Project 2 - Orange Project

Ioannis Manousaridis, Anastasios Tzelepakis

In this project two multivariate datasets for classification were used from UCI repository, the **Diabetic Retinopathy Debrecen Data Set**[1] and the **Breast Cancer Wisconsin (Original)**[2] datasets[3][4]. The Breast Cancer and the diabetic Retinopathy Debrecen datasets are multivariate and used for classification problems. The first one consists of 699 instances and has 10 attributes. The second one consists of 1151 instances and has 20 attributes.

Table 1: Classification results for the Breast Cancer Wisconsin Dataset. The Metric used for the evaluation is the Accuracy.

	Without outliers handling and all the features	Without outliers handling but applying 10-PCA	After outliers handling and selecting best 2 features	After outliers handling and selecting best 5 features	Mean Value
Knn Classifier:	96.70%	93.80%	93.00%	93.70%	94.30%
SVM (Linear):	96.80%	95.10%	**94.60%**	**95.20%**	95.43%
SVM (RBF):	**97.70%**	**95.60%**	94.40%	94.30%	**95.50%**
SVM (Polynomial):	91.20%	86.50%	84.20%	90.60%	88.13%
SVM (Sigmoid):	96.00%	89.10%	89.60%	92.80%	91.88%
Mean Value	**95.68%**	92.02%	91.16%	93.32%	

Table 2: Classification results for the Diabetic Retinopathy Debrecen Dataset. The Metric used for the evaluation is the Accuracy.

	Without outliers handling and all the features	Without outliers handling but applying 10-PCA	After outliers handling and selecting best 2 features	After outliers handling and selecting best 5 features	Mean Value
Knn Classifier:	**58.60%**	**68.40%**	**65.80%**	**68.30%**	**65.28%**
SVM (Linear):	51.30%	51.90%	45.30%	41.50%	47.50%
SVM (RBF):	45.70%	57.90%	45.30%	41.50%	47.60%
SVM (Polynomial):	56.00%	52.60%	45.30%	41.50%	48.85%
SVM (Sigmoid):	53.90%	52.60%	56.20%	54.80%	54.38%
Mean Value	53.10%	**56.68%**	51.58%	49.52%	

[1] More info here: https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set
[2] More info here:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
[3] Diabetic Retinopathy in csv format: https://drive.google.com/open?id=113-S08A7O_KQHjv_0fjAtw2WQXWjJIoG
[4] Wdbc in csv format: https://drive.google.com/open?id=1M3J6nERbk-VHYcRQJ9wpL1J-sfoYh4cc

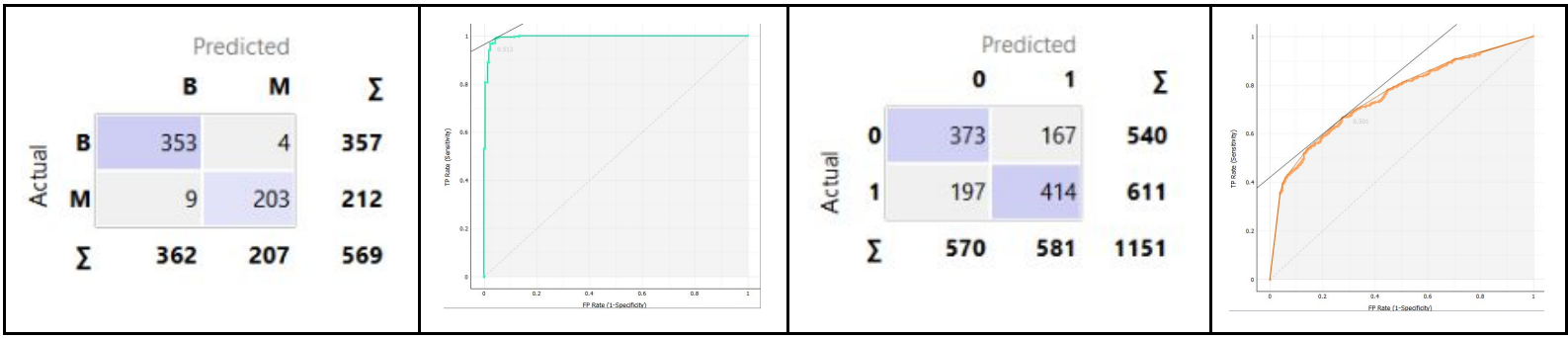

Figure 1: The images which are displayed here are the Confusion Matrix and the ROC plot of the best model for each dataset. On the left it is the wdbc dataset and the best model is the SVM model with the RBF kernel. On the right is the diabetic Retinopathy dataset and the best model is the KNN classifier.

Conclusions:

- It is clear that the Wdbc Dataset classes can be distinguished more accurate than than the classes Diabetic Retinopathy Debrecen. In almost every experiment performed on the Wdbc Dataset a score of 90% is achieved, whereas on the Diabetic Retinopathy Dataset there is not a score higher than 69%.
- In the case of Wdbc Dataset, selecting all features would produce the best results for each classifier, even if there was no Outliers Handling before. On the other hand, when performing experiments on the Diabetic Retinopathy Debrecen Dataset, better results would be observed when less features were used (either with Feature Selection or PCA). This probably has to do with the way that the data was sampled for the two different Datasets. Wdbc data seem to be more important and have a bigger impact to the final scores.
- The features of Breast Cancer Wisconsin Dataset present higher importance because they are uncorrelated. This is the main reason that, when keeping the original number of the features, better scores were achieved. When all features were kept and no Outliers handling was used the mean score of all classifiers was 95.68%.
- On the other hand, features of Diabetic Retinopathy Debrecen Dataset seem to be more correlated. This is the main reason that, when reducing the number of the features, better scores were achieved. When PCA with 10 components and no Outliers handling was used the mean score of all classifiers was 56.68%.
- Experiments on the Diabetic Retinopathy Debrecen Dataset showed that non-linear Classifiers achieved the best scores. The best scores were achieved by the Knn Classifier in each occasion. The mean value of the scores that Knn achieved is 65.28% and the highest value was achieved by applying the PCA and it is 68.40% .
- Experiments on the Breast Cancer Wisconsin Dataset showed that SVM Classifier achieved the best scores. The best scores were achieved by SVM Classifier with RBF kernel. The mean value of the scores that SVM with RBF kernel is 95.50% and the highest value was achieved by using all features without handling the outliers and it is 97.70%.