# MA334 Individual Assignment

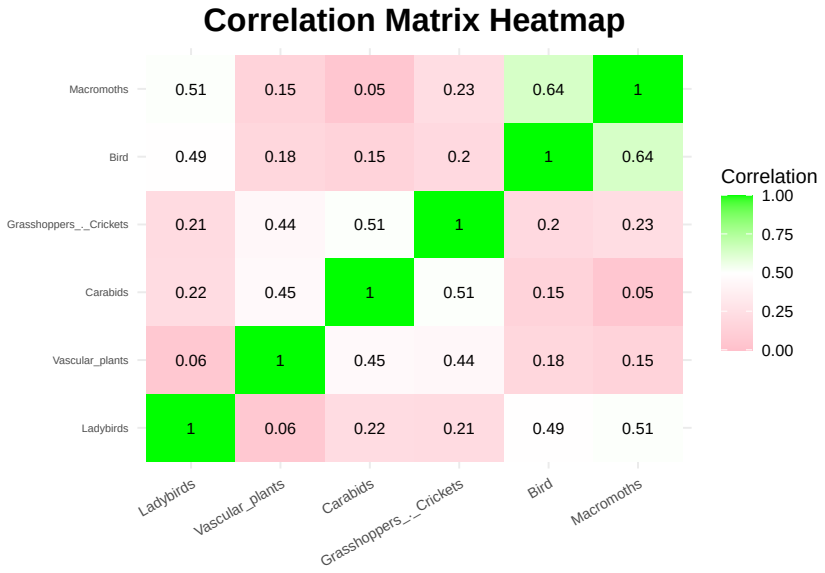
Sakthivel Vinayagam (2400589)

## Univariate analysis and basic Statistics:

| | Summary Statistics | | | | | | |
|---|---|---|---|---|---|---|---|
| Taxonomic_group | Min | 1st Q | Median | Mean | 3rd Q | Max | T25_win_mn |
| **Ladybirds** | 0.0614 | 0.1837 | 0.3853 | 0.4439 | 0.6357 | 1.8400 | 0.4067 |
| **Vascular_plants** | 0.4179 | 0.7244 | 0.7975 | 0.7804 | 0.8508 | 1.0000 | 0.7901 |
| **Carabids** | 0.0115 | 0.2871 | 0.5125 | 0.4884 | 0.6712 | 1.0000 | 0.4891 |
| **Grasshoppers_._Crickets** | 0.1290 | 0.4259 | 0.6042 | 0.5888 | 0.7660 | 1.0943 | 0.5974 |
| **Bird** | 0.2415 | 0.7634 | 0.8580 | 0.8287 | 0.9208 | 1.0541 | 0.8469 |
| **Macromoths** | 0.0895 | 0.6791 | 0.8355 | 0.7932 | 0.9254 | 1.2604 | 0.8119 |

The dataset gives an overview of biodiversity metrics for six taxonomic groups in Scotland: Ladybirds, Vascular Plants, Carabids, Grasshoppers & Crickets, Birds, and Macromoths. Vascular Plants are found to have the greatest stability; the median value is 0.7975, and the 25% Winsorized mean is 0.7901. In contrast, with a maximum of 1.84, Ladybirds display the most variability. An interesting point for Birds is to note their fairly high 3rd Quartile (0.9207) and fairly steady mean (0.8287), indicating its ecological relevance. Carabids show the supposed lowest minimum value (0.0115), which indicates little biodiversity in some areas. Grasshoppers & Crickets, while moderate in variability, show that macromoths have a rather wide range: mean of 0.7932 and maximum of 1.2604. This analysis stresses the biodiversity and variation of the taxonomic groups in Scotland.

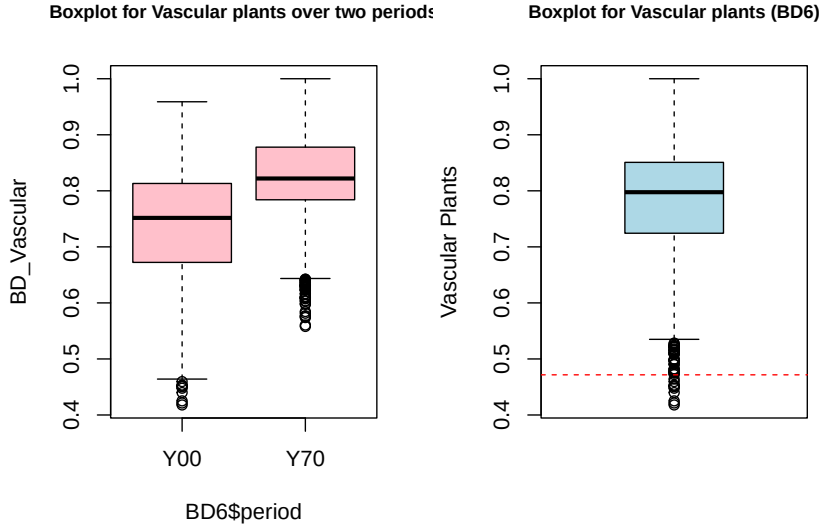
## Correlation:




