Group 12 Assignment 7
Reagan Hill and Ehsan Kahrizi

Introduction

This report focuses on the application of Support Vector Machines (SVM) with different kernels for predictive modeling on two datasets: maternal health risk assessment and apple quality classification. The maternal health risk assessment model benefits mothers, healthcare professionals, and policymakers by aiding early detection of health risks. The apple quality classification model serves apple farmers, supply chain managers, and consumers by ensuring consistent fruit quality.

The SVM models demonstrated effectiveness for both datasets. For the Maternal Health Risk Dataset, the linear SVM provided balanced classification results, emphasizing the importance of SystolicBP and DiastolicBP in determining risk levels. High recall for mid/high-risk cases was achieved, minimizing the risk of undetected health concerns. For the Apple Quality Dataset, the SVM model with a linear kernel effectively distinguished between good and bad quality apples, that providing reliable performance metrics like precision, recall, and F1 scores. This benefits stakeholders involved in quality control and agricultural production by optimizing sorting processes and reducing waste.

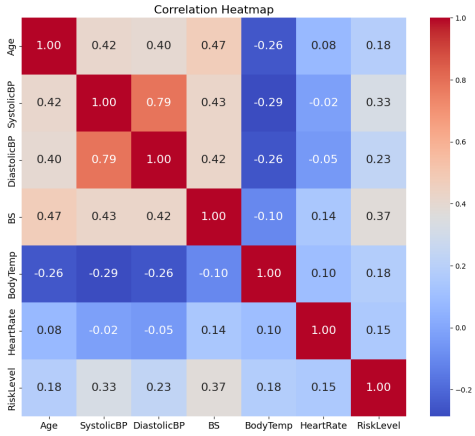
In the second part of the study, we used logistic regression to predict possum gender and population. The gender prediction model showed moderate performance, useful for wildlife researchers and conservationists. The population model performed exceptionally well, aiding biologists and ecologists in understanding possum distribution and traits.

Overall, the study compared SVM and logistic regression models to determine the most effective predictive techniques for each dataset's requirements and stakeholders. The results are available online through Google Slides and Github.

Part I

Dataset

In the first part of our project, we utilized two distinct datasets for analysis. The Maternal Health Risk Dataset consisted of 1014 samples with six numerical features that include ‘Age’, ‘SystolicBP’, ‘DiastolicBP’, ‘BS (Blood Sugar)’, ‘BodyTemp’, and ‘HeartRate’. The target column, `’RiskLeve’, originally had three classes (low, mid, and high risk) but was simplified to two classes: low risk (0) and mid/high risk (1). DiastolicBP and SystolicBP were identified as the most correlated features, guiding their use in SVM modeling.



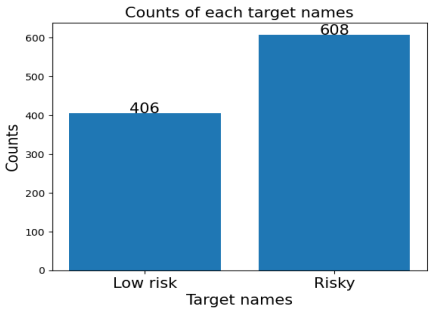
The second dataset that we studied was the Apple Quality Dataset which included 4001 observations (cleaned to 4000). The futures of it were ‘Size’, ‘Weight’, ‘Sweetness’, ‘Juiciness’, and ‘Ripeness’. The aim was to predict fruit quality. The target column was binary: good quality (1) and bad quality (0). To prepare for modeling, the dataset was cleaned, and standardized. Also, we removed the non-informative columns like IDs. Both datasets were well-suited for testing various SVM kernels to evaluate predictive performance effectively.

