1. Introduction

Bird diversity plays a key role in maintaining balanced ecosystems. This study examines 95 bird species across eight zones in a region of Australia, identified by SITE_ID, with data on flight behaviors (Height_max, Height_min), weather conditions, habitat types, and location coordinates. By analyzing these factors, the study aims to understand influences on flight patterns and habitat choices, which are critical for species survival and ecological stability. Utilizing this data provides an efficient approach to predicting bird species and behaviors, reducing the need for extensive field observation.

Research Questions:

a. Predict bird species using flight height and habitat?
In this question on the base of two exploratory variables Height_Max and HABITAT, we predict bird belong to which species.

b. Is there a link between maximum flight height, species, and habitat?
This question checks if certain species fly higher or lower depending on their habitat.

c. Does adding geographic coordinates to flight height and site ID improve species prediction?
We want to see if using location data with flight height and site ID helps identify species more accurately.

Motivation: This study helps us better understand bird behaviors and their preferred habitats. By predicting species and understanding their flight patterns, we can improve conservation efforts and ecological planning to protect bird diversity

2. Methods and Analysis

2.1. Overview Of Data and Statistical Analysis/ EDA

Analysis of Maximum Height Normalization:

We used a boxplot, density plot, and histogram to examine the distribution of Maximum Height, assessing data behavior and normality.

Histogram:

The histogram shows Height_max values clustered at low levels with a spike near zero and a long right tail, indicating a right-skewed distribution.

BoxPlot:

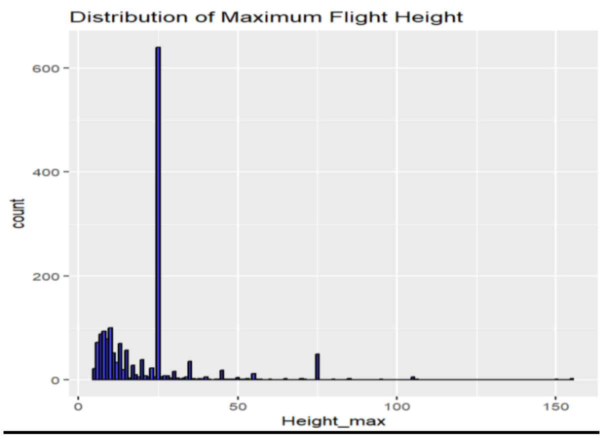
The box plot for Height_max shows a lower range cluster with many high outliers, indicating a right-skewed pattern where a few birds fly much higher than most.

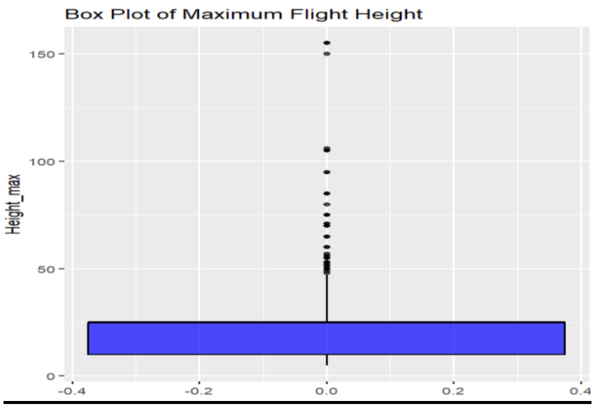
Density Plot:

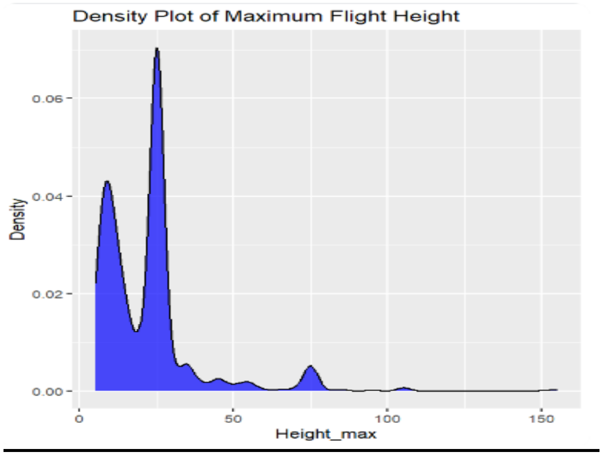
The density plot for Height_max peaks at low values, with a right tail showing a few higher altitudes, indicating right skew.

