**Assignment #3**

**Support Vector Machine**

**Written Responses**

1. In you written response writes a paragraph explaining your findings about each column.

Many of the columns are int64, meaning that almost all the columns are numerical, the bare column is stored as an object instead of a numerical value, because it contains non-numerical values like ‘?’, so, Pandas assumes that the whole column is a String. The Max

No missing Values are reported, the ID column has a very large range, it has 8 columns, the range 1 to 10 suggest they might be ordinal features (ratings or severity levels).

1. what are the key insights and findings from the plots

A graph of different colors

AI-generated content may be incorrect.

This bar chart represents the distribution of the classes in this case the 10 columns in the dataset, the similar values of each features means that there are no missing values or imbalances, The different color represent the different categories, providing a different distinction

A graph of a number of bars

AI-generated content may be incorrect.

This histogram displays the distribution of the Bare Nuclei Class, this one is highly right-skewed, so, most of the values are concentrated on the lower end (between 1 and 2), the there Is another increase in the 10, suggesting there is high amount of bare nuclei values which may correlate with malignant tumors, the rest of the values are imbalanced, indicating there are few bare nuclei counts.

A screenshot of a computer screen

AI-generated content may be incorrect.

Finally, this graph represents the distribution of the classes in the dataset, where: Class 2 = Benign Tumors and Class 4 = Malign Tumors.

In this graph we can observe that there is a high count in Benign Tumors than Malign ones.

1. Print out two accuracies score one for the model on the training set i.e. X\_train, y\_train and the other on the testing set i.e. X\_test, y\_test. Record both results in your written response.

A number on a blue background

AI-generated content may be incorrect.

The accuracy in this case is 95% mean that the SVM classifier is correctly predicted, it has a high accuracy and is well distinguishing between classes. This suggest the model is very effective.

1. Generate the accuracy matrix. Record the results in your written response.

A screenshot of a computer

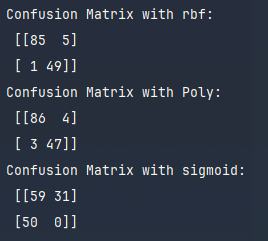
AI-generated content may be incorrect.

Well, the Matrix is structured as [[True Negative (TN) False Positive (FP)]

[[False Negatives (FN) True Positives (TP)]

In my matrix the 85 represents that the model is correctly predicting the negative class, the 5 is incorrectly predicting the False Positive class, the 1 is incorrectly predicting the False Negative class, and finally the 491 is the number of True positives, the model is correctly predicting the Positive Class.

1. By now you have the results of four SVM classifiers with different kernels recorded in your written report. Please examine and write a small paragraph indicating which classifier you would recommend and why?



Well, let’s start with the first confusion matrix, which shows a good performance, High TP and TN, low FP and FN, the second Confusion matrix performs a poor performance with a slightly higher TP and lower FP compared to the first one, and the last one with sigmoid indicates a poor performance, especially in the TN and FN, making it non suitable for this dataset.

The best kernel is the RBF and Poly, based on its confusion matrix, its performance metric and application requirements are what makes them suitable, The sigmoid kernel and Poly.

1. Take a screenshot showing your num\_pipe\_firstname object and add it to your written report.



Representation of a machine learning preprocessing pipeline using scikit-learn’s pipeline.

1. Take a screenshot showing your grid search parameter object and add it to your written report.



Fitting the training Data:

A screen shot of a computer code

AI-generated content may be incorrect.

15. Print out the best parameters and note it in your written response



These are the parameters used in the grid search, o find the best combinations of SVM for the model.

16. Printout the best estimator and note it in your written response

A computer code on a dark background

AI-generated content may be incorrect.

This is the best estimator, is a pipeline that includes all the preprocessing data steps and a SVM classifier.

17. Fit the training data to the best model. Printout the accuracy score and note it in your written  
response

A close up of a blue background

AI-generated content may be incorrect.

The accuracy score of the best model is 95% this indicates that this model is very optimal and perform very well on the test data, so 95% of the predictions made by the model are correct.

19. Finally, in your written response Compare the results and write your conclusions. As part of conclusions indicate the main difference between exercise #1 and exercise #2.

So, as we can see both exercise works similar but has different approaches, the first exercise is using manual tunning hyperparameter as an example we are using C, gamma, kernel, degree to do SVM Models, having a Accuracy score of 0.9571% being a very good performance for this model, efficient and optimal. On the other hand, in the second exercise we use a pipeline with transformers as simpleImputer, StandarScaler, to preprocess the data before using it in the SVM model, having also an Accuracy score of 95% which is effective and make the predictions correctly the 95% of the time. So, in conclusion their Accuracy Score is similar between both (#1. 0.9571 and #2.0.95), In this case this result suggest that the First exercise is the most optimal for this model just for few decimals, but, while Exercise #1 achieved a slightly higher accuracy, the Exercise #2 offers a more scalable and reproducible approach, making it a better choice for real world applications