CSE 574-D: Introduction to Machine Learning, Fall 2023

Team:

Pandey Ashmita, pandey7@buffalo.edu

Assignment 1: Classification and Regression Methods

Introduction

This report outlines the methodologies, results, and analyses conducted in the assignment focused on classification and regression methods. The main objective of the assignment was to understand and implement data preprocessing, logistic regression, linear regression, and ridge regression from scratch without the use of machine learning libraries [1][2].

Part I: Data Analysis & Preprocessing

Introduction:

The provided code aims to process, analyze, and visualize two datasets: "penguins.csv" and "diamond.csv". The datasets undergo various stages of processing, including handling missing data, standardizing strings, removing outliers, and visualizing different data patterns.

1. Data Loading and Description:

1.1 Penguins Dataset:

- Total samples: 344

- Features:

  - Calorie requirement: Range (3504 - 7197)

  - Average sleep duration: Range (7 - 14 hours)

  - Bill length (mm): Range (32.1 - 124.3)

- Bill depth (mm): Range (13.1 - 127.26)
- Flipper length (mm): Range (10 - 231)
- Body mass (g): Range (882 - 6300)
- Year: Range (2007 - 2009)

1.2 Diamond Dataset:

- Total samples: 53940

- Features:

- Average US salary: Range ($30,000 - $48,999)
- Number of diamonds mined (millions): Range (0.6 - 5.2)

2. Data Cleaning:

2.1 Handling Missing Data:

- Missing numeric values were filled using the mean of their respective columns.

- Missing non-numeric values were replaced by the mode of their respective columns.

2.2 Standardizing Strings:

- The "species" column in the penguins dataset (if present) was standardized by capitalizing its values.

2.3 Handling Outliers:

- Outliers were identified and removed using the IQR method for each numeric feature.

3. Data Visualization:

For both datasets, the following visualizations were generated:

1. Diamond: Average US Salary (Boxplot)

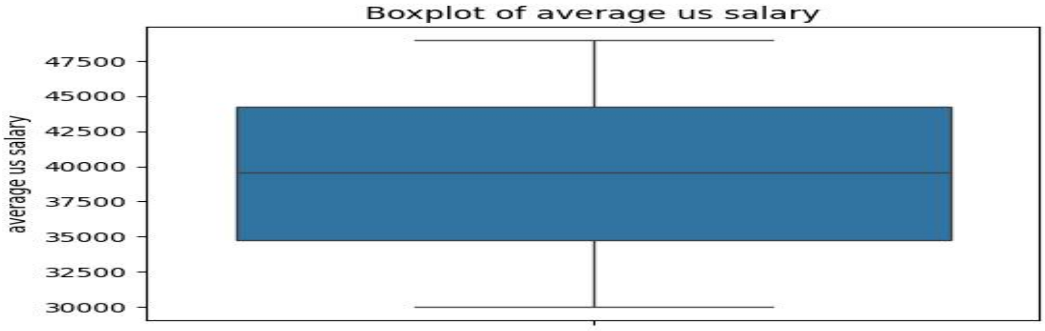

fig: diamond_average us salary_boxplot.png

Interpretation:

The boxplot represents the distribution of average US salaries in the diamond dataset. The median (Q2) seems to be around $39,000. The lower quartile (Q1) and the upper quartile (Q3) show the range where the central 50% of the data lies. Outliers, if any, would be displayed as dots outside the whiskers of the boxplot.

2. Diamond: Average US Salary (Distribution)

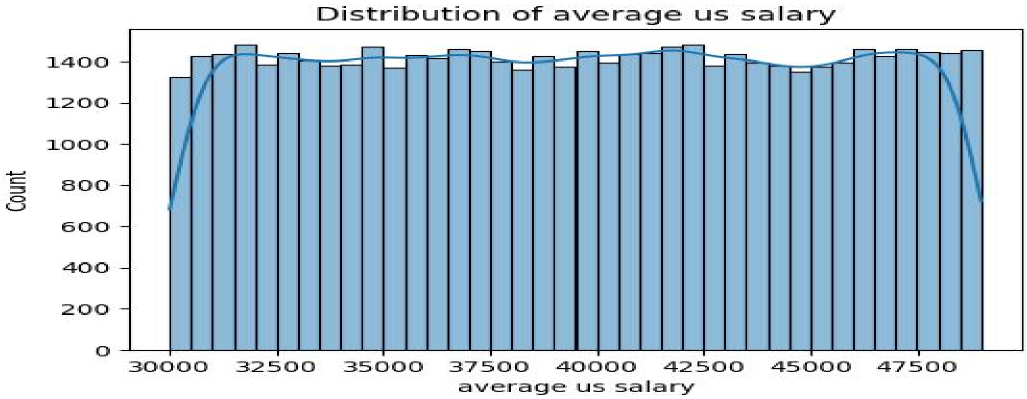

fig: diamond_average us salary_distribution.png

Interpretation:

This distribution plot displays the frequency of different average US salaries in the dataset. The shape of the distribution can provide insights into the central tendency and spread of the data.

3. Diamond: Barplot

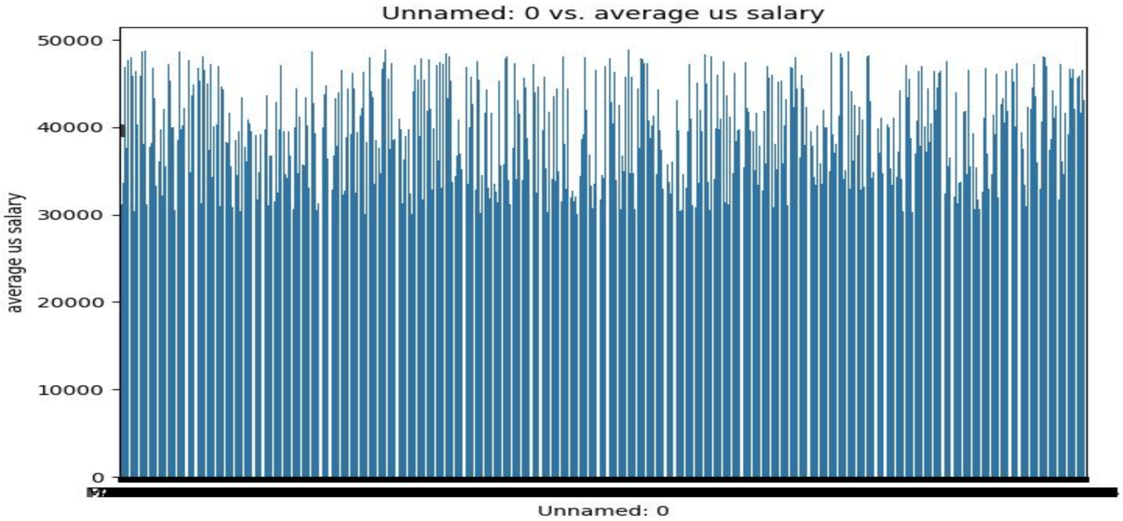

fig: diamond_barplot.png

Interpretation:

The barplot visualizes categorical data with rectangular bars. The height of each bar indicates the count or frequency of each category. This plot helps in comparing different categories in terms of their counts or frequencies.