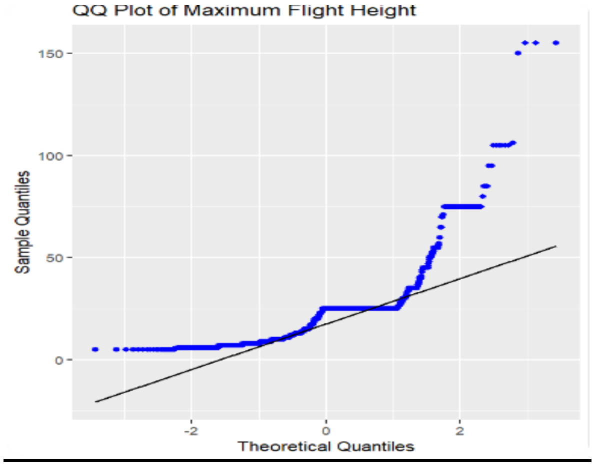
QQ Plot:

The QQ plot for Height_max shows strong deviation from normal, especially in upper quantiles, indicating a right skew with high outliers.









The plots indicate that Height_max is skewed and not normally distributed, so models or tests should adjust for skewness or transform the data to meet normality.

Analysis After Transformation

We used the same visualizations—histogram, density plot, QQ plot, and boxplot—to check the distribution of the transformed bird data after applying to remove the outliers from the Height_max data.

Box Plot:

The transformed Height_max box plot shows a compact, balanced distribution without outliers, indicating reduced skewness.

Histogram:

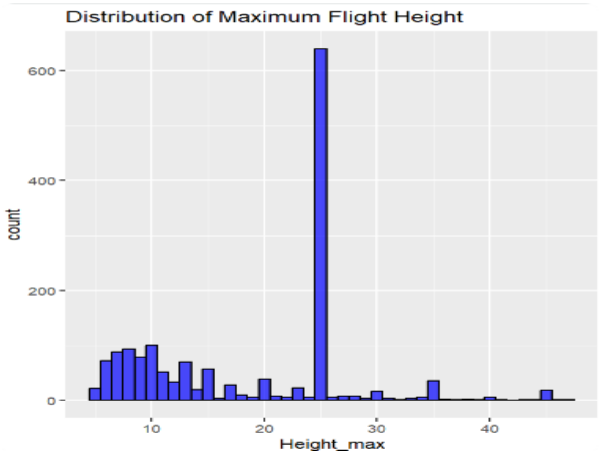
The histogram for Height_max shows a large concentration of values around a low range, with a peak near 15. There are fewer counts as height increases, creating a right-skewed distribution.

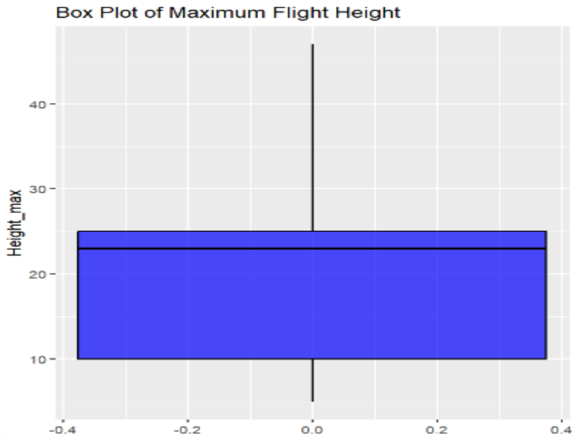
Density Plot:

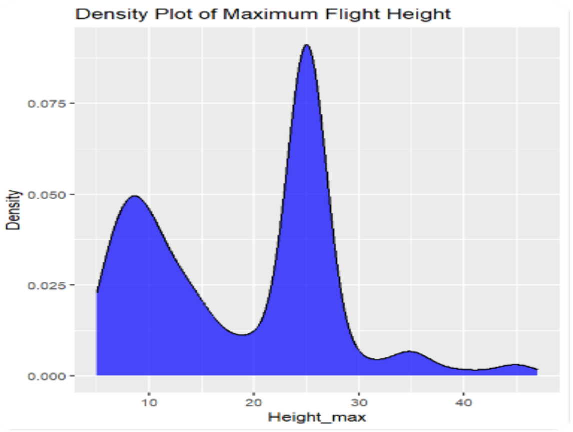
The density plot for Height_max shows two main peaks, with a strong concentration of around 15 and a smaller peak of around 10. This suggests a bimodal distribution with most values in the lower range.

