



Universiteit
Leiden
The Netherlands

Opleiding Informatica

Analysing time series data of children speaking in public
to learn more about social anxiety

Marlo Brochard

Supervisor:
Matthijs van Leeuwen & Hugo Manuel Proença

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

19/06/2019

Abstract

Marlo Brochard, Informatica&Economics, Leiden University

Abstract of bachelor's Thesis, Submitted 19 June 2019: Analysing time series data of children speaking in public to learn more about social anxiety

The aim of this thesis is to extract information from mixed time series and static data, in order to learn more about the development of stress responses during public speaking by adolescents. This is done by a descriptive data mining technique called subgroup discovery. This will result in subsets of the entire data set described by conditions, also called rules. The time series data consists of ECG data collected during a public speaking task. The static data consists of data collected by means of a questionnaire.

In the first part, results of an experiment whether we can tune the quality measure such that we obtain a desirable trade-off between subgroup size and the deviation of the target are presented. These show that it is possible to tune the quality measure such that we can make a desirable trade-off between the subgroup size and the deviation of the target. The expert, with whom we are collaborating, indicated to tune the quality measure such that it find subsets that have at least 60 participants. The results of the second experiment which compares a tuned version of the quality measure to a default version of the quality measures, show that the tuned version is the only quality measure that returns subgroups with a subgroup size around 60. So it is indeed better, for the purpose of this thesis, to use a tuned version instead of a default version of the quality measures.

In the second part, results of an experiment whether there is an observed effect or correlation that indicates that the obtained subgroups are unlikely to be based on luck are presented. These show that at least the top-100 subgroups with nominal targets were significant. The results of the second experiment, show that a higher heart rate, amplitude and age is associated with a higher social anxiety score. A lower heart rate variability and increasing BMI is also associated with a higher social anxiety score.

This aim of this thesis is to offer useful insights so that psychologists can get a better understanding of adolescents. Finally, I recommend further research in order to improve and obtain new insights.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Thesis overview | 3 |
| 2 | Conceptual framework | 4 |
| 2.1 | ECG data | 4 |
| 2.2 | Heart Rate Variability | 5 |
| 2.3 | Amplitude | 6 |
| 2.4 | Social anxiety | 7 |
| 3 | Problem statement | 7 |
| 3.1 | Subgroup discovery example | 8 |
| 3.2 | Quality measure | 9 |
| 4 | Related work | 9 |
| 4.1 | Case studies | 9 |
| 4.2 | Methods | 10 |
| 5 | Method | 11 |
| 5.1 | The data | 11 |
| 5.1.1 | Static data | 12 |
| 5.1.2 | Time series data | 13 |
| 5.2 | Preprocessing | 13 |
| 5.2.1 | Data cleaning | 13 |
| 5.2.2 | Peak detection | 13 |
| 5.3 | Feature extraction | 15 |
| 5.3.1 | HRV | 15 |
| 5.3.2 | BPM | 16 |
| 5.3.3 | Amplitude | 16 |
| 5.3.4 | Delta variables | 16 |
| 5.4 | DataFrames | 16 |
| 5.5 | Subgroup discovery | 17 |
| 5.6 | Statistical measurement: p-value | 19 |
| 6 | Results | 22 |
| 6.1 | Experiment setup | 22 |
| 6.2 | Tuning alpha | 23 |
| 6.3 | Tuned vs default measures | 24 |
| 6.4 | Filtering the results: Statistical tests | 25 |
| 6.5 | Interpreting the results | 26 |
| 7 | Discussion | 34 |
| 7.1 | Small sample size | 34 |
| 7.2 | Data influenced by environment | 34 |
| 7.3 | Remove outliers | 35 |

| | | |
|-----------|--|-----------|
| 7.4 | Missing important data properties | 35 |
| 7.5 | Depending on the correctness of Cortana | 35 |
| 8 | Conclusion | 35 |
| 9 | Future research | 37 |
| 9.1 | Make use of the tone of the message | 37 |
| 9.2 | Develop a data mining program in Python | 37 |
| 9.3 | Make predictions about the psychological condition of a person | 37 |
| 9.4 | Collecting data about the environmental condition in which the adolescents are presenting | 37 |
| | References | 39 |
| 10 | Appendix | 40 |
| 10.1 | Peak detection function | 40 |
| 10.2 | Obtained subgroups in T1 | 41 |
| 10.3 | Obtained subgroups in T2 | 42 |
| 10.4 | Obtained subgroups in delta variable | 43 |
| 10.5 | The top-100 subgroups used for the statistical test. | 44 |

1 Introduction

At this moment in time, you would think that data mining is a new technology which becomes more and more popular. However, data mining is an umbrella term for techniques that are used to extract interesting knowledge from the data. Some of these techniques are already invented in the year 1700. In this period, it all started with the Bayes theorem and the regression analysis. In the last two decades, data mining has become very popular. One of the reasons for this is that evolutionary steps were taken between 1960 and 1990 with regard to storing data and making data accessible. At the moment these steps were taken, the use of data in order to support the decision making process becomes a lot more important than before [Sau17]. Nowadays, data is everywhere and they estimate that the amount of data in 2020 will be 40 million petabytes [Tho15]. Because of this, it will be more important to use data mining techniques in order to discover new insights and detect hidden patterns in this huge amount of data. For humans, it is almost impossible to detect these useful insights and to see the coherence of the data.

Nowadays, data is not only generated by machines. Thanks to improved equipment, human data can also be collected with high precision. An example of data generated by humans is ECG data. ECG data is data that is collected by continuously monitoring the changes in the electrical signals that are sent by our heart. This results in time series data which is a series of data points collected and stored in time order. This data can be used to extract valuable features such as Heart Rate Variability (HRV) and Beats Per Minute (BPM). Using ECG data we are able to gain insights in the psychological conditions of a person such as stress [Lam15]. A psychological condition that is being investigated is social anxiety, the way in which we define social anxiety is explained in the next chapter. Research has shown that people with a high social anxiety score have a low Heart Rate Variability [AQK⁺13].

The social and development institute of Leiden University is researching social anxiety of children between 8 and 17 years old. This institute is interested in how social anxiety develops over time in adolescents. My bachelor thesis is in collaboration with this institute and I will use two types of data, time series data and static data, to discover patterns that say something about the social anxiety score of adolescents. The goal of these patterns is to understand the adolescents better, and to customize treatments so that it is specially created for adolescents who satisfy these patterns. In this way, they can treat these adolescents in a more accurate way because they have more insight into what kind of a person is in front of them. This will give them the possibility to adjust the treatment.

