Table of Contents

1. Introduction:

This report aims **to classify the sex of Drosophila (fruit flies)** using various machine learning classification techniques. The dataset used for this study contains wing traits and asymmetry measures for different populations of Drosophila. The primary objectives of this study are to classify the sex of Drosophila using various machine learning techniques, to identify key features differentiating males from females, and to compare the performance of different classifiers.

2. Data Cleaning and Analysis

2.1 Initial Checks

The dataset '84_Loeschcke_et_al_2000_Wing_traits_&_asymmetry_lab_pops.csv' contains various wing traits and asymmetry measures for different populations of Drosophila. Upon initial check to understand the structure and content of the data, and potential issues with the data, it has been noted that the dataset comprises **of 1731 entries and 16 columns.** We then perform an initial check of the data types and identified the following columns with missing values.

Upon inspection, I found that the 'sex' column is complete, ensuring our data's integrity in terms of the target variable for our classification task. The columns with missing values are characteristics of the flies. Notably: Wing_shape (19), Wing_vein (6), Asymmetry_wing_area (26), Asymmetry_wing_shape (26), Asymmetry_wing_vein (14).

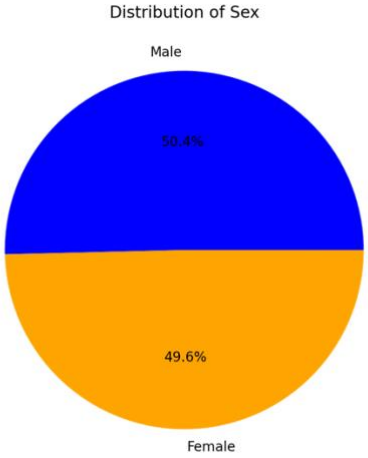
I noticed that the proportion of missing values in these columns compared to the total number of rows is very small. Given the minimal amount of missing data, I have decided to **impute the missing values by mean** rather than drop them. Imputation by mean strategy was chosen for its simplicity and effectiveness, ensuring that the small proportion of missing values does not significantly impact the analysis.

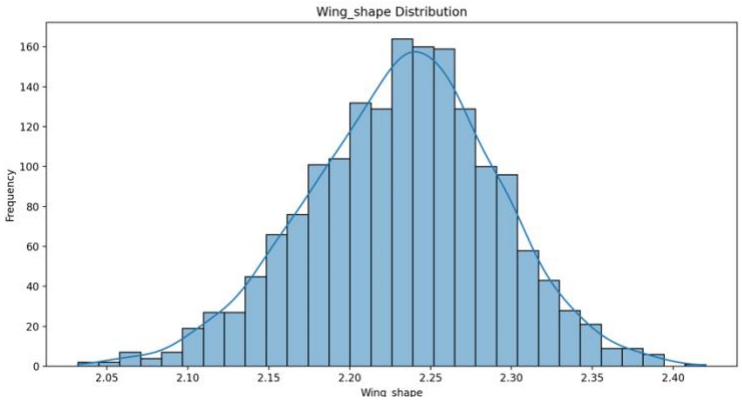
Columns that are irrelevant to our study such as 'Latitude', 'Population', 'Longitude', 'Year_start', 'Year_end', 'Vial', and 'Replicate' are also removed to simplify our analysis.

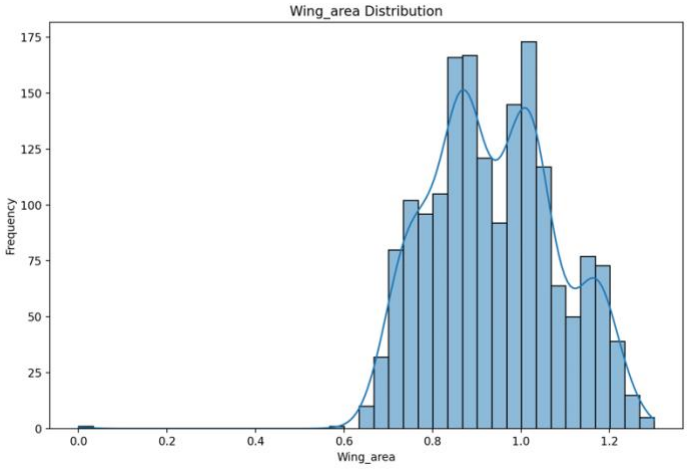
2.2 Exploratory Data Analysis (EDA)

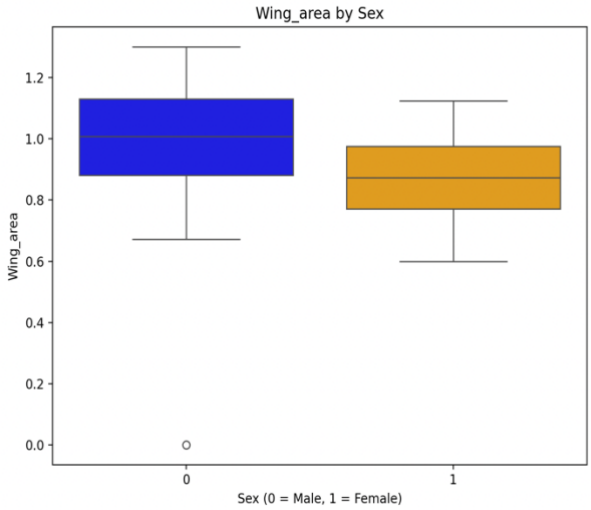
In order to predict the sex of Drosophila, it is essential to identify characteristics that will be selected for our classification models. This would involve understanding the relationships between the features and the target variable (sex) to determine which features have the most impact for our classification models.

