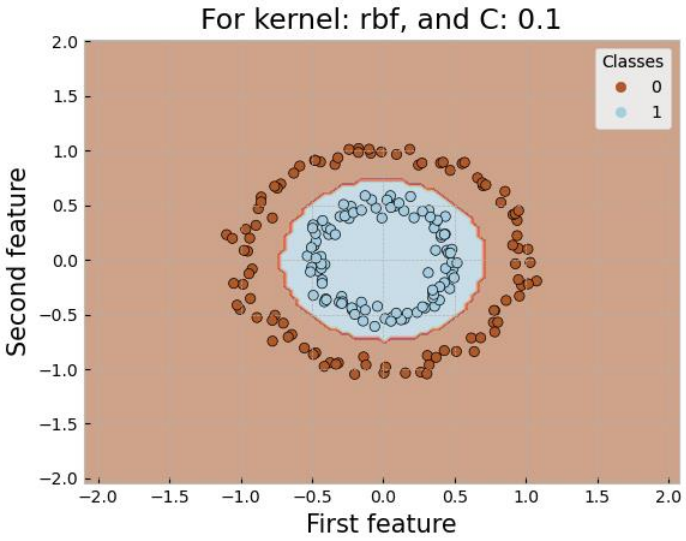
CNG 409 Assignment 3

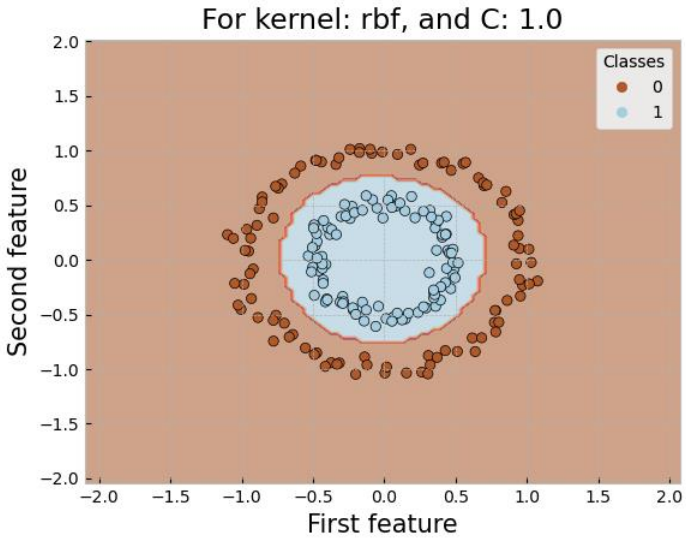
Talal Shafei 2542371

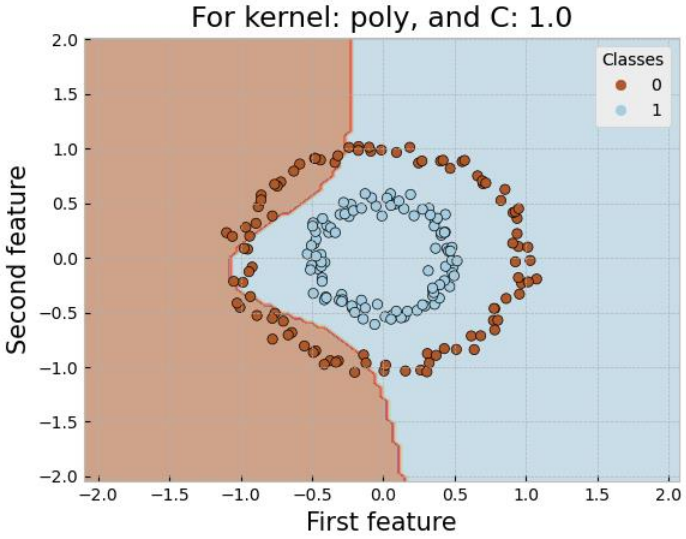
Part 2:

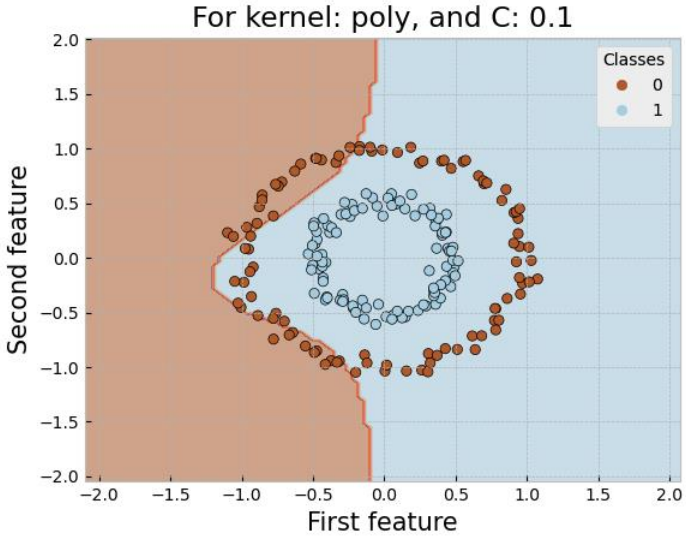
Dataset1:

