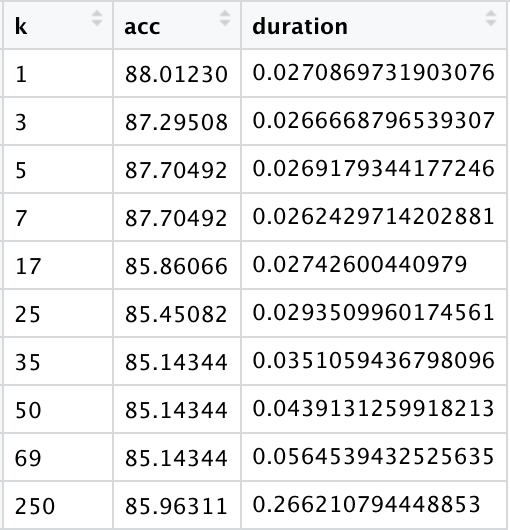
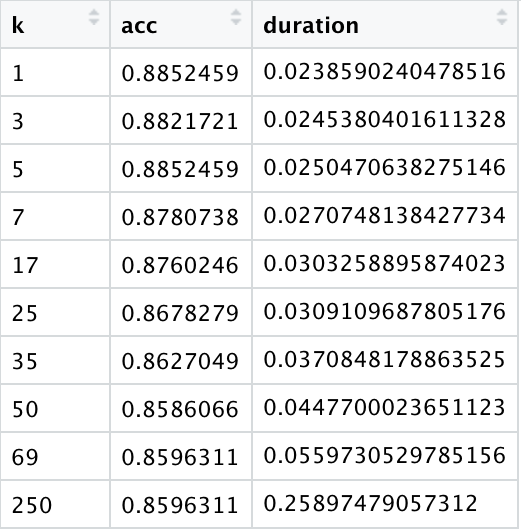
**Machine Learning – Exercise 1**

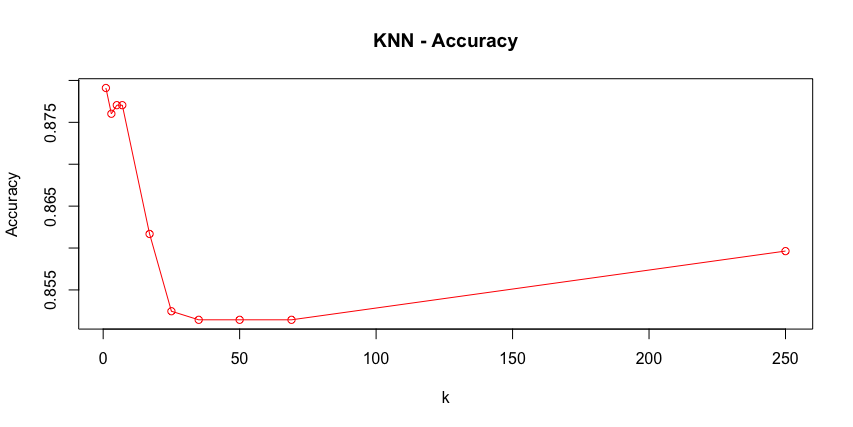
**Classification**

**Group 7**

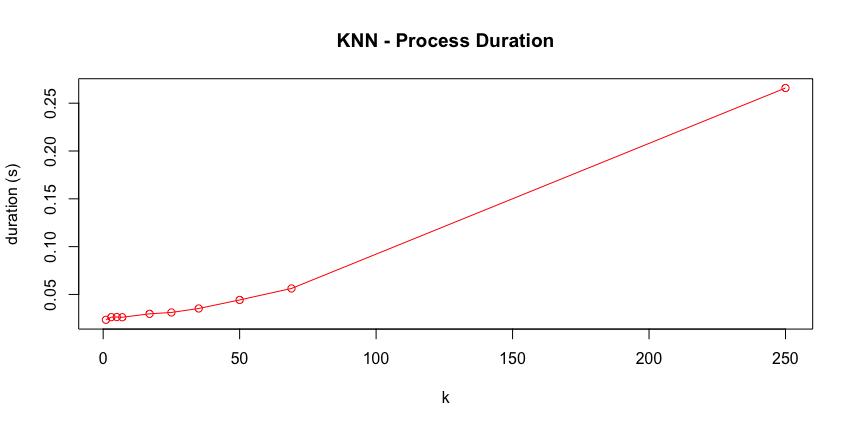
1. **Classification using K-NN**
   1. **What is K-NN?**  
      The k-nearest neighbors algorithm is a non-parametric method used for classification and regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors with the object assigned to the class most common among its k nearest neighbors. K is a positive number, typically small, in order to reduce the number of calculations, and thus the process duration. As the experiments will show, the accuracy of the algorithm is not increasing with the increasing of k. If k=1, then the class of the object would the class of the single neares neighbour.
   2. **Classification on small datasets**  
      1. **Crime-Mapping dataset**  
         The evaluation of the K-NN classifier on this small datasets is made by using the algorithm to predict if there is an official hospital record about the crime, i.e., if the variable phxcommunity should have value of Yes or No. Since the dataset contains also nominal categorical variables, the transformation of categorical data to numerical is needed. The experiment is made using multiple values of parameter k. In order to get most accurate results, the numeric variables need to be normalized.  
         The evaluation of the accuracy of the algorithm, the presumption, that the greater k does not imply better results, is confirmed. As the figure states, there is no connection between accuracy and k. For example, the algorithm showed better results with k=17 as by using the k=25, but the classification using k=25 made less correctpredictions as by using k=250.