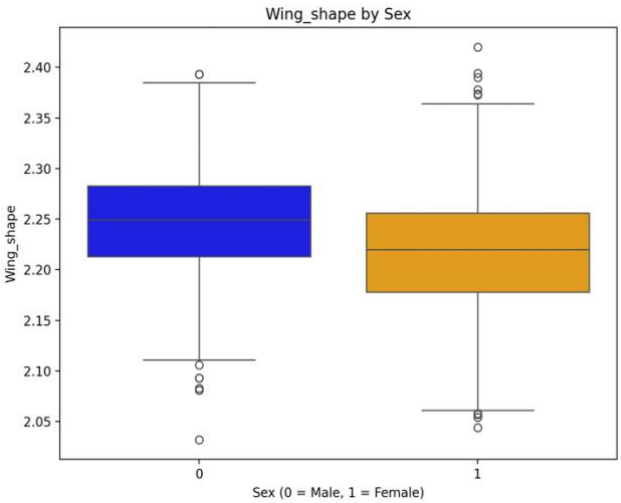
An EDA has been done on the data to discover patterns or anomalies that will help guide the selection of features for building the models later on.

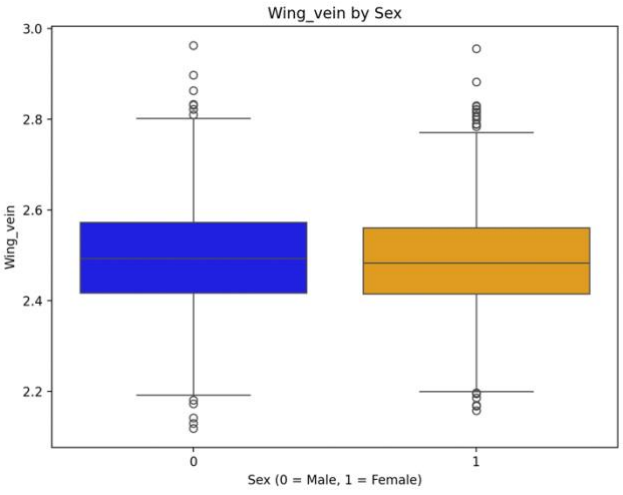

Distribution of Sex
Male
50.4%
49.6%
Female

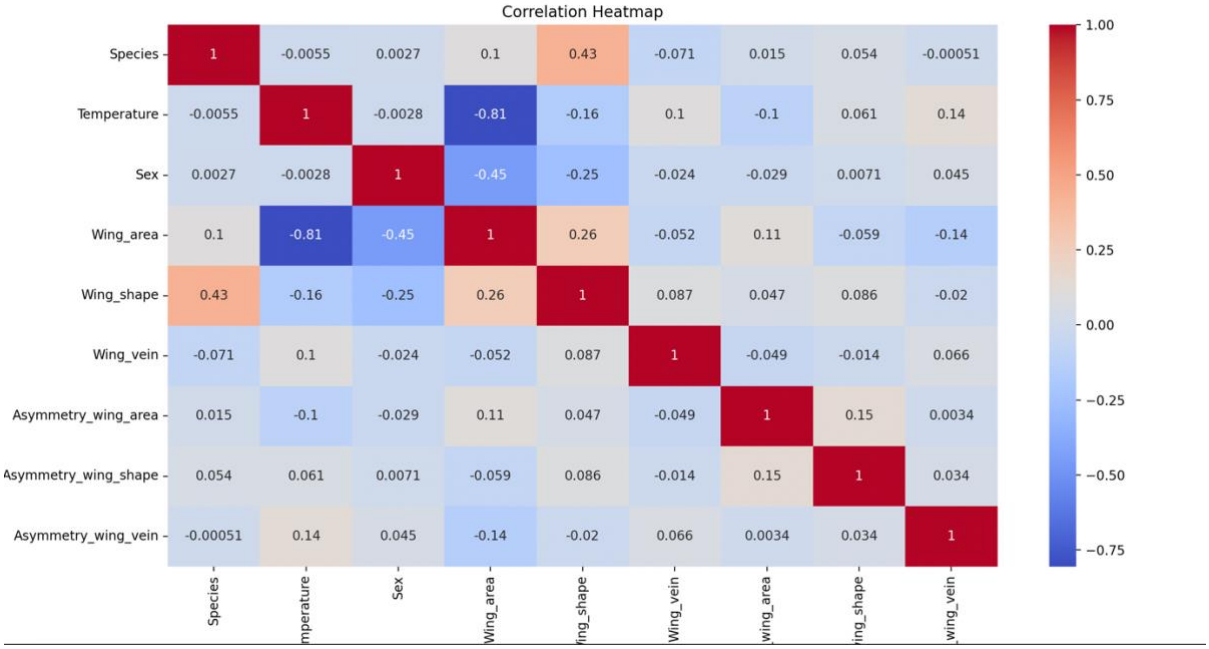

Wing_shape Distribution
Frequency
Wing_shape


Wing_area Distribution
Frequency
Wing_area


Wing_area by Sex
Wing_area
Sex (0 = Male, 1 = Female)


Wing_shape by Sex
Wing_shape
Sex (0 = Male, 1 = Female)


Wing_vein by Sex
Wing_vein
Sex (0 = Male, 1 = Female)



2.3 EDA Findings

Distribution of sex

The dataset is nearly balanced with 50.4% males and 49.6% females. This balance ensures that our models will not be biased towards one class.