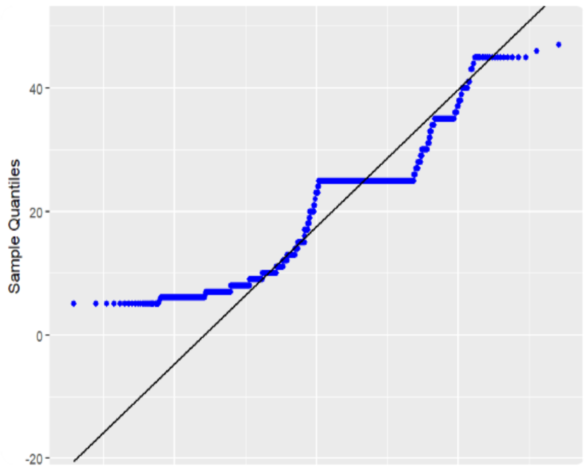
QQ Plot:

The QQ plot for Height_max after handling outliers shows some improvement toward normality, but deviations from the line remain, especially in the upper quantiles. This indicates that the data still has some non-normal characteristics.



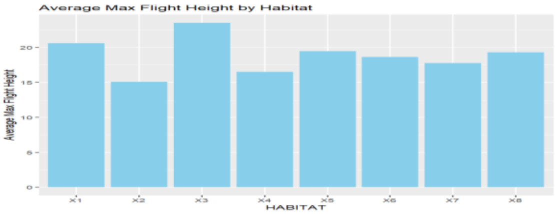






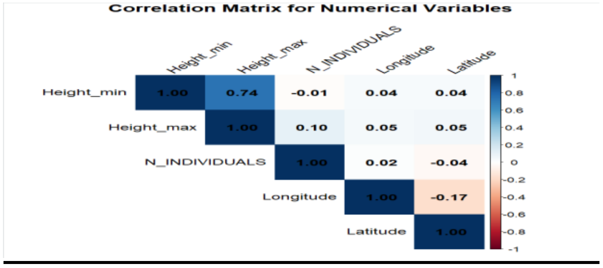
After handling outliers and applying transformations, Height_max shows a more compact and balanced distribution, reducing skewness. However, QQ plot analysis reveals the data still deviates from normality, particularly in higher values.

Comparison:



The chart compares average maximum flight heights across habitats X1 to X8, with X3 showing the highest and X2 the lowest. Other habitats have comparable heights, suggesting habitat may slightly influence flight height.

Correlation:



The matrix reveals a strong positive correlation (0.74) between Height_min and Height_max, while other variables show weak correlations. Latitude and Longitude have a slight negative correlation (-0.17), indicating minimal overall linear relationships.

2.2. Method/Model Selection For Question A

To predict bird species based on flight height and habitat, we selected a **classification decision tree**. This approach is suitable because our response variable (**SPECIES**) is categorical, and decision trees are designed to predict categorical outcomes effectively, we have two other explorary variables **HABITAT(**Categorical**)** and **Height_max(**Numeric Continues**)**. Additionally, the **non-linearity** of our data aligns well with the decision tree's capability to handle complex patterns without requiring a linear relationship. This method allows us to make accurate predictions about species based on these specific characteristics.

To find the best size of the tree we used the K folds method and through cross-validation we got a minimum error at 6, so we will have 6 terminal nodes. We can see in Figure 2.1.

