Concerning the preprocessing of data, due to having a clean dataset it didn’t require to much transformation of the dataset. Therefore, to improve the performance of the models, specially when dealing with the linear models, the data was scale using the min max method. Furthermore, as result of the analysis regarding the number of hits and flops that each artist had on the dataset, it was created the feature \*’hit\_artist’\*. Additionally, it was proceeded with the binning using a decision tree to find the optimal splitting points for each feature. Finally, it was done the process of feature selection.

Regarding the results of the linear models Logistic Regression and SVC, respectively recorded a test accuracy score of approximating 0.74 and 0.81. Therefore, SVC did perform better than the Logistic Regression. It is argued that this outcome is due to the data isn’t linearly separated and with the SVC it is possible to use the kernel functions. Allowing to project the data points into a hyperspace enabling the separation. Nonetheless, for the SVC it was use grid search to find the optimal value for the hyperparameters C and gamma. This revealed to be a slower fine tunning process, due to the nature of the SVC and how the grid search is performed.

In respect to the multi-classifier of the \*’DISCRETE\_VALENCE’\*, there was the problem of class imbalanced. Thus, it was selected 4 types of resampling methods and experimented on the models developed. Across the linear models used, the TomekLinks was what achieved a better accuracy training score. Therefore, that was the selected resampling strategy to be applied to fix the imbalanced dataset. Nonetheless, the scores of the multi-classifiers were lower when comparing with the scores achieved on the binary classifier