**Figure 1.1 - Tables showing the percentage of correct predictions (acc) and process duration in seconds for each chosen value of k – with/without normalization of variables**

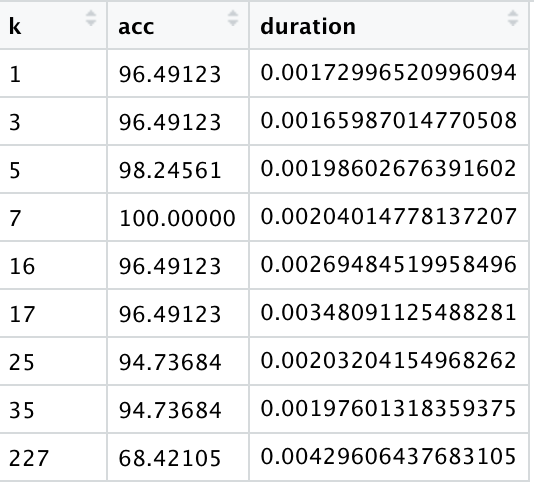
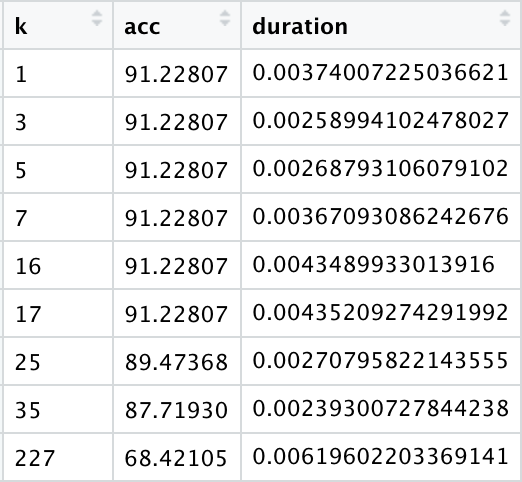


**Figure 1.2 - KNN accuracy on Crime-Mapping dataset**

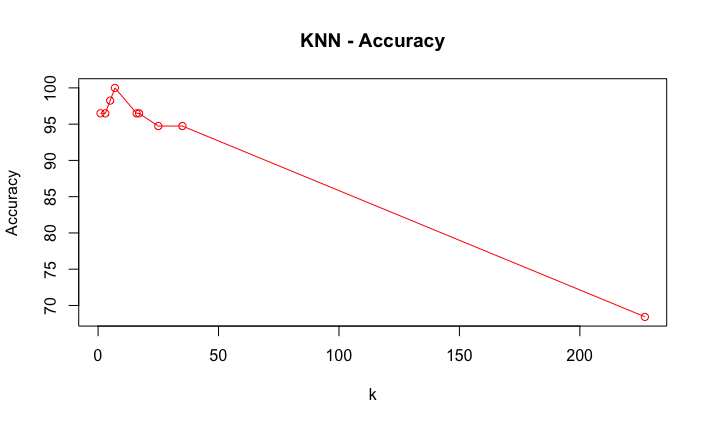


**Figure 1.3 - KNN - Process duration**

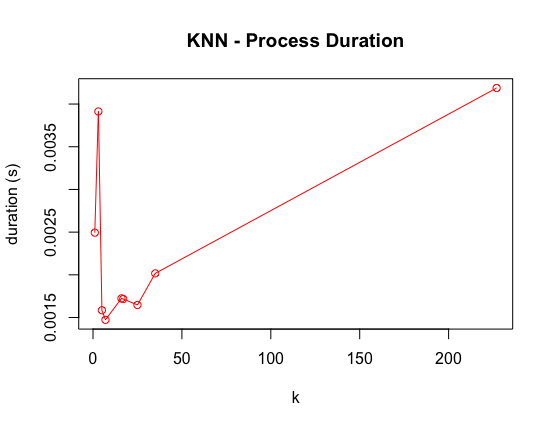
* + 1. **Breast-Cancer dataset**  
       The evaluation of the K-NN classifier on this small datasets is made by using the algorithm to predict if the cancer is benign or malignant, i.e., if the variable class should have value of B or M. The experiment is made using multiple values of parameter k. In order to get most accurate results, the numeric variables need to be normalized.  
       The evaluation of the accuracy of the algorithm, the presumption, that the greater k does not imply better results, is again confirmed. The classification using k=7 on the normalized dataset results with 100% accuracy.  
       As the diagramm on figure 2.3 states, the process duration increases with big values of k. However, it is not the statement, as the process duration of the classification using the k=5 is less then when using k=7.