Data mining techniques can be applied from two perspectives [Tak19]:

- Descriptive, which is used to extract interesting knowledge from data. This also includes subgroup discovery.
- Predictive, which is used to discover knowledge that can be used to make predictions. This includes classification and regression.

We want to extract interesting knowledge from the data without making predictions. A possible tool to apply the descriptive type of data mining is subgroup discovery. Subgroup discovery (SD)

is one of the possible methods to discover interesting patterns in the data. Subgroup discovery is a technique to detect subsets within the population who behave differently than the population. The output has the following form: Females of the age below 15 and a BMI above 24 have a higher average social anxiety score relative to the entire population. This technique will be explained in more detail in section 3 and 5.

In order to rank the obtained subgroups, we have to choose a so called quality measure. The results of the subgroup discovery task will be based on the CWRAcc quality measure. The formula used to calculate the CWRAcc is as follows:

$$CWRAcc = G^\alpha * (m_G - m_D)$$

Where G is the subgroup size which is the number of participants satisfying the pattern of the subgroup. Furthermore, m_G is the average social anxiety score of the subgroup and m_D is the average social anxiety score of the data set. In the default version of the CWRAcc formula α is set to 0.5, but we are going to investigate whether we can tune α so that we obtain a desirable trade-off between subgroup size and the deviation of the target

During this bachelor thesis we will analyse time series and static data in order to gain insights which provide us with information about the social anxiety of adolescents. The time series data consists of ECG data that is collected during a public speaking task. The static data consists of information about the participants that is collected by letting the participant fill in a questionnaire. The main goal is to discover patterns that gives information about the social anxiety of a person with the use of these two data types. The evaluation of how valuable each pattern is for the social and development institute falls outside the scope of this research. During this bachelor project the quality of the patterns found in the data is based on quality measurements.

My research question is formulated as follows:

Using data mining techniques, what information can be extracted from the data, time series data and static data, in order to learn more about the development of the stress responses during public speaking by adolescents?

The insights obtained from this study are intended to help the psychological department to better understand social anxiety in children between 8 and 17 years old.

Furthermore, this bachelor project will give answers to the following sub-questions:

1. Can we tune alpha such that we obtain a desirable trade-off between subgroup size and the deviation of the target?
2. Is it better, in the purpose of this thesis, to use a tuned version than a default version of the quality measures?
3. Statistically, are the subgroups significant or just a coincidence, meaning is there an observed effect or correlation that indicates that the obtained subgroup is unlikely to be based on luck.
4. Which insights can be derived from the subgroups so that psychologists can get a better understanding of adolescents?

1.1 Thesis overview

The current chapter contains the introduction to the subject that will be discussed in this bachelor thesis. Section 2, the conceptual framework, will be used to provide the most important concepts with a definition. Section 3, the problem statement, will explain the problem addressed in this research in more detail. Furthermore, it will dive into the SD technique that is being used in this research. Section 4, related work, will discuss studies that have been previously done in this field. Section 5, methods, will explain how the experiments are performed. Section 6, results, will present and describe the obtained results. Section 7, the discussion, will address some aspects that have to be taken into account. Last but not least, Sections 8 and 9 draw a conclusion and indicate which possible next steps can be taken.

This bachelor thesis is performed at LIACS with Matthijs van Leeuwen as supervisor. We do this in collaboration with Esther van den Bos of the Developmental and Educational Psychology institute at Leiden University.

2 Conceptual framework

During this section we will give a theoretical explanation of the terms ECG, HRV, amplitude and social anxiety.

2.1 ECG data

Many people have already seen an electrocardiogram(ECG). For example, when you are in the hospital your heart rhythm can be measured. This data is then plotted in a graph called an electrocardiogram, which can often be seen on the monitor in the hospital. But, it is often not clear where this data comes from and what there is being measured.



Figure 1: ECG monitor. Source: www.dhresource.com.

Our muscle cells communicate with each other through electrical and chemical signals, also called impulses. The electrical signals send by our heart determines our heart rate. These signals come from a group of cells that are settled in our heart, and cause the ventricles to contract and relax again. An ECG measures the changes in these electrical signals, which are then plotted in a graph, also called an electrocardiogram. This data can be measured from the surface in several ways. Figure 2 shows how the ECG data is collected for this experiment. During this bachelor project, the results are only based on the CH1 measurements [Inf19].

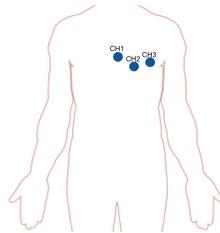


Figure 2: Setup that is used for the collection of the ECG data we are using during this project. Channel 1 is the ECG, Channel 2 is the inter beat interval and Channel 3 is the skin conductance level.

The data obtained, by continuous monitoring the changes in electrical signals, is a series of data points collected and stored in time order, also called time series data. This time series data is very valuable for this research since we can derive three useful features from it:

1. Heart Rate Variability
2. Beats Per Minute
3. Amplitude

These features will be explained in more detail in the following subsections. These features will be used to detect abnormal behavior in the social anxiety score of a person.

Research has been done in which ECG data is used to explain certain behavior. This has shown that the heart rhythm of boys and girls was higher during the speaking task than after the speaking task. Furthermore, the heart rhythm of boys and girls was higher during the speaking task than before the speaking task. This bachelor project uses the same data as the data used to make the previous explained findings [vdBW15]. Furthermore, research has been done in which ECG data was used to study the social psycho-physiological compliance of collaborating students. Among other things, it was found that the correlation of HRV is greater for signals coming from collaborating students than individuals [ACT+16].

2.2 Heart Rate Variability

Heart Rate Variability(HRV) can be derived from ECG data. It measures the variability in duration between two consecutive heartbeats, which is shown in the figure below.

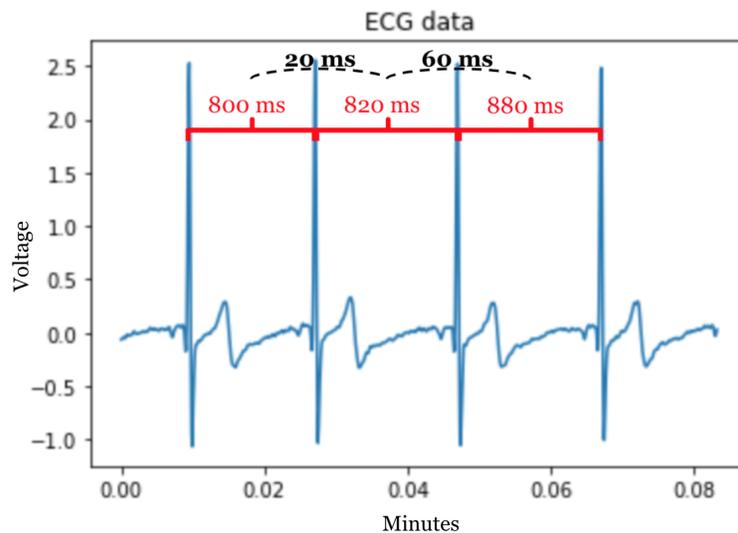


Figure 3: This figure shows how the HRV calculation works. We see a sample of how ECG can look like, with the time in minutes on the x-as and the voltage(Electrical signals) in mV on the y-as. The distance in milliseconds between two consecutive peaks is shown in red. The variation between the distance in milliseconds is shown in black. The variation, which is shown in black, is used to calculate the HRV.

