PCS5024 - Statistical Learning
Classification with the Adult dataset

Vinicius Bueno de Moraes — NUSP 10256432

June, 2022

1 Introduction

The work presented here aims to show the results obtained from the analysis carried out on an available database, which can be downloaded here. The entire script developed to analyze her can be found here, it was adapted from Priyanka Sharma's development, available here. In addition, other references were used and follow at the end of this one.

2 Analyze features, missing data, and overall characteristics of the dataset. Select features, discretize numerical features, handle missing data.

To start analyzing the data, one of the first things to do was to index the adults.csv database classes sequentially. Afterwards, in order to deal with the missing data in the Dataset, the value that appeared the most in each category (mode) was placed in these empty positions, thus preventing empty fields from appearing in the database, thus achieving more reliable variables for training and testing of the classifiers applied later in this work. Moreover, the data presented in the Dataset was coded and normalized with the StandardScaler function from the Python sklearn library. Then we analyzed how the variables were related to each other, in order to have a brief intuition of how well two different classes could be separated (with Dataset converted to numerical values). For this, the correlation between them was used, as shown in the figure below:

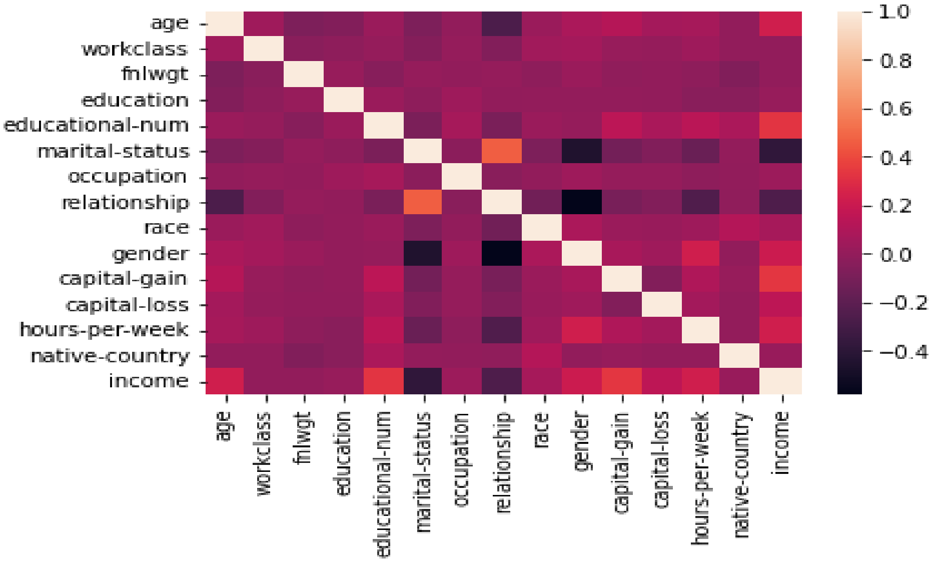
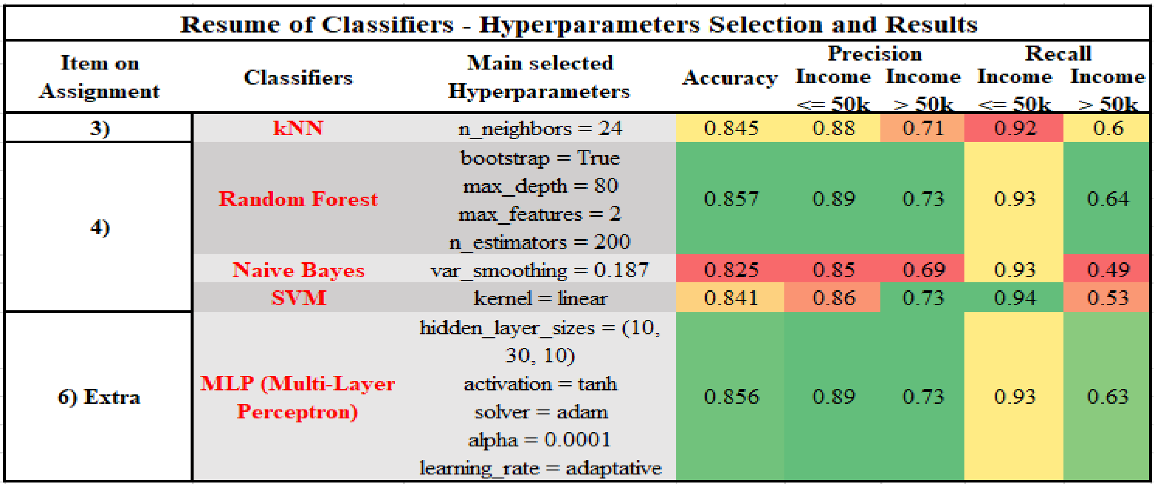


Figure 1: Correlation between all Features of Dataset.

Considering the previous correlation matrix, we can analyze how big or small the correlation between the classes is. With this in mind, in the next section, the classification with different methods to predict if a person earns more than 50,000.00 dollars per year or not, based on the data in the table is done - if you chose to use the complete Dataset for such analysis, regardless of the correlation of the categories with the object of study, income.

3 Analysis and comparison of results between the Classifiers

After executing the classification introduced in the previous item, with some methods, a table is shown below that summarizes Accuracy, Precision, Recall and the main Hyperparameters chosen for each model like this the respective values assigned to them. These values are selected within a range of possible simulated values, using the GridSearchCV function of the sklearn library.



Resume of Classifiers - Hyperparameters Selection and Results

Item on Assignment	Classifiers	Main selected Hyperparameters	Accuracy	Precision		Recall	
				Income <= 50k	Income > 50k	Income <= 50k	Income > 50k
3)	kNN	n_neighbors = 24	0.845	0.88	0.71	0.92	0.6
4)	Random Forest	bootstrap = True max_depth = 80 max_features = 2 n_estimators = 200	0.857	0.89	0.73	0.93	0.64
	Naive Bayes	var_smoothing = 0.187	0.825	0.85	0.69	0.93	0.49
	SVM	kernel = linear	0.841	0.86	0.73	0.94	0.53
6) Extra	MLP (Multi-Layer Perceptron)	hidden_layer_sizes = (10, 30, 10) activation = tanh solver = adam alpha = 0.0001 learning_rate = adaptative	0.856	0.89	0.73	0.93	0.63

Figure 2: Table with the consolidation of the results of the classifiers for items 3 (kNN), 4 (Random Forest, Naive Bayes, SVM) and extra 6 (Multi-Layer Perceptron).

3.1 Analysis and conclusions

With the previous table in hand, and saying that all classifiers had a training data set applied to themselves and then another part of the sample being the test itself, with its best hyperparameters and 5-fold cross-validation (when applied), we arrived at The conclusion is that MLP showed the best results for this dataset in question, not guaranteeing the same behavior for a new database to be analyzed, as stated in the theorem about classifiers. In general terms, the results were similar among them, having a small advantage for one or the other depending on the object of investigation (accuracy, precision, recall).

4 References

To create the Python script that provides the above analysis, these sources of information were useful:

https://www.kaggle.com/datasets/wenruliu/adult-income-dataset
https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/
https://www.springboard.com/blog/data-analytics/naive-bayes-classification/
https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn