# Assignment 1: Wine quality classification with K-NN

Student: Hanshi Tang 1266337; Daksh Agrawal 1340113

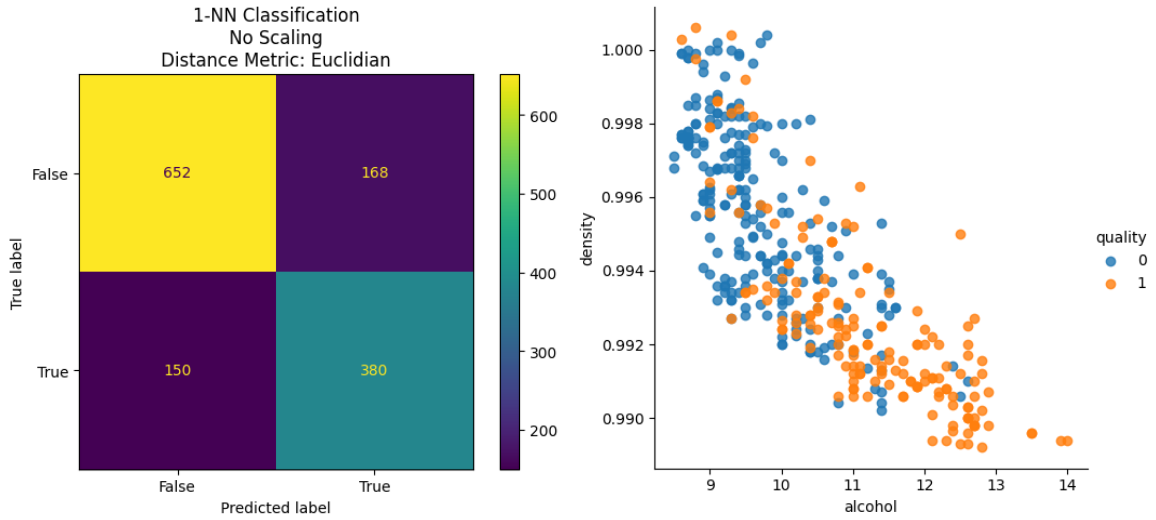
**Q2**

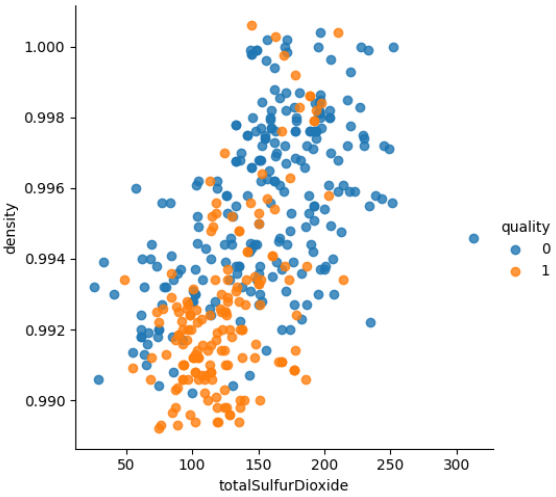
The 1 Nearest Neighbour (1-NN) classifier yielded a moderately well accuracy of 76%, which provides an intuitive understanding of the correct classification on wine qualities. This suggests that our dataset contains enough features to identity instances into correct classes. 1-NN model may demonstrate a less robust outcome if the high quality and low-quality classes share a significant overlapping region. This is because 1-NN only relies on the closet training sample by distance.

Extracting scatter plots from the group of paired plots, Alcohol and Density exhibit clear separations. Most high-quality wine samples share high alcohol and low-density values. In such case, 1-NN classifier has a stronger opportunity to perform well. After calculating Euclidean Distance, it would be uncommon for the NN to be categorised in a false class.

Randomness also impacts 1-NN's performance. The random picking mechanism for tie-breaking purpose would lead to inconsistency of prediction results in each run.

The issue of having imbalances on the scales for paired features is prevalent. The most extreme case could be Total Sulphur Dioxide which spans across 0 to 300, and this feature would dominate the prediction process. In contrast, those with smaller scales like density could barely contribute.