The distance between two consecutive heartbeats is a so called RR-interval. The RR-interval is used in order to calculate the HRV using the following formula:

$$HRV = \sqrt{\frac{((RRinterval1 - RRinterval2)^2 + (RRinterval2 - RRinterval3)^2 + \dots)}{n}}$$

Where n is the number of RR-intervals used to calculate the HRV

Using the HRV we cannot only say something about well-being, but also about psychological conditions [Far17].

In every healthy heart there is a natural variation in the RR interval. This variation is caused by our body, which has to find a balance in the two states of the ANS, the Autonomous Nervous System. The Autonomous Nervous System is part of the human nervous system and affects the functioning

of internal organs. Research has shown that when people experience stress, the sympathetic system is activated (Fight modus). As a result of this, the heart rate will rise and the distance between two heartbeats becomes smaller. This ensures that the difference in RR-intervals becomes larger. Furthermore, it has been shown that when the human is in the recovery phase, the parasympathetic system is activated(Flight modus). As a result of this, the heart rate will drop and the distance between two heart beats will increase. Furthermore, it has been indicated that low HRV is associated with stress [Tet18].

2.3 Amplitude

The amplitude can also be derived from ECG data. It measures the average difference between the maximum and the minimum of a heartbeat, which is shown in the figure below:

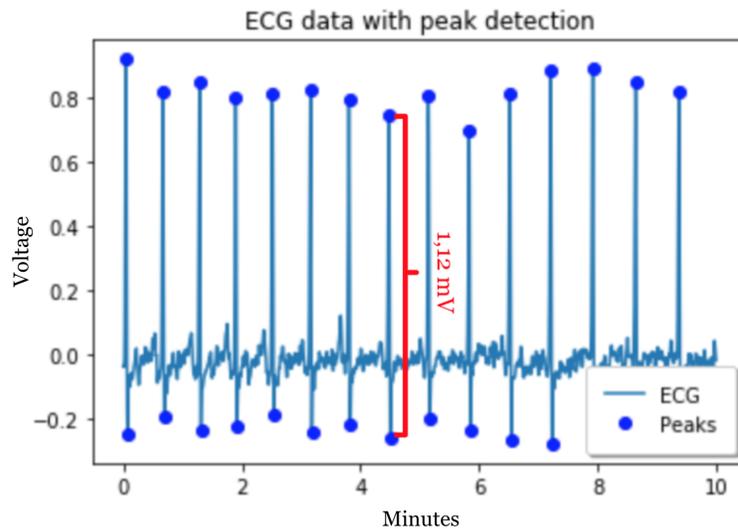


Figure 4: This figure shows how the amplitude calculation works. We see a sample of how ECG, after peak detection, can look like with the time in minutes on the x-as and the voltage(Electrical signals) in mV on the y-as. The distance in milliseconds between the maximum and minimum of a peak is shown in red. The distance between the maximum and minimum, which is shown in red, is used to calculate the amplitude.

The following formula is used to calculate the amplitude:

$$Amplitude = \frac{((maxPeak1 - minPeak1) + (maxPeak2 - minPeak2) + \dots)}{n}$$

Where n is the number of peaks used to calculate the amplitude

The amplitude can be used to indicate the degree of effort, and therefore stress. Research has shown that when people make a physical effort, the amplitude of this person increases significantly. Furthermore, it has been shown during this study that when people stop exercising physically, the heart rate remains high for 5 minutes but the amplitude gradually declined [HKYM95]. With the help of the amplitude we could, for example, see if a certain average amplitude over time reflects relevant information with regard to social anxiety.

2.4 Social anxiety

Social anxiety is a very subjective concept which cause that everyone interprets it in his or her own way. What do we mean by social anxiety?

Social anxiety means that a person is afraid of being judged negatively by other people, which lead to unpleasant feelings such as depression. How much anxiety does a person experience in social situations on a scale from 1 to 5?

We measured this value on the basis of a questionnaire that is globally approved for measuring the social anxiety score of a person: the Pubertal Development Scale [PCRB88].

In this project we want to explain the social anxiety score using ECG data of children between 8 and 17 years old. The obtained ECG data has been measured in two situations, during a speaking task in front of an audience and during a situation where they were not confronted with social contact. One of the common situations where people with a high social anxiety score tend to have trouble with is speaking in public. Because of this, the ECG data we are using is a good source for discovering patterns associated to social anxiety. The following will support the previous mentioned statement that ECG data can be used to say something about social anxiety:

“In school, I was always afraid of being called on, even when I knew the answers. I didnt want people to think I was stupid or boring. My **heart would pound** and I would feel dizzy and sick. When I got a job, I hated to meet with my boss or talk in a meeting. I couldnt attend my best friends wedding reception because I was afraid of having to meet new people. I tried to calm myself by drinking several glasses of wine before an event and then I started drinking every day to try to face what I had to do.” [nat]

3 Problem statement

This bachelor project has the goal to discover patterns in the data, which can provide the developmental and educational psychology with information about the development of the stress responses during public speaking by adolescents. By answering the questions, mentioned in section 1, we can provide the Developmental and Educational Psychology unit of Leiden University with information about the social anxiety of adolescents. This information will contain subgroups that behave differently than the population with regard to social anxiety.

Subgroup discovery is a data mining technique which is used to discover subgroups in the dataset with regard to the target attribute. Subgroup discovery was first introduced by Kloesgen [Fay96] and Wrobel [K97]. When using subgroup discovery we assume that the two following aspects are given:

1. There is a population of individual data points, within this bachelor project these are the different participants.
2. These individual data points have a property that is interesting, within this bachelor project this is the social anxiety score.

The main goal of subgroup discovery is to discover subsets of the population that are the most interesting from a statistical point of view, meaning that the subsets are interesting according to

the statistics of the subsets. For example, that a certain subgroup has a much higher average social anxiety score than the population. This involves searching for subgroups that behave differently than the population. A subgroup is formed by a certain rule, which describes the relationship between certain properties and the target. This can be formally represented as follows:

$$R : Rule \rightarrow Target_{value}$$

The $Target_{value}$ is the value of the attribute which is interesting for the subgroup discovery task, for example a social anxiety score of 3.43. The rule represents the properties of the discovered subgroup, for example $gender=man$ AND $age<12$.

