Homework 1

Lawrence Leymarie

Table of contents

1 Analysis

The dataset provided in the assignment is cmposed of 8 predictor variables labeled X1 through X8, all of which are continuous variables. The range of values over all of the predictors the maximum is -3.22 and the max is 4.12 which shows the dataset is very balanced with most of the values being around 0 + or - 4. The outcome variable of the data labeled as "Group" is a categorical, binary value containing either an 'A' or a 'B'. The dispersion of outcomes is fairly similar with 40.5% of all outcomes being 'A' and 59.5% being 'B', this should tell us that the model should not be very bisased towards either outcome although we should expect to see a better score for predicting outcome 'B' in comparison.

2 Methods

The method for this assignment is to compare the performance of three classification/clustering algorithms in order to assess each models relevance to the specific data. First, the dataset was split into training and testing sets for model validation, each model was treated with the same preprocessing pipeline which simply scaled the data using sandard scaler. The first model that was used is an SVM which uses the Keras package. In order to obtain the optimal performing SVM model, I used a grid containing a range of numerical values to train with for the C,

gamma and used linear and rbf kernels, with a cross validation of 5, the model chose the following parameters as the best performing.

Params	Best
C	5
Gamma	0.001
Kernel	rbf

The next model was a logistic regression using the Keras package. This model was implimented quite simply by passing through a pipeline with the scaled data.

Finally, the last model I used was a KNN which received a similar preprocessing as the previous two. For hyperparameter tuning I used a range between 1 and 19 for n neighbors which by using gridsearch, the grid chose n_neighbors as 1 as the best performing model. The KNN was also trained with a cross validation with 5 folds.

3 Results

Overall the performance of each model was very similar, the accuracy score for the SVM model was of 0.75 for the training set and 0.77 for the testing set which means our model predicts 3 out of every 4 datapoints which is fairly good. The ROC AUC score went from 0.84 in training to 0.87 in testing which indicates our model is predicting outcomes confidently as the ideal value should be 1. Finally, the confusion matrix indicated a strong model where true positives and negatives have a high predicion rate. The model exhibits a sensitivity of approximately 78.62% on the training set and 83.05% on the testing set, indicating it has a strong ability to correctly identify positive instances across both datasets. Specificity is about 69.04% for the training set and 71.95% for the testing set, demonstrating a good capacity to correctly identify negative instances. These results suggest the model is effectively balanced in predicting both positive and negative outcomes, with a slightly better performance in identifying true positives and true negatives in the testing set compared to the training set.

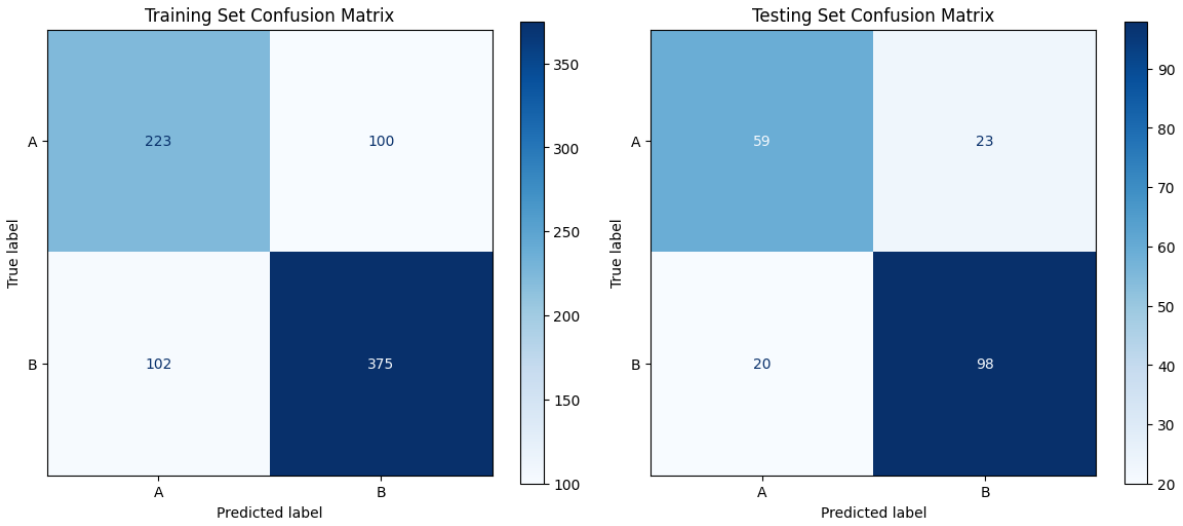

Figure 1: Confusion Matrix for SVM model

The logistic regression model had a very similar performance with a training accuracy of 0.765 and an ROC/AUC of 0.84. The testing performance was very close to the training with an accuracy of 0.765 and an ROC/AUC of 0.856. The accuracy score for both of the training and testing set being equal indicates a strong model. Furthermore, a sensitivity of approximately 82.49% for the training set and 76.03% for the testing set, indicating a solid capability in identifying positive cases. However, specificity was lower, with about 69.94% for the training set and 65.82% for the testing set, suggesting room for improvement in correctly classifying negative cases. These results hint at a slight overfitting issue, where the model performs better on the training data compared to unseen testing data, indicating a need for further adjustments to enhance generalization and balance between sensitivity and specificity.