1.000
0.998
0.996
0.994
0.992
0.990
density
50
100
150
200
250
300
totalSulfurDioxide
quality
0
1

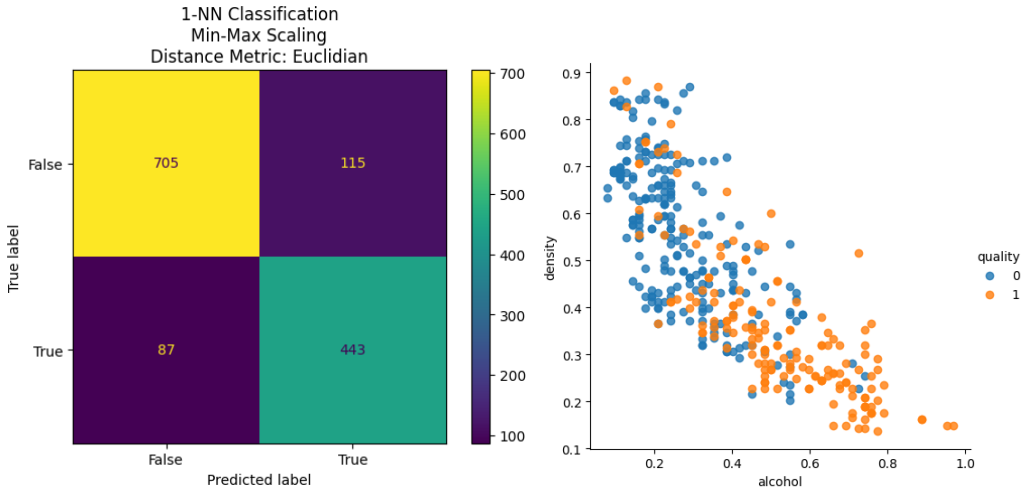
**Q3**

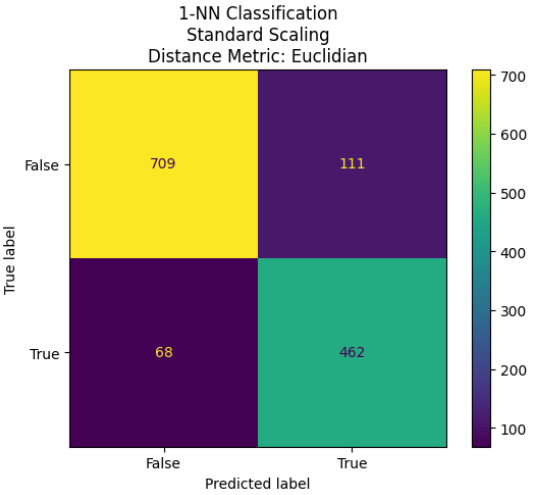
Normalisation serve as a pre-processing step for distanced based Machine Learning algorithms. The Min-Max scaling adjusts the features under the same scale from 0 to 1 while preserving its relative distance within its own class range. Standardization provides a similar effect, but distributions will share the same mean of 0, and standard deviation of 1.

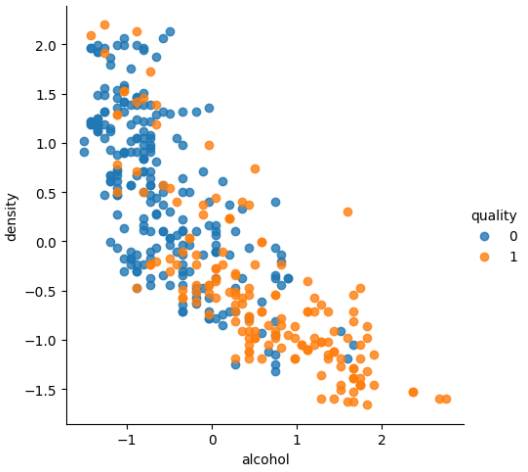
K-NN model is biased when the results are generated from the few features with significantly larger scales. Moreover, normalisation may specifically improve 1-NN model, since it only relies only on 1 closet data sample. The information utilization rate could be relatively low. After normalisation, more features could contribute to the result, even if they were on smaller scales.

Prediction accuracy has been improved overall according to the distance metrics. Under the method of Min-Max Scaling, F1-score rises to 0.87 for Class 0 and 0.81 for Class 1. Comparing the Min-Max Scaled scatter plot with original one, the scale on axes have been normalised, and both axes are in a fair position to contribute. However, this process preserves the distributions of the dataset, and the spread on the plot remains identical. The equality in scaling optimises the performance for K-NN models, and now similarity between samples could be measured consistently across all 11 dimensions.

