AGH University of Science and Technology

Ioannis Manousaridis
Anastasios Tzelepakis

Project 1 - Orange Project

In this project two multivariate datasets for classification were used from UCI repository, the wine[1] and the Breast Cancer Wisconsin (Original)[2] datasets[3][4].

In order to find the best features in both datasets, the distributions and the “Informative Projections” of the scatter plot tool were used.

In figure 1 on the left upper side it is the scatter plot of the wine dataset for the best features with the outliers and on the right upper side it is the scatter plot without the outliers. On the down left side it is the scatter plot of the Breast Cancer Wisconsin (Original) dataset for the best features with the outliers and on the right upper side it is the scatter plot without the outliers. From this figure, the best features in each occasion with the help of the “Informative Projections” are visible.

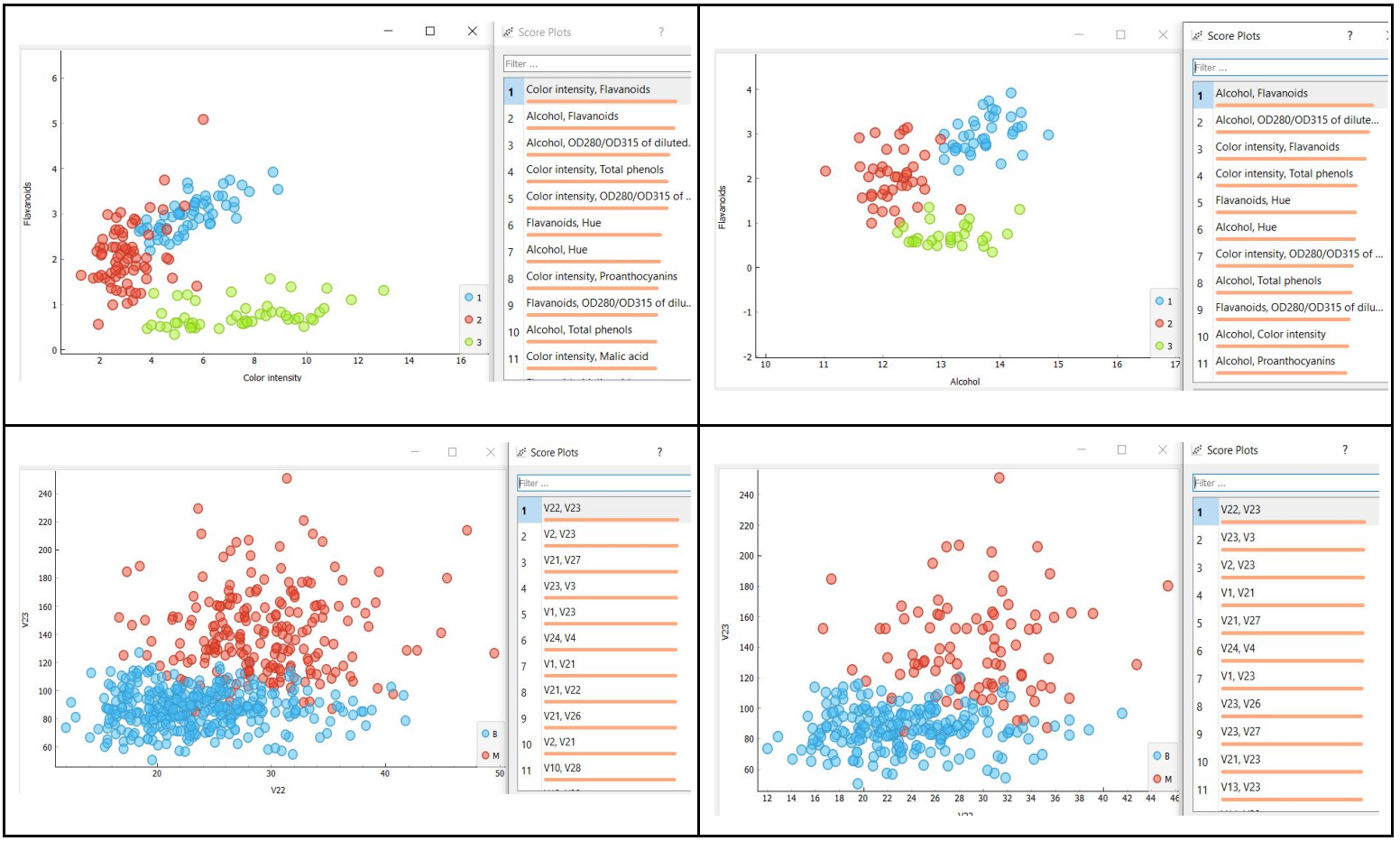

Figure 1: The scatter plots of the two datasets with outliers on the left side and without on the right.

Wine

With outliers

The best features with the outliers are the color intensity and the flavonoids. The K-NN classifier for 5 neighbours for all features and with euclidean distance produced the results which can be seen on the upper left side of figure 2.

[1] More info here: https://archive.ics.uci.edu/ml/datasets/Wine
[2] More info here:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
[3] Wine in csv format: https://drive.google.com/open?id=1zCHtkgRQg0n6IkRoz0cOg2yigcQX8NQa
[4] Wdbc in csv format: https://drive.google.com/open?id=1M3J6nERbk-VHYcRQJ9wpL1J-sfoYh4cc

Without outliers

The best features without the outliers are the alcohol and the flavonoids. The K-NN classifier for 5 neighbours for the best features and with euclidean distance produced the results which can be seen on the upper right side of figure 2.

Breast Cancer Wisconsin (Original)

With outliers

The best features with the outliers are the V22 and the V23. The K-NN classifier for 4 neighbours for all features and with euclidean distance produced the results which can be seen on the down left side of figure 2.

Without outliers

The best features without the outliers are the V22 and the V23. The K-NN classifier for 4 neighbors for the best features and with euclidean distance produced the results which can be seen on the down right side of figure 2.

Model	AUC	CA	F1	Precision	Recall
kNN	0.878	0.719	0.720	0.721	0.719

Model	AUC	CA	F1	Precision	Recall
kNN	0.985	0.963	0.963	0.963	0.963

Model	AUC	CA	F1	Precision	Recall
kNN	0.963	0.928	0.928	0.928	0.928

Model	AUC	CA	F1	Precision	Recall
kNN	0.973	0.941	0.941	0.941	0.941

Figure 2: The results of the 2 datasets with the kNN classifier

Conclusions

Wine

The selection of the best features in combination with the elimination of the outliers lead to a high improvement of all the metrics which were taken.

Breast Cancer Wisconsin (Original)

The original data of all the features included the outlier had high metrics. The selection of the best features in combination with the elimination of the outliers improved slightly the metrics. This issue occurs because the original dataset with the outliers has very good features already and there not many bad outliers.