Data analysis and prediction with classification trees

A teaching concept for upper secondary computing education

Teacher version

Dr. Andreas Grillenberger
Freie Universität Berlin
Computing Education Research Group
Contact: `andreas.grillenberger@fu-berlin.de`

Created with support by *StRin RS Anne-Katrin Jäger*

Inhaltsverzeichnis

Introduction for the teacher

Goals

In class, students will be given the opportunity to see how today's ubiquitous correlation-based data analysis works. They try to gain information from a large mountain of data without having clear knowledge about the concrete way of analysis.

The aim is to give students a critical perspective on data analysis and to become aware of the limitations of this analysis.

The following learning objectives will therefore be pursued:

The students…

- use an example to explain the difference between causality and correlation in relation to data analysis.
- describe the process of a typical correlation-based data analysis (possibly with the aid of a diagram).
- describe the concept of a „classification tree“ and create one for given rules.
- use a „classification tree“' to create a prediction for a data record.
- assess analyses with regard to their quality on the basis of the misalignments that occur.
- perform simple correlation-based data analyses with a suitable tool on the computer itself.
- assess real and fictitious examples of correlation-based data analyses with regard to their benefits and risks.

About the material

All materials are licensed under Creative Commons *CC BY-NC-SA 4.0*[1] and may be redistributed under the terms of this License. For use in class and distribution to students, the naming required by the license may be explicitly waived. You can request the source files of this material from the author at any time. At the end of the concept, all the tasks mentioned below are also summarised in a student version in worksheets.

Overview

Time estimate: depending on the level of detail, approx. three to four double hours are estimated.

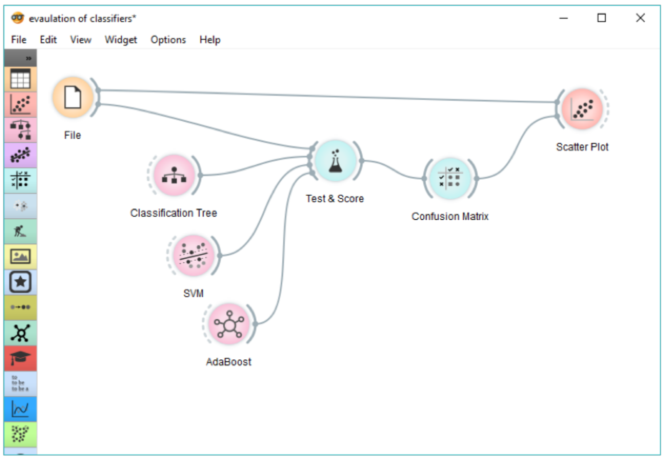
- getting insight into the use of data analysis using a real world example; attempt to explain this by discussing possible causal relationships
- conceptuality: causality vs. correlation as „explainable with common sense“' vs. „unexplainable, but appearing correct based on data“'; discussion of the associated danger due to misclassifications
- overview of the data analysis process and determination of where causality vs. correlation is central
- manual conduction of a simple correlation-based analysis to learn the principles
  - finding rules in the data set

[1] https://creativecommons.org/licenses/by-nc-sa/4.0/

  - representation the rules as a classification tree
  - using the classification tree to predict attributes of other data sets
  - discussion of quality
- conduction of an analysis at the computer to recognize the potential and limits when using larger datasets
  - establish an assumption as to which attributes of the data set are relevant for the determination of the attribute to be predicted
  - use of a data analysis tool to create a classification tree; verification of previous hypotheses about relevant attributes
  - generate an automated prediction based on the generated classification tree; initial assessment of results
  - using of confusion matrix to systematically detect errors in predictions; assessment of analysis quality
  - review of ways to improve the quality of analysis
- discussion of various fictitious and real examples of the use of data analysis for their usefulness and dangers

Data analysis tool used

The data analysis tool Orange3, developed at the University of Ljubljana, is used. This allows an easy access, because the data analysis is not done by text based programming, but in a kind of data flow diagram. This means that all analysis options are directly visible, while the visual orientation also provides a better overview of the analysis process.



Worksheet 1 (concept): Logic or just observation?

As introduction, it is suggested to use a real-world example interesting for the students that also raises various questions. For example the following story can be used:

> *Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: Ïf we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?"*
> *[...]*
> *As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy predictionßcore. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy. One Target employee I spoke to provided a hypothetical example. Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug. There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August. What's more, because of the data attached to her Guest ID number, Target knows how to trigger Jenny's habits. They know that if she receives a coupon via e-mail, it will most likely cue her to buy online. They know that if she receives an ad in the mail on Friday, she frequently uses it on a weekend trip to the store. And they know that if they reward her with a printed receipt that entitles her to a free cup of Starbucks coffee, she'll use it when she comes back again.*
> *[...]*
> *Pole applied his program to every regular female shopper in Target's national database and soon had a list of tens of thousands of women who were most likely pregnant. If they could entice those women or their husbands to visit Target and buy baby-related products, the company's cue-routine-reward calculators could kick in and start pushing them to buy groceries, bathing suits, toys and clothing, as well. When Pole shared his list with the marketers, he said, they were ecstatic. Soon, Pole was getting invited to meetings above his paygrade. Eventually his paygrade went up.*
> *[...]*
> *About a year after Pole created his pregnancy-prediction model, a man walked into a Target outside Minneapolis and demanded to see the manager. He was clutching coupons that had been sent to his daughter, and he was angry, according to an employee who participated in the conversation.*
> *"My daughter got this in the mail!"he said. SShe's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"*
> *The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.*
> *On the phone, though, the father was somewhat abashed. Ï had a talk with my daughter,"he said. Ït turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."*
> *— Charles Duhigg, The New York Times Magazine, February 16 2012*

Fulltext available at: `https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html`

This introduction raises the question of how Target was able to guess the corresponding products and how such analyses generally work. This can be discussed in class.

While in this example it can still be assumed that clever people have considered the criteria for recognising pregnancy and checked them against customer data, the following second example at the latest shows that this cannot always work in this way:

For students, e. g. on a worksheet

In class, you've already seen an article about how data is used in retail today to present advertisements tailored to customers. However, online shops are now going further and trying to deliver many products to their customers as quickly as possible:

> *Anticipatory shipping may be closest that retail can come to a crystal ball. Amazon, which now has a patent for the algorithm-based system, could conceivably use the system to ship products before you even place an order.*
> *Amazon filed for the patent, officially known as "method and system for anticipatory package shipping,ïn 2012, and it was awarded on Christmas Eve of the following year. The patent summary describes a method for shipping a package of one or more items "to the destination geographical area without completely specifying the delivery address at time of shipment,"with the final destination defined en route. [. . .] So just imagine a company the size of Amazon making accurate supply and shipping decisions before a customer's decisions are finalized.*
>
> *— Lance Ulanoff, Mashable.com, January 21 2014*

Fulltext available at: `https://mashable.com/2014/01/21/amazon-anticipatory-shipping-patent`

A further contextualization can be done by examples like Same-Day-Delivery, as offered by different online merchants.