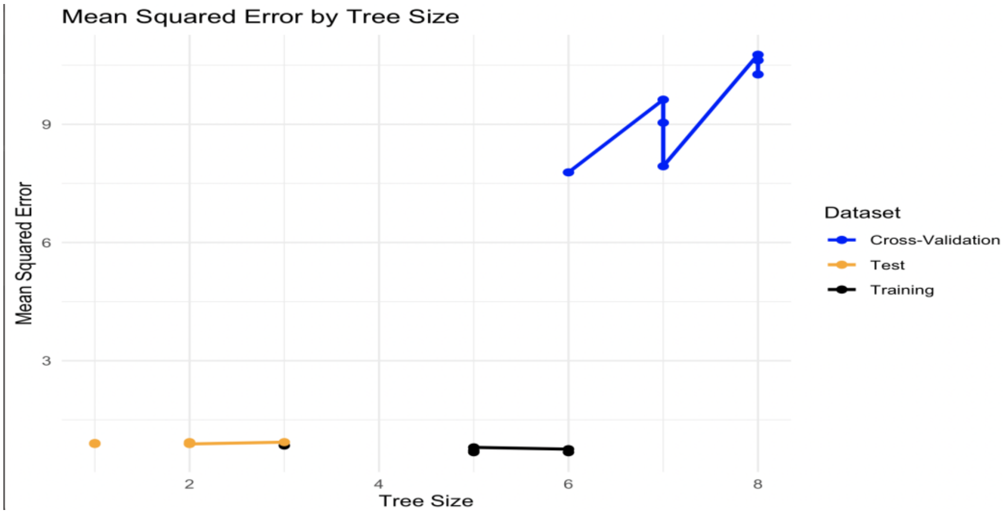


Figure 2.1

2.3. Method/Model Selection For Question B

To address the question of whether there is a clear link between maximum flight height, species, and habitat, we opted for **Multiple Linear Regression (MLR)**. This approach is well-suited to our analysis because our response variable, **`Height_max` (maximum flight height)**, is continuous, while the explanatory variables, **`Species` and `Habitat`**, are categorical. MLR allows us to explore how these categorical predictors influence a continuous outcome and is particularly effective when investigating relationships between multiple predictors and a single outcome.

$$Yi = \beta 0 + \beta 1X1i + \beta 2X2i + \epsilon i$$

where (Yi) denotes maximum flight height, (X1i) denotes species, and (X2i) denotes habitat type.

2.4. Method/Model Selection For Question C

We investigated whether adding geographic coordinates (latitude and longitude) to maximum flight height and site ID improves species prediction accuracy. Using **K-Nearest Neighbors (K-NN) classification**, we tested two models: one with only flight height and site ID, and another including geographic coordinates. We then compared hit and misclassification rates to assess if geographic data enhances species classification.

3. Results and Discussion

3.1. PART A

To predict bird species using flight height and habitat, we filtered the dataset to records with a maximum height of 30 meters, simplifying the analysis to focus on key species. We used a **Classification Decision Tree model**, suited for categorical outcomes and non-linear relationships. The initial model, with 8 terminal nodes, showed limited effectiveness in distinguishing species based on the selected variables .

To enhance the model's performance, we implemented cross-validation, specifically using **K-fold methods** to assess how well the model would generalize to an independent dataset. Cross-validation also helped identify the **optimal size of the tree (6)**. Following this, we pruned the tree based on cross-validation results to reduce overfitting, resulting in a pruned tree with **6 terminal nodes** showing in Figure 3.2**.**