The correlation matrix reveals insights regarding the relationships between six ecological variables Ladybirds, Vascular Plants, Carabids, Grasshoppers & Crickets, Birds, and Macromoths. There was a strong positive correlation between Macromoths and Birds (0.64), suggesting a notable association, and also a negligible correlation between Ladybirds and Vascular Plants (0.06). Carabids and Grasshoppers & Crickets (0.51) also had a moderate positive correlation. For most other pairwise correlations, there is some variety in weak or moderate correlation, reflecting the fact that our

variables are associated to different extents. This analysis therefore hints at some dependencies among the variables that may need further exploration.

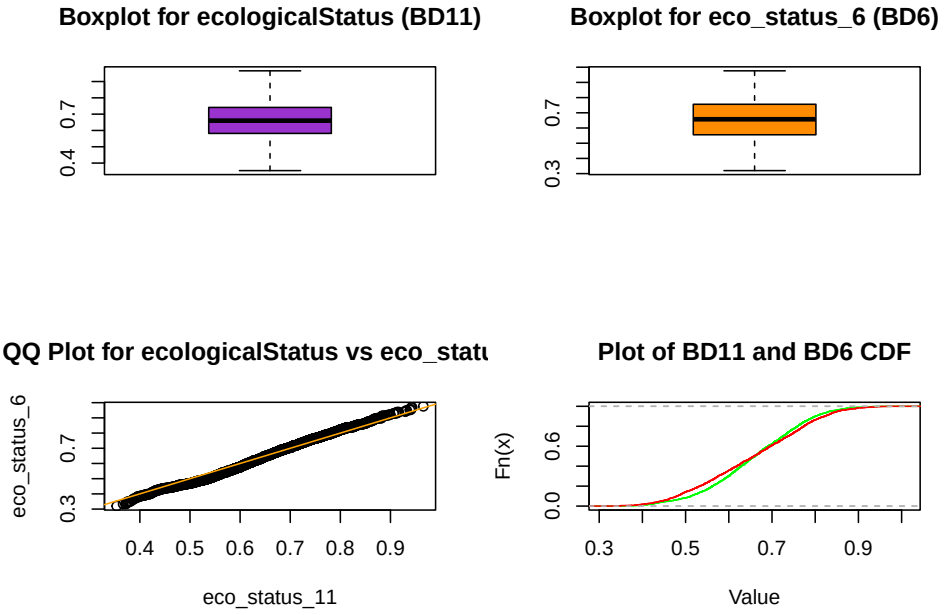
## Boxplot:




The boxplots illustrated a two-period comparison between vascular plants, say Y00 and Y70, would postulate the latter to have a higher median compared to the former, generally depicting an improvement or increase in the measured ecological status. However, Y70 also has more outliers, indicating higher variability and some extreme observations. Further, the BD6 box plot reinforces this trend by being consistent with regard to the median but having many outliers below 0.5, thus showing some very unusual values within the dataset. Identifying outliers, a range of 0.42 to 0.46 reveals deviations that could signal specific anomalies or unique ecological conditions. Taken altogether, these results suggest improvement in the vascular plants status with time but increasing variability, possibly worthy of deeper examination.

## Hypothesis Tests:

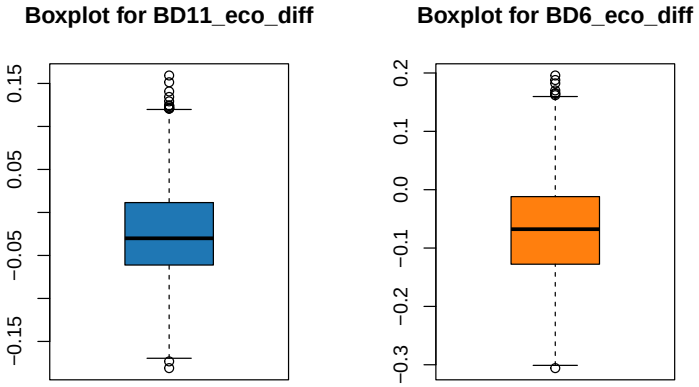
## Kolmogorov-Smirnov (KS) test:




The Kolmogorov-Smirnov test was carried out in order to compare the distributions of eco_status_11 and

eco_status_6. It returned a test statistic =0.072432, which is the maximum difference between their cumulative distribution functions (CDFs). The very low p-value =0.0001218 indicates that the two distributions are significantly different; hence, the null hypothesis that they are identical can be rejected. The 92% confidence interval for the D statistic is also (0.01344205, 0.1314228). It corroborates the reliability of this difference. All these results indicate that the ecological status values for eco_status_11 and eco_status_6 are not identically distributed, which points to remarkable differences between the two datasets.

## One sample T-Test:




The results of one-sample t-test on BD6_eco_diff clearly showed that the mean, which was taken as the hypothesis (-0.02278728), differed significantly, with a t-statistic of -15.213 and the very highly significant p-value (<2.2e-16). The average sample (-0.0666) showed a 92% confidence interval (-0.07167 to -0.06157), further at the same time confirming the differentiation. The boxplots comparing BD11_eco_diff and BD6_eco_diff display symmetric distributions but highlight outlier patterns uniquely: upper-range outliers for BD11_eco_diff and lower-range outliers for BD6_eco_diff. These findings suggest that there may be meaningful ecological trends or differences, and with the outliers' presence, further investigations should seek to understand their causation and implications.

# Contingency table/comparing categorical variables:

Table 1: Contingency Table

| | DOWN | UP | Sum |
|---|---|---|---|
| DOWN | 621 | 17 | 638 |
| UP | 113 | 174 | 287 |
| Sum | 734 | 191 | 925 |

This contingency table compares "UP" and "DOWN" transitions between two variables (BD11UP and BD6UP). It provides counts for each combination, row and column totals, and a grand total of 925.

Table 2: Independent Table

| | DOWN | UP | Sum |
|---|---|---|---|
| DOWN | 506 | 132 | 638 |
| UP | 228 | 59 | 287 |
| Sum | 734 | 191 | 925 |

The independent contingency table categorizes "UP" and "DOWN" states, the total cell count being 925. The outcome of the summation is that "DOWN" is shown in the rows' 506 and 132, whereas "UP" is found on fewer occasions (228 and 59). The table implies a potential disproportion in the number of the two states.

## Likelihood-ratio:

The analysis includes results from a log-likelihood ratio (G-test) for independence and a two-sample test for equality of proportions. The G-test shows a highly significant result ($p < 2.2 \times 10\hat{}-16$), which strongly rejects the null

hypothesis of independence between the variables in the contingency table. The calculated proportions of "UP" states are 0.3102703 for BD6 and 0.2064865 for BD11, indicating a notable difference. A two-sample test for equality of proportions, with continuity correction applied, further confirms this difference with a significant p-value of 4.519 × 10^−7. This supports the alternative hypothesis that the proportions are not equal. The confidence interval for the difference in proportions ranges from 0.06608197 to 0.14148560, reinforcing the observed disparity. The estimated proportions for the two groups (prop 1 = 0.3102703, prop 2 = 0.2064865) show a higher proportion of "UP" states in BD6 compared to BD11. These results suggest a systematic difference in the behavior of the two groups, which is statistically significant and unlikely to be due to random chance. The analysis strongly supports the presence of an association between the variables.

## Odds-ratio:

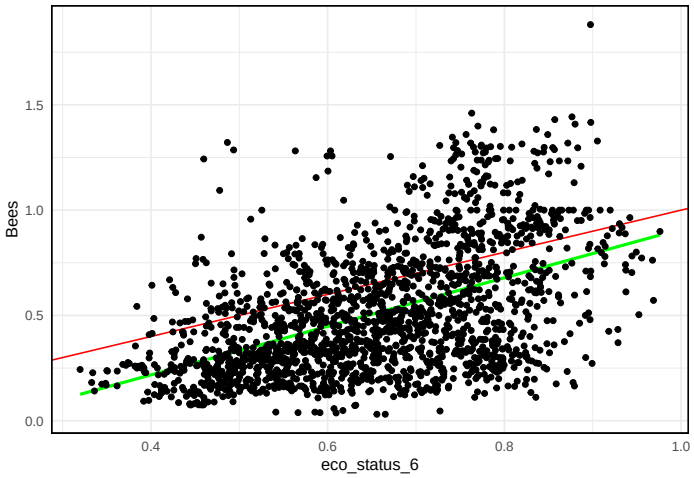
The odds ratio was calculated using two different methods, both resulting in a value of 56.24883. This odds ratio suggests a strong link between the two variables examined, indicating that the odds of the "UP" state in BD11 are roughly 56 times greater than in BD6 when taking into account the respective probabilities. The agreement between the two methods reinforces the reliability of the calculation. This significant odds ratio underscores a marked difference in the distribution of "UP" and "DOWN" states between the two categories.