The students can now be given a first task in which the goal should be to recognize that this kind of "predictionöf customer behavior can no longer be done purely on logical conclusions. The task could therefore be as follows:

Aufgabe 1

To find out what a customer might order next, retailers need to collect and analyze extensive data about their customers.

a) What do online retailers know about their customers? Where do they get the information from?

Information about the customer	**Source**
first name and surname	*registration*
address	*registration*
date of birth	*registration*
favorite articles	*orders*
friends	*shipping addresses*
locations	*shipping & IP addresses*
...	...

b) Probably not all the information that an online retailer has about its customers is also important when it wants to find out which item the customer could order next. In the table above, mark the rows you think are important for this purpose by making a $+$ next to the important rows.

The results cannot be checked for correctness, but they give the students the opportunity to think about the idea of data and to realise that there is actually very little data that seems really helpful. Only by looking at the total amount of data can conclusions be drawn from these seemingly relevant data. The aim is to discuss possible attributes. This is good, because it shows that no logical conclusions can be drawn and thus leads to correlative analyses. Correlative analyses can then be discussed on this basis. The results can be saved in the gap text:

Aufgabe 2

Fill in the following gap text: In data analysis, there are two ways we can make predictions:

1. If we already know something about the data to analyze, then we can explain how something works and draw conclusions from it. So there are logical connections, so called causal connections, which we can use for prediction.
   Example:
   `IF a customer has bought chips in the last 5 purchases, THEN he/she will probably also purchase some next time.`

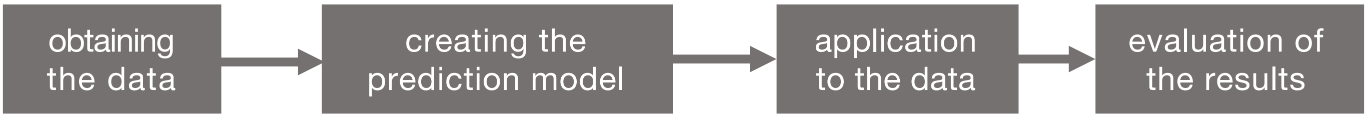
2. In other areas we do not recognize any logical connections. Instead, we can look for pattern in the data. These also provide us with information, but we often cannot explain them to ourselves. Such connections can be called correlative.
   Example:
   `IF a customer has bought item X and he/she lives in Y and is at least 35 years old, THEN he/she will also buy Z.`

causal connections help us to understand things, but they are often relatively uninteresting for data analysis: They are rather obvious and well-known, so that they produce only little new information. But we can logically explain that they are correct. Instead, correlative correlations are often more exciting as they give us new information. But they have the disadvantage that they are not necessarily logically comprehensible: How exactly place of residence and age shape purchasing behavior, we can usually not explain logically. Besides, we first have to find them, which is relatively difficult.

At the end of this lesson, the students have already gained a first insight into the goals and problems of data analysis. On this basis, a first manual data analysis is carried out in the next lesson.

Worksheet 2 (concept): Manual Data Analysis (part I)

On this basis, the process of data analysis can be discussed with the students and a model of the process can be presented, which can also serve as Advance Organizer and thus structures the further learning process. This can be worked out together on the blackboard, for example using index cards for the four sub-processes:

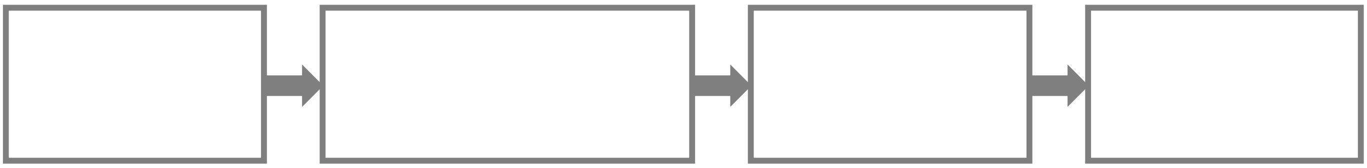


In order to gain a basic understanding of how the analyses work, a first sample dataset is provided, which is analysed manually in one step through the entire process. In the prediction model development phase, an attempt is made to address both causality and correlation-based data analysis and to highlight their differences.

Aufgabe 1

In class it has already been worked out how a data analysis works. Complete the following gap text and the following process model:

As a first step of data analysis, the data must be acquired/collected and stored. From this data, usually a small subset is chosen as a basis to create the forecast model, i.e. to find rules that allow the prediction of the desired properties. These rules can then be used to create forecasts/predictions. The last step of any data analysis should be the evaluation of the results in order to ensure the best possible quality of the results.



The first step for the students is to create a suitable prediction model, this is done in the following task:

Aufgabe 2

An online shop has collected various data about its customers and now wants to encourage its customers to make further purchases by sending individual vouchers. Therefore, he wants to find out which product category is particularly interesting for each customer.

The shop has already collected the following data about each of its customers: `age, marital status, number of children, preferred payment method, categories of the last four products purchased (film, sports, software, electronics, clothing, music, books or car accessoires).`

In order to send each customer a voucher that they are likely to redeem, the online shop wants to find out which category is particularly interesting for the buyer. Which if-then-rules do you suspect might help the online shop? *Note: of course you can combine several conditions with „and" e.g. „purchase1 = electronics and children = 0".*

- `IF XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX THEN the next purchase is from the category XXXXXXXXXXXXXXXXXXXXX`
- `IF XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX THEN the next purchase is from the category XXXXXXXXXXXXXXXXXXXXX`
- `IF XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX THEN the next purchase is from the category XXXXXXXXXXXXXXXXXXXXX`

Students may come to different rules, but the overall conclusion is that too little information is available to identify meaningful and valid causal relationships. The next task can therefore be the following:

Aufgabe 3

After the company's data scientists had recognized that there are no valid correlations in the information currently available, it was decided to try a different approach: The online shop asked some of its customers what was of interest to them next. The following data came out. What correlations do you see in the table below?

Beispiel:

IF purchase1 is a movie and has been paid with debit card, THEN the customer will buy items from the car category next.

