Data Exploration and Preparation

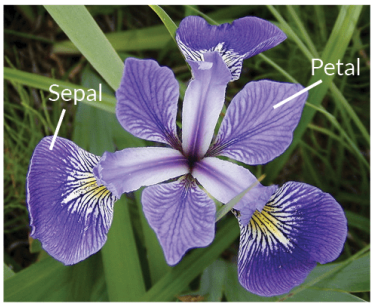
Exercise Sheet #1

Exercise 1.0: Install RapidMiner Studio

Go to https://rapidminer.com/educational-program/, sign up with your TU Darmstadt mail address for the RapidMiner Educational License Program, and download RapidMiner Studio.

As soon as you have installed RapidMiner Studio, sign in using your just created RapidMiner account which should unlock the educational version of RapidMiner. Now you are ready to explore RapidMiner Studio!

Exercise 1.1: *Iris Flowers*



Iris Versicolor

Iris Setosa

Iris Virginica

Imagine you are a biologist working at a national research institute for botany. As part of your last field trip, your research group was able to discover three new flower species which you decided to name:

- "Iris setosa",
- "Iris virginica", and
- "Iris versicolor".

In order to understand the characteristics that distinguish the species from another, your team gathered data samples of a set of representative flowers by measuring four attributes:

- sepal width,
- sepal length,
- petal width, and
- petal length.

As you are specialized in data science, your team has asked you to explore the data in order to derive decisive characteristics for each of the three species based on the measured attributes.

⇨ Please explore the iris dataset visually in RapidMiner Studio and define value ranges for some attribute combination that can be used to distinguish the different species[1].

[1] You can find the iris dataset in RapidMiner here: *Samples/data/Iris*.

Exercise 1.2: Restaurant Orders

After your biologist career did not really work out for you, you start to work as a business analyst for the well-known restaurant chain "*Out-and-In*". Thus far, business decisions were made rather by intuition than on actual facts. Now, the chain wants you to improve their business understanding through analyzing their sales data. The managers are betting on you to improve their decisions.[2]
You start to explore their sales data and, first of all, focus on understanding the basics of their daily business:

a. Which was the day with the highest revenue?
b. Which items had the highest revenue in 2019?
c. To get a better overview, you add the sum of ordered items per order to the result set. You keep using the data extended with this extra feature in the following.
d. You noticed that for some orders, the same product is listed multiple times but the "total products" feature is filled with a 1 which is wrong based on the restaurant's process for creating this data. You exclude respective orders.
e. Afterwards, you create a new feature which indicates the full price for each order.
f. Then, you have a look at different scatter plots of the data. Can you identify any outliers? If yes, how could you exclude them?
g. Finally, you visualize the following KPIs to understand the performance of the current business:
    - Total Price per order
    - Average price per order
    - Total price per month and year

   Can you identify any trends?

Exercise 1.3: Titanic

Last night, you've watched the blockbuster movie "*Titanic*" which is based on a true dramatic incident. As you've noticed that certain people were able to survive while others died in the movie, your inner Sherlock starts to wonder: Were there certain groups of people that were more likely to survive?
After weeks of research, you finally stumble across a dataset that reports on all passengers of the Titanic incident. Driven by your spirit of discovery, you start to investigate the incident by analyzing the data.[3]

a. First of all, you are trying to comprehend the quality of the data. Is there any missing data in the dataset? If yes, how much?
b. After checking the data quality, you want to understand the age distribution of the passengers. How many people are 1-19, 20-45, 46-65 and >65 years old?
c. Furthermore, you want to understand how many people died and survived. Thus, you create a tabular output that indicates for each of the above age categories the actual number of people who died and who survived as well as the survival ratio of each age category. Please explore and explain the result.
d. The age categories already showed some interesting results, but you want to explore further possible features that may provide further insight into the survival determinants. Which features appear to explain the chances of survival best? Modify your operators to test changes in the survival rates.
e. What kind of model would you choose to predict the chances of a person's survival? Why?

---

[2] You can find the restaurant orders dataset in our git repository.
[3] You can find the titanic dataset in RapidMiner here: *Samples/data/Titanic*.

Exercise 1.4: *Customer Credit Risk Analysis*

Imagine you are working as a data scientist at the bank "*N42*". A few minutes ago, your manager came into your office and said:

"Yesterday, I tested the online service of one of our competitors and I couldn't believe it: they offer an AI for automatically granting credits…we need to offer this as well! I want you to provide me with an initial prototype tomorrow morning."

After taking a shocked zip from your coffee, you are starting to work on this unexpected task.

Without having a look at the available data:

a. Into which type of machine learning problem may this be translatable into?
b. What kind of data do you need to create such an AI?

You identified a suitable set of data which you were able to export as a CSV file named "*customer_credit.csv*" from your SAP system which is used for handling credits at N42. After importing it into your development system, you start to explore the data. You start to answer the following questions to develop a good data understanding[4]:

c. How many examples are contained in the dataset?
d. Does the dataset contain a feature which you could use as a label for training your AI? If yes, which one is it?
e. How many requests of *non-foreign workers* are contained in the data? How many are unknown?
f. You decide to exclude examples for which it is not known whether the person is a foreign worker or not. Filter your dataset accordingly. Keep using this filtered dataset to answer the remaining questions.
g. How many requests for granting a credit were made to buy a car? How many were successfully granted for buying a *used* car?
h. You decide to train an initial AI solution by conducting the following steps:
    i. First, you split your filtered data into a training set (containing 70% of the data) and a test set (containing 30% of the data) using *stratified sampling*. What does stratified sampling do?
    ii. Then, you train a *decision tree* based on the training data set. Why would be a good idea to apply an algorithm like a *decision tree* for this use case?
    iii. Finally, you *apply* your trained decision tree to your test data and *compute its performance*. Is your trained model a good model?
i. You decide to train the same AI that only focuses on credit requests for buying new cars. Is the resulting decision tree better or worse than the previous one?

[4] You can find the customer credit dataset in our git repository.