Correlation Heatmap

To identify relationships between different features and the target variable, a correlation heatmap was created

Key insights:

- **'Wing_area'** has a moderate negative correlation with 'Sex' (-0.45).
- **'Wing_shape'** also shows a moderate negative correlation with 'Sex' (-0.25)
- Wing_vein, Asymmetry_wing_area, Asymmetry_wing_shape, and Asymmetry_wing_vein show weak correlations with Sex.

These insights show that **Wing_area** and **Wing_shape** are more likely to have a significant impact on prediction of sex.

Boxplots

- Reveals that males tend to have larger wing areas and shapes compared to females, indicating they could be important for our classification models.

3. K-Nearest Neighbour Classifier (k-NN)

3.1 Training (k-NN classifier)

K-Nearest Neighbours is a simple, non-parametric algorithm used for classification and regression. It classifies a data point based on the majority class among its k nearest neighbors. The value of k determines the number of neighbours considered and is the crucial hyperparameter that can significantly affect the performance of the model.

A KNN classifier was trained and experimented with different values of k to the normalized data. The data set was split into training and testing sets using **an 80-20 split.** The training set is used to train the model whereas the training set is used to evaluate the model's performance.

To determine the best value of k, k values ranging from 0 to 30 are tested and the performance of the k-NN classifier was evaluated using cross validation to select the k value that gives the highest accuracy.

To improve visualization and potentially the performance of the model, I have **applied Principal Component Analysis (PCA)** to reduce the dimensionality of the dataset. This is particularly helpful in making the decision boundary visualization more interpretable.

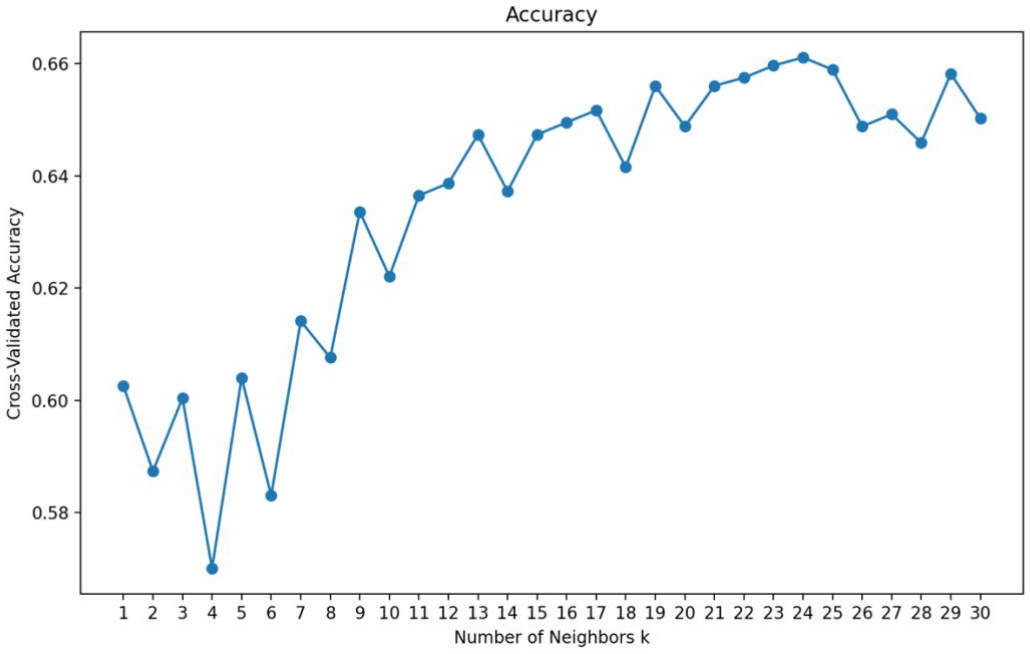


Figure 3.1.1

3.2 Results (k-NN Classifier)

The k value of 24 was chosen as it had the best accuracy based on our **cross validation**. The performance of the k-NN classifier with the best k value is as follows:

Best k value: 24
Accuracy: **0.6772**
Precision: 0.6816
Recall: 0.6893
F1 Score: 0.6854

When looking at the precision-recall graph in figure 3.3.1, we can see that our model maintains high precision for low recall values but struggles to maintain precision as recall increase. This indicates that the model is accurate in its predictions for certain classes but fails to capture all relevant classes completely.

When looking at the confusion matrix (Figure 3.3.2), we can observe:

- True Positives: Correctly identified as Female = 122
- True Negatives: Correctly identified as Male = 113
- False Positives: Incorrectly identified as Female = 57
- False Negatives: Incorrectly identified as Male = 55

This indicates that while the model can accurately identify many instances of both classes, it still has a moderate number of incorrect predictions, which affects its overall performance. The reason for it might be there could be overlapping feature values between the sexes, consequently making it challenging for the model to distinguish between them.

Overall, the KNN classifier with PCA performed reasonably well with an accuracy of **65.43%**.

3.3 Figures (k-NN classifier)

To visually interpret the performance of the k-NN classifier, several plots were generated. This includes decision boundaries, confusion matrix and precision-recall graph.