4. Diamond: Correlation Matrix

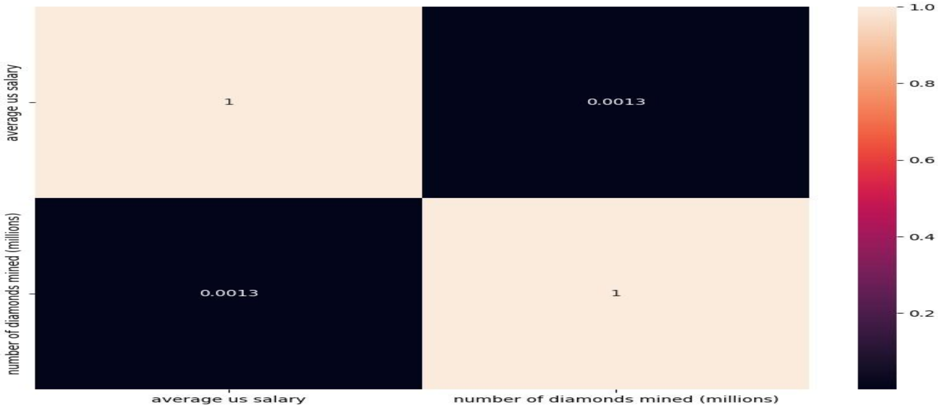

fig: diamond_correlation_matrix.png

Interpretation:

The correlation matrix heatmap visualizes the pairwise correlation between different numeric variables in the dataset. A darker color, either towards blue or red, indicates a stronger correlation, with blue indicating negative correlation and red indicating positive correlation.

5. Diamond: Number of Diamonds Mined (Boxplot)

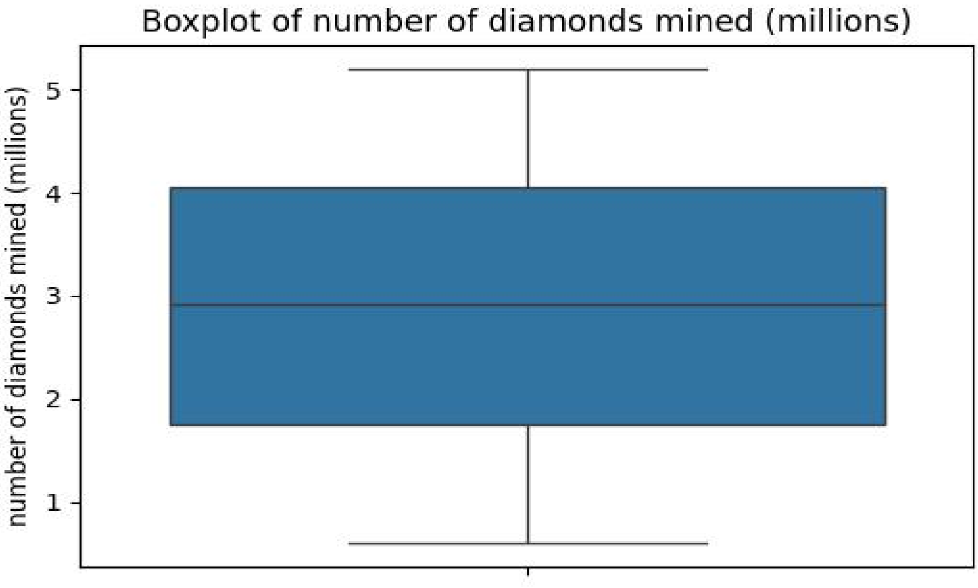

fig: diamond_number of diamonds mined (millions)_boxplot.png

Interpretation:

This boxplot showcases the distribution of the number of diamonds mined (in millions). Like the previous boxplot, it provides insights into the median, quartiles, and potential outliers in the data.

6. Diamond: Number of Diamonds Mined (Distribution)

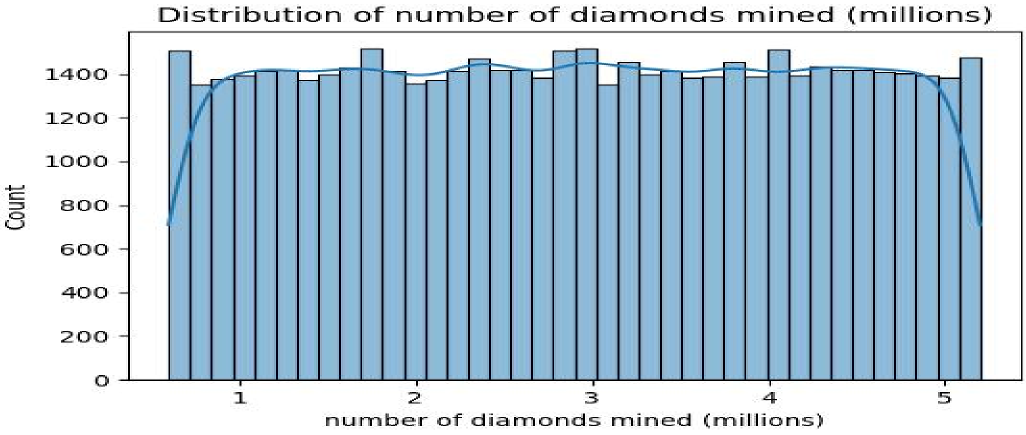

fig: diamond_number of diamonds mined (millions)_distribution.png

Interpretation:

The distribution plot displays the frequency distribution of the number of diamonds mined in millions. It helps in understanding the spread and skewness of the data.