**Figure 2.1 - Tables showing the relation correct predictions (acc) and process duration in seconds for each chosen value of k – with/without normalization of variables**

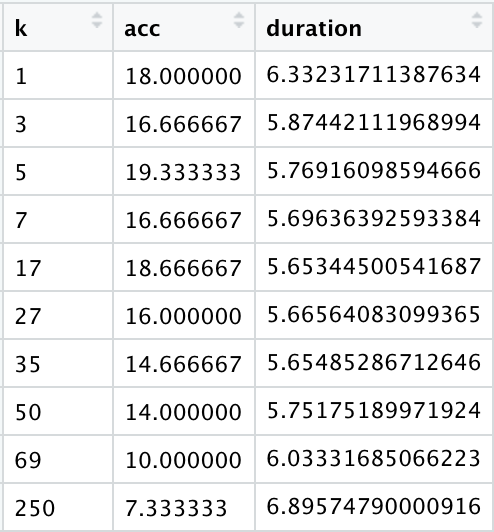
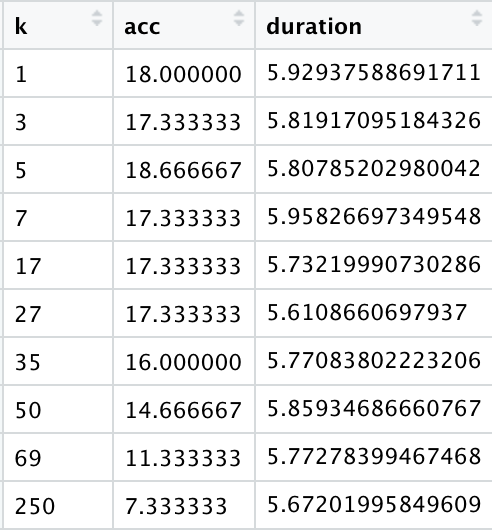


**Figure 2.2 - KNN accuracy on Breast-Cancer dataset**

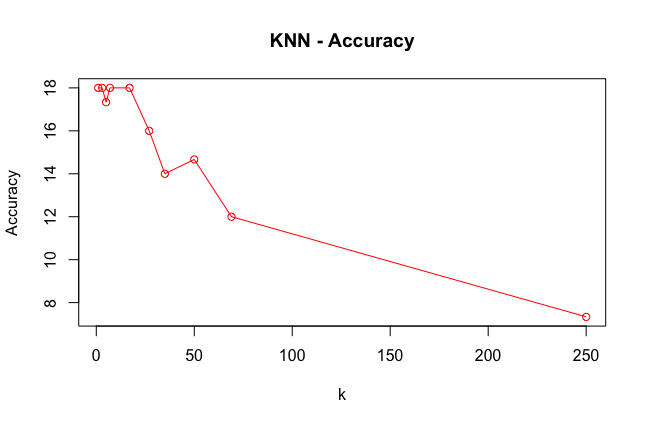


**Figure 2.3 - KNN - Process duration**

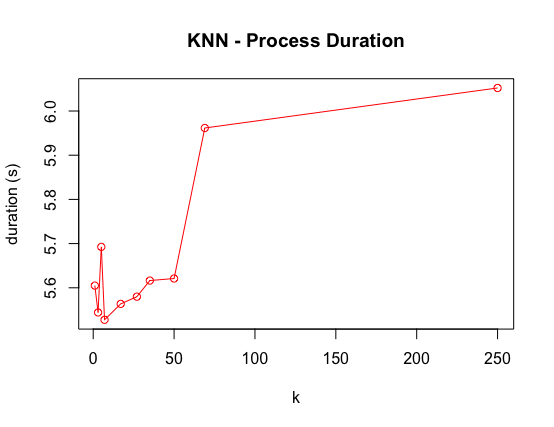
* 1. **Classification on large datasets**
     1. **Amazon reviews**  
          
        The evaluation of the K-NN classifier on this large dataset is made by using the algorithm to predict which user hase made the review. The experiment is made using multiple values of parameter k. In order to get most accurate results, the numeric variables need to be normalized.  
        The presumption, that the greater k does not imply better results, is again confirmed. However, due to huge amount of classes (50), the algorithm shows really bad results (see figure 3.2)   
        As the diagramm on figure 3.3 states, the process duration increases with big values of k.   
        The normalization of the data showed no significat impact in this example.

**Figure 3.1 - Tables showing the relation correct predictions (acc) and process duration in seconds for each chosen value of k – with/without normalization of variables**



**Figure 3.2 - KNN accuracy on Amazon-Reviews dataset**



**Figure 3.3 - KNN - Process duration**

* + 1. **Synthetic Financial Datasets For Fraud Detection**  
         
       The evaluation of the K-NN classifier on this large dataset is made by using the algorithm to predict if a financial transaction is a fraud or not . The experiment is made using multiple values of parameter k. In order to get most accurate results, the numeric variables need to be normalized. However, since the dataset is pretty large, the K-NN algorithm did not manage to process the classification using k=1 in 12h.