Standardisation also provides crucial improvement with its accuracy increased to 87%, compared to 76% before pre-processing. This is even higher than 85% with Min-Max scaling. The scatter plot represents the number of standard deviations away from the mean. Without a confined upper or lower bound, this method is particularly useful as it retains the representation of outliers, without suppressing the general distribution.

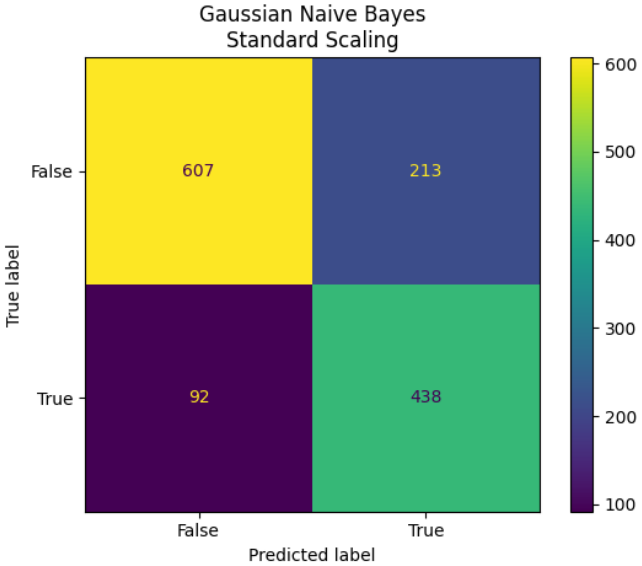



1-NN Classification
Standard Scaling
Distance Metric: Euclidian
709
111
68
462
False
True
True label
Predicted label
700
600
500
400
300
200
100


density
alcohol
quality
0
1
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2

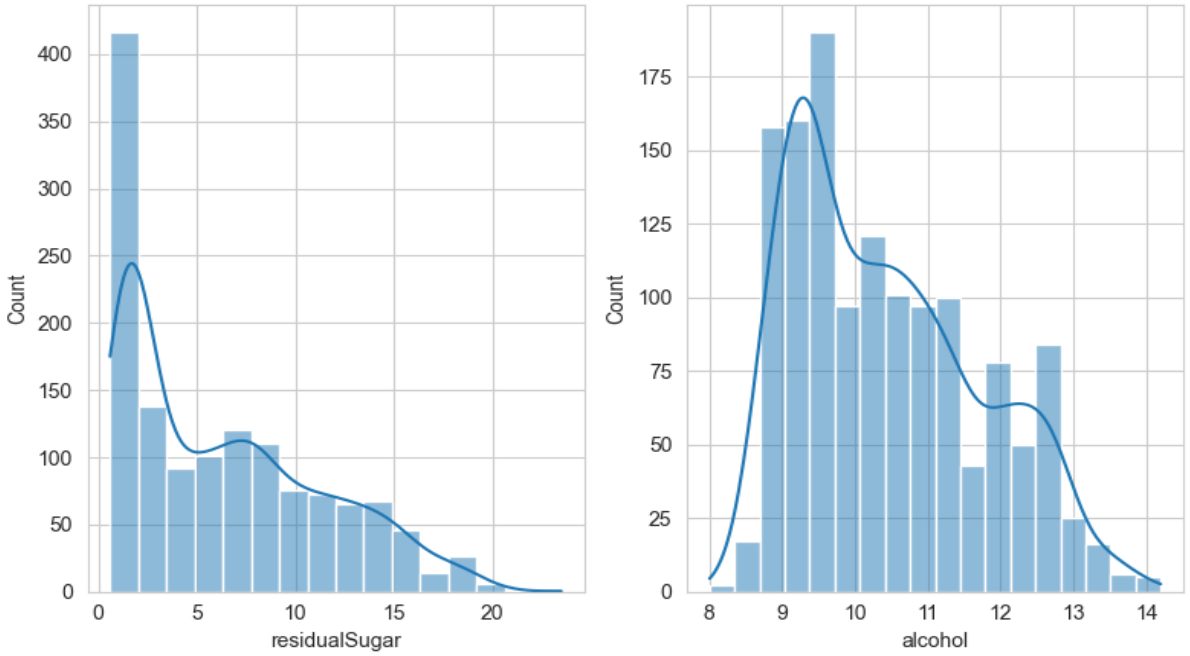
**Q4.1**

1-NN models yielded with a much more robust accuracy of 87% compared to Gaussian Naïve Bayes (GnB) model after deploying normalisation. Some discrepancies for the effectiveness could be attributed to the fundamental assumptions between models.