The social and development institute of Leiden University has not enough insights in social anxiety of children between 8 and 17 years old. Using subgroup discovery we will provide the institute with rules that specify subsets of participants. This ensures that the institute obtain information in which characteristics are associated with higher social anxiety. With this information, special treatments can be made because they can better understand the children. In order to obtain the right subgroups a desirable trade-off between subgroup size and the deviation of the target has to be made. The subgroup size is the amount of participants satisfying the rule of the subgroup. The deviation of a numeric target is the difference between the average target value of the subgroup and the dataset. For a nominal target it is the difference between the number of ones in the subgroup and the dataset.

3.1 Subgroup discovery example

Let A be a dataset with four different attributes that can be used to define a subgroup. The four attributes are as follows:

- Gender = $\langle M, F \rangle$ (Binair)
- Age = $\langle 8 - 17 \rangle$ (Numeric, for example 12.48)
- BMI = $\langle 10 - 28 \rangle$ (Numeric)
- BPM = $\langle 50 - 180 \rangle$ (Numeric)

There is also an attribute which is used as a target attribute, the attribute which we are investigating. The target attribute is as follows:

- Social Anxiety Score (SAS) = $\langle 1 - 5 \rangle$ (Numeric)

Some of the possible rules that can be formed using this dataset are as follows:

$$R_1 : (Gender = M \text{ AND } Age \leq 12.48) \rightarrow SAS = 4.32$$

$$R_2 : (Gender = F \text{ AND } Age \leq 10.20 \text{ AND } BMI \geq 24.32) \rightarrow SAS = 4.11$$

Where rule R_1 represents a subgroup of male participants with an age of less than or equal to 12.48 years whose social anxiety score differs from that of the population. Rule R_2 represents a subgroup of female participants with an age of less than or equal to 10.20 years and a BMI greater than or equal to 24.32 whose social anxiety score differs from that of the population.

3.2 Quality measure

The ranking of the subgroups that have been discovered is done on the basis of a so-called quality measure. The quality measure provides the researcher with information about the importance and interest of each subgroup. There are many different types of quality measures to use for subgroup discovery tasks. During this bachelor project we use the CWRAcc and Z-score, which are quality measures for numerical targets. What these two different measurements entail can be found in [HGCJ11]. As mentioned above, we have to make a desirable trade-off between subgroup size and the deviation of the target. This requires a quality measure in which we can set the importance of these two properties. Section 1 introduced the tuneable version of the CWRAcc quality measure. In this formula, we are able to choose α which specifies how important the subgroup size is.

Last but not least, a certain search strategy must be chosen, examples of search strategies are beam search and depth first search. During this bachelor project we will make use of the depth first search strategy [HGCJ11]. This is a search strategy that is chosen in most cases when exhaustive search is possible. Exhaustive search is possible when the data set is of moderate size. Depth first search directly mines the top k subgroups in a short period of time. The disadvantage is that all attributes are considered in a fixed order with the reason to limit the search space size. This being said, the depth first search strategy is a possibility when the target is not complex [LK12]. Furthermore, the number of bins for numeric attributes must be specified. This indicates how many groups the values of the numerical attributes can be divided in. For example, a numeric attribute with a range from 10 to 20. When there are two bins specified, the data could be divided into 10 to 15 and 15 to 20. But when there are three bins specified, the data could be divided into 10 to 14, 15 to 17 and 18 to 20. This could cause for more redundant subgroups which is briefly explained later on.

4 Related work

In this chapter we will discuss studies that are relevant for this bachelor project. Analyzing ECG data to say something about the psychological state of that specific person is a subject that has already been investigated.

4.1 Case studies

HRV and cognitive collaboration PLoS One published a study in 2016 in which ECG data was studied in order to say something about cognitive cooperation in cardiac physiology. This study has shown that the correlation of the Heart Rate Variability is greater for signals coming from persons working in a team than for people who do not work in teams. This research has also indicated that HRV is indeed influenced by mental workload and stress [ACT⁺16]. During this bachelor project we look at the Heart Rate Variability of an individual, which we will use to say something about the social anxiety score of a person.

Stability of individual differences In 2015 a study was published in which the stability of individual differences in heart rate, parasympathetic, sympathetic and HPA axis responses in adolescents was investigated. It was shown that the heart rate was higher by boys and girls during the speech than before. Furthermore, it was shown that the Heart Rate Variability by girls was lower during the speech than before. During this research they came to the conclusion that age,

gender and puberty development are three very interesting areas [vdBW15]. During this bachelor project we use the same ECG data, and we will use this research as the basis.

Reduced HRV in Social anxiety disorder PLoS One published a study in 2013 in which Heart Rate Variability was studied in order to say something about social anxiety disorder. It was shown that social anxiety disorder is associated with a reduced HRV. Furthermore, it was demonstrated during this study that there is no significant correlation between age, BMI and HRV in the social anxiety group [AQK+13]. During this bachelor project it is investigated, among other things, whether BMI, HRV and age contain relevant information, which can then be used to say something about a person's social anxiety score.

4.2 Methods

Statistical test: p-value Research has been done to exploit False Discoveries using the p-value as the statistical validation method. During this experiment they formulated the null-hypothesis as follows: The subgroup s is generated by the distribution of false discoveries (DFD). This means that the obtained subgroup is a false discovery. A false discovery is a subgroup that is found by chance which means that the subgroup is not an actual subgroup. They validated the subgroups which were obtained using the WRAcc quality measure. For each data set, in total 20, they performed a SD run and reported the best 1000 subgroups. The properties of the data sets can be found in the paper [DK11] on page 155. After that, they calculated the p-value for each obtained subgroup and put it in a table with the WRAcc value of that subgroup.

Using the table containing the p-value and the WRAcc value, they found that with significance level $\alpha = 10\%$, which means that the significance threshold is 0.1, the subgroups needs to have a WRAcc of at least 0.054 to have a p-value below 0.1, resulting in a rejection of the null hypothesis. With a significance level $\alpha = 5\%$ a WRAcc value of at least 0.068 is needed to have a p-value below 0.05 resulting in a rejection of the null hypothesis and with a significance level $\alpha = 1\%$ a WRAcc value of at least 0.093 is needed to have a p-value below 0.01, resulting in a rejection of the null hypothesis. A higher confidence level requires a higher WRAcc value to obtain enough evidence against the null-hypothesis.