Figure 3.3.3 shows how the k-NN classifier divides the classes based on the nearest neighbours. The plot confirms the regions classified as male and female based on the principal components PC1 and PC2. Green dots represent male class and red dots represent female instances. The green and red shaded areas indicate the decision regions for each class.

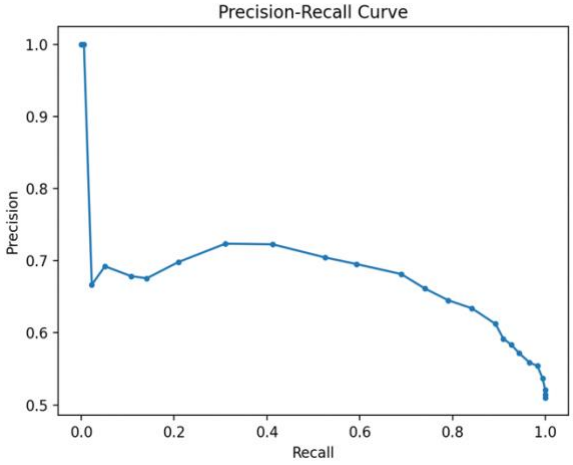


Figure 3.3.1

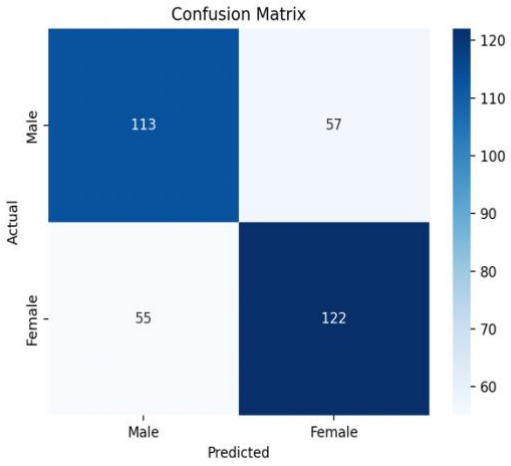


Figure 3.3.2

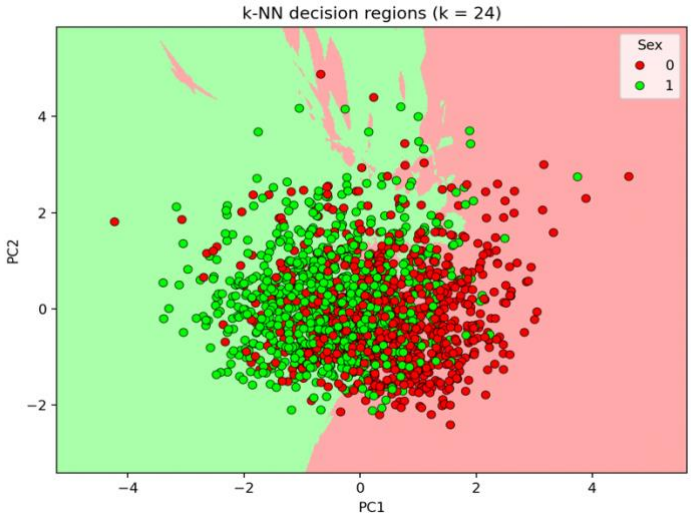


Figure 3.3.3

4. Logistic Regression

4.1 Training (Logistic Regression)

Logistic Regression is an algorithm for binary classification, designed to predict the probability of a binary outcome based on one or more input features. It utilizes a logistic function to calculate the likelihood of a specific class label. For our study to classify sex of Drosophila, an 80-20 split of the PCA-transformed data that were used earlier were once again utilized to train the model.

We will **compare logistic regression with and without regularization** to see if there are any improvements to our model by utilizing regularization.

For regularization, I have utilized L2 regularization and in order to find the best hyperparameter "C" for our regularization, a search was done with cross validation to find the best regularization parameter for our model.

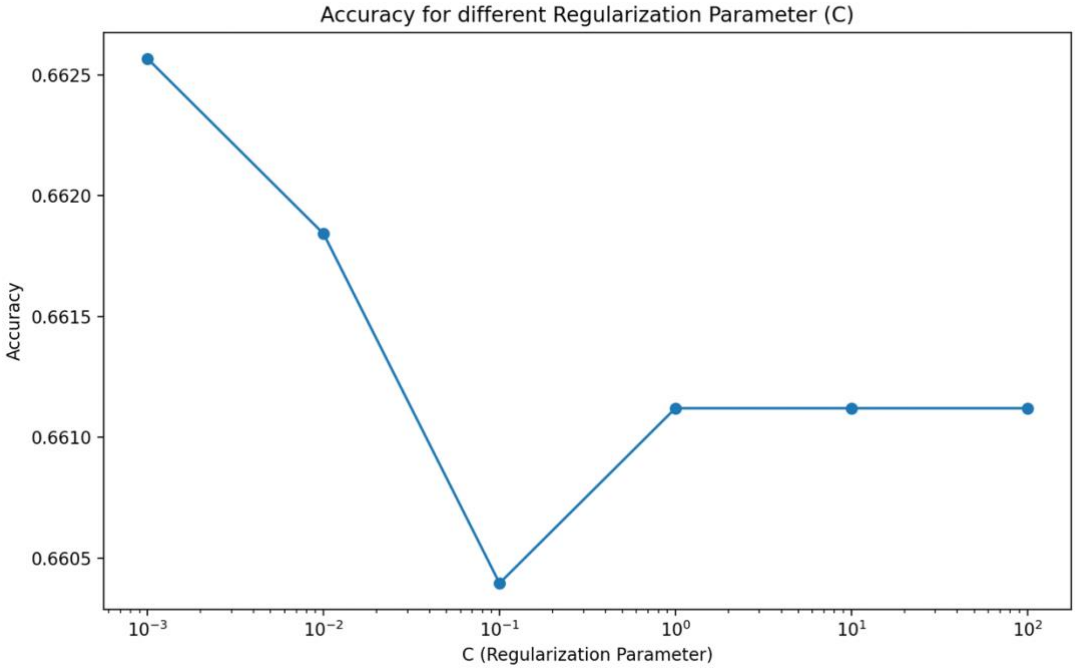


Figure 4.1.1

From the figure above, we can see that 0.001 has the highest accuracy score and therefore has been chosen for our model of 'With Regularization'.