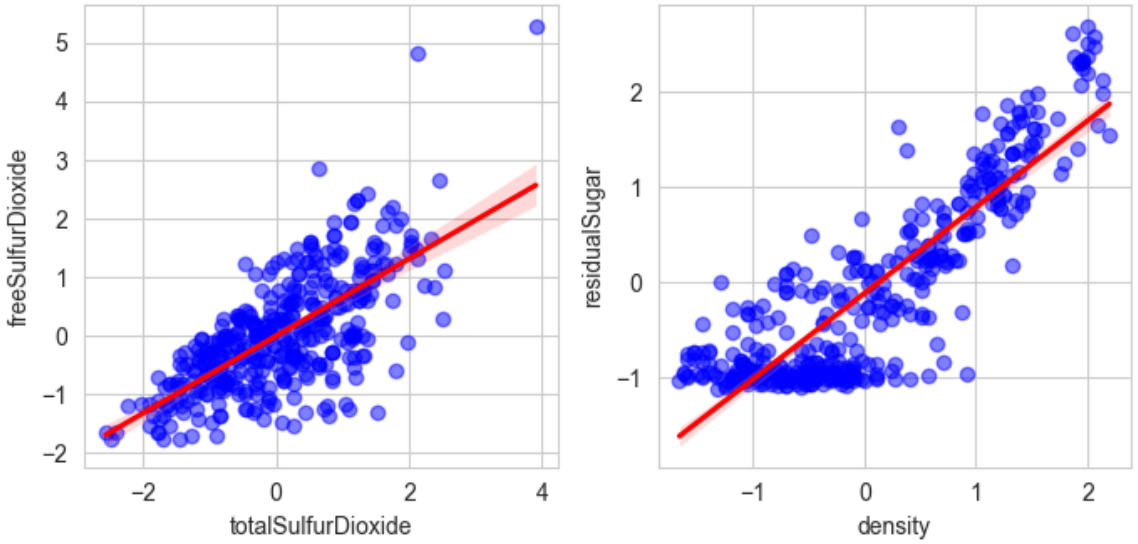



The GnB model requires a prerequisite assumption of Normal Distribution. Under a real-world scenario such as wine quality classification, it would be ambitious to assume all attributes strictly follow normal distributions. This leads to risks of having misclassifications.

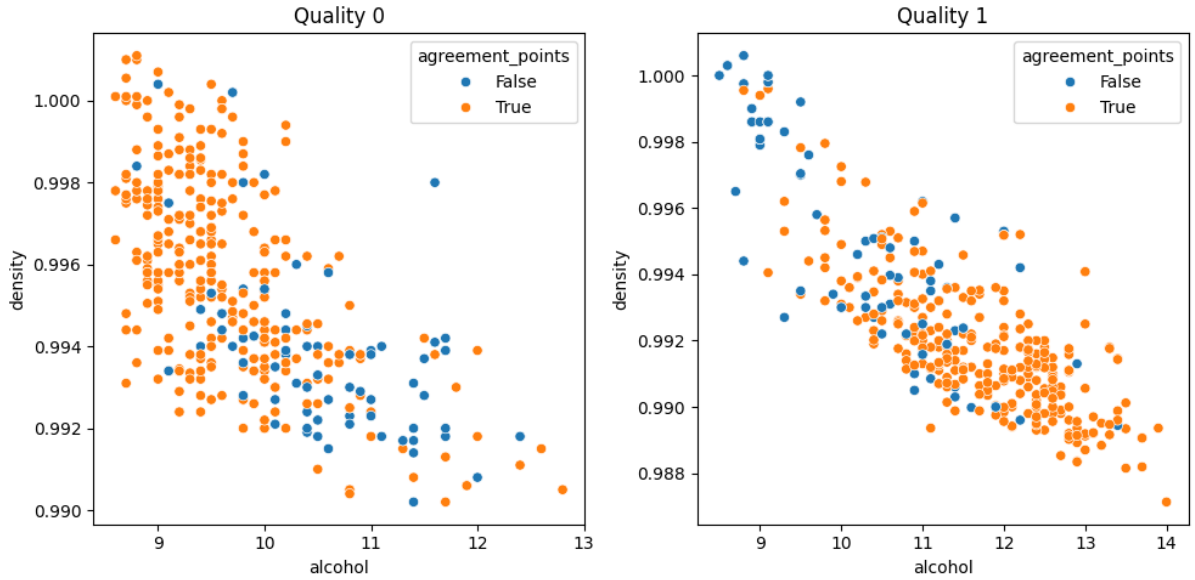
Residual sugar and alcohol levels shown below are examples that contradict with the assumption. Residual sugar has a strong right skewness, with a long tail for high residual sugar values. This indicates that majority of wine samples contains low residual sugar values. For alcohol levels in wine, the major peak of the histograms is located slightly more centred while the Kernel Density Estimate line demonstrates a smaller peak at the value of 12.5. Both features fail to meet the ideal normal assumption and lead to negative impacts on predictions.