In short: „purchase1“=„movie“ and „payedWith“=„debit“ ⇒ „nextPurchase“=„car“

age	purchase1	purchase2	purchase3	purchase4	married	children	payedWith	nextPurchase
25-50	movie	software	movie	sports	yes	1	debit	car
25-50	electronics	music	movie	software	no	1	VISA	books
<18	movie	electronics	sports	sports	no	0	debit	car
<18	movie	music	clothes	sports	no	0	debit	car
18-25	books	music	movie	household	no	1	debit	books
<18	books	movie	movie	books	no	0	VISA	books
25-50	movie	movie	sports	sports	yes	1	debit	car
25-50	music	movie	movie	toys	no	1	debit	books
25-50	music	music	movie	household	no	1	debit	books
25-50	electronics	music	books	software	yes	1	Mastercard	electronics
25-50	software	electronics	movie	toys	no	1	Mastercard	books
25-50	movie	movie	sports	sports	yes	0	Mastercard	electronics
25-50	music	electronics	books	electronics	yes	1	Mastercard	electronics

- IF XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX THEN the next purchase is from XXXXXXXXXXXXXXXXXXXXX
- IF XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX THEN the next purchase is from XXXXXXXXXXXXXXXXXXXXX
- IF XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX THEN the next purchase is from XXXXXXXXXXXXXXXXXXXXX

The students can manually read out rules from the data, which seem arbitrary. Therefore, these rules cannot be explained logically, as is often the case for correlation-based analyses.

At this point it can be discussed whether rules like „no child“ ⇒ „nextPurchase“ = „electronics“ should be included – from the data available this seems valid, but it can be assumed that a rule based on only one data set is not very valid.

To make these rules easier to use, they are usually visualized as a classification tree. In technical terms, this is a (not necessarily binary) decision tree whose leafs represent the decisions made. A task for converting the rules into a tree could look as follows:

Worksheet 3 (concept): Manual Data Analysis (part II)

Aufgabe 1

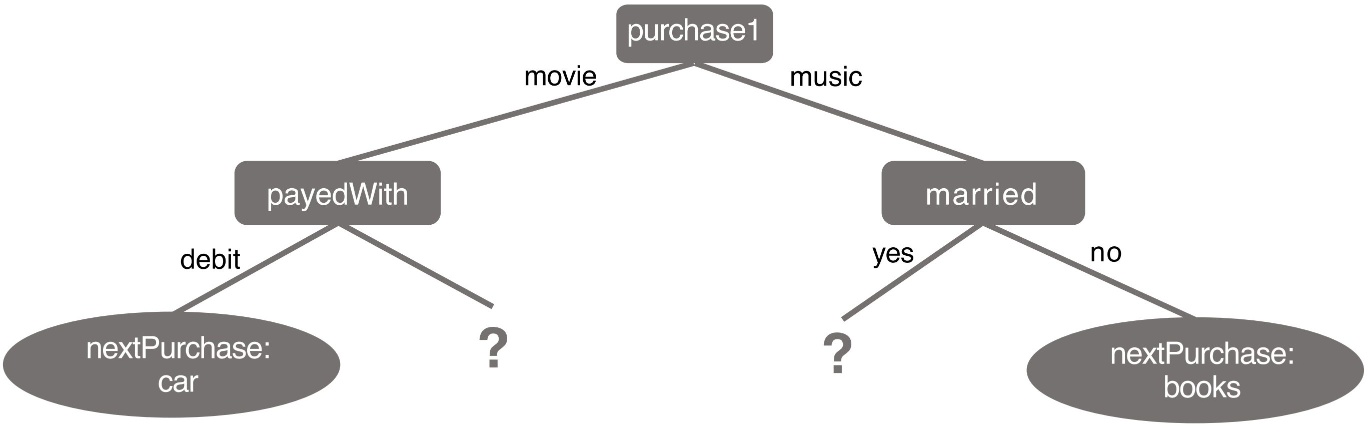
If the online shop now wants to make predictions, it sorts the customers into different categories - this is known as „classification".

In order to make a larger set of rules easier to understand and apply, they are presented as a „classification tree". This tree symbolizes the decisions made according to the rules.
Example: The rules

1. *„purchase1" = „movie" and „payedWith" = „debit" ⇒ nextPurchase = „car"*
2. *„purchase1" = „music" and „married" = „no" ⇒ nextPurchase = „books"*

can be visualized as a tree as shown below.

- In the tree, mark the path you need to take if you want to find out what a customer who bought a film as „urchase1" and paid it with a debit card might purchase next.
- We also know the following rule: *purchase1 = „movie" and „paidWith" = „mastercard" and „married" = „yes" ⇒ nextPurchase = „electronics"*
  To take this into account in the tree, you must add another decision. Think about where this makes sense and complete the decision tree.
- Check your supplement by color marking the path you need to go through the tree to find out what a customer will buy next who bought a movie as „purchase1" and paid with „mastercard".



With this classification tree it is now easy to describe which rules prevail in the previously known data. To make a forecast, this known information must be used to predict information about a customer (to generate a forecast), which can only be checked afterwards (if at all). For this purpose, the following tasks can be presented to the students:

Aufgabe 2

Use the classification tree above to decide which product category the following customers are likely to purchase next. If this decision cannot be made based on the rules from the tree, write a **?** in the field „nextPurchase (prediction)“.

age	**purchase1**	**purchase2**	**purchase3**	**purchase4**	**married**	**children**	**payedWith**	**nextPurchase (prediction)**
25-50	movie	movie	sports	sports	yes	1	debit	car
25-50	music	electronics	sports	sports	no	1	debit	books
>50	movie	music	clothes	sports	no	1	debit	car
<18	movie	music	movie	Haushalt	yes	1	Mastercard	electronics
>50	books	software	movie	sports	yes	1	VISA	?

In the example, an ambiguous assignment has been intentionally inserted: it is not possible to make an unambiguous prediction for this customer. This can be used to initiate a discussion about what should happen in such cases – and how good this analysis can be at all: What causes a single contrary data set to be added? How can we improve the analysis? How important is the best possible analysis in this case? . . .

Worksheet 4 (concept): Data Analysis at the Computer (part I)

Of course, data analyses are typically not carried out manually, but with the aid of a computer. Therefore, the students should be given the opportunity to try out how to analyze data using a suitable tool and also to work with larger data than before.

For example, a data set can be chosen that contains the student data of more than 600 students from Portugal, and in which the final grade/score of the students can be predicted. This context, how the students' work might be graded, directly affects the students and that is likely to be very controversial among them in order to demonstrate that the students themselves could in principle also be directly affected by such analyses.

For students, e. g. on a worksheet

It would certainly be a very practical thing for your teacher to transfer the idea of the online shops to your school grades: then it would be enough to just correct a few works each time and "predict"the grades of all the others. But how (well) does that really work?

