5.1. Who should get a bank credit? – let’s get serious

The credit data set was already introduced in the last exercise. In the following we will use the outcome of the previous exercise as baseline. Hint: You can easily reuse parts of your process and simply replace classifiers and/or pre-processors.

1. Apply a Naïve Bayes classifier to the data set using X-Validation (10-fold). How can you interpret the precision and recall of the “bad customer” class? Try to improve the situation by increasing the number of “bad customers” in the training set. To increase the number of “bad customer” examples, first filter all bad customers from the data set and then append these examples to the original set. How does precision and recall change if you apply this procedure twice?

2. To model a use case specific evaluation, as observed in the previous example, replace the Performance (Classification) operator by the Performance (Costs) operator. Set up your cost matrix by assuming that you will lose 1 Unit if you refuse a credit to a good customer, but that you lose 100 Units if you give a bad customer a credit. Rerun the experiments from 1 and evaluate the results.

3. As Naïve Bayes classifiers might not be the best possible solution, you should now try to find a good/better classifier for the data set. To do so, use the Compare ROCs operator. What classification algorithm would you prefer? Start with the classifiers Decision Trees, k-NN and Naïve Bayes.

4. As the creation of training data is mostly a manual task as humans tend to be fallible training data might include noise. Simulate this behavior by using the Add Noise operator and change the parameter “label noise” from 0% to 10% to 20%. Is the Decision Tree still the most promising alternative? How does the performance of the other classifiers evolve?
4.2. Open Competition: Finding rich Americans

The Adult data set from the UCI data set library (http://archive.ics.uci.edu/ml/datasets/Adult) describes 48842 persons from the 1994 US Census. The data set is provided as adult.arff file in ILIAS.

Your task is to find a good classifier for determining whether a person earns over 50.000 $ a year. Beside of being accurate, your classifier should also have balanced precision and recall.
To evaluate your classifiers use split validation (split ratio=0.8, linear sampling).

In order to find the best classifier, you may experiment with:
1. different algorithms
2. different parameter settings
3. the balance of the two classes in the data set
4. the set of attributes that are used or not used
5. other preprocessing techniques

People are described by the following 14 attributes:

- **age**: continuous.
- **fnlwgt**: continuous.
- **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num**: continuous.
- **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Han-

**relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

**race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

**sex**: Female, Male.

**capital-gain**: continuous.

**capital-loss**: continuous.

**hours-per-week**: continuous.

**native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.

In order to increase your understanding of the data set, you might want to visualize different attributes or attribute combinations. All are distributed equally over the two classes.