CSC 177 Report

# Problem Statement

Text here

# Methodology

Preprocessing is a critical step in any data analysis or machine learning workflow, as it ensures that the data is clean, consistent, and ready for analysis. Here’s a comprehensive guide to preprocessing the exoplanet dataset from the NASA Exoplanet Archive, structured in a logical order.

1. Steps in Data Preprocessing
   1. Import Libraries
      1. Import necessary libraries, including NumPy for numerical operations, pandas for data manipulation, and Scikit-learn for preprocessing techniques.
   2. Load Data
      1. Load the data into a DataFrame while handling comments and missing values.
   3. Remove Unnecessary Row
      1. Remove metadata rows to focus on the actual dataset.
   4. Remove Duplicate Data
      1. Identify and drop duplicates to ensure each row is unique, which is critical for accurate analysis.
   5. Handle Missing Values
      1. First, replace empty strings with NaN, and then fill missing values with the mean of their respective columns to maintain data integrity.
   6. Data Transformation
      1. Encode Categorical Variables
         1. Convert categorical variables into numerical format using Label Encoding to prepare them for machine learning algorithms.
      2. Normalize Numerical Features
         1. Normalize numerical features to ensure they are on a similar scale, which is particularly important for algorithms sensitive to the scale of input data.
      3. Handle Outliers
         1. Identify and remove outliers using the Interquartile Range (IQR) method to prevent them from skewing analysis and model performance.
2. Importance of Order
   1. The order of these steps is essential for effective preprocessing. Each step builds on the previous ones to ensure that the data remains consistent and meaningful. For example, handling duplicates should occur before filling missing values, as duplicates can skew the mean calculations. Similarly, outlier removal should happen after handling missing values and transformations to prevent excessive data loss.

By following these steps in the specified order, we can effectively preprocess the exoplanet dataset, ensuring that it is clean, consistent, and suitable for further analysis or machine learning tasks. Each step is critical, as it helps maintain the integrity of the dataset, prepares it for modeling, and ultimately leads to more accurate and reliable results.

# Experimental Results and Analysis

Text here

# Task Division

Text here

# Project Reflection

Text here