4.2 Results (Logistic Regression)

Regularization vs Without Regularization

Results:

No Regularization:
Accuracy: 0.7089
Precision: 0.7111
Recall: 0.7232
F1 Score: 0.7171

L2 Regularization:
Accuracy: 0.7147
Precision: 0.7191
Recall: 0.7232
F1 Score: 0.7211
Best C: 0.001

ANALYSIS

Accuracy

- The model with L2 regularization showed just a slight improvement in accuracy (0.7147) compared to the model without regularization (0.7061).
- This improvement, although small, suggests that regularization helps in generalizing the model better to unseen data.

Confusion Matrices

- It can be seen in figure 4.2.2(a) and 4.2.2(b) that the performance of the regularized model in the confusion matrix is just slightly better at classifying both classes (male and female).
- The regularized model has fewer false positives (52 vs 50) and true negatives (118 vs 120) compared to the non-regularized model, which explains its slightly higher precision and recall.

Precision-Recall graph

- The curve for the regularized model is slightly favorable as it maintains higher precision at different recall values. This suggests that the regularized model would make more accurate predictions even when the recall increases compared to the non-regularized model.

Conclusion

The addition of regularization, although **not as drastic** as I would have hoped, would still help my model perform better and be more effective against unseen data.
The decision boundaries and precision-recall curves indicate that regularization did help in making the model slightly more robust and generalizable, preventing it from overfitting to the training data.

The confusion matrices show a similar pattern of misclassification, but the regularized model performed slightly better. It is also important to note that **regularization rarely improve the performance on the same dataset that the model is trained** and will most probably improve performance on new data, which is the goal.

Additionally, the relatively small size of data and also the possibility of our non-regularized logistic regression might not be overfitting significantly might be the reason there is not much difference on the dataset.

Overall, **logistic regression displays a solid performance** in classifying the sex of Drosophila, whether regularization was applied or not. This is proven by the accuracy (71.47% and 70.61%), precision, recall and F1 score in both scenarios, indicating its reliability for this classification task.

4.3 Figures (With vs Without Regularization)

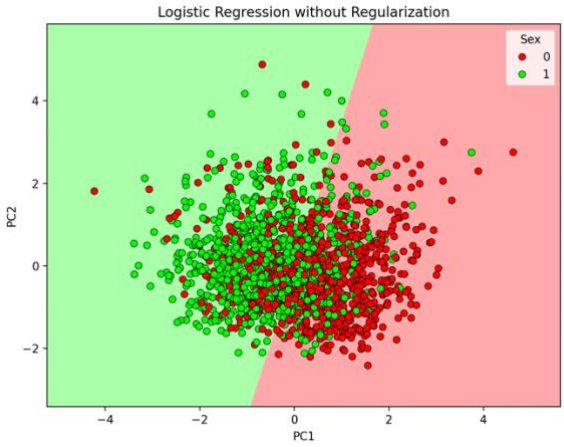


Figure 4.2.1(a)

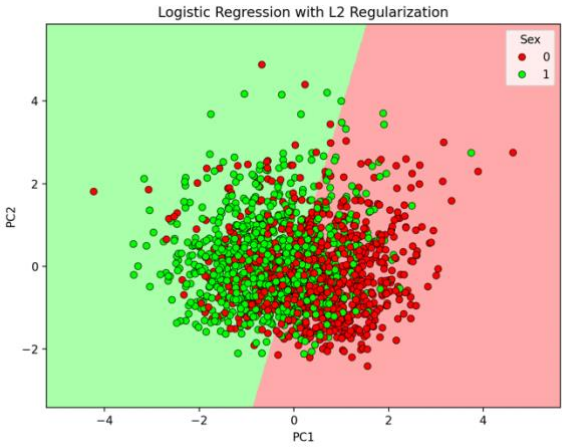


Figure 4.2.1(b)

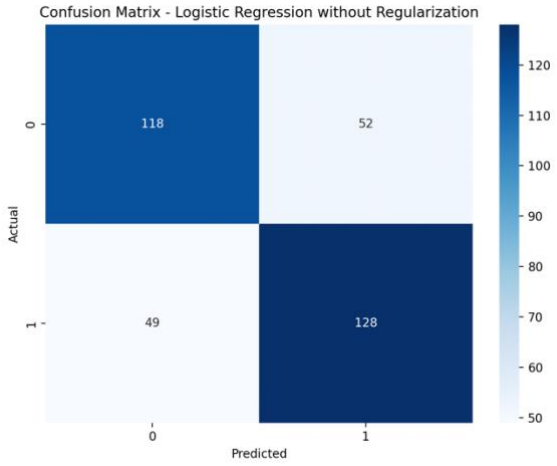


Figure 4.2.2(a)