7. Diamond: Pairplot

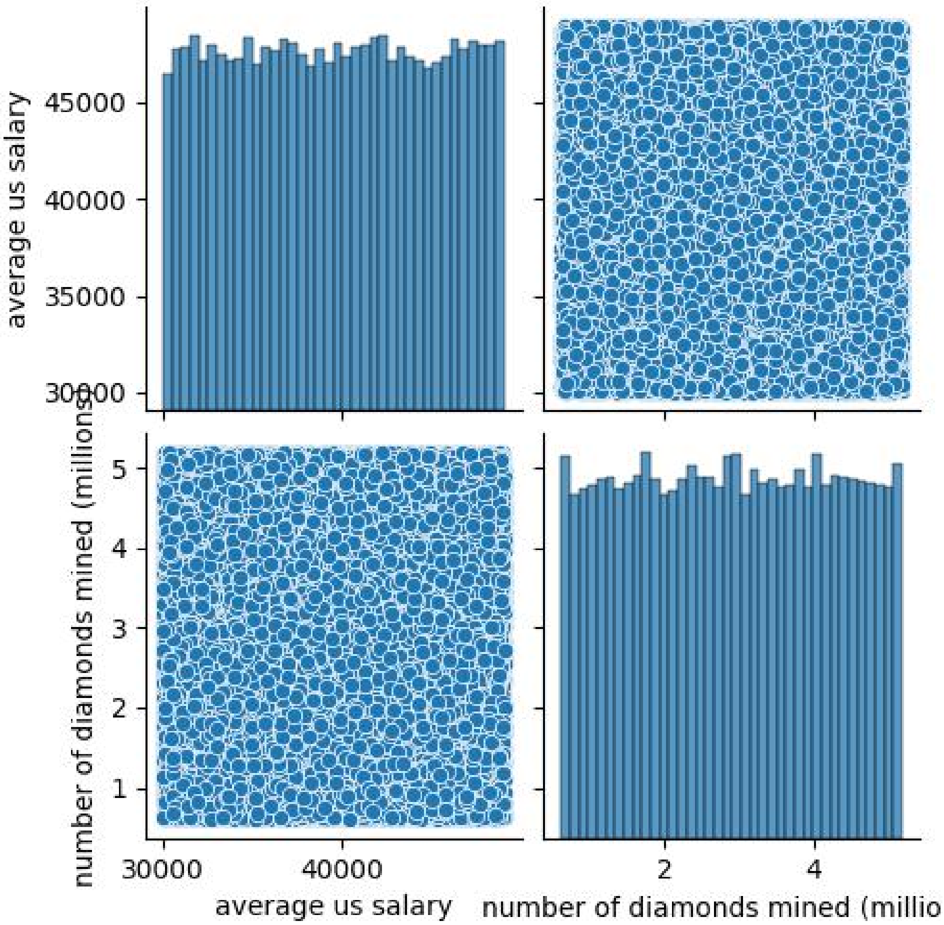

fig: diamond_pairplot.png

Interpretation:

The pairplot provides scatter plots for pairwise relationships in the dataset and histograms for univariate distributions. It gives a snapshot of relationships between multiple variables simultaneously.

8. Penguins: Average Sleep Duration (Boxplot)

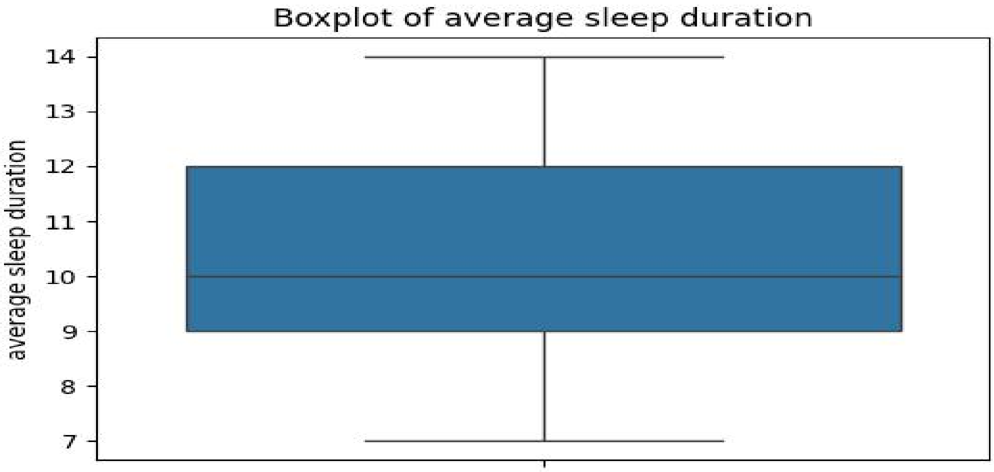

fig: penguins_average sleep duration_boxplot.png

Interpretation:

This boxplot visualizes the distribution of average sleep duration among penguins. It provides insights into the median sleep duration, its spread, and potential outliers.

9. Penguins: Average Sleep Duration (Distribution)

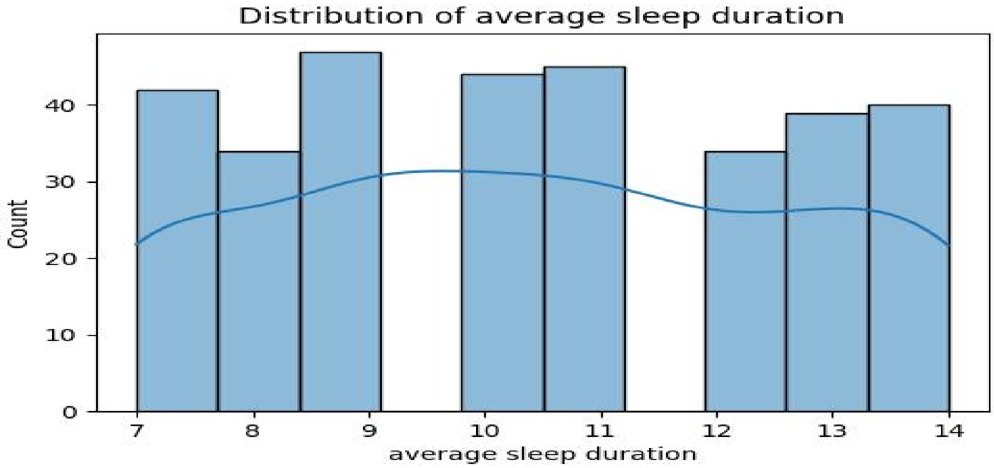

fig: penguins_average sleep duration_distribution.png

Interpretation:

The distribution plot for the penguins dataset showcases the frequency distribution of average sleep duration. It helps in understanding the central tendency and distribution of sleep durations among penguins.