Unfortunately, this result can not be used during this bachelor project because we have used a tuned version of the WRAcc quality measure. We calculate the WRAcc value with a power of 1.1 instead of the default 0.5 which represent the importance of the coverage [DK11].

Coronary heart disease risk group detection Research has been done to the detection of heart diseases by patients. With the use of subgroup discovery they want to find relevant and interesting risk groups, meaning groups with an increased risk of heart diseases. They collected data in three different stages of 238 patients, which is almost equal to the amount of participants during this research. One of the stages is the collection of ECG data which is collected during rest. They derived different features from the ECG data including heart rate. Furthermore, they had some other parameters available such as BMI. They performed the subgroup discovery task using heuristic beam search as search strategy and the following quality measure:

$$q = \frac{TP}{FP + g}$$

Where g is the coverage of the subgroup.

One of the interesting subgroups found is the following rule: Body mass index over 25 kgm^{-2} AND age over 63 years. People satisfying this rule are part of a risk group with regard to heart diseases [LCGF04]. In this bachelor project ECG data, BMI and heart rate are also used for the subgroup discovery task. Subgroup discovery is used to obtain subgroups with an increased chance of having social anxiety instead of an increased chance of heart diseases.

5 Method

During this section we will briefly explain which steps are taken in order to perform this research. First of all, we will explain the data that is used. After that, we will state which preprocessing steps has been performed in order to improve the quality of the data. Then we will give a description of how the features are extracted from the data. Last but not least, we will explain how we used subgroup discovery and evaluate the obtained subgroups.

5.1 The data

The data has been provided by the Developmental and Educational Psychology of the Institute of Psychology at Leiden University. The data of 327 children (167 boys), with an age between 8 and 17 years, is collected during seven phases. These seven phases are the following:

1. The participants first watched a nature movie while seated (20 min)
2. Secondly they watched the movie while standing (5 min)
3. Next they received instructions which reminds them of their presentation and that it will be recorded and evaluated (3 min)
4. After that, they get the time to practice the presentation (5 min)
5. Next they went to an empty classroom. At the moment they were in the classroom the audience walks into the classroom and take their seats (1 min)
6. When the audience find their places, the participant was allowed to start his/her presentation about a movie they like the most or dislike the most (5 min)
7. After the presentation there was a recovery phase (30 min), in which they watched a nature film for 10 minutes.

After two years this task was repeated, where 243 participants (125 boys) of the previous task show up again. The data contains two types of data: time series data which is recorded during the seven different phases, and static data which is collected a week before the actual task.

5.1.1 Static data

The static data consists of ten properties, which are:

- A unique personal number: Every participant who came back for the second experiment were given the same personal number as in the first experiment.
- Gender: a binary number, 1 is a female and 2 is a man.
- Age: a numeric value between 8 and 17 years old.
- SocialAnxietyScale: a numeric value between 1 and 5. We measured this value on the basis of a questionnaire that is globally approved for measuring the social anxiety score of a person: the Pubertal Development Scale [PCRB88].
- PersonalReportPublicSpeakingAnxiety: a numeric value between 1 and 5. This is a value that indicates how much anxiety the participants have when talking in public. We also measured this value on the basis of a questionnaire that is globally approved for measuring the anxiety score of a person: the Pubertal Development Scale [PCRB88].
- Medication: a binary number, 1 is using medication and 2 is not using medication.
- BMI: a numeric value. The body mass index is used to display the ratio between height and weight. This value is often used to indicate whether a certain person is overweight or underweight.
- VisualAnalogueScale_nervous: a numeric value between 1 and 100. Every participant was asked to indicate how nervous, they think, they were by drawing a line on a ruler of 10 centimeter. A line drawn at 5.5 centimeters corresponds to a score of 55.
- VisualAnalogueScale_heartrate: a numeric value between 1 and 100. Every participant was asked to indicate what, they think, their heart rate was by drawing a line on a ruler of 10 centimeter.
- VisualAnalogueScale_sweating: a numeric value between 1 and 100. Every participant was asked to indicate how much, they think, they sweat by drawing a line on a ruler of 10 centimeter.

This data was collected one week before the participants had to perform the actual task. The data was stored as follows:

| geschlecht | AgeT1 | T1PreSASMean |
|------------|------------------|------------------|
| 1 | 13,8350444900753 | 1,22222222222222 |
| 1 | 16,9281314168378 | 2,16666666666667 |
| 2 | 9,4052019164956 | 4,11111111111111 |

Figure 5: This is an example of how the static data is formatted. This is only a fraction of all the attributes stored. In total there are 51 attributes.

5.1.2 Time series data

During the seven different phases, explained in section 5.1, the ECG data of every participant was collected. Because this time series data was stored in .txt format, it was very friendly to work with. Using this data several features are derived such as HRV, HR and amplitude.

5.2 Preprocessing

Before we could start looking for interesting subgroups, a number of preprocessing steps had to be performed. This was necessary to ensure that the data was of the right quality, so that the experiments became as reliable as possible.

5.2.1 Data cleaning

One of the most important steps within data mining is to ensure that the data you are using is of the right quality. To ensure this, it is important that the data is clean and that wrong fields do not occur in the data. During this bachelor project, three different types of incorrect fields were removed from the data.

Cleaning the numeric values While processing the ECG data, we discovered that there were some unexpected letters mixed up with the numerical values, as shown in figure 6. Because of this, python could not process these values as numerical values, so we had to remove these letters from the data.

```
53815.5 -0.234e536 0.59 11.6683
53815.5 -0.24072 0.59 m11.666
```

Figure 6: Characters within the time series data.

Replacing commas While putting the ECG data in a data frame, we found out that the numeric data in the data frame was split by a comma. This ensured that calculations using this numerical data could not be performed. For this reason, we had to change the commas by dots, which enabled Python to process it correctly.

Deletion of empty fields The moment the data was loaded into the data frame, we could analyze the data in a more easier way. Rows containing empty fields were detected during this process. These rows ensures that the entire process is influenced negatively, which forced us to remove these rows. This caused that only data from 237 of the 327 participants was in the data frame.