Configurations was formed to test two C values (0.1 and 1) and two kernels (Radial Basis Function, Poly: degree 3 by default)









As we can see from the figures above RBF was much better to find hyperplane to classify the data, also since there is no overlapping between the classes C value didn't affect the margin that much, it slightly bigger for C equals one.

Dataset2:

Configuration was formed to tune C and kernel of SVM, by using exhaustive search over C values (1.0, and 5.0) and kernel values (RBF, Poly: default degree 3)

Results obtained from cross-validation with 10 splits repeated 5 times:

Index	C	Kernel	Accuracy (%)	STD	95% Confidence Intervals
1.	1.0	Poly	88.000	7.303	[80.843, 95.157]
2.	1.0	RBF	92.267	6.016	[86.371, 98.163]
3.	5.0	Poly	91.733	6.33	[85.527, 97.940]
4.	5.0	RBF	94.133	5.439	[88.803, 99.464]

Based on the results the best configuration was the fourth one with C: 5.0 and kernel: RBF, with the highest Mean accuracy 94.267 %, and the lowest Standard deviation, therefore it is the most reliable.

Part 3:

In this part we will use nested cross-validation to compare four algorithms (KNN, SVM, Decision Tree, and Random Forest), we begin by showing all the results based on the Accuracy score, then do the same based on the F1 score.

The values of the hyper parameters we are going to test for each model are:

KNN:

Number of neighbors: 10, 15
Metric: Manhattan, Cosine

SVC:

C: 10, 15
Kernel: RBF, Sigmoid

Decision Tree:

Cost complexity parameter alpha: 0.01, 0.02
Criterion: Gini, Entropy

Random Forest:

Number of Estimators: 100, 150
Criterion: Gini, Entropy

Accuracy:

Inner Results:

Below, we show the best configurations obtained from the inner cross-validation for each model.

KNN:

Index	N_neighbours	Metric	Mean	STD	95% Confidence Interval
1.	Manhattan	10	0.735	0.030	[0.705,0.764]
2.	Manhattan	10	0.726	0.026	[0.701,0.751]
3.	Cosine	10	0.709	0.028	[0.681,0.737
4.	Manhattan	15	0.724	0.021	[0.703,0.745]
5.	Manhattan	10	0.712	0.038	[0.676,0.749]
6.	Manhattan	15	0.722	0.025	[0.697,0.746]
7.	Cosine	15	0.720	0.024	[0.696,0.744]
8.	Manhattan	10	0.728	0.024	[0.705,0.751]
9.	Manhattan	10	0.734	0.038	[0.697,0.772]
10.	Cosine	15	0.725	0.019	[0.706,0.744]
11.	Cosine	10	0.733	0.026	[0.708,0.758]
12.	Cosine	15	0.720	0.027	[0.693,0.747]
13.	Cosine	10	0.723	0.027	[0.697,0.749]
14.	Manhattan	10	0.740	0.028	[0.713,0.768]
15.	Manhattan	15	0.712	0.026	[0.687,0.738]

From the table we can see that Manhattan and cosine nearly the same amount of times but 10 neighbours was getting better results than 15 in general.

SVM:

Index	C	Kernel	Mean	STD	95% Confidence Interval
1.	10	RBF	0.731	0.029	[0.703,0.759]
2.	10	RBF	0.730	0.034	[0.697,0.763]
3.	10	RBF	0.704	0.036	[0.669,0.739]
4.	10	RBF	0.734	0.029	[0.705,0.763]
5.	10	RBF	0.709	0.034	[0.676,0.742]
6.	10	RBF	0.739	0.030	[0.709,0.768]
7.	10	RBF	0.713	0.032	[0.682,0.744]
8.	10	RBF	0.730	0.039	[0.692,0.768]
9.	10	RBF	0.737	0.037	[0.701,0.773]
10.	15	RBF	0.737	0.035	[0.703,0.771]
11.	10	RBF	0.721	0.026	[0.696,0.746]
12.	10	RBF	0.708	0.044	[0.665,0.751]
13.	10	RBF	0.724	0.036	[0.689,0.759]
14.	10	RBF	0.744	0.027	[0.718,0.770]
15.	10	RBF	0.713	0.037	[0.677,0.749]

For the SVM we can see that C = 10 and RBF kernel was the best configuration in all the inner cross validation, except at in the 10^{th} time.