Before Pruning

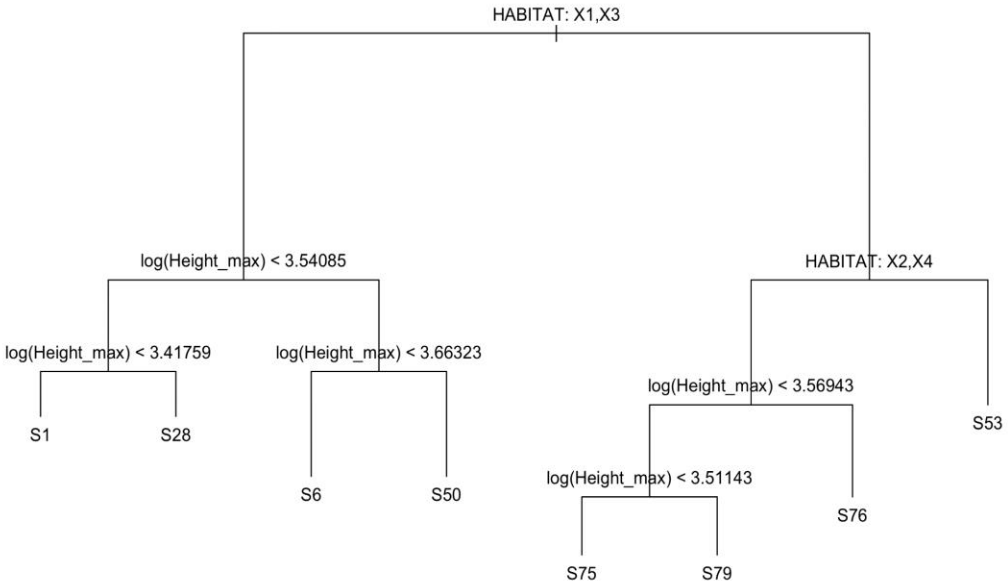

HABITAT: X1,X3
log(Height_max) < 3.54085
HABITAT: X2,X4
log(Height_max) < 3.41759
log(Height_max) < 3.66323
log(Height_max) < 3.56943
S53
S1
S28
S6
S50
log(Height_max) < 3.51143
S76
S75
S79

Figure 3.1

After Pruning

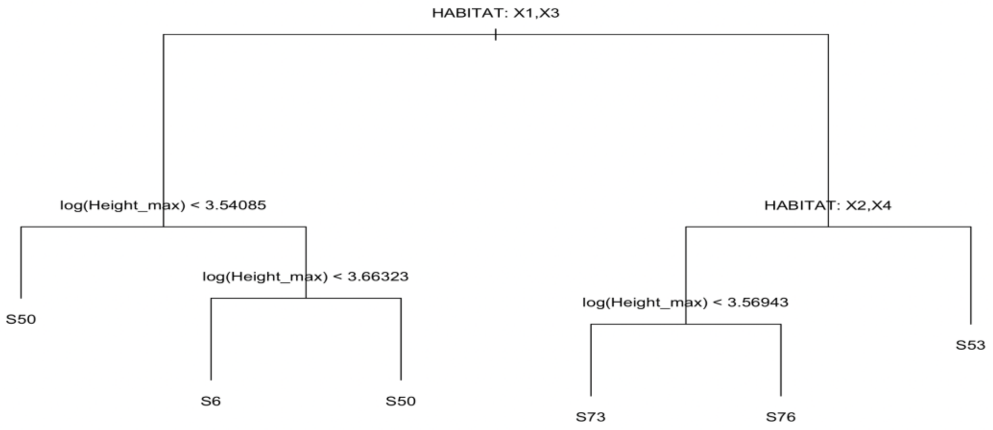

HABITAT: X1,X3
log(Height_max) < 3.54085
HABITAT: X2,X4
log(Height_max) < 3.66323
log(Height_max) < 3.56943
S50
S53
S6
S50
S73
S76

Figure 3.2

We evaluated the model's performance on a test set and computed the accuracy, which was only about 13.63%. This low accuracy suggests that while Height_max and habitat are important factors, they may not be sufficient alone for accurately predicting bird species. It implies that additional predictors.

3.2. PART B

To examine the link between maximum flight height, species, and habitat, we applied Multiple Linear Regression (MLR), finding no significant relationship among these variables. The high p-value (F = 0.9581, p = 0.5445) led us to fail to reject the null hypothesis, indicating that species and habitat do not significantly affect flight height. Additionally, confidence intervals for many species and habitat predictors, such as SPECIESS13 [-0.27, 0.39] and HABITATX2 [-0.18, 0.61], crossed zero, suggesting unreliable effects. These results indicate a weak, inconsistent association between response variable [Height_max] and the explanatory variables[HABITAT and SPECIES], confirmed by the non-significant regression findings.

Overall, the analysis does not support a clear link between maximum flight height, species, and habitat.

Check Assumptions Of Model

Normality of Residuals:

The Q-Q plot shows that residuals approximately follow a normal distribution, with most points near the diagonal, indicating the normality assumption is reasonably met.