5.2.2 Peak detection

To ensure that the feature extraction step can be performed correctly, it is necessary to detect the peaks in the ECG data. Using these peaks, calculations can be performed to calculate the desired features. The maxima and minima are detected during this preprocessing step. This is achieved through the Python "find_peaks" library, where we had to set the "prominence" factor [com19]. This parameter indicates what the prominence of the peaks is allowed to be. The prominence indicates what the difference between the lowest and highest peak can be. Figure 7 will explain it in more detail. But since the peaks in the ECG data of different people have a different kind

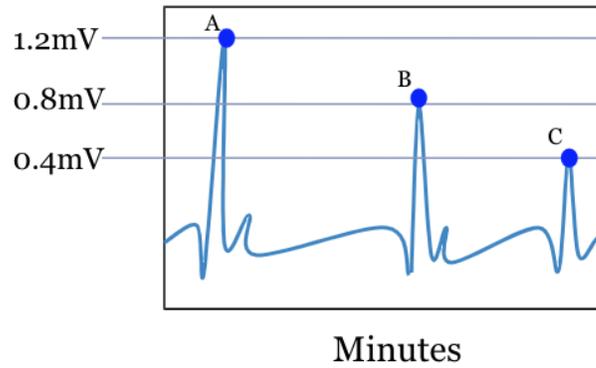


Figure 7: This figure is an example of how ECG can look like. When the prominence is 0.4, the distance between the current peak and the highest peak needs to be lower or equal to 0.4. This will result in the detection of peak A and B but not C. A prominence of 0.8 will result in the detection of peak A, B and C.

of shape, as shown in figure 9, a prominence of 1 is not always the optimum. For this reason a function has been written that checks which prominence is the most appropriate with the given ECG data. This function set the prominence to a default value of one, because in most cases this was the best value. After that, we calculate the average distance between two peaks in order to set a baseline. This baseline indicates whether some of the peaks remain undetected. We iterate through peaks to find a distance between two peaks which is greater than 1.5 times the average calculated distance. If the distance is greater than 1.5 times the average calculated distance we assume that there is probably a peak between these two peaks. At the moment we find such a distance, we decrease the prominence with 0.1 and repeat the previous explained steps. The moment no such distance is detected, we stop and use the current prominence. The complete code of this function can be found in the appendix.

After running the customized function, the peaks in the ECG data were stored per person in a data frame, which is visualized in figure 8. In this way the peaks that were found could be used to derive the desired features.

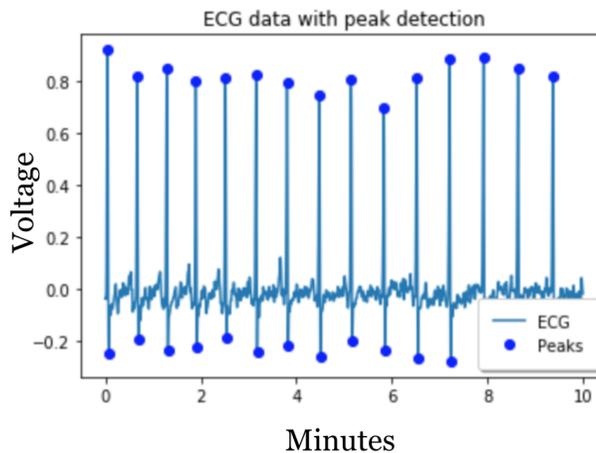


Figure 8: This is an example of how the peak detection works on ECG data. The blue dots are the maxima and minima of the peaks.

5.3 Feature extraction

Besides cleaning up the data, it is important to derive the right features. Because the features must contain the correct information so that there is enough information available. The results that are obtained are derived from these features, making it very important that these features contain the correct information. In this project four different types of features were calculated.

5.3.1 HRV

The heart rate variability has been explained in section 2. This feature represent the variability in the distance between two peaks. This is calculated as follows:

- First of all we iterate over `minValue` which is a variable that contains the timestamp of each peak in milliseconds.
- After that we calculate the distance between two consecutive peaks in milliseconds and append it to a variable called `HRV`.
- Then we iterate over `HRV` which contains all distances between to peaks in milliseconds.
- During this loop we add all differences between two consecutive distances, using the formula denoted in section 2.
- Last of all, we took the square root to get the RMSSD of the average variability. RMSSD is the root of the mean of the squared differences between two consecutive RR intervals.

While deriving the HRV feature, we came to the conclusion that the ECG data of some people was completely incorrect, so an extra check was added. This check verifies the length of the variable `HRV`. If the length is zero, this function returns 0 for all the features mentioned in section 2. The function ensures that the program does not crash if no peaks are found. The moment that this appears, the number 0 is returned for all features. Subsequently, these participants were also removed from the data frame since these data points negatively influence the results.

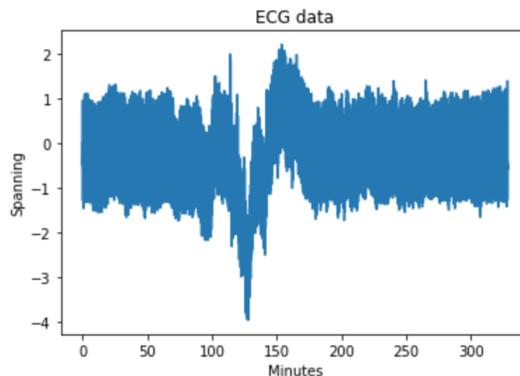


Figure 9: Example of ECG data that is completely incorrect. This shows that we had to remove data of participants in order to improve our results by removing measurement errors.

5.3.2 BPM

The Beats Per Minute shows the average number of heartbeats per minute. This is calculated as follows:

- First of all we iterate over `minValue` which is a variable that contains the timestamp of each peak
- After that we add the distance between two peaks in milliseconds to the distance variable
- Last of all we calculate the average distance between two peaks and the number of times the average distance fits into 60 seconds.

5.3.3 Amplitude

The amplitude has been explained in section 2. This is the average distance between the maximum and minimum of a peak. This is calculated as follows:

- First of all we iterate over `minValue` which is a variable that contains the timestamp of each `Maxpeak`. Then we iterate over `minValueminimum` which is a variable that contains the timestamp of each `Minpeak`.
- After that we check whether the maximum and minimum corresponds to the same peak. If so, the difference between the maximum and minimum will be added to the difference variable.
- Last of all, we divide the total difference by the total number of differences added.

5.3.4 Delta variables

As mentioned in section 5.1 two identical experiment has been performed two years after each other. During these two experiments the same static and time series data has been collected. Also the same features are derived for these two experiments. The delta variables represent the difference of a feature between the two measurements, measurement-1 part of the first experiment and measurement-2 part of the second experiment. The difference between the two measurements is calculated for all available features such as HRV, BPM, BMI, Medication etc. All the derived features for this project can be found in figure 12. With the use of these delta variables we can look how the different features are developed over a period of two years. In this way we hope to gain more insight into the development of the social anxiety score among adolescents.