Another core assumption of GnB model is the conditional independence among all features. In the context of wine quality classifications, attributes do share natural correlations. The following features exhibit certain degrees of positive correlations. A moderate positive correlation of 0.646 were observed for free Sulphur Dioxide and Total Sulphur Dioxide. At the same time, a strong positive correlation at 0.830 were observed for residual sugar and density of wines. These positive correlated features together have an unfair advantage over their independent ones and compromises the prediction capability.




When observing instances where both models disagree, they happen to disagree mutually on similar instances. Since Alcohol and Density are major contributing factors, both models tend to disagree when the test instance is present in the opposite quality cluster. This creates challenges for classifiers regardless of models. The assumption here might be both models are biased to predict and cluster samples into predominant class.




**Q 4.2**

Three distance metrics including were deployed for comparison against three different normalisation techniques, which helps analysts optimise the executions in practise.

Euclidean Distance metric is sensitive to the scales of features, Min-Max Scaling restricts all features by the scale of 0 to 1, and Standardisation refines features with the same mean of 0 and standard deviation of 1. For Cosine Similarity, it is technically scale-invariant since the angle between two vectors dominate the prediction calculation. However, the accuracy still improved, as normalisation achieves a more balanced influence of features on distance computation.

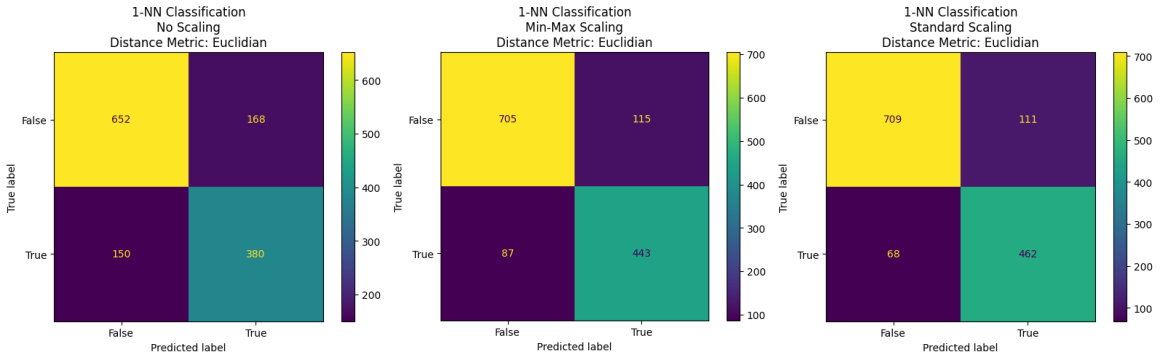
Mahalanobis Distance itself contains standardisation logic embedded behind, and this feature diminishes the meaning of extra scaling process. Specifically, Mahalanobis distance incorporates the variance and co-variance of samples into calculation, and essentially becomes scale-invariant. This is critically beneficial for datasets with huge difference in scales. The experiment also supports the hypothesis since all three yielded the same accuracy of 0.86. Without scaling, Mahalanobis achieves much better accuracy than others, 10% higher than Euclidean and 9% higher than Cosine Similarity.

In comparison, results for both Euclidean Distance and Cosine Similarity method have been significantly improved after Min-Max scaling and standardisation. Especially for Euclidean Distance, 11% enhancement were observed. Scaling mitigates their disadvantages, and their accuracy eventually ties with Mahalanobis.