10. Penguins: Bar Plot

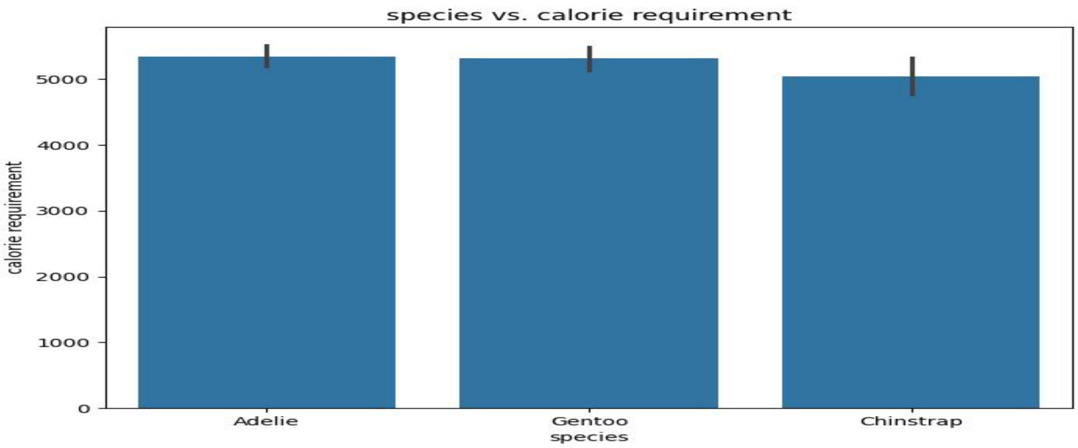

fig: penguins_barplot.png

Interpretation:

The bar plot represents a categorical attribute of the penguins dataset, comparing the frequency or count of categories. The vertical axis likely represents counts or another numerical measure, while the horizontal axis displays distinct categories.

11. Penguins: Bill Depth (mm) - Box Plot

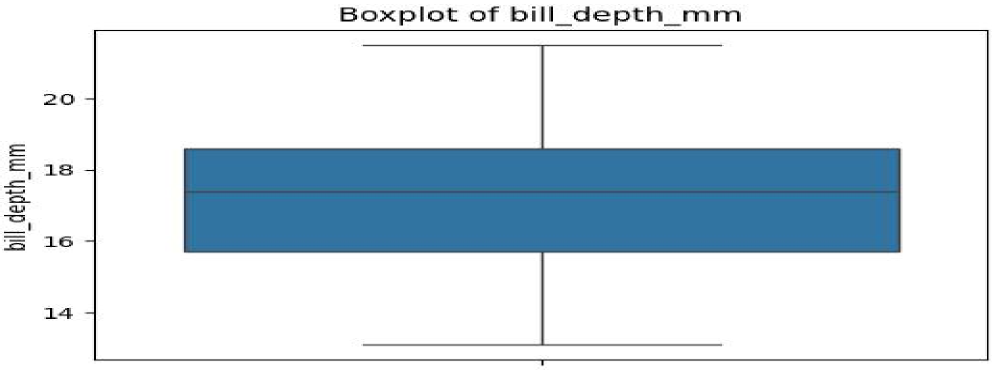

fig: penguins_bill_depth_mm_boxplot.png

Interpretation:

The box plot visualizes the distribution of bill depth for penguins, revealing the median, quartiles, and potential outliers. Separate boxes might represent different species or categories of penguins.

12. Penguins: Bill Length (mm) - Box Plot

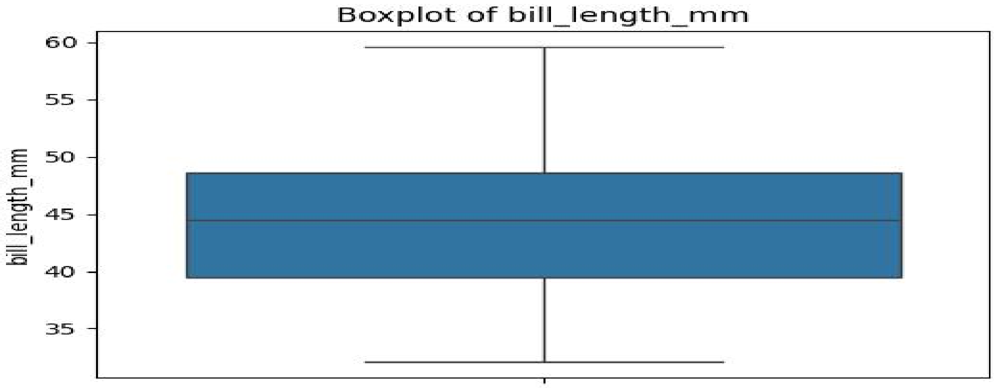

fig: penguins_bill_length_mm_boxplot.png

Interpretation:

This plot displays the distribution of bill lengths among penguins. Multiple boxes might suggest variations based on species, region, or other criteria.