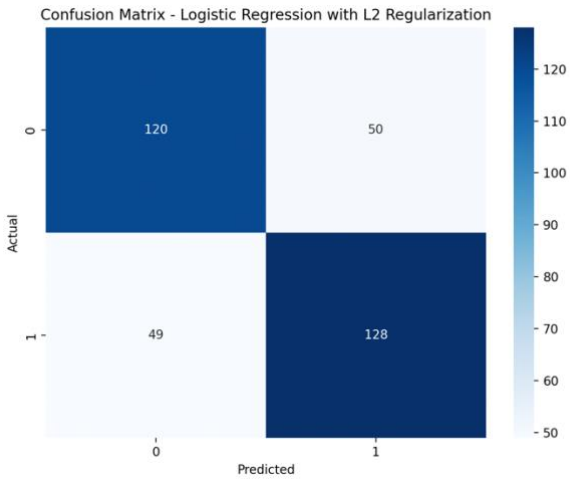


Figure 4.2.2(b)

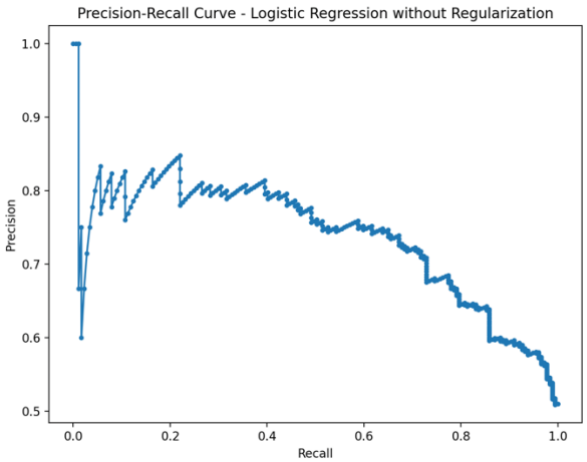


Figure 4.2.3(a)

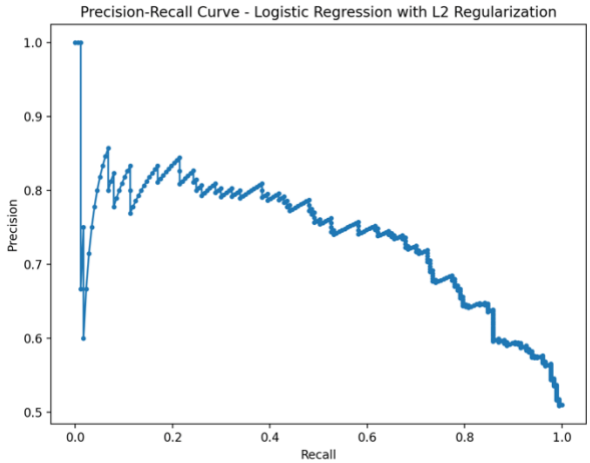


Figure 4.2.3(b)

5. Random Forest classifier

5.1 Training (Random Forest)

Random forest classifier works by creating multiple decision trees during training and outputting the class that is the mode of the classes (for classification) or mean prediction (regression) of the individual trees. This method is famous for its robustness to overfitting, the ability to handle high-dimensional data and is also very accurate since it uses a large number of decision trees to make predictions.

For this method, we will also **compare Random Forest with PCA and without the use of PCA** on our dataset. Therefore, I have trained the Random Forest classifier using both the original feature and dataset with features that has been transformed by PCA.

The hyperparameters considered included:

- Number of trees (n_estimators)
- Maximum depth of the trees (max_depth)
- Minimum number of samples required to split a node (min_samples_split)
- Minimum number of samples required at each leaf node (min_samples_leaf)
- Number of features to consider at each split (max_features)

After I get the models for the original feature and pca, I have decided **to find and plot the features with the most importance scores** before visualizing them. While it could be beneficial to create a new classifier using only the most important features in datasets with a large number of features. However, in our case, since we only have six relevant features, we utilized all six in our models.

5.2 Results (Random Forest)

Here are the results for our non-pca model vs pca model for Random Forest:

Best Parameters by importance (Figure 5.3.1):

1. Wing_area
2. Wing_shape
3. Wing_vein
4. Asymmetry_wing_vein
5. Asymmetry_wing_shape
6. Asymmetry_wing_area

Best parameters for Random Forest with PCA:

- max_depth: 10
- max_features: log2
- min_samples_leaf: 1
- min_samples_split: 5
- n_estimators: 50

Best parameters for Random Forest with Original Features:

- max_depth: None
- max_features: sqrt
- min_samples_leaf: 1
- min_samples_split: 5
- n_estimators: 100

Scores for Random Forest with PCA:

- Accuracy: 0.6686,
- Precision: 0.6632,
- Recall: 0.7119,
- F1 Score: 0.6866

Scores for Random Forest with Original Features:

- Accuracy: 0.7349
- Precision: 0.7225
- Recall: 0.7797
- F1 Score: 0.7500

ANALYSIS

Feature Importance

Wing_area and Wing_shape were the most significant features, which **aligns with our EDA findings** and further validates their importance in predicting the sex of Drosophila.

Confusion Matrix

Random Forest with PCA:

- The model correctly identified 111 instances of True Positives and 98 instances of True Negatives.
- However, it misclassified 72 False Positives and 66 False Negatives.