Analysis Technique

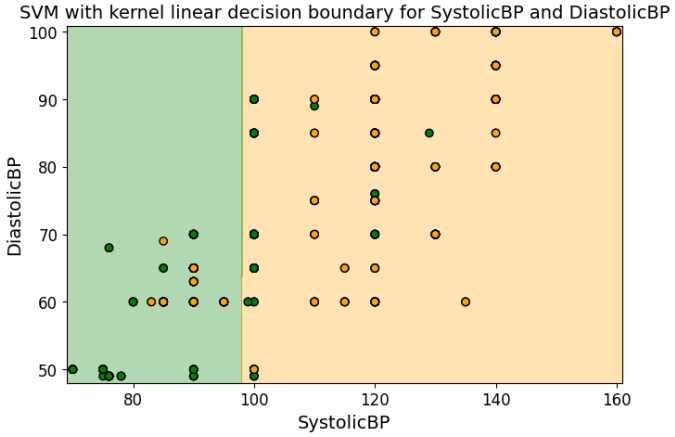
The analysis technique was done using different kernels for both datasets to assess their predictive performance. For the Maternal Health Risk Dataset, an initial feature selection process identified DiastolicBP and SystolicBP as the most correlated variables with the target, guiding the primary analysis. SVM models were trained using linear, polynomial, and RBF kernels to compare their effectiveness. The data was split into 80% training and 20% testing sets. The metrics for checking the performance of each method included precision, recall, and F1-scores were computed. Class weights were adjusted (0.9 for class 0 and 1 for class 1) to prioritize accurate detection of mid/high-risk cases. Lastly, Execution time for each model was also tracked to evaluate computational efficiency.

Also, in the Apple Quality Dataset, we used the SVM with a linear kernel. Firstly, the dataset was standardized to ensure uniform feature scaling. The performance metrics like the previous datasets were precision, recall, and F1 scores. Also, accurately predicting considered 1 for good quality and 0 for bad quality. Lastly, the confusion matrices were used to illustrate the models’ classification results, and decision boundaries were analyzed to assess model performance visually. This technique ensured comprehensive model evaluation and highlighted the optimal SVM kernel for each dataset's predictive needs.

Results

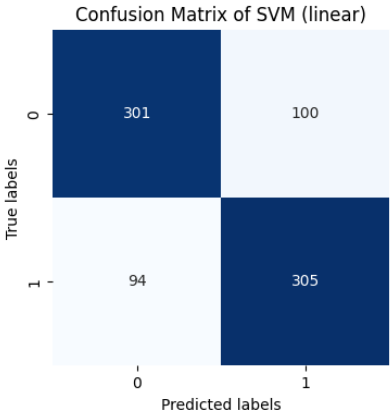


This plot displays the distribution of target classes in the Maternal Health Risk dataset, following the simplification of the “RiskLevel” column into two categories: "Low risk" and "Risky" (mid/high risk). The count for "Low-risk" cases is 406, while "Risky" cases total 608. This class distribution shows a moderate imbalance, with the "Risky" category having a higher count than "Low risk." This imbalance was considered in the modeling process by adjusting class weights to 0.9 for "Low risk" and 1 for "Risky" to prioritize accurate detection of higher-risk cases. The focus on the "Risky" group aligns with the project’s goal of effectively identifying potential health risks, as misclassifying higher-risk individuals could lead to more significant consequences.



This plot shows the decision boundary generated by SVM with a linear kernel, using “SystolicBP” and “DiastolicBP” as features to classify maternal health risk levels. The green and orange shaded regions represent the decision boundary that separates the two classes, with green indicating "Low risk" and orange representing “Risky” The linear kernel was chosen here due to the apparent linear separability of the two classes based on these blood pressure variables. The model effectively distinguishes the two classes along this boundary, suggesting a strong correlation between these blood pressure features and risk levels. This separation indicates that “SystolicBP” and “DiastolicBP” play significant roles in risk classification, providing a simple yet interpretable model. However, some overlap near the boundary suggests that further analysis or additional features may enhance the model's accuracy.

This confusion matrix illustrates the performance of the SVM model with a linear kernel in classifying maternal health risk levels. The matrix shows 301 correct classifications for “Low risk” (true negatives) and 305 correct classifications for “Risky” (true positives). There are 100 false positives (misclassified as “Risky”) and 94 false negatives (misclassified as “Low risk”). These results indicate that the model performs reasonably well, with a balanced classification of both classes. However, there are some misclassifications, particularly with 100 false positives, which suggests a slight bias toward predicting the “Risky” class. This bias aligns with the project’s goal of prioritizing higher-risk cases, as the cost of missing a true “Risky” case could be more significant. The overall balance between true positives and true negatives demonstrates that the linear SVM provides a straightforward and interpretable model, effectively capturing the underlying risk levels in the dataset.



Technical

To prepare the datasets, we removed irrelevant columns, handled missing values, and standardized features for consistency. For analysis, we used SVM due to their strong performance in binary classification tasks, testing linear, polynomial, and RBF kernels. We prioritized DiastolicBP and SystolicBP in the Maternal Health dataset for interpretability. Initial runs with default class weights led to misclassifications in high-risk cases, prompting weight adjustments to improve recall. While SVM proved effective, logistic regression could provide a simpler alternative.