This question will be discussed in detail with the students in the following by carrying out an analysis on the basis of the available data set, which will make it possible to draw fundamental conclusions about the quality and the possibilities of such data analyses. It is a good idea to go through the same process manually as before and pick up again at this point. Thus, the first step is to create classification rules and a classification tree. In order to think their way into the given data set, the students are first given the following task:

Aufgabe 1

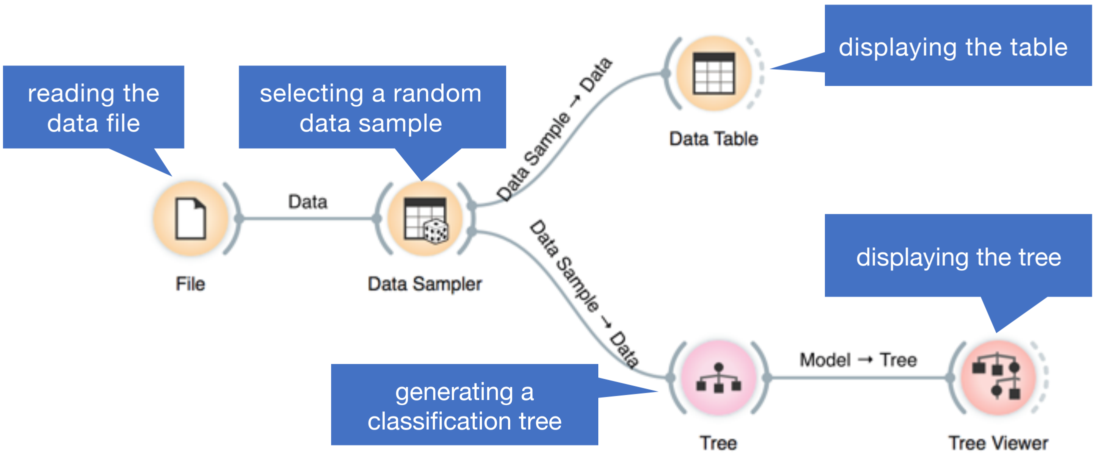
First without a computer: Which of the student attributes shown in the table below do you think are important for "predicting"the final grades? Mark them in the following table.

attribute	**description**	**relevant for prediction?**
school	abbreviation of school name: “GP” or “MS”	
gender	“M” or “F”	
age	number	
neighborhood	“urban” or “rural”	
family size	“≤3” or “>3”	
education mother	primary; lower_secondary; upper_secondary; higher	
education father	see above	
job mother	healthcare; teacher; at_home; civil_services; other	
job father	see above	
travel time	travel time to school: “<15min”; “15-30min”; “30-60min”; “>60min”	
learning time	time the student spends learning outside of class per week: “<2h”; “2-5h”; “5-10h”; “>10h”	
failures	number of class failures in the past: 0; 1; 2 or 3	
support	is the student getting support from his family in education purposes: “yes” or “no”	
extra classes	is the student participating in paid extra classes: “yes” or “no”	
extracurricular	is the student participating in extracurricular activities: “yes” or “no”	
internet	can the student use internet at home: “yes” or “no”	
family relationship	how well are the familiar relationships of the student: “very bad”; “bad”; “medium”; “good”; “very good”	
spare time	how many spare time has the student: “very little”; “little”; “medium”; “much”; “very much”	
go out	how important it is for the student to go out with friends: “very important”; “important”; “medium”; “hardly important”; “not important”	
health	the student’s health is: “very bad”; “bad”; “medium”; “good”; “very good”	
absences	how often was the student absent?: number	
result 1	points in the first test: 0 to 20	
result 2	points in the second test: 0 to 20	
result 3	points in the third test: 0 to 20	

After the students have made their assumptions based on the task and got to know the data set, an analysis can now be carried out on the PC. In order to reduce the time required, it is a good idea to provide a prepared analysis file on which the students can build. A first task can then be to work with the classification tree:

Aufgabe 2

Now we'll try it all out on the computer. Start the program „Orange3" on the computer and load the project „studentGrades". A part of the analysis is already prepared there:



The program loads the file with the student data (*File*). From this a small part of the data (default: 30 %) is selected (*Data Sampler*) as a sample, which in real would correspond to the "corrected works". Based on this sample, the computer automatically searches for rules describing these data stores them as a classification tree (*Tree*). In order to display this tree, it is passed to the *Tree Viewer*.

Now take a look at the classification tree using the *Tree Viewer* (double click it). The tree looks a bit more complicated than the one in the last worksheet. Can you see any differences to the attributes you expect? Do the criteria by which a student gets a certain grade seem logical and meaningful to you?

Note: The tree may look different to your neighbor. This is because the 30 % of student data on each computer is randomly selected. You can also select a new sample for yourself using the Data Sampler command „Sample Data. A new tree will then be created automatically.

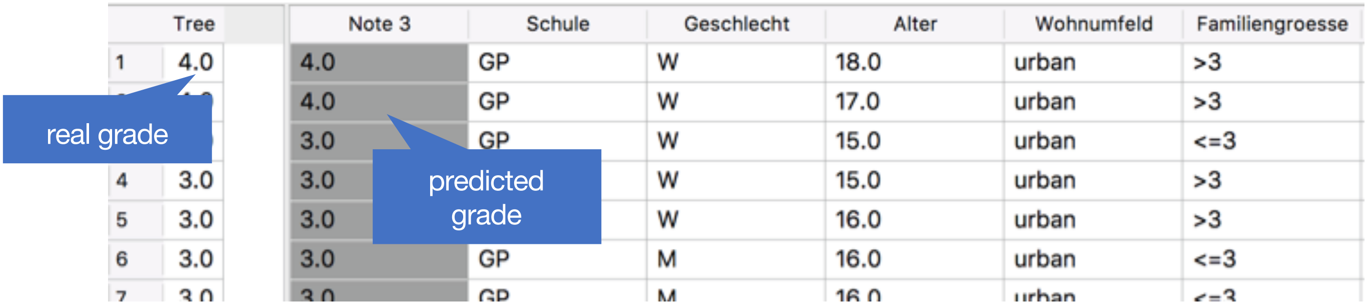
The exercise gives students a basic understanding of the program used and enables them to understand the structure of the analysis and the resulting classification tree, which forms an important basis for the further procedure in which the students are now to extend the analysis and ultimately evaluate it.

The next step is to use the generated model to predict the grades of all students.

Worksheet 5 (concept): Data Analysis at the Computer (part II)

Aufgabe 1

Of course we want to use the classification tree to automatically predict the students' points. You can do this by dragging the *Prediction* symbol from the library on left to the programming area on the right. The prediction function requires two inputs: The *tree* it should use to make predictions, and the *data* it should predict something about. Therefore, draw a connection from the semicircle to the right of the *Tree* (this semicircle corresponds to the output/return value of this function) to the input of the *Prediction* function and from the output of the *File* to the input of the *Prediction*.
To see what predictions Orange3 made, we can double-click on the *Prediction*. You will then see a table that looks like this:



	Tree	Note 3	Schule	Geschlecht	Alter	Wohnumfeld	Familiengroesse
1	4.0	4.0	GP	W	18.0	urban	>3
[illegible]	4.0	4.0	GP	W	17.0	urban	>3
[illegible]	[illegible]	3.0	GP	W	15.0	urban	<=3
4	3.0	3.0	[illegible]	W	15.0	urban	>3
5	3.0	3.0	[illegible]	W	16.0	urban	>3
6	3.0	3.0	GP	M	16.0	urban	<=3
7	3.0	3.0	GP	M	16.0	urban	<=3

What does it look like - was your prediction perfect? How good would you rate it in school grades?

