Introduction: In the field of machine learning, data sets aren't always perfectly assembled to gain meaningful predictions. Thus, it is important to analyze data sets and their properties before applying machine learning algorithms.Issues such as null values, unequal values across data columns, and outliers are just a few examples of situations where machine learning can benefit from certain methods of preprocessing.

To assess the abilities of preprocessing, I have chosen two data sets to compare. The first data set provides specifications regarding penguin size. My goal is to predict penguin body mass based on a variety of parameters listed in the data set. The second data set lists properties of red wine. My goal with this data set is to predict the quality of wine based on the given parameters. Among analyzing each of these data sets, I noticed some of the data included null values, unequal observations, and outliers. To assess these issues, I am going to use some preprocessing techniques.

Results (Penguin Size): In this data set, penguin size is measured by seven features including species, island, culmen length, culmen depth, flipper length, body mass, and sex. For each of these features, the number of observations was unequal. Thus, to equalize the number of observations between all features, I used a drop null values function ”drop null values” to asses this. After

performing this function, the number of observations for each feature was 333. Next, I used the information command to asses the data type. Here, data type was not consistent throughout each feature with species, island, and sex features showing the object data type. Here, I used the replace function to replace each object term with a number instead. For species, Adelie=1, Chinstrap=2, and Gentoo=3. For island, Torgersen=1, Biscoe=2, Dream=3. For sex, male=1 and female=2. Once this was complete, I generated a heat map to show correlation (Figure 1). From the heat map, body mass and flipper length showed high correlation, so the feature flipper length was dropped from the data set. Next, species and body mass showed high correlation, so the feature species was dropped from the data set. The new heat map after removing these correlated features is seen in Figure 2. Next, I used the describe function to check for outliers. There were no obvious outliers, so I did not use any function for outlier removal. The linear regression score showed a score of 0.7099 after preprocessing compared to 0.8372 before preprocessing.

Results (Wine Quality): In this data set, wine quality is measured by twelve features including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. For each of these features, the number of observations was 1599.

The first preprocessing measure I used was to check for null values using a isnull().values.any() function. There were not any null values, so a heat map was generated to assess correlation (Figure 3). From this, the features free sulfur dioxide, density, volatile acidity, and citric acid were removed after displaying high correlation. With these removed, a new heat map showed no highly correlated features (Figure 4).

Next, outliers were removed from the remaining features using a function that filters out any data point ± 3 STDVs away from the mean. Boxplots of before and after outlier filtering are seen in Figure 5.

After these preprocessing measures, the linear regression score was 0.3143 compared to 0.3606 before preprocessing.

Reflection: After reviewing and using different preprocessing methods, I have noticed that even though the data becomes ”cleaner” for data analysis, the regression scores in both cases decreased after preprocessing. This tells me that choosing preprocessing methods is difficult. Learning AutoML techniques will hopedully help take the guessing game out of the equation a little bit.

Figures:

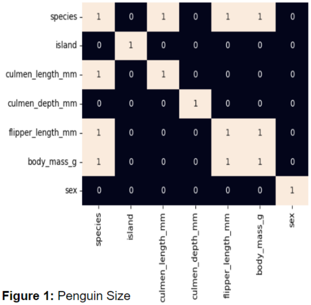

Figure 1: Penguin Size Heat Map before Preprocessing

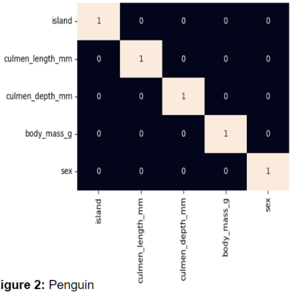

Figure 2: Penguin Size Heat Map after Preprocessing

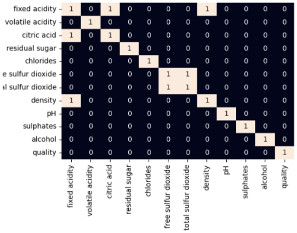

Figure 3: Wine Quality Heat Map before Preprocessing

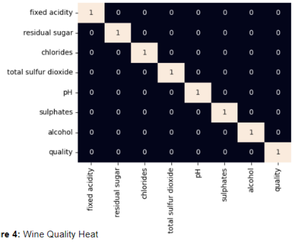

Figure 4: Wine Quality Heat Map after Preprocessing

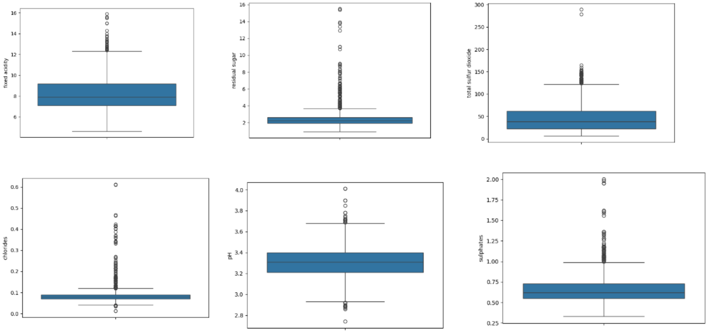

Figure 5: Wine Quality Box Plots after Outlier Removal

References:

1. “How to Replace Strings with Numbers in Python Pandas Dataframe.” Saturn Cloud Blog, 27 Dec. 2023, saturncloud.io/blog/how-to-replace-strings-with-numbers-in-python-pandas-dataframe/.

2. “Pandas Dataframe.Dropna() Method.” GeeksforGeeks, GeeksforGeeks, 31 Mar. 2023, www.geeksforgeeks.org/python-pandas-dataframe-dropna/.

3. “How to Check If Any Value Is Nan in a Pandas DataFrame.” Chartio, chartio.com/resources/tutorials/how-to-check-if-any-value-is-nan-in-a-pandas-dataframe/. Accessed 16 Feb. 2024.

4. Goyal, Chirag. “Outlier Detection Removal: How to Detect Remove Outliers (Updated 2024).” Analytics Vidhya, 8 Jan. 2024, www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/.

5. “Wine Quality Prediction - Machine Learning.” GeeksforGeeks, GeeksforGeeks, 12 Sept. 2022, www.geeksforgeeks.org/wine-quality-prediction-machine-learning/.