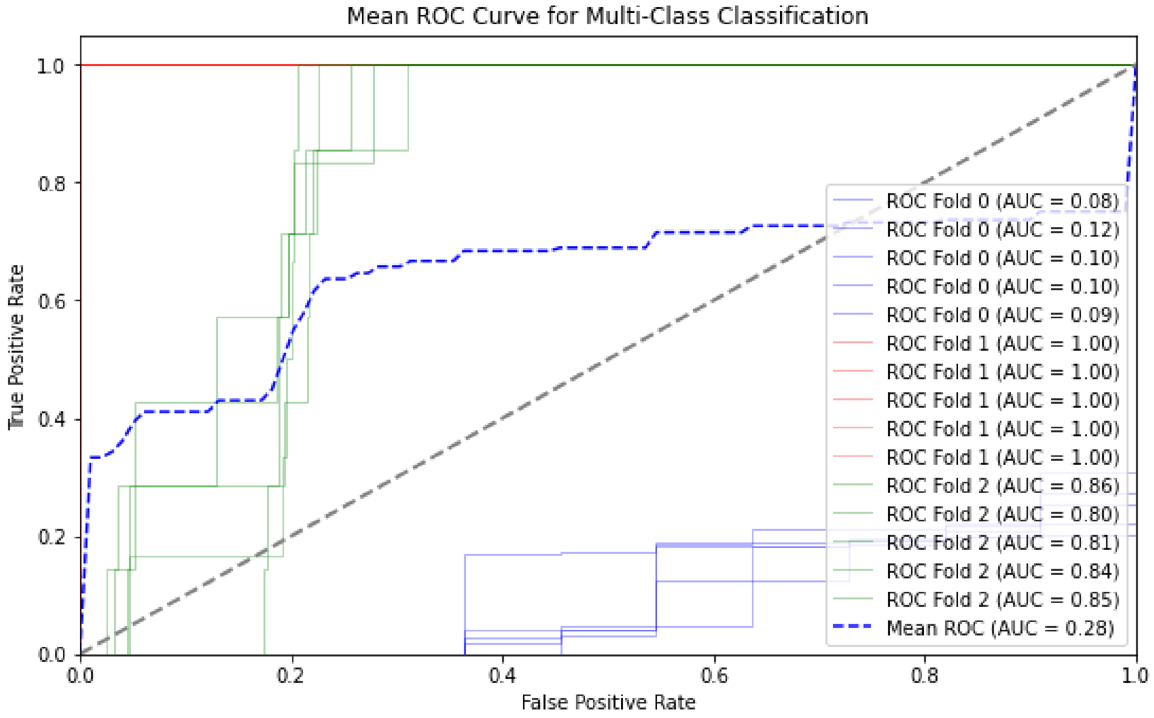




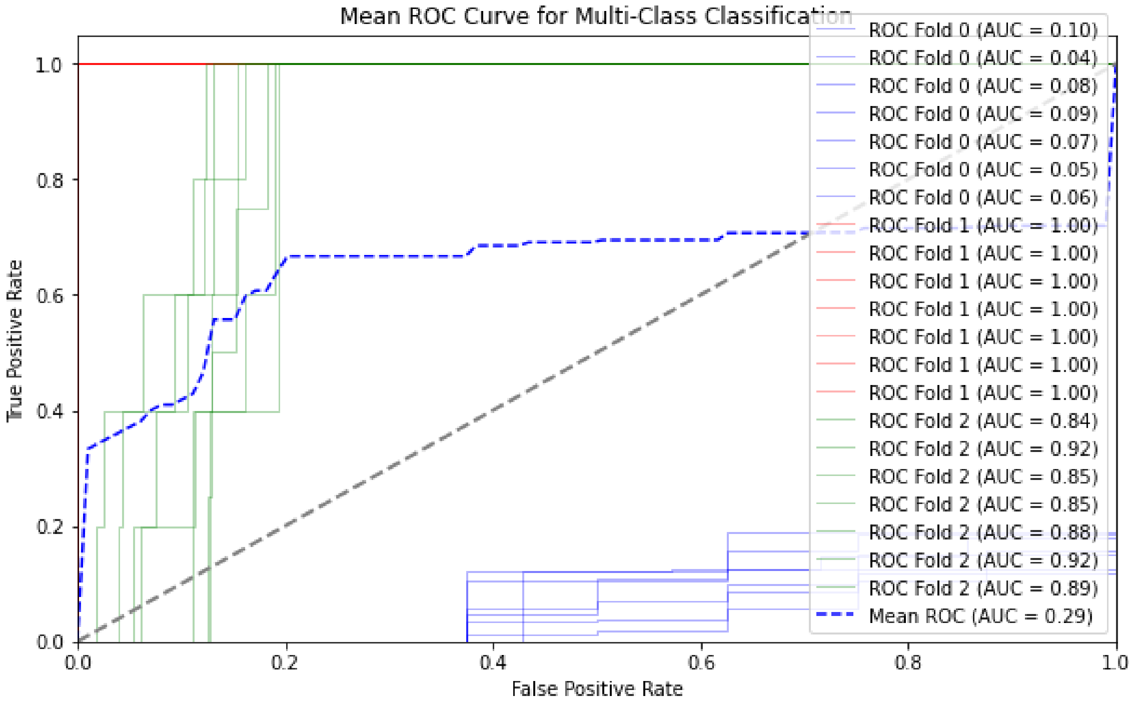
ROC Curve

Receiver Operating Characteristic (ROC) curves are vital tools in machine learning, serving to assess the performance of classification models. These curves depict the trade-off between Sensitivity (True Positive Rate) and Specificity (True Negative Rate) across various classification thresholds. A model's ROC curve is considered better when it deviates farther from the diagonal line, which represents random guessing. The Area Under the ROC Curve (AUC-ROC) quantifies overall performance, with 0.5 indicating randomness and 1.0 denoting perfection. ROC curves are particularly valuable in imbalanced datasets, helping to strike the right balance between false positives and false negatives by choosing an appropriate threshold. However, they lack insight into the cost of misclassification and assume adjustable thresholds, limiting their application in certain real-world scenarios. Nevertheless, understanding ROC curves is essential for informed model selection and optimization.

For XGBoost:



For RandomForest:



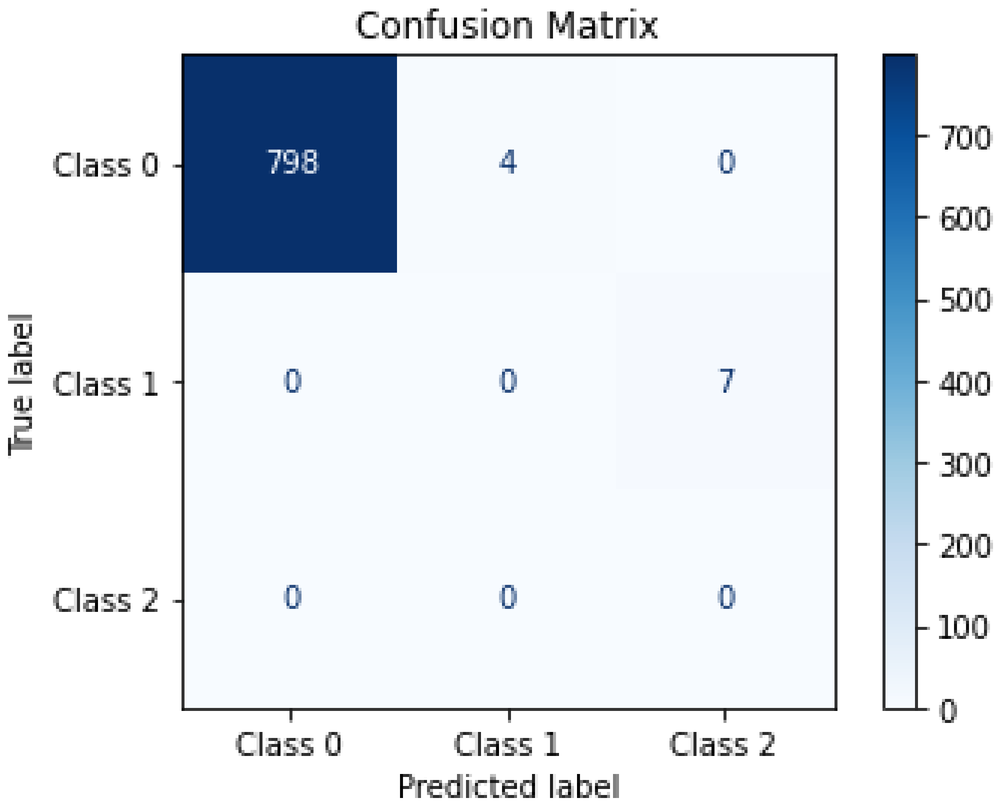
Confusion Matrix

In the context of machine learning, a Confusion Matrix is a critical tool for assessing classification model performance. It consists of four components:

1. True Positives (TP): Correctly predicted positives.
2. True Negatives (TN): Correctly predicted negatives.
3. False Positives (FP): Incorrectly predicted positives (Type I error).
4. False Negatives (FN): Incorrectly predicted negatives (Type II error).

Interpreting a Confusion Matrix provides insights into a model's accuracy, errors, and various metrics like Accuracy, Precision, Recall, Specificity, and the F1 Score. It aids in fine-tuning models for specific applications and informs decisions regarding model deployment. This concise overview underscores the importance of the Confusion Matrix in evaluating classification model efficacy and its practical applications across domains like healthcare, fraud detection, and sentiment analysis.

Below shown is the Confusion Matrix for the best model for particular Classification



Hyperparameter Tuning

Hyperparameter optimization was done using **GridSearchCV** on the XGBoost model. The hyperparameters tuned were:

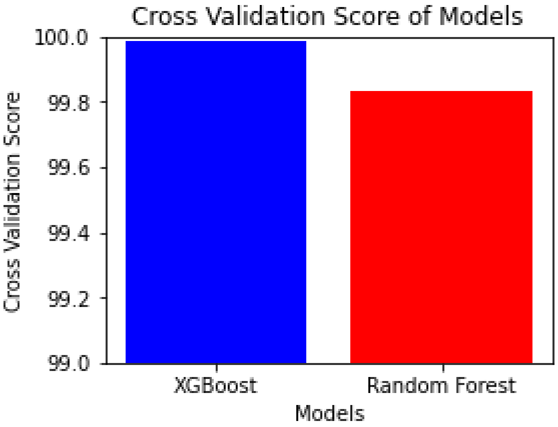
- max_depth - maximum tree depth
- min_samples_split - min observations to split node
- n_estimators - number of trees

GridSearchCV evaluates all combinations of provided hyperparameters systematically using cross-validation. It is an efficient way to find the optimal hyperparameters for the model. Tree-based models like XGBoost have several important tuning parameters that impact performance and GridSearchCV tries different configurations to find the best values.

The optimal parameters found were **max_depth=5**, **min_samples_split=2**, **n_estimators=100** which improved accuracy to 99.65% from 99.80237% with default parameters. So GridSearchCV helped boost model performance.

RESULTS

CLASSIFIER	**CROSS VALIDATION SCORE (7- fold)**	**ACCURACY**
RandomForest	99.8330 (+/- 0.3389%)	99.7913%
XGBoost	99.9833% (+/- 0.0818%)	99.9583%
XGBoost with PCA (80 components)	99.8664% (+/- 0.1204%)	99.9583%


Cross Validation Score of Models
Cross Validation Score
XGBoost
Random Forest
Models