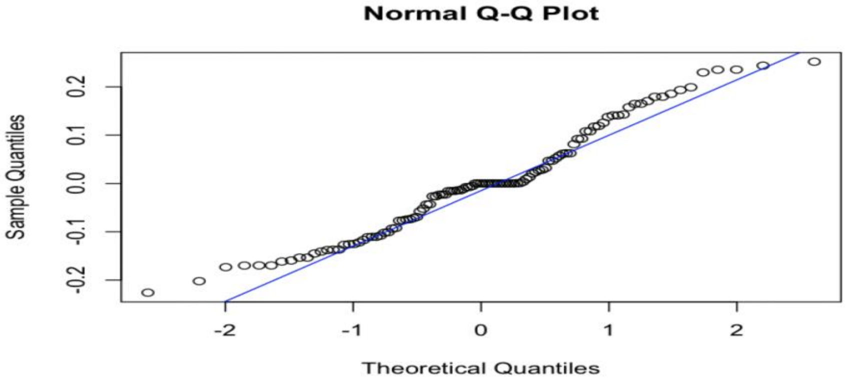


Figure 3.3

Homoscedasticity (Constant Variance of Residuals):
The residuals vs. fitted values plot shows a mostly random scatter around zero, suggesting that constant variance is generally met, though slight patterns may warrant further investigation..

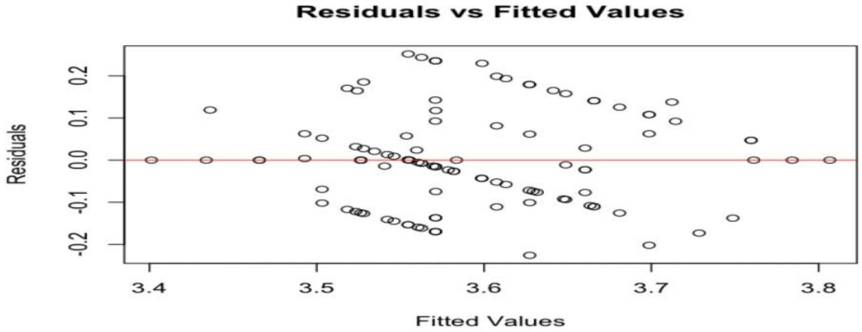


Figure 3.4

Homoscedasticity (Constant Variance of Residuals):
The Variance Inflation Factor (VIF) values are all below the threshold of 5, indicating that multicollinearity is not a concern in this model, ensuring the stability and reliability of the coefficients.

3.3. PART C

Yes, combining geographic coordinates with maximum flight height and site ID slightly improves the model's ability to predict species, as shown by the increase in accuracy from 4.47% to 4.55%. However, this improvement is minimal, indicating that while geographic variables contribute some predictive value, they do not significantly enhance the model's accuracy for species prediction in this dataset.

Using **K-Nearest Neighbors (K-NN) classification**, we tested two models: one with only Height_max and SITE_ID, and another that also included geographic coordinates (Longitude and Latitude). The model without geographic data achieved a low accuracy of 4.47%, with a high misclassification rate of 95.53%. Adding geographic coordinates improved accuracy slightly to 4.55% and reduced misclassification to 95.45%. This modest increase suggests that while geographic data offers a minor improvement in predictive accuracy, both models show limited effectiveness in accurately classifying bird species in this dataset.