Decision Tree:

Index	Ccp_alpha	Criterion	Mean	STD	95% Confidence Interval
1.	0.02	Entropy	0.721	0.034	[0.688,0.754]
2.	0.02	Gini	0.699	0.005	[0.695,0.704]
3.	0.02	Gini	0.696	0.014	[0.682,0.710]
4.	0.01	Entropy	0.721	0.021	[0.700.0.742]
5.	0.02	Gini	0.700	0.001	[0.699,0.701]
6.	0.01	Entropy	0.705	0.044	[0.662,0.749]
7.	0.01	Entropy	0.713	0.041	[0.673,0.753]
8.	0.01	Gini	0.707	0.023	[0.688,0.714]
9.	0.02	Entropy	0.701	0.013	[0.688,0.714]
10.	0.01	Entropy	0.719	0.034	[0.686,0.752]
11.	0.02	Gini	0.700	0.001	[0.699,0.701]
12.	0.01	Entropy	0.702	0.035	[0.668,0.736]
13.	0.01	Entropy	0.703	0.030	[0.674,0.733]
14.	0.01	Gini	0.715	0.024	[0.691,0.738]
15.	0.01	Gini	0.708	0.030	[0.678,0.737]

In Decision Tree the ccp_alpha was 0.02 most of the times when criterion was Gini, and vice versa it was 0.01 when criterion was entropy.

Random Forest:

Index	N_estimators	Criterion	Mean	STD	95% Confidence Interval
1.	150	Entropy	0.765	0.027	[0.755,0.776]
2.	150	Entropy	0.745	0.022	[0.736,0.753]
3.	150	Entropy	0.750	0.022	[0.741,0.758]
4.	150	Entropy	0.765	0.025	[0.755,0.775]
5.	150	Gini	0.735	0.027	[0.724,0.746]
6.	100	Entropy	0.753	0.028	[0.742,0.764]
7.	150	Gini	0.743	0.015	[0.737,0.750]
8.	150	Entropy	0.754	0.027	[0.743,0.764]
9.	150	Entropy	0.761	0.026	[0.750,0.771]
10.	150	Gini	0.760	0.020	[0.752,0.768]
11.	100	Entropy	0.751	0.021	[0.742,0.759]
12.	150	Entropy	0.747	0.027	[0.737,0.758]
13.	100	Entropy	0.753	0.22	[0.744,0.761]
14.	150	Entropy	0.761	0.026	[0.751,0.771]
15.	100	Entropy	0.742	0.022	[0.733,0.751]

In Random forests 150 estimators and entropy criterion was the best configuration most of the times.

Outer Results (Models Comparison):

Here, we evaluate the models on the best configurations they obtained from the tables above and the model that gets highest reliable score we choose to use for the final project on the data.

Stats / Models	Mean (Accuracy)	STD	95% Confidence Interval
KNN	0.720	0.014	[0.713,0.727]
SVM	0.741	0.025	[0.728,0.754]
Decision Tree	0.703	0.015	[0.695,0.710]
Random Forest	0.764	0.012	[0.757,0.770]

Based on the Accuracy Random Forest was the better Model for this problem, with the highest score 0.764 and the lowest standard deviation 0.012

F1 Score:

Inner Results:

Below, we show the best configurations obtained from the inner cross-validation for each model.

Due to using Repeated Stratified K-Fold all the partitions have balanced classes and none of the models use stronger weight to one class in the account of the other, therefore F1 score will be equal to Accuracy in the non-stochastic models, and nearly the same in stochastic ones.

KNN:

Index	N_neighbours	Metric	Mean	STD	95% Confidence Interval
1.	10	Manhattan	0.735	0.030	[0.705,0.764]
2.	10	Manhattan	0.726	0.026	[0.701,0.751]
3.	10	Cosine	0.709	0.028	[0.681,0.737]
4.	15	Manhattan	0.724	0.021	[0.703,0.745]
5.	10	Manhattan	0.712	0.038	[0.676,0.749]
6.	15	Manhattan	0.722	0.025	[0.697,0.746]
7.	15	Cosine	0.720	0.024	[0.696,0.744]
8.	10	Manhattan	0.728	0.024	[0.705,0.751]
9.	10	Manhattan	0.734	0.038	[0.697,0.772]
10.	15	Cosine	0.725	0.019	[0.706,0.744]
11.	10	Cosine	0.733	0.026	[0.708,0.758]
12.	15	Cosine	0.720	0.027	[0.693,0.747]
13.	10	Cosine	0.723	0.027	[0.697,0.749]
14.	10	Manhattan	0.740	0.028	[0.713,0.768]
15.	15	Manhattan	0.712	0.026	[0.687,0.738]

