Classification: K-nearest-Neighbors, Naïve Bayes

The data to be analysed includes information from a telecommunications operator. A problem in this business is customers who leave the operator for a competitor (customer churn), since the cost of attracting a new customer is much higher than the cost of retaining customers.

The data set to be analysed includes some demographic data of customers as well as information about the contract established.

1. Start by loading the data (”Churn_DataSet.csv”).
2. Remove duplicate lines from the data set, if any.
3. Check if there are missing values in the attributes and if there is proper processing of the missing values.

4. Graphically explore the attributes according to the goal attribute: churn, through boxplots and histograms.
5. Make the selection of attributes correlated with the target attribute by applying the appropriate tests (ANOVA and Chi-Square).

6. The K-nearest-Neighbours is an algorithm that relies on the calculation of distances between the values of the attributes, so some transformations to the data are needed:
   a. Convert the binary nominal variable to numeric using 1/0 mapping, and the nominal variable to numeric using *get_dummies*().
   b. Check if there are numerical attributes with different value ranges and normalise their values.

7. Split the data into training and testing samples (80% / 20%), making a stratified division, i.e. with the original distribution of the target attribute (Churn: Yes/No) in both sets.

KNeighbors Classifier

8. Train the KNeighbors classifier using its default parameters on the train set and get the confusion matrix and the precision, recall, and F1-score measurements for both classes (Churn: Yes/No).
9. Train the KNeighbors classifier for several uneven K values in the range [1..151] using cross-validation (cv = 5) with the train set and save for each K value the average recall of the 5 cross-validation iterations.
10. Plot the recall against K and determine the best value of K.

Naïve Bayes Classifier

11. Train the GaussianNB classifier using cross-validation (CV = 5) with the train set and determine the average and standard deviation of the recall for the various iterations.

12. Train the KNeighbors algorithm with the best K and the GaussianNB algorithm using the train set and get the final performance of both models using the test set.

13. Visualise the ROC curves of both models on the same chart and get the AUC of each model.

14. Which of the models is more suitable for the churn forecast?