5.4 DataFrames

Two types of data tables were used. A data table that contains the exact values as obtained. And a data table where the delta variables has been converted to binary values. Here, "0" stands for a decrease (-) in the value and a "1" for an increase (+) in the value. In this way we can see whether a fall / rise of a certain variable causes the social anxiety score to fall or rise. An example can be:

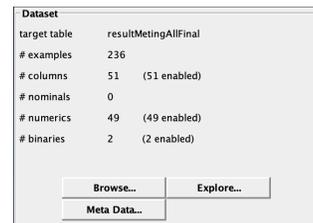
$$R_1 : (Gender = M \text{ AND } Age \leq 12.48 \text{ AND } \Delta BMI = 1) \rightarrow SAS = 1$$

An increase of the BMI of a man with an age below 12.49, we see an increase of the social anxiety score. We do not take into account the size of the change. This cause that a minimal change of 0.1 has the same affect as a change of 34.

5.5 Subgroup discovery

Cortana was used to execute the SD task. This is an open source program that was developed by Leiden University. Cortana is a data mining tool for finding local patterns in the data. This program can handle different types of data types, the input attributes and the target attributes can be both nominal, numerical and binary. Cortana has been used since this program has a wide range of settings, which depends on the type and number of target attributes. Furthermore, Cortana has a lot of settings that can be changed, for example, the search strategy and quality measure used for the subgroup discovery task [MK11]. In this subsection we will explain which settings are used and why these settings are used.

First, the obtained data frame is loaded into Cortana. The box Dataset shows information about the data frame that has been loaded. For example, we see the number of events in the data frame, namely 236. Furthermore, there is information about how many columns there are, namely 51, and how many of them are nominal, numerical or binary, namely 0, 49 and 2 respectively. In this part of the framework it is indicated which columns are and which are not included when searching for subgroups. Furthermore, the type of each feature can be adjusted to the desired type. During section 6, it will be indicated which features were and which were not used when finding the relevant subgroups.



| Dataset | |
|--------------|----------------------|
| target table | resultMetingAllFinal |
| # examples | 236 |
| # columns | 51 (51 enabled) |
| # nominals | 0 |
| # numerics | 49 (49 enabled) |
| # binaries | 2 (2 enabled) |

Buttons: Browse..., Explore..., Meta Data...

Figure 10: "Dataset" box in Cortana

In the box "Target Concept" it must be indicated which quality measure is used to rank the subgroups. During this assignment we use the Weighted Relative Accuracy. The formula used to calculate the CWRAcc is as follows:

$$CWRAcc = G^\alpha * (m_G - m_D)$$

In this formula m_D can have two meanings. When using a numeric target it means the average of the target in the entire data set. But when using a nominal target it means the number of ones in the entire data set. When using a numeric target m_G means the average of the target in the subgroup. But when using a nominal target it means the number of ones in the subgroup. Furthermore, G means the coverage of the subgroup.

The α as the power of G is a value that indicates how important the coverage of a subgroup is. The bigger α , the more important the coverage is. This will cause that subgroups with a high coverage, have a high quality measure because the difference between m_D and m_G is multiplied by a bigger number. The exact value of α will be indicated and explained in section 6. The exact value is determined by tuning α such that the average coverage of the subgroups meets the coverage mentioned by the expert. Furthermore, it must also be indicated which attribute is the target attribute. During this assignment, the target attribute is the social anxiety score.

It is also investigated which quality measure, a tuned version of the CWRAcc, a default version of the CWRAcc or the z-score, is the best in this project. For this research we use the average coverage of the subgroups of the previous mentioned quality measures. Secondly, we use the explanation of the subgroup size threshold mentioned by the expert. This will be briefly discussed

during section 6.

In the box "Search Conditions" it must be indicated what the search conditions are. The refinement depth can be adjusted here, this indicates how many splits are allowed. For example, a refinement depth of 3 looks like this:

$$R_1 : (Gender = M \text{ AND } Age \leq 12.48 \text{ AND } \Delta BMI = 1) \rightarrow SAS = 1$$

$$R_2 : (Gender = F \text{ AND } Age \leq 12.48 \text{ AND } Age \geq 10.48) \rightarrow SAS = 3.34$$

During this bachelor project we set the refinement depth to 8. This particular value was set because we did not get a subgroup in the top-100 with a refinement depth over 8.

Furthermore, it must be specified how many events must be covered by the relevant rule. This value was set to default because tuning α in the CWRAcc formula handles this aspect. Finally, you can indicate how many subgroups you want to get back and how much time Cortana has to find the subgroups. We set these values to 0, which means that no maximum is set, in order to not disturb the subgroup discovery task.

In the box "Search Strategy" the search strategy must be specified. One of the parameters that can be adjusted here is the strategy type. The top-down strategy is often used to find high-quality subgroups. This is because it first of all looks at simple rules, which are subsequently refined further and thereby become more specific. During this assignment top-down with Depth first search is used. This means that we start in the root, and then go further into the tree by choosing one line.

Even though there are multiple rules that can be applied at one time, we continue with one possibility. The moment we fail, we continue with the alternatives. You can also indicate how many "bins" must be created per attribute. The moment you have many numeric values, it is important to select the correct number of bins. If there are too many bins, it can happen that many redundant subgroups are created. This looks like this:

$$R_1 : (Gender = M \text{ AND } Age \leq 12.48) \rightarrow SAS = 1$$

$$R_2 : (Gender = M \text{ AND } Age \leq 12.48 \text{ AND } Age \leq 11.48) \rightarrow SAS = 1$$

$$R_3 : (Gender = M \text{ AND } Age \leq 12.48 \text{ AND } Age \leq 10.48) \rightarrow SAS = 1$$

Reducing the number of bins ensures that there are fewer possible split values of an attribute. During this bachelor project we made use of 4 bins, which caused for a minimization of redundant subgroups. Last but not least, we can indicate how many threads it will use. During this experiment we found out that setting the thread value above 1, deviation is caused in the subgroups that were found. Because of this, we set the thread value to 1.

By pressing subgroup discovery, subgroups in the data are searched using the specified columns/features. The result is a list of subgroups that have been found, as shown in figure 11. The following is indicated for each subgroup:

- Depth: How many splits have been made resulting in the given rule.