13. Penguins: Body Mass (g) - Box Plot

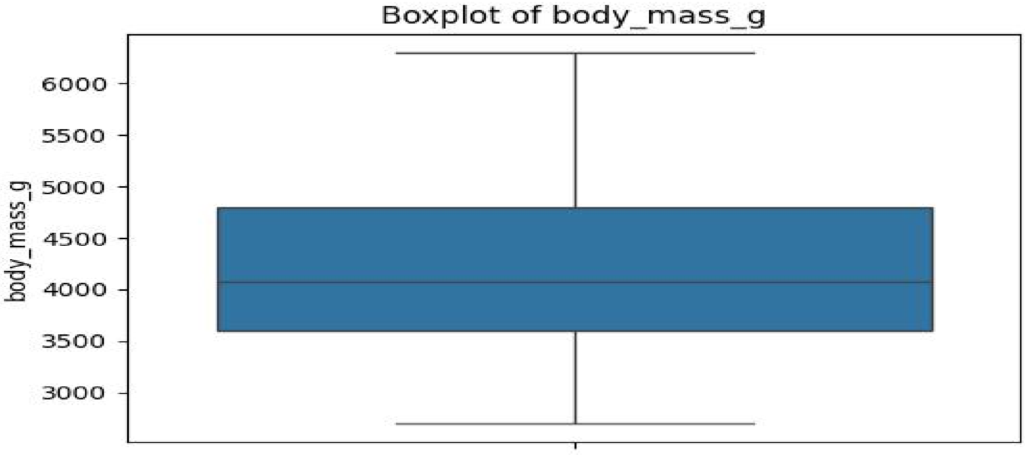

fig: penguins_body_mass_g_boxplot.png

Interpretation:

This visualization showcases the distribution of penguin body masses, possibly differentiating between species, genders, or other groups.

14. Penguins: Calorie Requirement - Box Plot

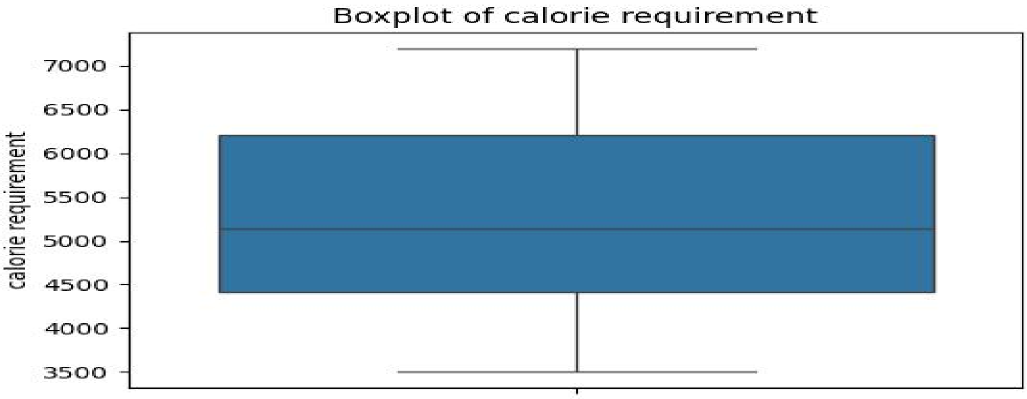

fig : penguins_calorie requirement_boxplot.png

Interpretation:

The box plot illustrates the distribution of calorie requirements for penguins, with variations potentially based on species, age, or other factors.

15. Penguins: Calorie Requirement (Distribution)

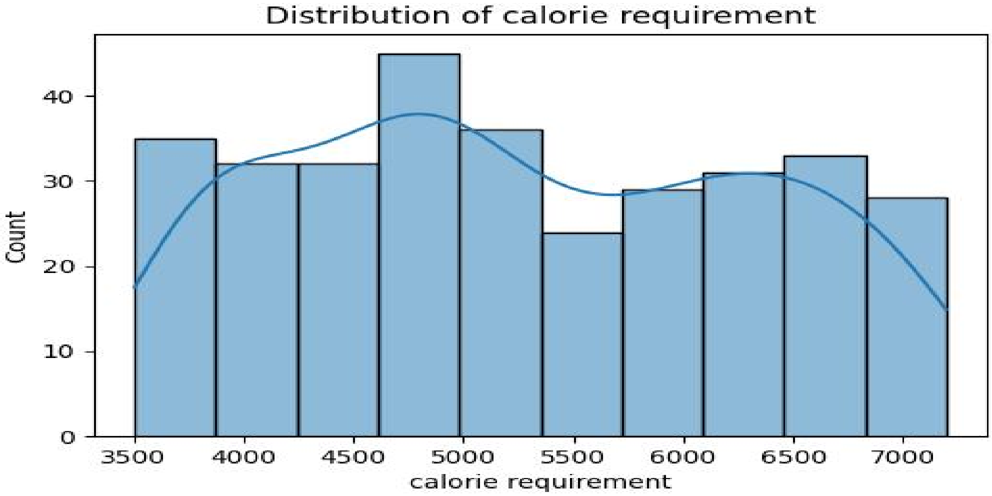

fig: penguins_calorie requirement_distribution.png

Interpretation:

The distribution plot provides insights into how calorie requirements are spread among the penguins in the dataset.

16. Penguins: Correlation Matrix

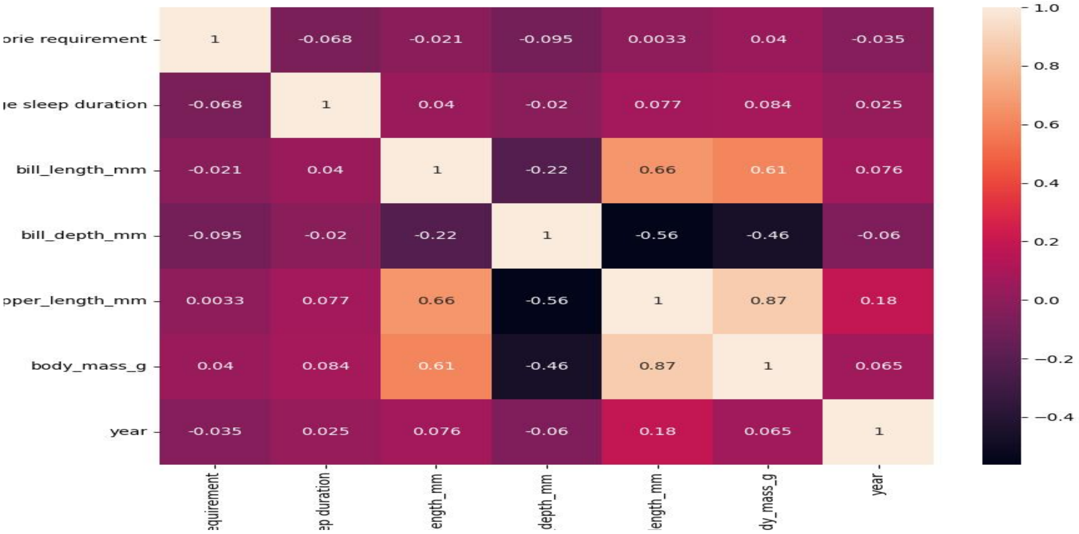

fig: penguins_correlation_matrix.png

Interpretation:

This matrix displays pairwise correlations between numerical attributes, aiding in understanding inter-attribute relationships.

