# Results Model Adult Set

The train and test datasets are quite well distributed, as it is possible to see in summary table below:

**Train**

| | Numeric Columns | | | | | | String Columns | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| idx | age | fnlwgt | education-num | capital-gain | capital-loss | hours-per-week | idx | workclass | education | marital-status | occupation | relationship | race | sex | native-country | Target |
| nulls | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | nulls | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| count | 32,561.00 | 32,561.00 | 32,561.00 | 32,561.00 | 32,561.00 | 32,561.00 | count | 30725 | 32561 | 32561 | 30718 | 32561 | 32561 | 32561 | 31978 | 32561 |
| mean | 38.58 | 189,778.37 | 10.08 | 1,077.65 | 87.30 | 40.44 | unique | 8 | 16 | 7 | 14 | 6 | 5 | 2 | 41 | 2 |
| std | 13.64 | 105,549.98 | 2.57 | 7,385.29 | 402.96 | 12.35 | top | PRIVATE | HS-GRAD | MARRIED-CIV-SPOUSE | PROF-SPECIALTY | HUSBAND | WHITE | MALE | UNITED-STATES | <=50K |
| min | 17.00 | 12,285.00 | 1.00 | 0.00 | 0.00 | 1.00 | freq | 22696 | 10501 | 14976 | 4140 | 13193 | 27816 | 21790 | 29170 | 24720 |
| 25% | 28.00 | 117,827.00 | 9.00 | 0.00 | 0.00 | 40.00 | | | | | | | | | | |
| 50% | 37.00 | 178,356.00 | 10.00 | 0.00 | 0.00 | 40.00 | | | | | | | | | | |
| 75% | 48.00 | 237,051.00 | 12.00 | 0.00 | 0.00 | 45.00 | | | | | | | | | | |
| max | 90.00 | 1,484,705.00 | 16.00 | 99,999.00 | 4,356.00 | 99.00 | | | | | | | | | | |

**Test**

| | Numeric Columns | | | | | | String Columns | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| idx | age | fnlwgt | education-num | capital-gain | capital-loss | hours-per-week | idx | workclass | education | marital-status | occupation | relationship | race | sex | native-country | Target |
| nulls | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | nulls | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| count | 16,281.00 | 16,281.00 | 16,281.00 | 16,281.00 | 16,281.00 | 16,281.00 | count | 15318 | 16281 | 16281 | 15315 | 16281 | 16281 | 16281 | 16007 | 16281 |
| mean | 38.77 | 189,435.68 | 10.07 | 1,081.91 | 87.90 | 40.39 | unique | 8 | 16 | 7 | 14 | 6 | 5 | 2 | 40 | 2 |
| std | 13.85 | 105,714.91 | 2.57 | 7,583.94 | 403.11 | 12.48 | top | PRIVATE | HS-GRAD | MARRIED-CIV-SPOUSE | PROF-SPECIALTY | HUSBAND | WHITE | MALE | UNITED-STATES | <=50K |
| min | 17.00 | 13,492.00 | 1.00 | 0.00 | 0.00 | 1.00 | freq | 11210 | 5283 | 7403 | 2032 | 6523 | 13946 | 10860 | 14662 | 12435 |
| 25% | 28.00 | 116,736.00 | 9.00 | 0.00 | 0.00 | 40.00 | | | | | | | | | | |
| 50% | 37.00 | 177,831.00 | 10.00 | 0.00 | 0.00 | 40.00 | | | | | | | | | | |
| 75% | 48.00 | 238,384.00 | 12.00 | 0.00 | 0.00 | 45.00 | | | | | | | | | | |
| max | 90.00 | 1,490,400.00 | 16.00 | 99,999.00 | 3,770.00 | 99.00 | | | | | | | | | | |

Data Cleansing made was:

1. Encoding the labels of target variable (>50K and <=50K). To achieve that, it was necessary to normalize the strings in all columns, including the target, to get rid of punctuations and any special character.
2. There are duplicates inside both data sets, however as it belongs to US Census, it is possible, even with a small odd, to be two or more individuals with the same features. Those duplicates have been left in data sets.
3. Null values (or “?” as in original data set) have been found in columns:
    a. *workclass* (1836 entries for Train data set and 963 entries for Test data set)
    b. *occupation* (1843 entries for Train data set and 966 entries for Test data set)
    c. *native-country* (583 entries for Train data set and 274 entries for Test data set)

    These nulls have been deleted from both data sets to lessen the variability of variables and help the model better perform. Since they only represent 5% of the entire dataset, it have not missed much information with this deletion.

4. Some aggregations have been made in columns *workclass*, *education* and *marital-status*. These aggregations had the goal to unify the minority and similar classes present in each columns in order to decrease the posterior size of matrix (after string codification with One-Hot).
5. To remove collinearity existent in features (it is possible to see that also in Correlations/Association matrix), it was chosen to remove the columns *fnlwgt*, *education-num*, *occupation* and *relationship* over to *education*, *workclass* and *marital-status*.
6. The target variable is imbalanced (approximately 24% of train and test set belongs to “>50K” class). This imbalance is not extreme, so it hasn’t been necessary any complex treatment to deal with it. Although some metrics like precision, accuracy and recall could be distorted.
7. For preprocess, first have been applied One-Hot encoder algorithm to encode each class of each column in a new column it 1 or 0, indicating if each entry belongs to the class or not.

Then the *MinMax scaler* was applied to numerical features, preventing any problem of different scales.

8. Also, the columns *capital-gain and capital-loss* were removed, since they are mostly composed of zeros.

After these preprocessing steps, the train set have been modeled in three different algorithms (KNearestNeighbors, RandomForestClassifier and Multi-layer Perceptron) using cross-validation, with 3 folds, to tune the hyperparameters and then tested on test set.

After fitting, tunning and verifying each predictor, have found the best threshold in Receiver Operating Curve, then the metrics were also measured with that threshold up to date.

Those algorithms were implemented in Python 3.7, in Google Colab platform, using python libraries as pandas, numpy, scikit learn and tensorflow. The results are shown below:

| | KNearestNeighboors | | | RandomForest | | | Multi-Layer Perceptron | | |
|---|---|---|---|---|---|---|---|---|---|
| idx | Train | Test (threshold=0.5) | Test (best threshold) | Train2 | Test (threshold=0.5)3 | Test (best threshold)4 | Train5 | Test (threshold=0.5)6 | Test (best threshold)7 |
| ROC AUC | 0.79 | 0.71 | 0.76 | 0.69 | 0.69 | 0.80 | 0.74 | 0.74 | 0.80 |
| Accuracy | 0.72 | 0.79 | 0.77 | 0.82 | 0.82 | 0.79 | 0.83 | 0.82 | 0.78 |
| Precision | 0.85 | 0.59 | 0.52 | 0.73 | 0.72 | 0.55 | 0.69 | 0.67 | 0.53 |
| Recall | 0.80 | 0.55 | 0.74 | 0.44 | 0.44 | 0.81 | 0.57 | 0.56 | 0.82 |

The parameters of each predictor are:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                     weights='uniform')

RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=10, max_features='auto',
                       max_leaf_nodes=72, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=200,
                       min_weight_fraction_leaf=0.0, n_estimators=50,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)

Multi-Layer Perceptron(first layer = 40 neurons
                       Drop-out = 30%
                       Activation function = ReLu,
                       Second layer = 20 neurons
                       Drop-out = 30%
                       Activation function = ReLu,
                       Third layer = 1 neurons
                       Activation function = Sigmoid
                       )
```