--- Assignment 2 ---

Data Classification

Created by:

1. Armando Jacquis Federal 1301183626
2. Jonas de Deus Guterres 1301183615
3. Inacio Campos 1301183625

BACHELOR OF INFORMATICS

SCHOOL OF COMPUTING

TELKOM UNIVERSITY

BANDUNG

2021

Data Classification

After the first assignment about Data Pre-processing, where it involved removing the noise and treatment of missing values, this second assignment is focusing on data analysis. Moreover, this assignment is analysing Classification process. The classification models predict categorical class labels from a given dataset. That means to accurately predict the target class for each case in the data.

In this assignment, we collect two datasets: *Census Income* and *Labor Relations*, which have been cleaned from the previous assignment. To perform the Classification, we selected Naïve Bayes algorithm on those two datasets. The reason of choosing Naïve Bayes Algorithm is that it is intuitive and interpretable. That algorithm also is easily scalable to large numbers of predictor variables. It also works well with small datasets. Plus, it can handle both continuous data (Census Income data set) and discrete data (Labor Relations data set).

Note that for getting a better accuracy, it is needed to perform a good preprocessing of data and selecting the most efficient of training set and test set. In this assignment, to get the best model, we performed k-fold cross validation. The k subsets are used as test set and k-1 subsets is used as training set. Then, we selected the best result from the k-fold cross to validate the best classification.

1. First Dataset (Census Income data set)

To process the Classification for this dataset, we use 7-fold cross validation. It means we divided all records into seven subsets: six subsets are for training and one subset is for testing, with one subset is constituted by 14% of all records. This following is the output of the Classification with the percentage of Precision, Recall and the accuracy:

```
==============  1  FOLD ==================
True Positive : 2875  False Positive:  936  True Negative:  0  False Negative:  0
precision:  75.44 % Recall:  100.0 %
The accuracy is  75.44 %
==============  2  FOLD ==================
True Positive : 926  False Positive:  2885  True Negative:  0  False Negative:  0
precision:  24.3 % Recall:  100.0 %
The accuracy is  24.3 %
==============  3  FOLD ==================
True Positive : 955  False Positive:  2856  True Negative:  0  False Negative:  0
precision:  25.06 % Recall:  100.0 %
The accuracy is  25.06 %
==============  4  FOLD ==================
True Positive : 941  False Positive:  2870  True Negative:  0  False Negative:  0
precision:  24.69 % Recall:  100.0 %
The accuracy is  24.69 %
```

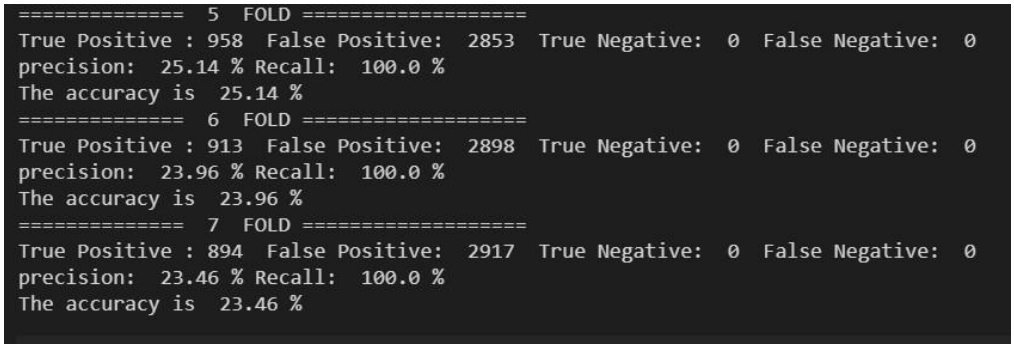

Fig 1. Result of 7-fold cross validation

Regarding to the above figure, the first fold gives the highest percentage value of Accuracy, which is 75,44%. It means the first fold is the best model to predict categorical class labels for Census Income dataset.

2. Second Dataset (Labor Relations dataset)

To process the Classification for this dataset, we use 5-fold cross validation. It means we divided all records into five subsets: four subsets are for training and one subset is for testing, with one subset is constituted by 20% of all records. The below is the output of the Classification with the percentage of Precision, Recall and the Accuracy:

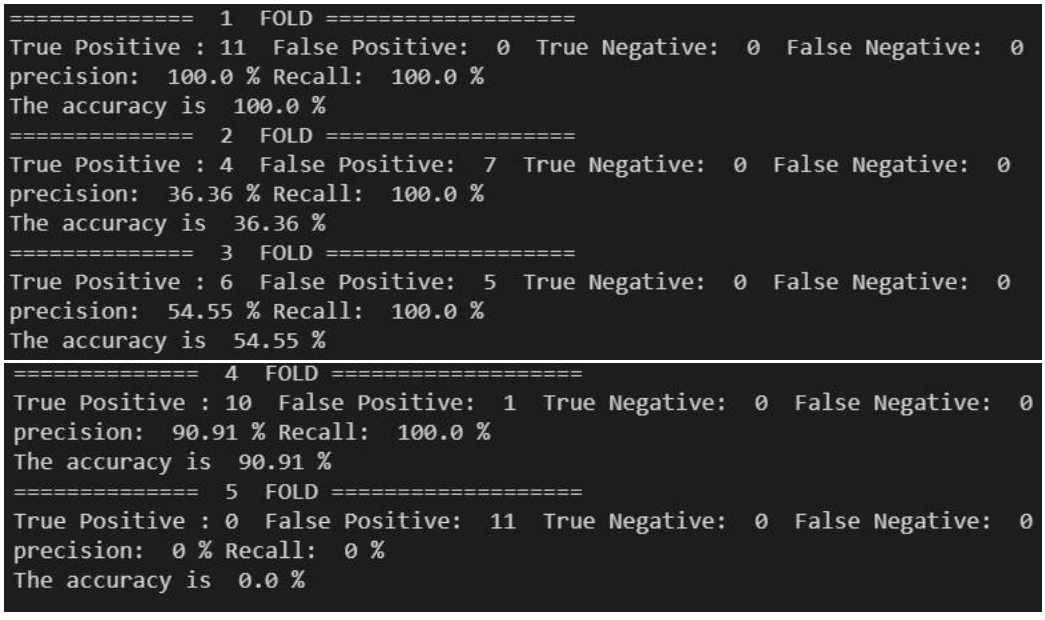

Fig2. Result of 5-fold cross validation

The above figure shows that from the 5 folds, the first one gives the highest percentage of Accuracy (100%). That means the first fold is the best model to predict categorical class labels for Census Income dataset.