Ⓐ — Ⓑ — Ⓒ — Ⓓ — Ⓔ — Ⓕ

If you compare the real scores and the predicted ones, how strong is the maximum deviation you find?
Bonus question: Here, as input for the prediction we used the original file again, not the output from the data sampler as before. Why would it be pointless to use the output of the data sampler as input of the prediction?

Since the evaluation based on the table is relatively laborious and inaccurate, it is a good idea to present the students with a way to make the whole thing more systematic. It is therefore a good idea to use a so-called „Confusion Matrix" as an aid. This two-dimensional matrix has as one dimension the actual value, as others the predicted value. Thus it allows to gain an insight into the validity of the prediction and to recognize outliers and their extent quickly and easily.

Aufgabe 2

If we want to judge our analysis, it is very time-consuming to look only at the table. Instead, we can use a *Confusion Matrix* that shows us how „confused“ the analysis was. You can (after dragging the icon from the list on the left to the workspace on the right) connect it directly to the prediction. If you double-click the Confusion Matrix it will show you a table with the real scores on the left and the predicted scores at the top. The table shows for each of these combinations how many grades have been sorted there:

- Mark the perfect predictions in the diagram. Where can you find them?
  On the diagonal from top left to bottom right
- How many students' grades did the analysis predict correctly??
  Varying depending on the student because of sampling
- How many students' grades were predicted only slightly wrong, i.e. only differing by one grade?
  Varying depending on the student because of sampling

Now that the students have probably seen that the analysis is by no means perfect, it makes sense to consider how it can be improved. This is done by giving the students an important adjustment screw: the size of the data set used to create the analysis model, i.e. the selected sample. In order to achieve the best possible results, the sample should of course be as large as possible (ideally 100 %). However, this often does not make sense, since part of the data is still needed to test the analysis model.

Aufgabe 3

We can improve the analysis a little. The number of students used to create the tree can be adjusted. Double click on the Data Sampler and change the percentage of student data.

- The analysis becomes better when the sample size is...

  □ increased

  □ decreased

- What percentage of student data is best for analysis?

  With as much data as possible, i.e. 100 %

- Does it make sense to use this percentage of data to create the model? What are the possible problems?

  No, because in this case (in the concrete example) all the exams would
  have to be corrected, hence the prediction does not make sense anymore.

- Would you feel comfortable if your teacher used this method to correct your work?

  □ Yes □ No

- If your teacher could manage to improve the quality of the analysis so that only a few students (10%) are wrongly assessed (better or worse), would that be a fair enough solution for you?

  □ Yes □ No

The last step in the automated analysis is of course to discuss what happens when we look at a single class with 30 students instead of the larger amount of students. As a teacher demo, the model can therefore be quickly rebuilt so that instead of all 600 students, only a sample of 30 students is included in the data set (of which only a small percentage is used to create the model). This makes it even clearer to the students that large amounts of data make sense for predicting, while small amounts of data can lead to enormously error-prone analysis results. This shows why any data-based business model depends on collecting a lot of data about its customers. At this point, depending on the prior knowledge of the students, a comparison with probability or the law of large numbers may also be useful.

Worksheet 6 (concept): Discussion of the results

The last part of the teaching sequence should be a discussion of the implications of what has been learned. It is important that students are aware of the following aspects of data analysis:

- As the number of data sets increases, a prediction typically becomes more accurate.
- In some places, even the smallest errors in the prediction are undesirable, while even many errors in other places are tolerable.
- Because we cannot logically comprehend the rules on which the predictions are based, such predictions often seem dangerous.
- Even under the assumption that we can only benefit from a data analysis because we behave perfectly, the prediction may classify us differently (for whatever reason) - so we should critically question how we stand about concrete usage scenarios.

For this purpose, the jigsaw method can be used, which is structured as follows:

Part I: Task for the home groups

There are five examples from the field of data analysis available to you, which you should get to know during this teaching phase.
Every member of the group should choose exactly one of the five examples that you would like to explore in more detail below. Then go to the expert groups where all those who deal with the same example come together.

Beispiele

- Analysis of credit card data/use:
  `https://bigdata-madesimple.com/how-to-use-big-data-to-successfully-fight-credit-card-fraud/`
  `https://www.govtech.com/fs/Machine-Learning-And-Big-Data-Know-It-Wasnt-You-Who-Just-Swiped-Your-Credit-Card.html`
- Data analysis and smart cars:
  `https://cars.usnews.com/cars-trucks/car-insurance/how-do-those-car-insurance-tracking-devices-work`
  `https://www.informationweek.com/big-data/big-data-analytics/big-data-drives-the-smart-car/d/d-id/1127767`
- Assessment of persons on the basis of data analyses: `https://www.theguardian.com/world/2019/mar/01/china-bans-23m-discredited-citizens-from-buying-travel-tickets-social-credit-system`
  `https://www.nytimes.com/2006/08/09/technology/09aol.html`
- Data analysis in the Smart Home:
  `https://www.scientificamerican.com/article/alexa-what-are-you-doing-with-my-familys-personal-info/`
  `https://www.cnet.com/g00/news/researchers-find-smart-meters-could-reveal-favorite-tv-shows`

Part II: Task for the expert groups

For one context of data analysis and prediction you will receive an example of how this could be used in a real or fictitious way.
Try to understand this example. Discuss the following questions and make notes. Prepare to briefly present and discuss your results in the home groups.

- Which data is used?
- How is the data analyzed?
- What is the purpose of this analysis?
- Is this example even practical? Why resp. why not?
- Are possible errors in the data analysis tolerable or not?
- Do you find such an analysis helpful/meaningful/useful or rather dangerous? Should it be permissible to use such an analysis?
- Would you willingly give up your data for this purpose?
- What could the data still be used for in the future - if they are already there?

Part III: Task for the home groups

Now that you have discussed the examples in the expert groups, you should compare them together. Now you have one expert in your group for each of the five examples.

So that everyone knows about the examples, you explain them to each other briefly. In particular, discuss the aspects discussed in the expert groups.

Discuss the similarities and differences between the examples: Are the analyses the kinds of data use you want and that make sense? If so, what are the advantages for you? If not, how can you possibly escape and prevent your data from being used in this way?

Worksheets for students

Worksheet 1: Logic or just observation?