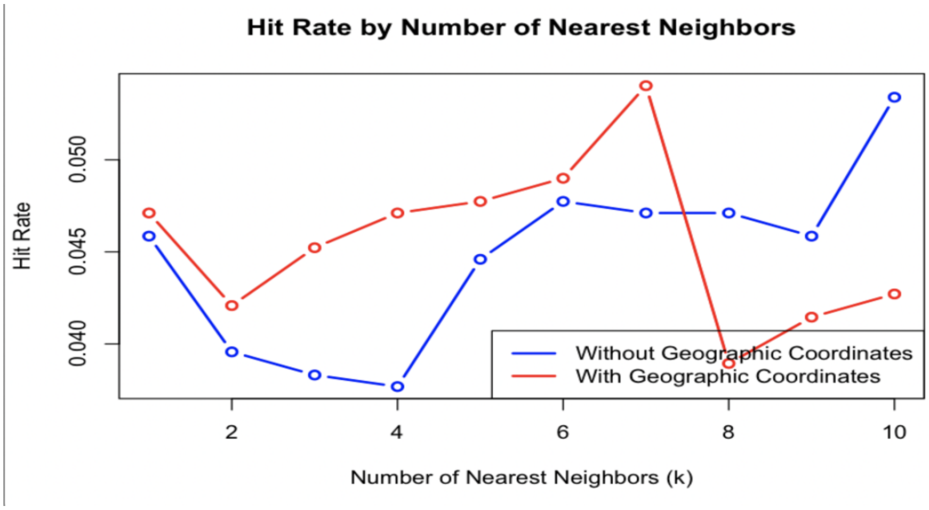


Figure 3.5

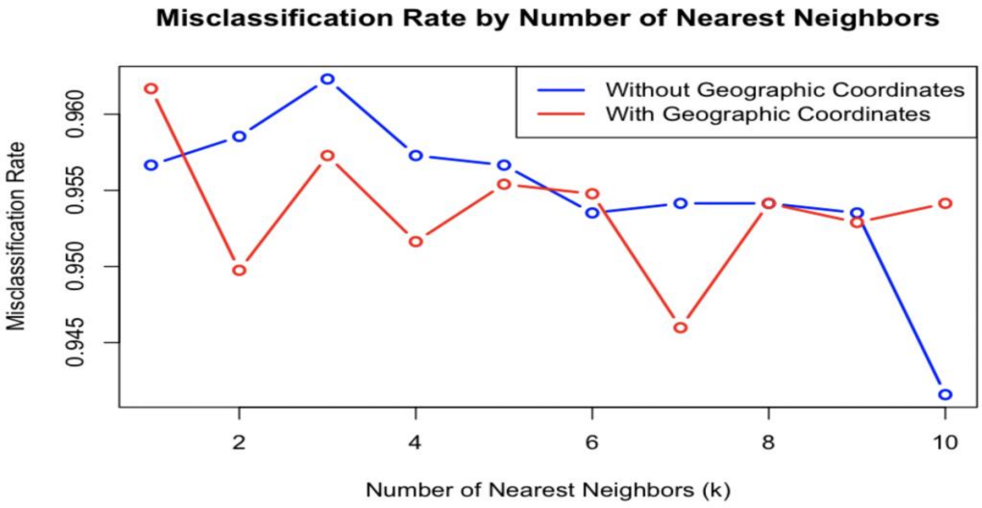


Figure 3.6

	Without_Geographic_Coordinates	With_Geographic_Coordinates
Minimum Hit Rate	0.0377 (3.77%)	0.0389 (3.89%)
Maximum Hit Rate	0.0534 (5.34%)	0.0540 (5.40%)
Overall (Average) Hit Rate	0.0447 (4.47%)	0.0455 (4.55%)
Minimum Misclassification Rate	0.9466 (94.66%)	0.9460 (94.60%)
Maximum Misclassification Rate	0.9623 (96.23%)	0.9611 (96.11%)
Overall (Average) Misclassification Rate	0.9553 (95.53%	0.9545 (95.45%)

Table 3.1

4. Conclusion

This study investigated bird species prediction and behavior across three key questions, employing **Decision Tree, Multiple Linear Regression, and K-Nearest Neighbors (K-NN) classification** models. First, we used flight height and habitat to predict bird species, achieving limited accuracy, suggesting that additional variables may be necessary for reliable classification. Next, we examined whether species and habitat influenced maximum flight height, finding no significant relationship in our **MLR** model. Lastly, we assessed the impact of adding geographic coordinates to species prediction using **K-NN**, while it slightly improved accuracy, the effect was minimal. Overall, these findings highlight the need for more robust predictive variables to effectively capture bird species behavior and habitat preferences.

Group Evaluation Form

Student (Student No.)	% Contribution	Signature
Muhamad Hammad Arshad(34721387)	**33.3**	Dhruvil
Hafiz Samran Elahi(34721242)	**33.33**	samran elahi
Owais Safdar(34721322)	**33.33**	Owais
Total	**100%**	