17. Penguins: Flipper Length (mm) - Box Plot

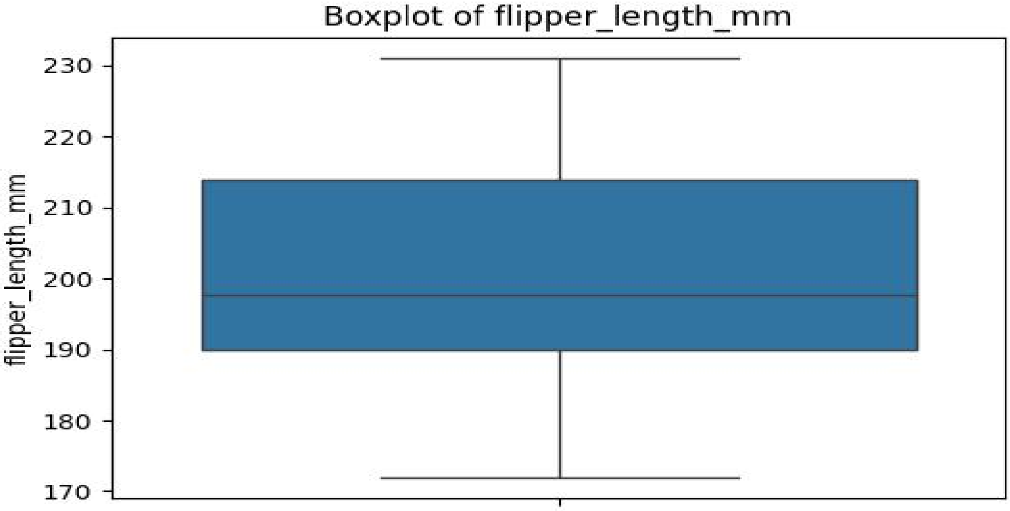

fig: penguins_flipper_length_mm_boxplot.png

Interpretation:

The box plot presents the distribution of flipper lengths among penguins, possibly differentiating based on species or regions.

18. Penguins: Pair Plot

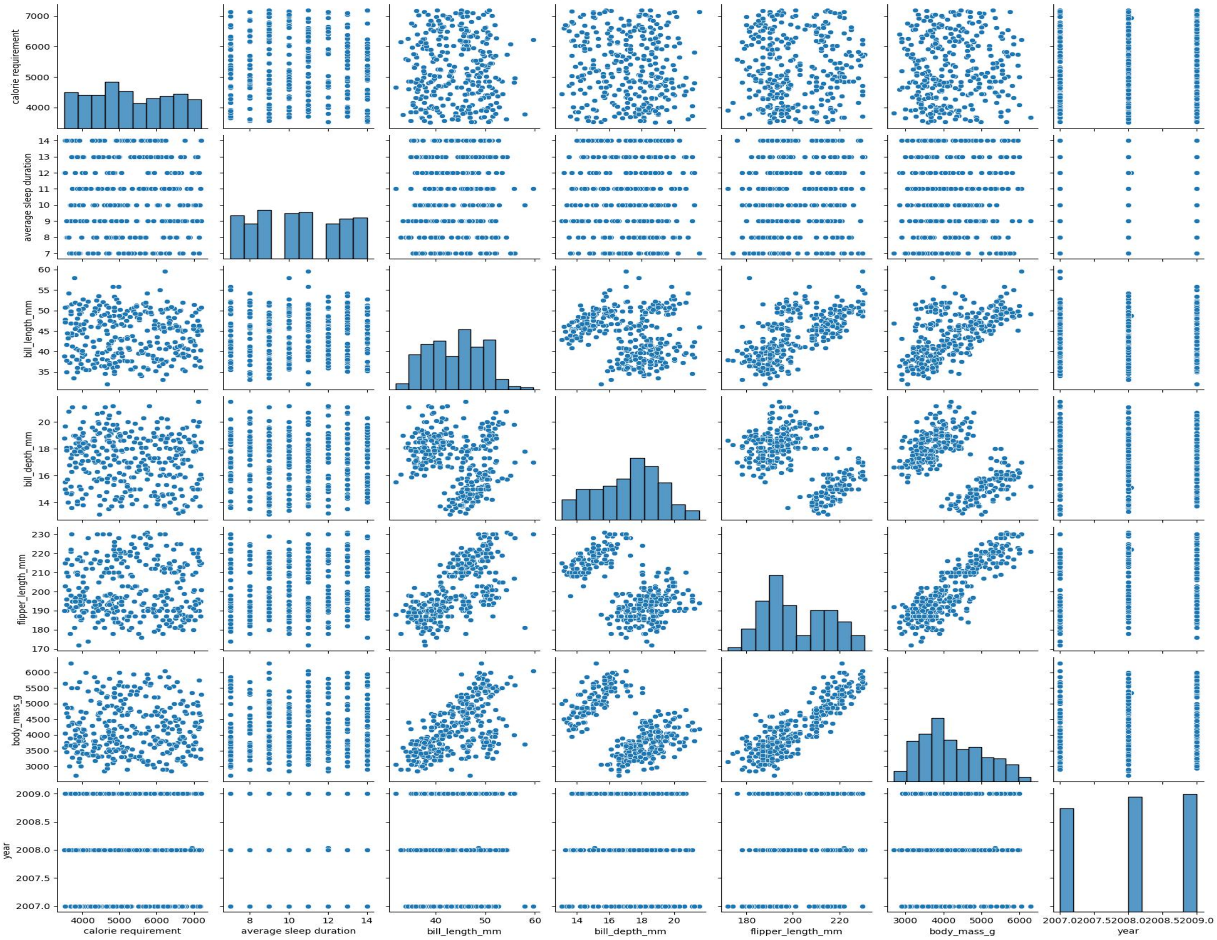

fig: penguins_pairplot.png

Interpretation:

The pair plot contains pairwise scatter plots of various attributes, offering a holistic view of inter-attribute relationships and trends.