## Sensitivity, Specificity and Youden's Index:

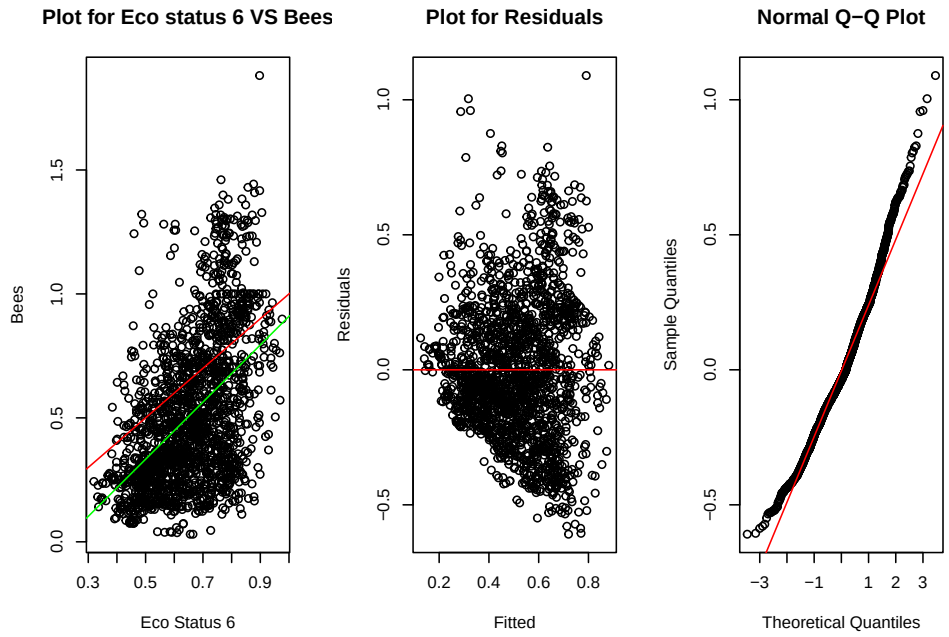
The performance of the classification system was evaluated using sensitivity, specificity, and Youden's index. Sensitivity, which indicates the true positive rate, was found to be 0.9109948, reflecting a strong ability to accurately identify positive cases ("UP"). Specificity, representing the true negative rate, was calculated at 0.846049, demonstrating a robust capability to correctly identify negative cases ("DOWN"). Youden's index, which combines sensitivity and specificity to assess the overall effectiveness of the classification, was computed to be 0.7570438. This value indicates that the system performs well, maintaining a good balance between sensitivity and specificity.

# Simple Linear Regression:

The linear regression analysis reveals a significant positive relationship between BD6_eco_status_6 and BD1_Bees, with a coefficient of 1.15252 (p-value < 2.2e-16), indicating that improvements in BD6_eco_status_6 are associated with an increase in BD1_Bees. The model is statistically significant overall, as evidenced by an F-statistic of 617.3 and a p-value < 2.2e-16. However, the R-squared value of 0.25 suggests that while BD6_eco_status_6 explains 25% of the variability in BD1_Bees, other factors may also influence BD1_Bees and warrant further investigation. The residuals are reasonably distributed, supporting the reliability of the model, though the moderate R-squared highlights the potential need for additional predictors to improve explanatory power.




The scatterplot illustrates the relationship between eco_status_6 (x-axis) and Bees (y-axis), where each point represents an observation. The red line depicts the fitted regression line, showing a positive correlation as Bees values tend to increase with higher eco_status_6. The green line may represent an alternative trend or a smoothed fit, highlighting variations in the data. While the overall trend supports the positive relationship observed in the regression analysis, the scatterplot reveals significant variability around the regression line, with some outliers deviating considerably. This suggests that while eco_status_6 is a strong predictor, other factors might also influence Bees, warranting further exploration or the inclusion of additional variables in the model.