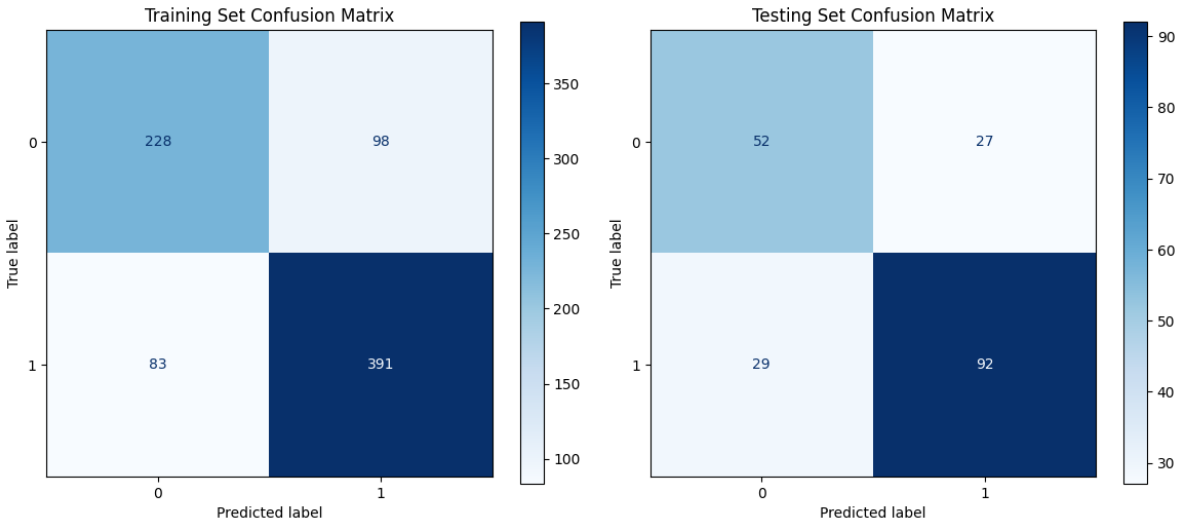

Figure 2: Confusion Matrix for Logistic Regression model

Finally, the KNN had the most puzzling results as I saw training accuracy and ROC/AUC scores at exactly 1, the testing score was at 0.75 accuracy and 0.739 ROC/AUC. these results indicate overfitting, meaning the model is too closely tailored to the training set and cannot understand the data from the test set as well. The results indicate perfect sensitivity and specificity (100%) on the training set, meaning the model accurately identified all positive and negative cases without any errors during training. However, in the testing set, the sensitivity dropped to approximately 68.64%, and specificity was about 75.61%. This shows that while the model was flawless in identifying both conditions in the training phase, its performance on unseen data was less consistent, particularly in identifying true positive cases. The reduction in sensitivity suggests that the model may struggle to generalize its predictions to new data, despite maintaining a relatively high specificity.

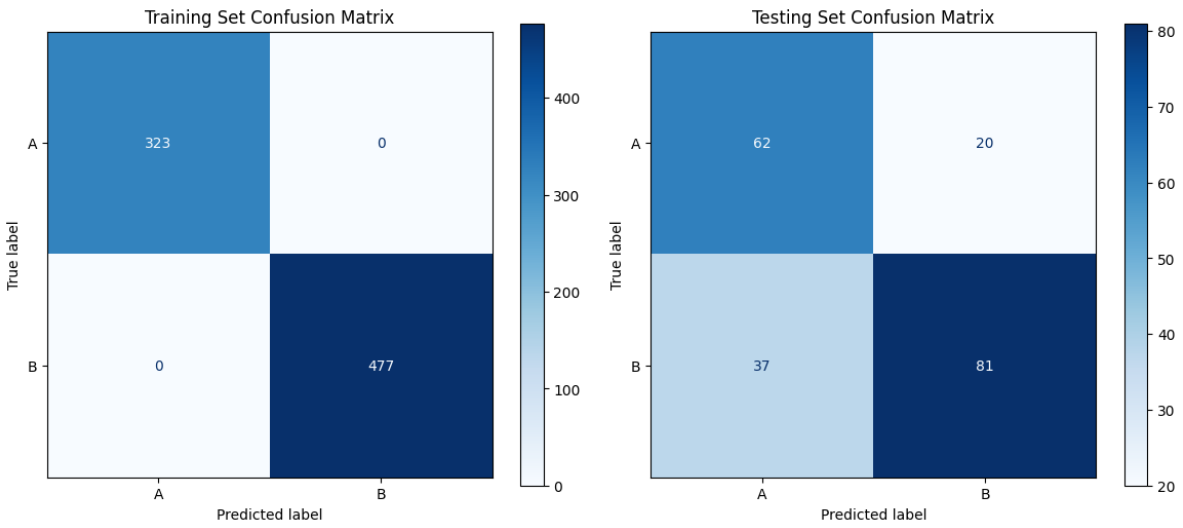

Figure 3: Confusion Matrix for KNN model

Overall, I would choose the logistic regression model to put into production because of its high performance and consitent metrics between training and testing. Despite having similar resuls as the SVM, the logistic regression seems to have generalized the data better than the SVM. This may be due to SVM being good for a high dimensionality dataset which is not the case for our example.

4 Reflection

In the process of completing this assignment, I realized that there is a place for every machine learning model, and despite learning a new model such as SVM, it doesnt imply that it will perform better than a simpler model like logistic Regression. Since the benefit of using an SVM is to translate data to a higher dimension for it to be split better, SVM would perform well on a more complicated dataset, in our example, the dataset was simple enough that a basic regression model is sufficient.