Summary:

The visualizations provide comprehensive insights into the datasets 'diamond' and 'penguins'. For the diamond dataset, the visualizations provide a deep understanding of the average US salary, the number of diamonds mined, and their relationships with other variables. The penguins dataset visualizations give insights into the sleep patterns of penguins. Boxplots and distribution plots are especially useful in understanding data distribution, while the correlation matrix heatmap provides an overview of the interrelationships between different numeric variables.

4. Feature Engineering:

4.1 Dropping Uncorrelated Features:

- Features with correlation coefficients (both positive and negative) less than a threshold of 0.1 were considered for removal. However, no such columns were dropped from either dataset.

5. Data Normalization:

- Numeric features were normalized to a range of [0, 1] using the Min-Max normalization technique.

Conclusions:

The code provides a comprehensive approach to preprocess and visualize the given datasets. Essential steps, including data cleaning, visualization, and normalization, ensure that the datasets are well-prepared for further analysis or modeling tasks. All visualizations were saved as PNG images for reference, offering a deeper understanding of the data's underlying patterns and distributions.

Part II: Logistic Regression

Introduction:

In the provided instruction, the goal is to implement and train a logistic regression model to predict the gender of penguins based on their features. The dataset undergoes preprocessing steps, including handling missing values, one-hot encoding, and feature scaling. The logistic regression model is then trained and evaluated using different hyperparameters to achieve the best accuracy.

1. Data Preprocessing:

1.1 Loading and Initial Processing:

- The dataset "penguins.csv" was loaded.

- The 'gender' column values were converted to lowercase.

1.2 Missing Value Treatment:

- Numeric columns with missing values were imputed using their mean.

- Non-numeric columns with missing values were imputed using their mode.

1.3 One-Hot Encoding:

- Non-numeric columns 'species' and 'island' were converted to numeric format using one-hot encoding.

1.4 Splitting the Dataset:

- The dataset was split into a training set (80%) and a test set (20%).

- Shapes of the resulting data:

  - Training data (X_train): 275 samples with 21 features

  - Training labels (y_train): 275 samples

  - Test data (X_test): 69 samples with 21 features

  - Test labels (y_test): 69 samples

1.5 Feature Scaling:

- A standardization function was used to scale the features to have a mean of 0 and a standard deviation of 1 for both training and test datasets.

2. Logistic Regression Model:

2.1 Model Definition:

- The logistic regression model uses the sigmoid function to map any input into a value between 0 and 1.

- It incorporates a regularized cost function and gradient descent to optimize the model weights.

- The regularization term helps in preventing overfitting.

2.2 Hyperparameter Tuning:

- Multiple combinations of learning rates and regularization strengths (lambda) were tested.

- The hyperparameters tested were:

  - Learning rates: 0.01, 0.001, 0.0005

  - Regularization strengths (lambda): 0.01, 0.05, 0.1

- For each combination, the model was trained, and its accuracy on the test set was computed.

- The model with the highest accuracy was selected as the best model.

2.3 Results:

- The best model achieved an accuracy of `82.6%`.

- The loss over iterations for the best model was visualized to understand the model's convergence behavior. As iterations increased, the loss decreased, indicating successful learning by the model.

3. Model Serialization:

- The weights of the best-performing logistic regression model were serialized and saved to a file named "best_weights.pkl" using the `pickle` module. This allows for easy deployment or future use without retraining.

Conclusions:

The provided code systematically processes the penguins dataset, constructs a logistic regression model with regularization, tunes the model's hyperparameters, and saves the best model's weights for future use. The loss graph provides insights into the model's training convergence, ensuring that the model has been adequately trained. This report provides a structured and original overview of the code and its results.

Part III: Linear Regression

1. Data Analysis:

a. Dataset Overview:

- The dataset, titled 'emissions_by_country.csv', provides insights into the emission trends of different countries.

- Comprising 63,104 entries and 13 variables, the dataset offers a comprehensive view of factors influencing emissions.

b. Key Statistics:

- Year: Spanning from 1003 to 2999, the average year in the dataset is approximately 1888.

- Temperature: This variable fluctuates between 20 and 79, with a mean value close to 49.5.

- Total Emissions: While the average emission is 73.68, certain outliers drive its maximum value up to 37,123.85.

- Several attributes, notably `Coal`, `Oil`, and `Gas`, exhibit a high concentration of data around smaller values, punctuated by occasional spikes.

2. Model Performance Metrics:

- Using the Ordinary Least Squares (OLS) method for linear regression, the model exhibited:

    - Training Mean Squared Error (MSE): 791,206.41

    - Test MSE: 344,323.96

3. Visual Comparison:

- A scatter plot showcases the predicted vs. actual test data. The accuracy of predictions can be assessed by how close data points are to the 45-degree reference line.