Diagnostic and scatter plots have been used to analyze the relationship between Eco Status 6 and Bees. The smooth regression line on the scatter plot is positive, indicating that with a rise in the values of Eco Status 6, Bee population also increases. The residual plot appears rather randomly scattered around zero, so there is no serious complaint of heteroscedasticity, though minor pattern issues do occur that necessitate inspection. The Q-Q plot shows that the residuals are not objectionable in most aspects of normality, except at the tail ends. Overall, the model effectively captures the relationship between Eco Status 6 and Bees, but slight residual deviations suggest room for model refinement to enhance predictive accuracy.

# Multiple Linear Regression:

The regression analysis indicates that the three significant predictors of the number of Bees are Ladybirds, Grasshoppers/Crickets, and Macromoths. Macromoths have the highest positive effect. The variables Vascular Plants, Carabids, and Birds are not significant and therefore contribute less to the model. These can be considered for exclusion in order to make the model simpler without major loss in explanatory power. The model explains 32.1% of the variance in Bees (R-squared = 0.3211), with an adjusted R-squared of 0.3189, indicating a good balance of explanatory power and model complexity. The residual standard error of 0.2449 and a correlation of 0.5666951 between observed and predicted values suggest moderate model performance. With an AIC of 53.43463, the model effectively balances fit and complexity, providing meaningful insights into factors affecting the Bee population.

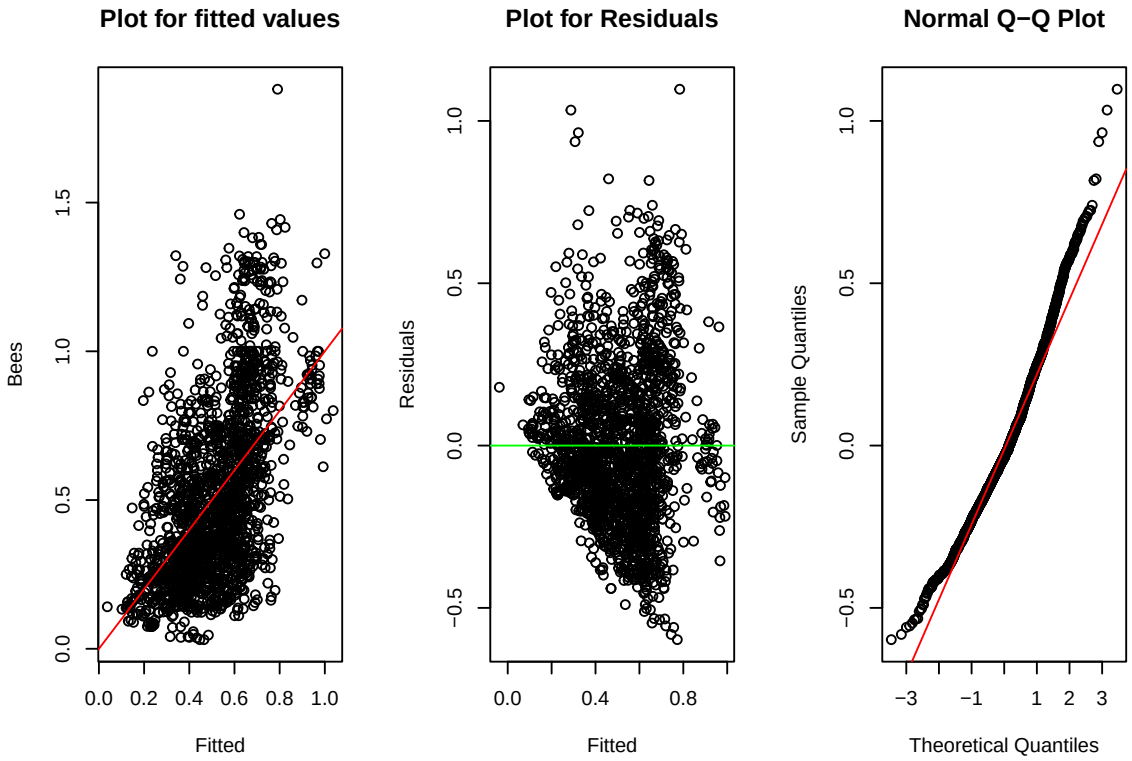
## Reduced Multiple Linear Regression:

The reduced multiple linear regression model identified Ladybirds, Carabids, and Macromoths as significant predictors of the number of Bees, with Macromoths having the strongest positive effect. Non-significant variables, such as Vascular Plants and Birds, contribute minimally to the model and could be excluded for simplicity. The model explained 31.3% of the variance (R-squared = 0.313) and showed a moderate positive correlation between predicted and observed values (correlation = 0.559). The AIC of the reduced model is 73.40625, making it more balanced between simplicity and performance and thereby more interpretable and efficient compared to the full model.