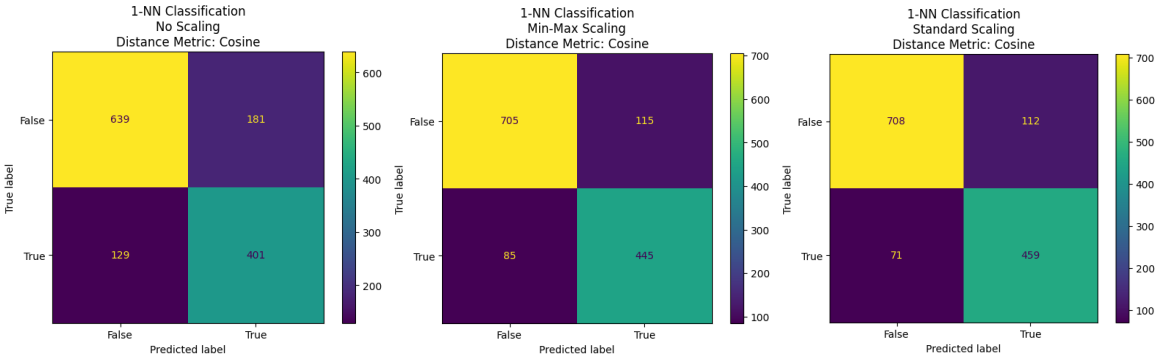
For this dataset, by only changing the distance metrics could not help improving the model's accuracy, and Euclidian Distance achieved the maximum accuracy overall, albeit only by a slim margin.

| **Accuracy Table** | | | |
|---|---|---|---|
| | **Euclidean Distance** | **Cosine Similarity** | **Mahalanobis Distance** |
| **No Scaling** | 0.76 | 0.77 | 0.86 |
| **Min-max** | 0.85 | 0.85 | 0.86 |
| **Standardisation** | 0.87 | 0.86 | 0.86 |