Random Forest with Original Features:

- The PCA model correctly identified 140 instances of True Positives and 117 True Negatives.
- It misclassified 53 False Positives and 37 False Negatives.

Surprisingly, we can see a clear difference in performance between the two models where the model with Original Features has outperformed the PCA model completely.

Precision-Recall Graph

- Graph with PCA have fluctuations, with precision values ranging from 0 to approximately 0.85 whereas graph of original features remains high (>0.9) initially and gradually decreases as recall increases.
- This shows that the graph of original features is more stable and higher precision making it more reliable than the one with PCA.

Performance Metrics:

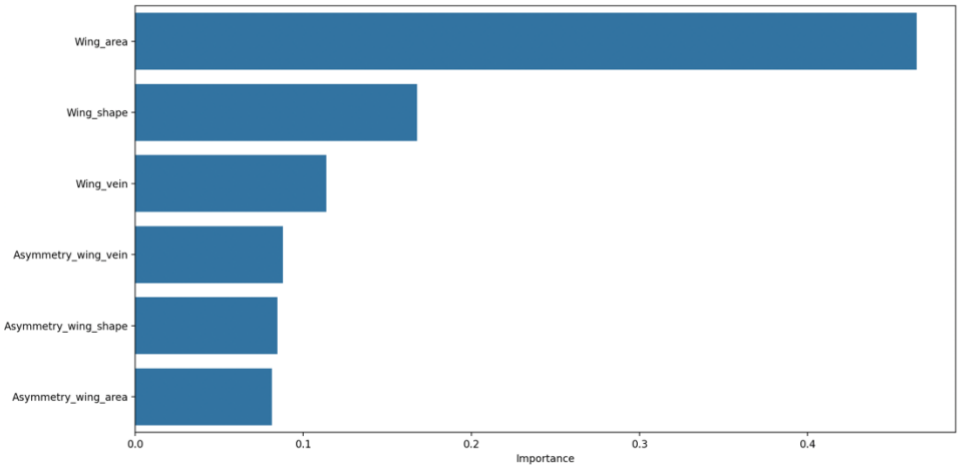
- The Original Feature performed way better than the PCA model across every metric. Higher Accuracy (73% vs 66%).

Conclusion

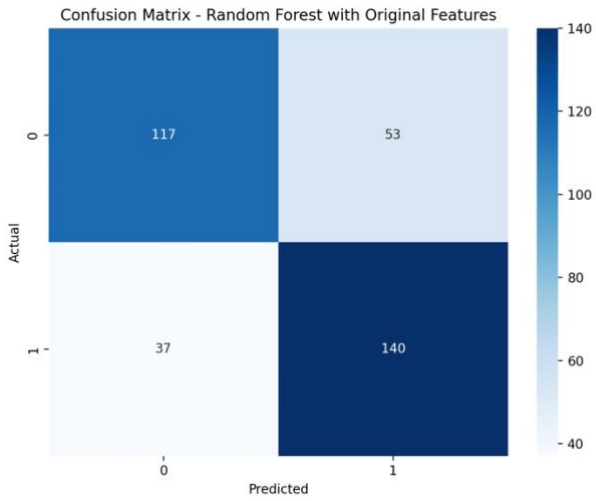
To my surprise, the model without pca (original features) is better than the one with pca. Before this, I have always thought that with PCA, the model should be better most of the time. A few reasons for this could be **the loss of information**, as pca reduces the dimensionality of our data and with our limited data, it is obvious that the reduced dimensionality has **taken a toll on the retention of information**. This is also due to our data having only 6 dimensions, so doing PCA would not be very helpful in our case.

It is also worth to note that, the random forest classier with Original Features has performed the best compared to all the different model we tested throughout this assignment.

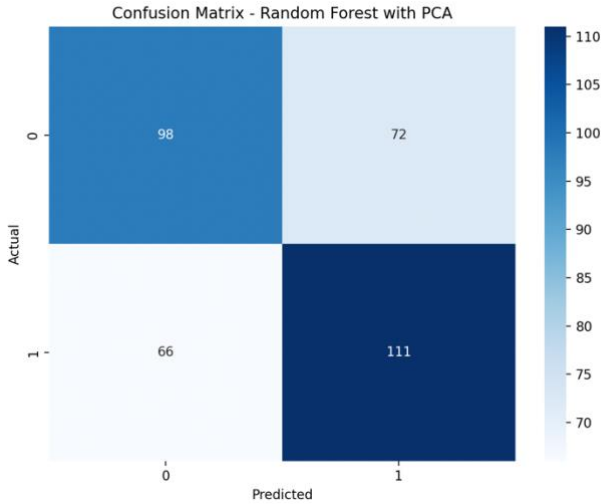
5.3 Figures (PCA vs Original Features)



