Analysis Report by

Sagar Parmar: 301271932

The Data Exploration indicates that we have 10 columns in our dataset excluding the “ID” column. Total number of records are 699.

Each column is the type of ‘int64’ but bare is the type of ‘object’. Other than ‘ID’ and ‘class’ column, all the other column has the value out of 10. Moreover, ‘bare’ column has some unrelated value ‘?’. ‘Class’ column has only two discrete values “2” & “4” indicating ‘benign’ and ‘Malignant’ cases respectively. Moreover, we can also see that the number related to class=4 in other words ‘Malignant’ cases are rather higher and more aggressive than ‘Benign’ cases in following columns: ‘thickness’, ’size’, ‘Epith’ and b1. While describing the Data and getting mean,std,count,etc.. values. I noticed that I didn’t get the ‘bare’ column which indicated again that data type is not int or float but an ‘object’.

A screenshot of a computer

Description automatically generated  
A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

By plotting graphs usually in the case of ‘Benign’ cells, the size and shape are almost between 1 to 4 maybe 5. In contrary to ‘Malignant cells’ where both the size and shape of the cells are larger and varies from 2-10 abnormally (irregular shape but size is larger). Blue Scatter plot is of Benign as red scatter plot is of Malignant. Moreover, the histplot suggests that comparatively in the benign cases, thickness of cells is much lower than in malignant cases. As you can see purple color represents the malignant case and yellow color represents the benign case.

A screenshot of a computer

Description automatically generated

By looking at the classification report of linear kernel, we can see that model performs well in identifying benign and malignant tumours, with good precision, recall, and F1-scores on both the training and testing sets. On the training set, it achieves 97% accuracy, with precision and recall rates close to 0.98 for benign tumours and slightly lower for malignant tumours, having F1-scores of 0.98 and 0.96, respectively. When focusing on testing set, the model makes minor adjustments but still gives us good performance, with an overall accuracy of 96%. Precision for benign tumours falls slightly to 0.95, but for malignant tumours it rises to 0.96, demonstrating the model's ability to generalise well to new data.

A screenshot of a computer

Description automatically generated

By looking at the classification report of RBF Kernel, we can see that model accurately classifies breast cancer tumours as benign (Class 2) or malignant (Class 4). In the training set, it gets an accuracy of 98%, with precision and recall better than before for benign tumours (0.99 and 0.98, respectively), and similarly high for malignant tumours (0.96 precision and 0.98 recall). This indicates that F1-scores of 0.98 for benign tumours and 0.97 for malignant tumours, suggesting It is the most reliable up to this point. On the testing set, the model makes minor adjustments but maintains high performance, with an overall accuracy of 96%. It has precision of 0.98 and 0.95 and recall of 0.97 and 0.96 for benign and malignant tumours, respectively, resulting in F1-scores of 0.97 and 0.95.

A screenshot of a computer

Description automatically generated

By looking at the classification report of Poly Kernel, we can see that model performs well in classifying breast cancer tumours, getting 98% accuracy on the training set and 93% on the testing set. other metrics on the training set are precision rates of 0.98 (benign) and 0.97 (malignant), as well as recall rates of 0.99 and 0.96. F1-scores are high, with 0.98 for benign and 0.96 for malignant tumours. On the testing set, precision for benign tumours is slightly lower at 0.91, with a recall of 0.98, whereas precision for malignant tumours is 0.96 with a recall of 0.85. It shows that this model is inferior to both rbf in linear model.

A screenshot of a computer

Description automatically generated

By looking at the classification report of Sigmoid Kernel, we can see that model's accuracy drops to 48% on both the training and testing sets. On the training set, it achieves a reasonable precision of 0.59 and recall of 0.71 for benign tumours, but both metrics drop to 0.01 for malignant tumours, showing that it is not reliable to identify these cases. This pattern continues in the testing set, with benign tumour accuracy at 0.55 and recall at 0.78, but malignant tumour identification is nearly nonexistent, with precision and recall at 0.00. This metrics shows us that this model is worse of all for this kind of dataset.

So by evaluating all kernels, I would say that rbf is by far the best kernel that I would recommend. It stands out from the rest in overall every metrics score for the dataset. It has higher accuracy and precision so there is very little chance of (False,True) scenarios. It identified malignant cases very precisely and gave overall accuracy 97% on training set and 96 on training set. It has proper balance of F1 score and recall as well. Unlike the sigmoid kernel which was the worst of all in identifying malignant cases. On the second place I would also recommend linear kernel cause polynomial kernel performed poorly in recall of malignant cases. As a result I would recommend RBF Kernel for the SVM classification based on the superior performance on the given dataset.

Exercise: 2

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

Train Data Accuracy Score = 97%

A screenshot of a computer

Description automatically generated

Test Data Accuracy Score = 96%

Exercise 1 covers basic machine learning concepts like data preprocessing, visualisation, and SVM modelling with various kernels, with a focus on basic skills. Exercise 2 builds on these concepts by adding pipelines for efficient preprocessing, using grid search for hyperparameter tuning, and serialising the model before deployment. The main difference is that Exercise 2 focuses on optimizing model performance for real-world applications, showing how to move from basic modeling techniques to a superior, deployment-focused workflow.