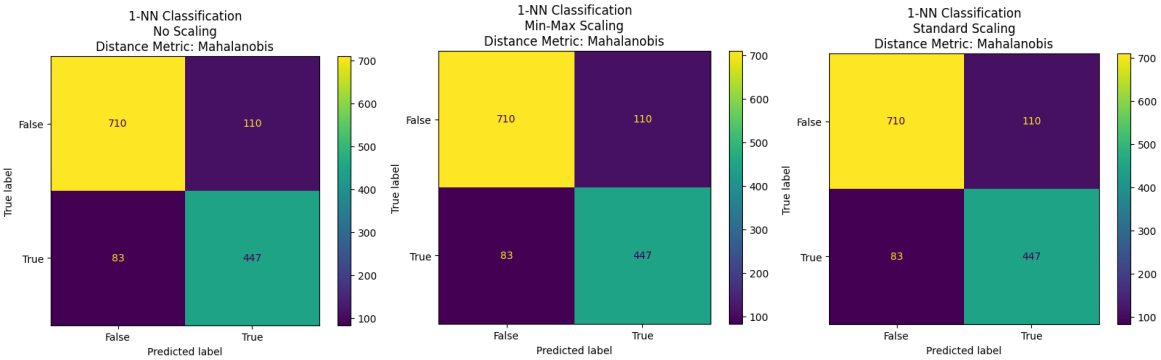
## Euclidean Distance




## Cosine Similarity



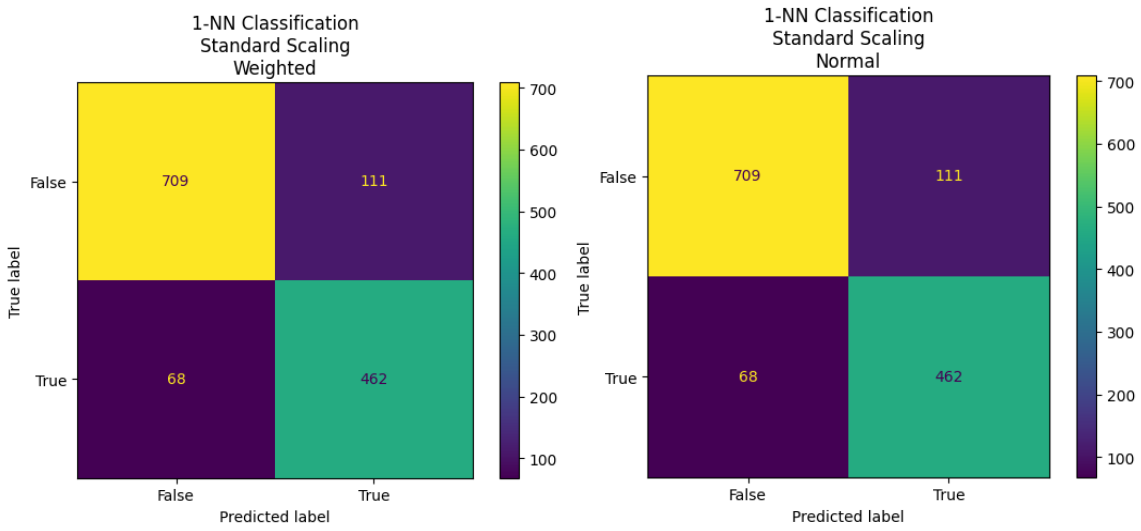

## Mahalanobis Distance



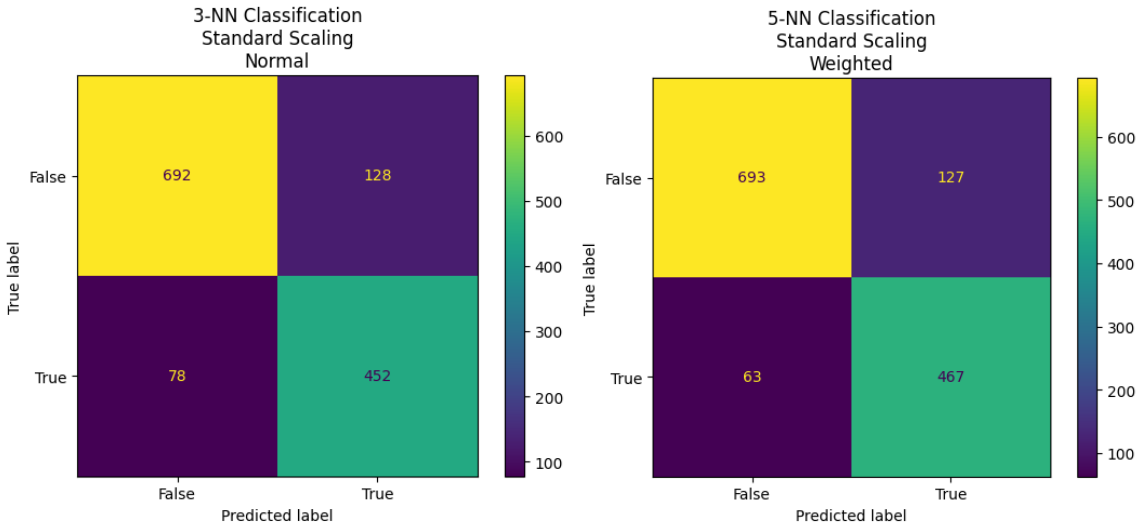
**Q 4.3**

The weighted k-NN classifier provides an innovative voting strategy compared with the traditional majority voting. Under such model, neighbours contribute differently to the voting according to their proximity to the query distance and becomes more robust and prepared for noise. As neighbours located further have a higher chance to be outliers, their value for prediction gets diminished.

It is crucial to finetune the hyperparameter of k when understanding the effect of weighted k-NN model. When k is small, weighted model demonstrates high sensitivity towards noise. In the most extreme case when k=1, weighting offers no effect as only 1 neighbour would be selected. However, overfitting might occur, and noise could be mis-captured as patterns.




As the number of k increases, both classifiers start considering more neighbours which helps smooth the impact of noise. However, now the risk of having the class boundary line blurred and the overlapping area increased also rises. The model could be underfitting and could not generalise new data samples well. Here, implementing weighting would be beneficial, as it would allow closer points to have priority votes than far away unrelated points, which might be considered neighbours due to high k value.



However, even with large k values, the models do not provide any critical improvement in accuracy rates. While a clustering pattern is present in the dataset, there are a significant number of out of cluster points which may continue being misclassified with increased k values. As the clustering effect becomes prevalent and boundaries start getting definite and strictly reflective of the cluster majority, overthrowing the impacts of any potentially helpful sparse training points with highly similar features and same labels present.