Group 12
Reagan Hill and Ehsan Kahrizi

Part II

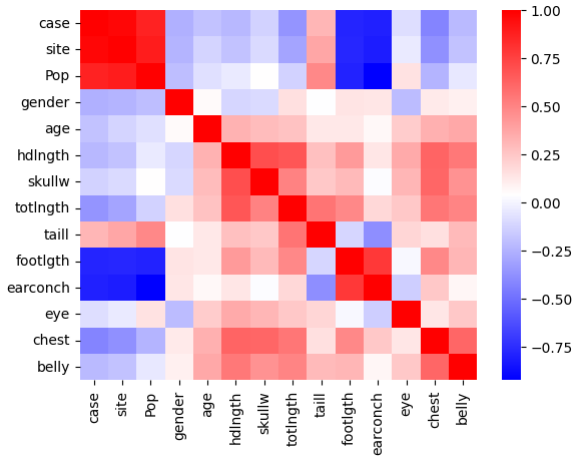
Dataset: Possums

Dataset

The dataset we chose to use for this section is the “Possum Regression” data set found on Kaggle. It is originally from the DAAG R package and has been used in a couple of books. The dataset consists of 14 columns and 104 rows of possum data collected in Australia. For each possum there is information such as the site where the possum was trapped, population (either Victoria or Other), sex, age, and measurements of ears, tails, and more. This was a very clean dataset that only needed a little bit of cleaning.

Analysis Technique

We focused on using logistic regression for this dataset. There were a few good candidates for target values, but we chose to look at predicting sex and population. Logistic regression worked well when trying to predict these values since there were only two possible ways to classify the possums. If we had tried to predict age, the recall and precision went down significantly. When trying to find which variables were predictive of the target value, we started with a heatmap. We were able to visually see where there were strong correlations. That didn’t necessarily mean they were great predictors though.



Results

Gender

For the modal to predict a possums gender we used the heat map above, it shows case, site, pop, and eye as having the most correlated values, but even with the correlation it has a hard time predicting consistently. We also found that using the location variable of “site” and all of the body dimension variables (eye, tail length, ear conch, etc) did not help the model at all.

Predicted based on just eye and site results in the following:

[male, female]

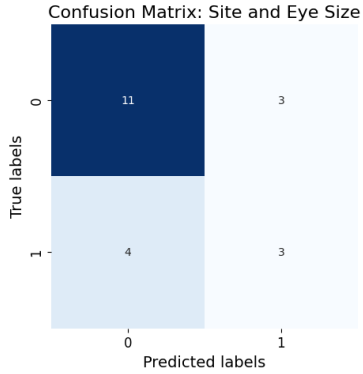
Precision [0.73, 0.5], Recall [0.79, 0.43], F-score [0.76, 0.46]

Predicting with site, head length, skull width, total length, tail length, foot length, ear conch, eye, chest, and belly results in the following:

[male, female]

Precision [0.69, 0.38], Recall [0.64, 0.43], F-score [0.67, 0.4]

It consistently performed mediocre for predicting males, but was always about as good as random for predicting females.



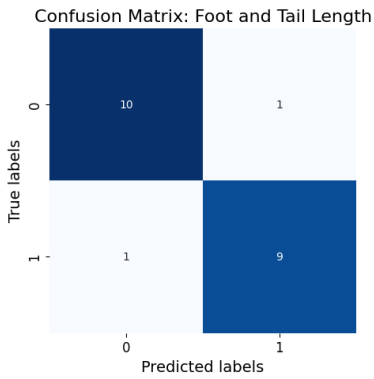
Population

The model to predict which population the possum belongs to was interesting in that the correlations shown in the heatmap did help find good predictive values. Using just foot length and tail length to predict if a possum is from Victoria or other (New South Wales or Queensland) we got the following results:

[Victoria, other]

Precision [0.91, 0.90], Recall [0.91, 0.90], F-score [0.91, 0.90]