SVM:

Index	C	Kernel	Mean	STD	95% Confidence Interval
1.	10	RBF	0.731	0.029	[0.703,0.759]
2.	10	RBF	0.730	0.034	[0.697,0.763]
3.	10	RBF	0.704	0.036	[0.669,0.739]
4.	10	RBF	0.734	0.029	[0.705,0.763]
5.	10	RBF	0.709	0.034	[0.676,0.742]
6.	10	RBF	0.739	0.030	[0.709,0.768]
7.	10	RBF	0.713	0.032	[0.682,0.744]
8.	10	RBF	0.730	0.039	[0.692,0.768]
9.	10	RBF	0.737	0.037	[0.701,0.773]
10.	15	RBF	0.737	0.035	[0.703,0.771]
11.	10	RBF	0.721	0.026	[0.696,0.746]
12.	10	RBF	0.708	0.044	[0.665,0.751]
13.	10	RBF	0.724	0.036	[0.689,0.759]
14.	10	RBF	0.744	0.027	[0.718,0.770]
15.	10	RBF	0.713	0.037	[0.677,0.749]

Decision Tree:

Index	Ccp_alpha	Criterion	Mean	STD	95% Confidence Interval
1.	0.02	Entropy	0.721	0.034	[0.688,0.754]
2.	0.02	Gini	0.699	0.005	[0.695,0.704]
3.	0.02	Gini	0.696	0.014	[0.682,0.710]
4.	0.01	Entropy	0.721	0.021	[0.700.0.742]
5.	0.02	Gini	0.700	0.001	[0.699,0.701]
6.	0.01	Entropy	0.705	0.044	[0.662,0.749]
7.	0.01	Entropy	0.713	0.041	[0.673,0.753]
8.	0.01	Gini	0.707	0.023	[0.688,0.714]
9.	0.02	Entropy	0.701	0.013	[0.688,0.714]
10.	0.01	Entropy	0.719	0.034	[0.686,0.752]
11.	0.02	Gini	0.700	0.001	[0.699,0.701]
12.	0.01	Entropy	0.702	0.035	[0.668,0.736]
13.	0.01	Entropy	0.703	0.030	[0.674,0.733]
14.	0.01	Gini	0.715	0.024	[0.691,0.738]
15.	0.01	Gini	0.708	0.030	[0.678,0.737]

We can see that KNN, SVM, and Decision Tree F1 score' tables are the same as the Accuracy tables above

Random Forest:

Index	N_estimators	Criterion	Mean	STD	95% Confidence Interval
1.	150	Entropy	0.766	0.028	[0.755,0.777]
2.	150	Gini	0.744	0.020	[0.736,0.752]
3.	150	Gini	0.748	0.025	[0.738,0.758]
4.	150	Entropy	0.764	0.024	[0.754,0.773]
5.	150	Gini	0.735	0.029	[0.724,0.746]
6.	150	Entropy	0.754	0.030	[0.742,0.765]
7.	100	Entropy	0.743	0.014	[0.738,0.749]
8.	150	Gini	0.752	0.023	[0.743,0.761]
9.	100	Gini	0.760	0.027	[0.759,0.770]
10.	150	Gini	0.761	0.021	[0.753,0.770]
11.	100	Entropy	0.750	0.021	[0.742,0.759]
12.	150	Entropy	0.748	0.026	[0.738,0.759]
13.	150	Entropy	0.755	0.21	[0.746,0.763]
14.	150	Entropy	0.763	0.027	[0.753,0.774]
15.	150	Entropy	0.743	0.024	[0.734,0.752]

A little different but still the 150 estimators and Entropy criterion appear more then the other configurations.

Outer Results (Models Comparison):

Here, we evaluate the models on the best configurations they obtained from the tables above and the model that gets the highest reliable score we choose to use for the final project on the data.

Stats / Models	Mean (F1 score)	STD	95% Confidence Interval
KNN	0.720	0.014	[0.713,0.727]
SVM	0.741	0.025	[0.728,0.754]
Decision Tree	0.703	0.015	[0.695,0.710]
Random Forest	0.761	0.011	[0.757,0.768]

Based on the F1 score the Random Forest was the better Model for this problem, with the highest score 0.761 and the lowest standard deviation 0.011