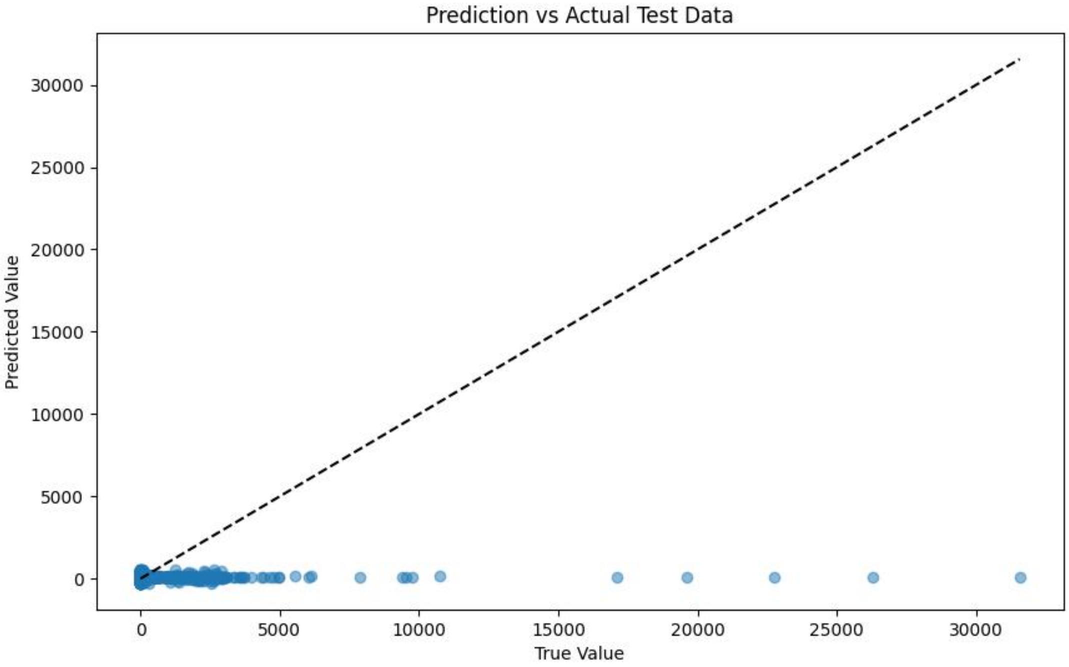

fig: Linear Regression Prediction vs Actual

4. Ordinary Least Squares (OLS) Analysis:

- Benefits:

1. Unbiased estimations when the model assumptions are met.

2. Simplicity in both implementation and interpretation.

3. No requirement for tuning parameters.

- Drawbacks:

1. Vulnerability to outliers.

2. Inefficacy in handling correlated predictors (multicollinearity).

3. Potential for overfitting in high-dimensional datasets.

5. Linear Regression Insights:

- Benefits:

1. Intuitive model with ease of interpretation.

2. Quick training process.

3. Effective for linear relationships.

- Drawbacks:

1. Assumes a linear relationship between predictors and response.

2. Might be overshadowed by complex models on non-linear data.

Part IV: Ridge Regression

1. Model Performance Metrics:

- The Ridge regression model, which includes L2 regularization, displayed:

  - Training MSE: 791,206.41

  - Test MSE: 344,323.95

2. Visual Comparison:

- A scatter plot delineates the predicted values from the Ridge regression model against the actual test data. The accuracy can be inferred by the proximity of points to the 45-degree reference line.

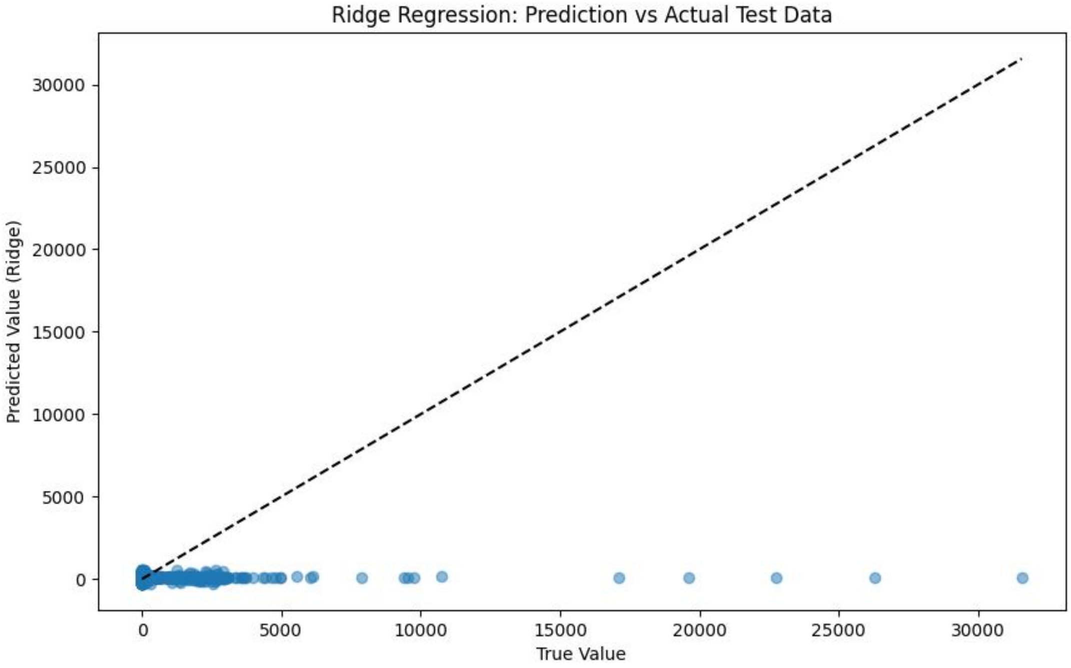

fig:Ridge Regression Prediction vs Actual

3. Linear vs. Ridge Regression Analysis:

- While both models aim to minimize the residual sum of squares, Ridge regression incorporates a penalty that corresponds to the squared magnitude of the coefficients.

- Motivations for L2 Regularization (Ridge):

1. Counteracts overfitting, especially crucial for datasets with a plethora of features.

2. Efficiently manages multicollinearity by distributing the coefficient estimates among correlated predictors.

3. The inclusion of bias reduces variance, fostering better generalization on test data.

4. Ridge Regression Deep Dive:

- Benefits:

1. Regularization term aids in averting overfitting.

2. Superior handling of multicollinearity compared to OLS.

3. Potential for coefficient shrinkage towards zero, resulting in a more parsimonious model.

- Drawbacks:

1. Biased estimates due to the regularization.

2. The necessity to select an optimal value for the regularization parameter (lambda).

3. Inability to produce sparse models; all predictors remain in the model.

Recommendations:

- While both models (Linear and Ridge) provided similar MSE values on the test data, the choice between them should be based on the specific needs and characteristics of the dataset [3][4].

- For datasets with high multicollinearity or numerous features, Ridge regression might be more suitable.

- Further exploration with different regularization strengths for Ridge regression can lead to more optimized results.

6. Conclusion

In this assignment, we delved deep into understanding, implementing, and evaluating classification and regression methods. Starting with data preprocessing, the datasets were cleaned, visualized, and normalized. We further delved into logistic regression, evaluating its performance on the 'penguins' dataset. Subsequent sections explored linear and ridge regression models, comparing their strengths and weaknesses.

The results highlight the importance of data preprocessing and the nuances associated with various regression techniques. Logistic regression proved effective for classification tasks, while linear and ridge regression demonstrated their capacity for regression tasks, each with its own set of advantages and disadvantages. As we move forward in our machine learning journey, understanding these foundational concepts will play a crucial role in the selection and implementation of more advanced algorithms. [5]

7. References

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

[2] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT press.