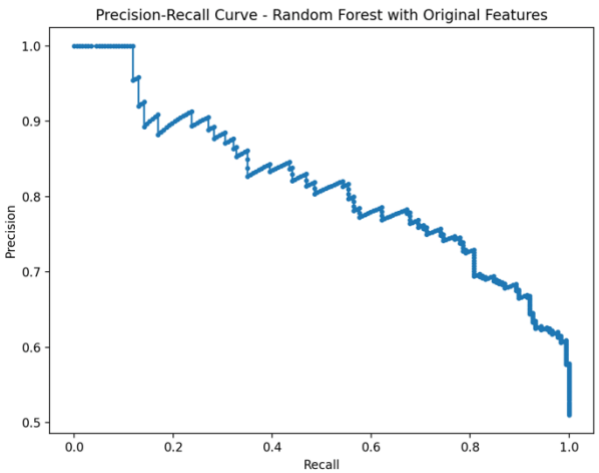
5.3.1



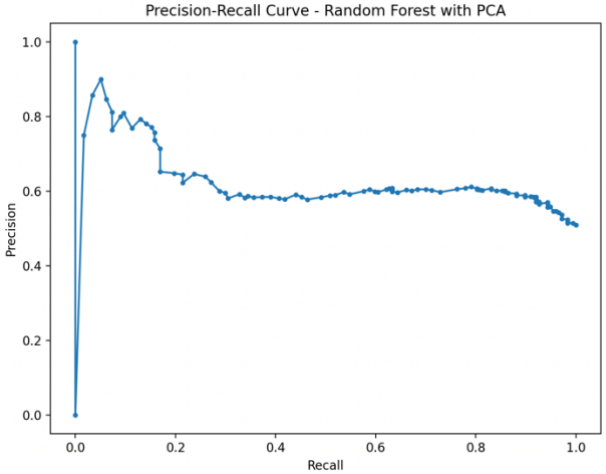
5.3.2(a)



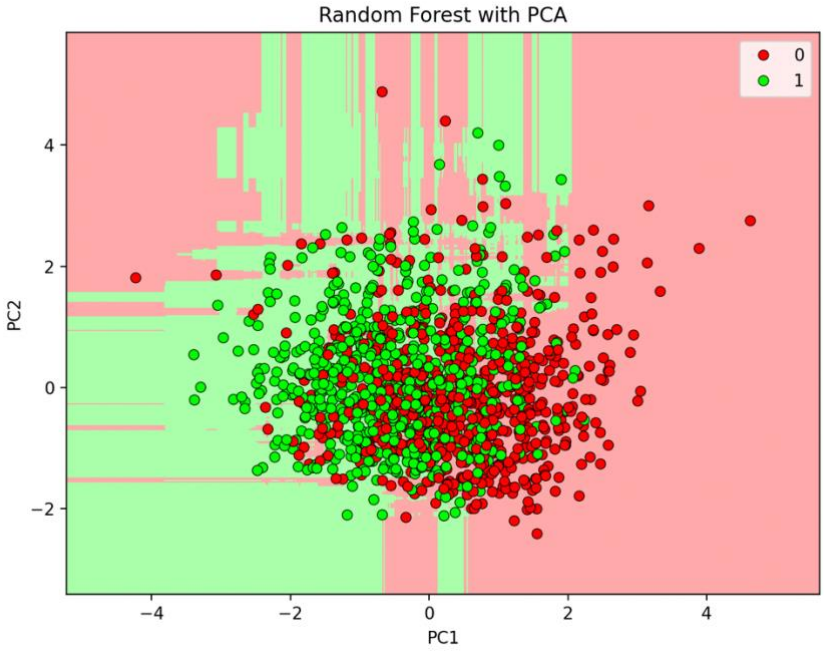
5.3.2(b)



5.3.3(a)



5.3.3(b)



5.3.4

6. Conclusion

In this study, we have done a comprehensive analysis to classify the sex of Drosophila using various machine learning classification techniques, in particular k-NN classifier, logistic regression and Random Forest classifier. Through detailed data cleaning, EDA and determining the best hyperparameters for each model, we aimed to identify the best model for this classification problem while also determining the features that has the most impact on our objective of classifying sex of the fruit flies.

Major Takeaways:

K-NN Classifier

- The best k value was 24 and the model achieved an accuracy of 67.72% with balanced performance metrics.

Logistic Regression

- Both regularized (L2) and non-regularized models were trained and compared.
- The regularized model showed a slight improvement in accuracy (71.47% vs 70.61%) and better generalization to unseen data.
- Overall, logistic regression performs reasonably well for this classification problem.

Random Forest Classifier

- Models were trained using both original features and PCA-transformed features.
- The model with original features significantly outperformed the PCA model and every other model tested throughout the study, achieving an accuracy of 73.49% compared to 66.86% for the PCA model.

Feature Impact

- 'Wing_area' and 'Wing_shape' are the two most significant features by as proven by our EDA findings and Random Forest Feature Importance Analysis.

Model Comparisons

Model	Accuracy	Precision	Recall	F1 Score
k-NN (k=24)	0.6772	0.6816	0.6893	0.6854
Logistic Regression	0.7089	0.7111	0.7232	0.7171
Logistic Regression (L2)	0.7147	0.7191	0.7232	0.7211
Random Forest (PCA)	0.6686	0.6632	0.7119	0.6866
Random Forest (Original)	0.7349	0.7225	0.7797	0.7500

Further Exploration:

Dimensionality Reduction

- While PCA can be beneficial for reducing dimensions and noise, in this study, it resulted in a loss of crucial information proven in our Random Forest Classifier study. Consequently, it would be interesting to find out how original features (Non-PCA) affect our first two models, k-NN neighbours and logistic regression as both of them utilized PCA datasets.

In summary, this study has provided me with hands-on experience in experimenting with datasets according to my curiosity and interests, applying various machine learning techniques. Through tasks such as data preprocessing, feature selection, and rigorous model evaluation, I have come to appreciate the significance of ML techniques and their potential when utilized effectively. Notably, the random forest classifier with original features emerged as the most effective model, emphasizing the importance of retaining valuable information.

(2635 Words)