## Multiple Linear Regression Analysis with Interaction Terms:

Table 3: AIC Values for Different Models

| | df | AIC |
|---|---|---|
| multi_lin_mod | 8 | 53.43463 |
| multi_lin_mod_reduced | 7 | 73.40625 |
| multi_lin_mod_interaction | 8 | 69.89372 |

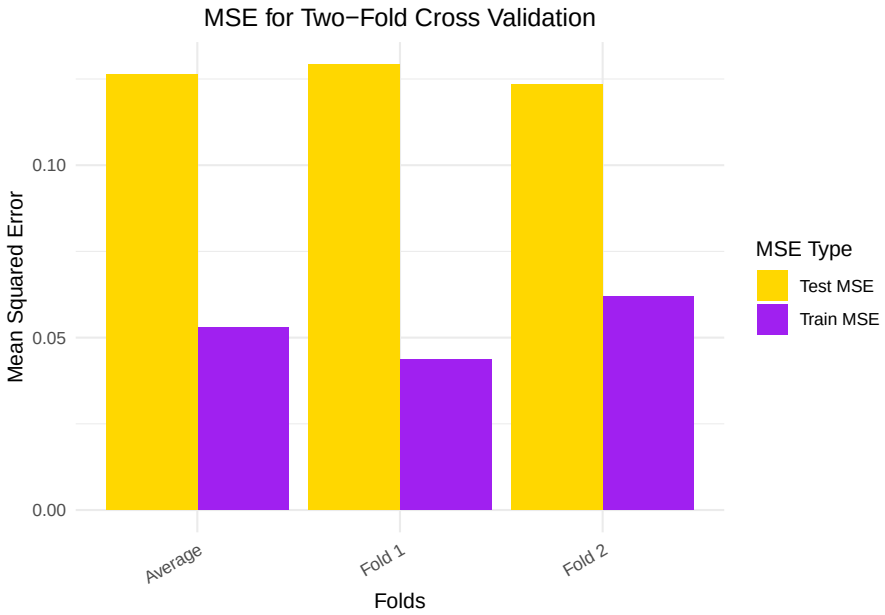



The relationship between the dependent variable Bees and the predictors Ladybirds, Vascular Plants, Carabids, Birds, Macromoths, and their interaction term Carabids × Macromoths is tested through multiple linear regression analysis. Significant predictors are Macromoths and the interaction term, which signifies the effect of the predictors on the response variable. The model explains 31.51% of the variance in Bees (R-squared = 0.3151), and the interaction model has the lowest AIC (69.89), which indicates it is the most efficient model. Diagnostic plots highlight some issues, such as residual deviations from normality and potential heteroscedasticity; hence, further refinement seems warranted. All in all, the interaction model brings forth some meaningful insights, but addressing the model assumptions may improve its robustness and predictive power.

## Two-Fold Cross-Validation:

Table 4: MSE Values for Each Fold and Average

| Metric | MSE Value |
|---|---|
| Fold 1 Train MSE | 0.0437515 |
| Fold 1 Test MSE | 0.1292456 |
| Fold 2 Train MSE | 0.0620575 |
| Fold 2 Test MSE | 0.1233161 |
| Average Train MSE | 0.0529045 |
| Average Test MSE | 0.1262809 |




The provided results and visualization highlight the performance of a model evaluated using two-fold cross-validation with Mean Squared Error being the performance metric. From the bar plot, the Train MSE and Test MSE for each fold are compared alongside their respective averages. The numerical values indicate that the Train MSE is consistently lower (0.04375149 and 0.06205749 for Fold 1 and Fold 2, respectively) compared to the Test MSE (0.1292456 and 0.1233161). This disparity suggests potential overfitting, where the model performs well on training data but struggles to generalize to unseen test data. The average Train MSE (0.05290449) and Test MSE (0.1262809) further confirm this trend. Although the Test MSE was relatively stable across folds with the present hyperparameters, the large gap between Train and Test MSEs evidently indicates room for improvement. In that, the use of regularization techniques, model simplification, or increasing sample size with more diversity might be considered to enhance generalization.