In class, you've already seen an article about how data is used in retail today to present advertisements tailored to customers. However, online shops are now going further and trying to deliver many products to their customers as quickly as possible:

> *Anticipatory shipping may be closest that retail can come to a crystal ball. Amazon, which now has a patent for the algorithm-based system, could conceivably use the system to ship products before you even place an order.*
> *Amazon filed for the patent, officially known as "method and system for anticipatory package shipping,ïn 2012, and it was awarded on Christmas Eve of the following year. The patent summary describes a method for shipping a package of one or more items "to the destination geographical area without completely specifying the delivery address at time of shipment,"with the final destination defined en route. [. . .] So just imagine a company the size of Amazon making accurate supply and shipping decisions before a customer's decisions are finalized.*
>
> *— Lance Ulanoff, Mashable.com, January 21 2014*

Aufgabe 1

To find out what a customer might order next, retailers need to collect and analyze extensive data about their customers.

a) What do online retailers know about their customers? Where do they get the information from?

Information about the customer	**Source**

b) Probably not all the information that an online retailer has about its customers is also important when it wants to find out which item the customer could order next. In the table above, mark the rows you think are important for this purpose by making a + next to the important rows.

Aufgabe 2

Fill in the following gap text: In data analysis, there are two ways we can make predictions:

1. If we already know something about the data to analyze, then we can explain how something works and draw ____________ from it. So there are logical connections, so called ____________________, which we can use for prediction.

   Example:

   `IF a customer has bought chips in the last 5 purchases, THEN he/she will probably also purchase some next time.`

2. In other areas we do not recognize any logical connections. Instead, we can look for ________ in the data. These also provide us with information, but we often cannot explain them to ourselves. Such connections can be called ___________.

   Example:

   `IF a customer has bought item X and he/she lives in Y and is at least 35 years old, THEN he/she will also buy Z.`

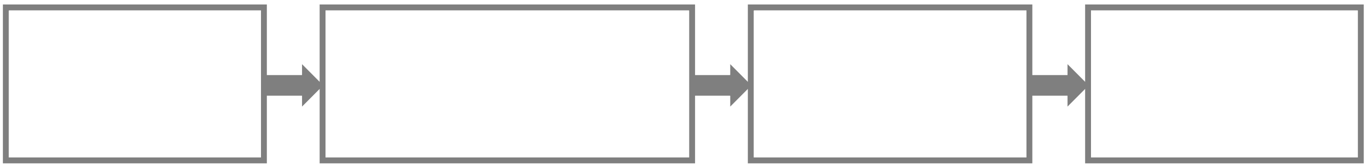
____________________ help us to understand things, but they are often relatively uninteresting for data analysis: They are rather ________ and ____________, so that they produce only little new information. But we can logically explain that they are _______. Instead, correlative correlations are often more exciting as they give us new information. But they have the disadvantage that they are not necessarily _________________________: How exactly place of residence and age shape purchasing behavior, we can usually not explain logically. Besides, we first have to find them, which is relatively _______.

Worksheet 2: Manual Data Analysis (part I)

Aufgabe 1

In class it has already been worked out how a data analysis works. Complete the following gap text and the following process model:

As a first step of data analysis, the data must be ____________________ and stored. From this data, usually a ______ subset is chosen as a basis to create the ________________, i.e. to find rules that allow the prediction of the desired properties. These rules can then be used to create ______________________. The last step of any data analysis should be the ____________ of the results in order to ensure the best possible _______ of the results.



Aufgabe 2

An online shop has collected various data about its customers and now wants to encourage its customers to make further purchases by sending individual vouchers. Therefore, he wants to find out which product category is particularly interesting for each customer.

The shop has already collected the following data about each of its customers: `age, marital status, number of children, preferred payment method, categories of the last four products purchased (film, sports, software, electronics, clothing, music, books or car accessoires)`

In order to send each customer a voucher that they are likely to redeem, the online shop wants to find out which category is particularly interesting for the buyer. Which if-then-rules do you suspect might help the online shop? *Note: of course you can combine several conditions with „and“ e.g. „purchase1 = electronics and children = 0“.*

- `IF` ____

  `THEN the next purchase is from the category` ____________________

- `IF` ____

  `THEN the next purchase is from the category` ____________________

- `IF` ____

  `THEN the next purchase is from the category` ____________________

Aufgabe 3

After the company's data scientists had recognized that there are no valid correlations in the information currently available, it was decided to try a different approach: The online shop asked some of its customers what was of interest to them next. The following data came out. What correlations do you see in the table below?

Beispiel:
IF purchase1 is a movie and has been paid with debit card, THEN the customer will buy items from the car category next.

In short: „purchase1“=„movie“ and „payedWith“=„debit“ ⇒ „nextPurchase“=„car“

age	**purchase1**	**purchase2**	**purchase3**	**purchase4**	**married**	**children**	**payedWith**	**nextPurchase**
25-50	movie	software	movie	sports	yes	1	debit	car
25-50	electronics	music	movie	software	no	1	VISA	books
<18	movie	electronics	sports	sports	no	0	debit	car
<18	movie	music	clothes	sports	no	0	debit	car
18-25	books	music	movie	household	no	1	debit	books
<18	books	movie	movie	books	no	0	VISA	books
25-50	movie	movie	sports	sports	yes	1	debit	car
25-50	music	movie	movie	toys	no	1	debit	books
25-50	music	music	movie	household	no	1	debit	books
25-50	electronics	music	books	software	yes	1	Mastercard	electronics
25-50	software	electronics	movie	toys	no	1	Mastercard	books
25-50	movie	movie	sports	sports	yes	0	Mastercard	electronics
25-50	music	electronics	books	electronics	yes	1	Mastercard	electronics

- IF ____________________

  THEN the next purchase is from ____________________

- IF ____________________

  THEN the next purchase is from ____________________

- IF ____________________

  THEN the next purchase is from ____________________

Worksheet 3: Manual Data Analysis (part II)

Aufgabe 1

If the online shop now wants to make predictions, it sorts the customers into different categories - this is known as „classification".