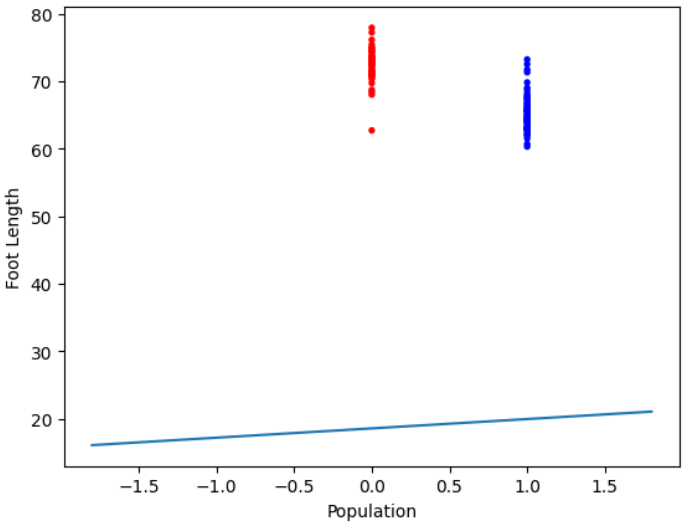
With populations classified as they are, the foot length and tail length are great predictors. The model consistently got values in the 90s and even gets 100% periodically. It does cause a little concern about overfitting, but it could also mean that tail and foot length are regional characteristics.

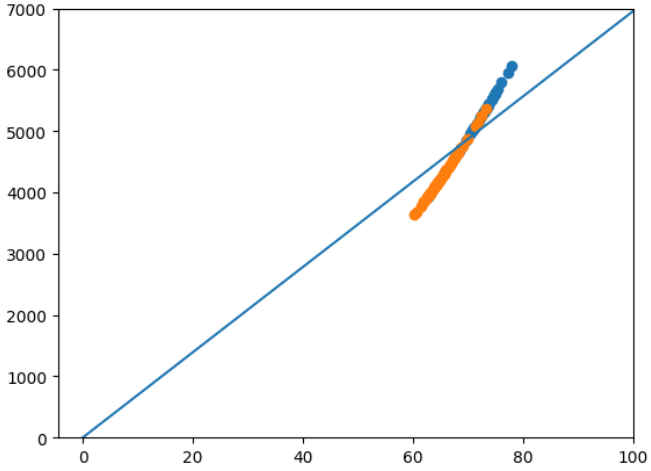


Technical

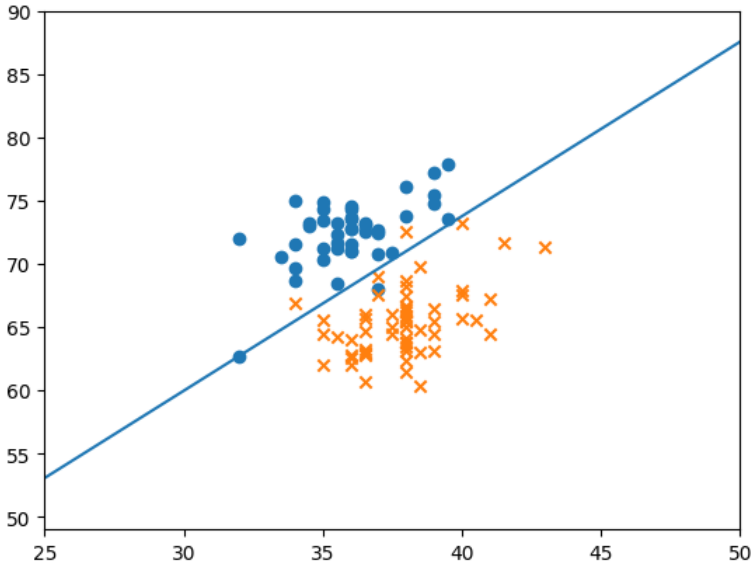
There wasn’t much cleaning needed for this dataset. There were 3 possums with NAN values, so we chose to drop them rather than try to make an inference. We also changed the Pop column into binary

values. 1 being Victoria and 0 being other. Finally, we created the gender column that turns “m” into 0 and “f” into 1.
For the Population graph, we tried visualizing the decision boundary using the kernel trick, since we couldn't get a very nice decision boundary using the individual variables, but as seen below, upping the dimensionality of Foot length by one didn’t have as nice an effect as we were hoping.





We got the best results visualizing the boundary line for the foot length and tail length.



Overall, we felt that the logistic regression was able to predict pretty well gender and especially well for population. I’m not sure how well these will generalize since the gender predictor has such high bias. Where the dataset isn’t the biggest, it might just be an issue of lacking data, but it also might just perform poorly on all data. The population predictor would generalize the best out of the 2 models since it is low variance and low bias.