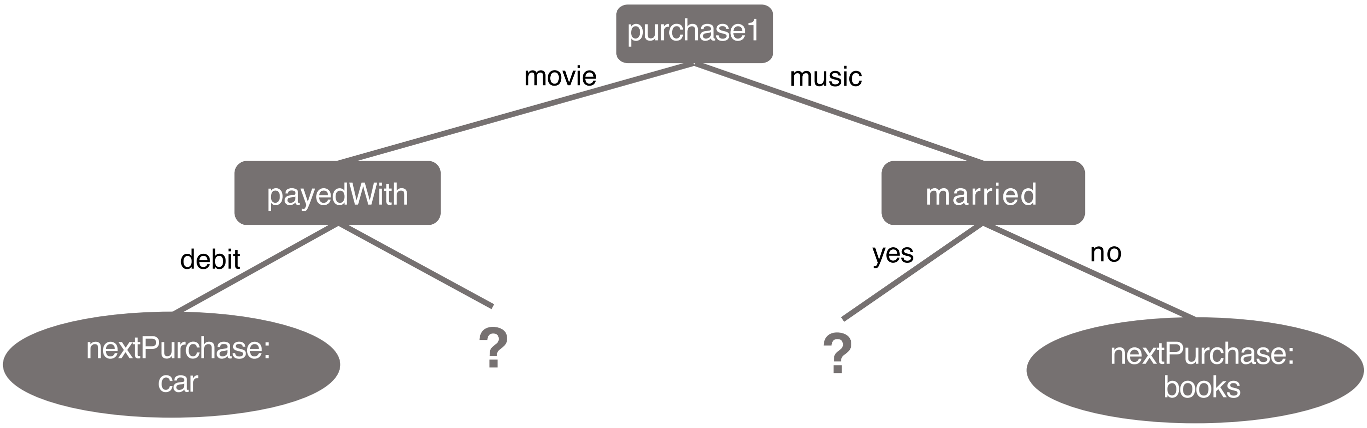
In order to make a larger set of rules easier to understand and apply, they are presented as a „classification tree". This tree symbolizes the decisions made according to the rules.

Example: The rules

1. *„purchase1" = „movie" and „payedWith" = „debit"* ⇒ *nextPurchase = „car"*
2. *„purchase1" = „music" and „married" = „no"* ⇒ *nextPurchase = „books"*

can be visualized as a tree as shown below.

- In the tree, mark the path you need to take if you want to find out what a customer who bought a film as „urchase1" and paid it with a debit card might purchase next.
- We also know the following rule: *purchase1 = „movie" and „paidWith" = „mastercard" and „married" = „yes"* ⇒ *nextPurchase = „electronics"*
  To take this into account in the tree, you must add another decision. Think about where this makes sense and complete the decision tree.
- Check your supplement by color marking the path you need to go through the tree to find out what a customer will buy next who bought a movie as „purchase1" and paid with „mastercard".



Aufgabe 2

Use the classification tree above to decide which product category the following customers are likely to purchase next. If this decision cannot be made based on the rules from the tree, write a

? in the field „nextPurchase (prediction)“.

age	purchase1	purchase2	purchase3	purchase4	married	children	payedWith	nextPurchase (prediction)
25-50	movie	movie	sports	sports	yes	1	debit	car
25-50	music	electronics	sports	sports	no	1	debit	books
>50	movie	music	clothes	sports	no	1	debit	car
<18	movie	music	movie	Haushalt	yes	1	Mastercard	electronics
>50	books	software	movie	sports	yes	1	VISA	?

Worksheet 4: Data Analysis at the Computer (part I)

It would certainly be a very practical thing for your teacher to transfer the idea of the online shops to your school grades: then it would be enough to just correct a few works each time and "predict"the grades of all the others. But how (well) does that really work?

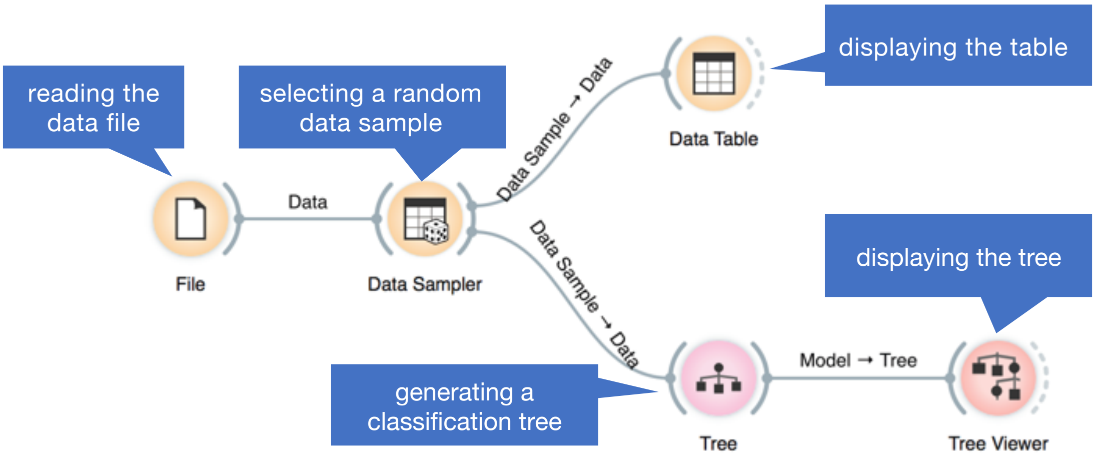
Aufgabe 1

First without a computer: Which of the student attributes shown in the table below do you think are important for "predicting"the final grades? Mark them in the following table.

attribute	**description**	**relevant for prediction?**
school	abbreviation of school name: “GP” or “MS”	
gender	“M” or “F”	
age	number	
neighborhood	“urban” or “rural”	
family size	“≤3” or “>3”	
education mother	primary; lower_secondary; upper_secondary; higher	
education father	see above	
job mother	healthcare; teacher; at_home; civil_services; other	
job father	see above	
travel time	travel time to school: “<15min”; “15-30min”; “30-60min”; “>60min”	
learning time	time the student spends learning outside of class per week: “<2h”; “2-5h”; “5-10h”; “>10h”	
failures	number of class failures in the past: 0; 1; 2 or 3	
support	is the student getting support from his family in education purposes: “yes” or “no”	
extra classes	is the student participating in paid extra classes: “yes” or “no”	
extracurricular	is the student participating in extracurricular activities: “yes” or “no”	
internet	can the student use internet at home: “yes” or “no”	
family relationship	how well are the familiar relationships of the student: “very bad”; “bad”; “medium”; “good”; “very good”	
spare time	how many spare time has the student: “very little”; “little”; “medium”; “much”; “very much”	
go out	how important it is for the student to go out with friends: “very important”; “important”; “medium”; “hardly important”; “not important”	
health	the student’s health is: “very bad”; “bad”; “medium”; “good”; “very good”	
absences	how often was the student absent?: number	
result 1	points in the first test: 0 to 20	
result 2	points in the second test: 0 to 20	
result 3	points in the third test: 0 to 20	

Aufgabe 2

Now we’ll try it all out on the computer. Start the program „Orange3“ on the computer and load the project „studentGrades“. A part of the analysis is already prepared there:



The program loads the file with the student data (*File*). From this a small part of the data (default: 30 %) is selected (*Data Sampler*) as a sample, which in real would correspond to the "corrected works". Based on this sample, the computer automatically searches for rules describing these data stores them as a classification tree (*Tree*). In order to display this tree, it is passed to the *Tree Viewer*.

Now take a look at the classification tree using the *Tree Viewer* (double click it). The tree looks a bit more complicated than the one in the last worksheet. Can you see any differences to the attributes you expect? Do the criteria by which a student gets a certain grade seem logical and meaningful to you?

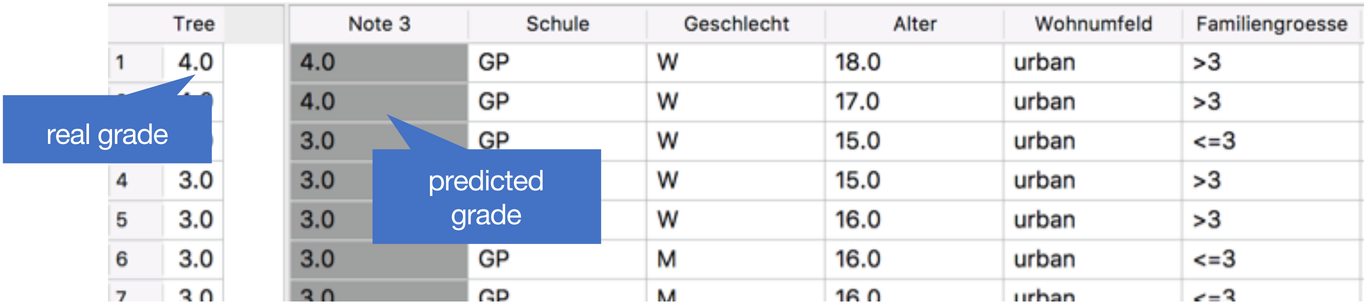
Note: The tree may look different to your neighbor. This is because the 30 % of student data on each computer is randomly selected. You can also select a new sample for yourself using the Data Sampler command „Sample Data. A new tree will then be created automatically.

Worksheet 5: Data Analysis at the Computer (part II)

Aufgabe 1

Of course we want to use the classification tree to automatically predict the students' points. You can do this by dragging the *Prediction* symbol from the library on left to the programming area on the right. The prediction function requires two inputs: The *tree* it should use to make predictions, and the *data* it should predict something about. Therefore, draw a connection from the semicircle to the right of the *Tree* (this semicircle corresponds to the output/return value of this function) to the input of the *Prediction* function and from the output of the *File* to the input of the *Prediction*.

To see what predictions Orange3 made, we can double-click on the *Prediction*. You will then see a table that looks like this:



	Tree	Note 3	Schule	Geschlecht	Alter	Wohnumfeld	Familiengroesse
1	4.0	4.0	GP	W	18.0	urban	>3
[illegible]	[illegible]	4.0	GP	W	17.0	urban	>3
[illegible]	[illegible]	3.0	GP	W	15.0	urban	<=3
4	3.0	3.0	[illegible]	W	15.0	urban	>3
5	3.0	3.0	[illegible]	W	16.0	urban	>3
6	3.0	3.0	GP	M	16.0	urban	<=3
7	3.0	3.0	GP	M	16.0	urban	<=3

What does it look like - was your prediction perfect? How good would you rate it in school grades?

Ⓐ — Ⓑ — Ⓒ — Ⓓ — Ⓔ — Ⓕ

If you compare the real scores and the predicted ones, how strong is the maximum deviation you find?

Bonus question: Here, as input for the prediction we used the original file again, not the output from the data sampler as before. Why would it be pointless to use the output of the data sampler as input of the prediction?

Aufgabe 2

If we want to judge our analysis, it is very time-consuming to look only at the table. Instead, we can use a *Confusion Matrix* that shows us how „confused“ the analysis was. You can (after dragging the icon from the list on the left to the workspace on the right) connect it directly to the prediction. If you double-click the Confusion Matrix it will show you a table with the real scores on the left and the predicted scores at the top. The table shows for each of these combinations how many grades have been sorted there:

- Mark the perfect predictions in the diagram. Where can you find them?

  ____

- How many students' grades did the analysis predict correctly??

  ____

- How many students' grades were predicted only slightly wrong, i.e. only differing by one grade?

Aufgabe 3

We can improve the analysis a little. The number of students used to create the tree can be adjusted. Double click on the Data Sampler and change the percentage of student data.

- The analysis becomes better when the sample size is. . .

  □ increased

  □ decreased

- What percentage of student data is best for analysis?

- Does it make sense to use this percentage of data to create the model? What are the possible problems?

- Would you feel comfortable if your teacher used this method to correct your work?

  □ Yes □ No

- If your teacher could manage to improve the quality of the analysis so that only a few students (10%) are wrongly assessed (better or worse), would that be a fair enough solution for you?

  □ Yes □ No

Worksheet 6: Discussion of the results

Part I: Task for the home groups

There are five examples from the field of data analysis available to you, which you should get to know during this teaching phase.

Every member of the group should choose exactly one of the five examples that you would like to explore in more detail below. Then go to the expert groups where all those who deal with the same example come together.

Beispiele

- Analysis of credit card data/use:
  `https://bigdata-madesimple.com/how-to-use-big-data-to-successfully-fight-credit-card-fraud/`
  `https://www.govtech.com/fs/Machine-Learning-And-Big-Data-Know-It-Wasnt-You-Who-Just-Swiped-Your-Credit-Card.html`
- Data analysis and smart cars:
  `https://cars.usnews.com/cars-trucks/car-insurance/how-do-those-car-insurance-tracking`
  `devices-work`
  `https://www.informationweek.com/big-data/big-data-analytics/big-data-drives-the-smart-car/d/d-id/1127767`
- Assessment of persons on the basis of data analyses: `https://www.theguardian.com/world/2019/mar/01/china-bans-23m-discredited-citizens-from-buying-travel-tickets-social-credit-system`
  `https://www.nytimes.com/2006/08/09/technology/09aol.html`
- Data analysis in the Smart Home:
  `https://www.scientificamerican.com/article/alexa-what-are-you-doing-with-my-familys-personal-info/`
  `https://www.cnet.com/g00/news/researchers-find-smart-meters-could-reveal-favorite-tv-shows`

Part II: Task for the expert groups

For one context of data analysis and prediction you will receive an example of how this could be used in a real or fictitious way.

Try to understand this example. Discuss the following questions and make notes. Prepare to briefly present and discuss your results in the home groups.

- Which data is used?

- How is the data analyzed?

- What is the purpose of this analysis?
- Is this example even practical? Why resp. why not?
- Are possible errors in the data analysis tolerable or not?
- Do you find such an analysis helpful/meaningful/useful or rather dangerous? Should it be permissible to use such an analysis?
- Would you willingly give up your data for this purpose?
- What could the data still be used for in the future - if they are already there?

Part III: Task for the home groups

Now that you have discussed the examples in the expert groups, you should compare them together. Now you have one expert in your group for each of the five examples.

So that everyone knows about the examples, you explain them to each other briefly. In particular, discuss the aspects discussed in the expert groups.

Discuss the similarities and differences between the examples: Are the analyses the kinds of data use you want and that make sense? If so, what are the advantages for you? If not, how can you possibly escape and prevent your data from being used in this way?