- Coverage: How many participants satisfy the rule of the subgroup.
- Quality: Based on the quality measure indicated
- Average: Average social anxiety score of the subgroup. This can be compared with the average of the entire population. For example, it is possible to see whether the subgroup has a higher social anxiety score or a lower social anxiety score compared to the entire population
- Standard deviation: Indicates how much it deviates from the entire population
- Conditions: This displays the subgroup rule

The subgroup discovery task will be performed three times with different attributes selected. The first task will be performed using the attributes of the first experiment on T1. The second task will be performed using the attributes of the second experiment on T2, two years after experiment 1. The third task will be performed using the binary delta attributes. For each SD task the top-5 subgroups are used to answer the questions mentioned in section 1. Last but not least, for the top-1 subgroups of each SD task will the expert give an interpretation of how they can use these subgroups in the psychology.

| Nr. | Depth | Coverage | Quality | Average | St. Dev. | p-Value | Conditions |
|-----|-------|----------|----------|----------|----------|---------|--|
| 1 | 2 | 93 | 0,205927 | 1,863799 | 0,559871 | - | VisualAnalogueScale_nervous4a <= 87.0 AND T1prePRPSAMean <= 2.44... |
| 2 | 1 | 119 | 0,20586 | 2,573763 | 0,731202 | - | T1prePRPSAMean >= 2.5789473 |
| 3 | 3 | 108 | 0,204525 | 1,937757 | 0,615942 | - | VisualAnalogueScale_nervous4a <= 87.0 AND T1prePRPSAMean <= 2.89... |
| 4 | 3 | 107 | 0,201563 | 2,484423 | 0,652356 | - | T1prePRPSAMean >= 2.2105262 AND T1prePRPSAMean >= 2.4210527... |
| 5 | 3 | 109 | 0,198163 | 2,62793 | 0,704479 | - | T1prePRPSAMean >= 2.2105262 AND T1prePRPSAMean >= 2.4210527... |
| 6 | 4 | 85 | 0,197675 | 2,546406 | 0,650616 | - | T1prePRPSAMean >= 2.2105262 AND T1prePRPSAMean >= 2.4210527... |
| 7 | 1 | 124 | 0,196723 | 1,978943 | 0,667547 | - | T1prePRPSAMean <= 2.5789473 |
| 8 | 3 | 112 | 0,196723 | 2,573909 | 0,734533 | - | T1prePRPSAMean >= 2.2105262 AND T1prePRPSAMean >= 2.4210527... |
| 9 | 2 | 94 | 0,196467 | 1,867021 | 0,634633 | - | T1prePRPSAMean <= 2.5789473 AND T1prePRPSAMean <= 2.368421 |
| 10 | 2 | 142 | 0,196467 | 2,5223 | 0,723697 | - | T1prePRPSAMean >= 2.2105262 AND T1prePRPSAMean >= 2.4210527 |
| 11 | 2 | 90 | 0,195668 | 2,558642 | 0,678464 | - | T1prePRPSAMean >= 2.5789473 AND HRVrust >= 36.701283 |
| 12 | 3 | 101 | 0,195128 | 2,588559 | 0,748187 | - | T1prePRPSAMean >= 2.2105262 AND Amplitudebefore >= 1.8172133 ... |
| 13 | 2 | 108 | 0,193855 | 2,573046 | 0,732385 | - | T1prePRPSAMean >= 2.5789473 AND medication = '0' |
| 14 | 4 | 84 | 0,193649 | 2,542328 | 0,653396 | - | T1prePRPSAMean >= 2.2105262 AND T1prePRPSAMean >= 2.4210527... |
| 15 | 2 | 91 | 0,193265 | 2,620879 | 0,753574 | - | T1prePRPSAMean >= 2.5789473 AND VisualAnalogueScale_hearttrate4b ... |
| 16 | 3 | 84 | 0,192669 | 1,864418 | 0,554718 | - | VisualAnalogueScale_nervous4a <= 87.0 AND T1prePRPSAMean <= 2.44... |
| 17 | 4 | 83 | 0,19247 | 1,845382 | 0,580373 | - | VisualAnalogueScale_nervous4a <= 87.0 AND T1prePRPSAMean <= 2.89... |
| 18 | 4 | 97 | 0,192151 | 1,945017 | 0,607717 | - | VisualAnalogueScale_nervous4a <= 87.0 AND T1prePRPSAMean <= 2.89... |
| 19 | 4 | 97 | 0,192151 | 1,945017 | 0,607717 | - | VisualAnalogueScale_nervous4a <= 87.0 AND T1prePRPSAMean <= 2.89... |
| 20 | 2 | 95 | 0,190653 | 1,918129 | 0,615737 | - | T1prePRPSAMean <= 2.5789473 AND VisualAnalogueScale_nervous4a <... |

Figure 11: A list of subgroups which is the result of the subgroup task. This gives an indication of how the properties of the subgroups are returned.

5.6 Statistical measurement: p-value

The p-value is a measurement that is used to find whether a given result is significant or not. This represent the probability of a given event, in this bachelor project the subgroups. The probability indicates if there is an observed effect or correlation that the subgroup is unlikely to be based on chance. The statistical test consists of a null-hypothesis and an alternative hypothesis. The smaller the p-value, the stronger evidence there is in favor of the alternative hypothesis. Most of the time a threshold of 0.05 is used. A p-value below 0.05 will lead to a rejection of the null-hypothesis and an acceptance of the alternative hypothesis. A p-value above 0.05 indicates that there is weak evidence against the null-hypothesis, the null-hypothesis is in this case not rejected nor accepted.

In this paper, the null-hypothesis is formulated as follows: The subgroup is a false discovery. A false discovery is an obtained subgroup that is not really a subgroup but just a coincidence. A p-value below 0.05 means that there is enough evidence against the null-hypothesis which means that the subgroup is significant and not only based on luck. A p-value above 0.05 means that there is weak evidence against the null-hypothesis which means that the subgroup is not significant [DK11].

Cortana has only the possibility to calculate the p-value for subgroups with a nominal target. Because of this, we cannot calculate the p-value of the subgroups with a numeric target. Our approach for this problem is as follows:

- First of all, we will calculate the p-value of the top-100 subgroups with a nominal target.
- After that, we will investigate till where the subgroups are significant, having a p-value below 0.05.
- Last but not least, we will use this result to say something about the significance of the subgroups with a numeric target. If, for example, the top-46 is significant, we assume that the top-5 of the subgroups with a numeric target are significant as well.

The Results section explains in detail how the settings are set to obtain the given results. All the derived features for this project can be found in figure 12.

| | Attribute | Cardinality | Type |
|--|-----------------------------------|-------------|---------|
| Attributes containing data of experiment 1 | ppn | 233 | numeric |
| | geschlecht | 2 | numeric |
| | AgeT1 | 220 | numeric |
| | T1PreSASMean | 56 | numeric |
| | T1prePRPSAMean | 62 | numeric |
| | Amplituderust | 205 | numeric |
| | BPMrust | 208 | numeric |
| | HRVrust | 208 | numeric |
| | Amplitudespeech | 219 | numeric |
| | BPMspeech | 219 | numeric |
| | HRVspeech | 219 | numeric |
| | Amplitudebefore | 213 | numeric |
| | BPMbefore | 214 | numeric |
| | HRVbefore | 214 | numeric |
| | medication | 2 | binary |
| Attributes containing data of experiment 2 | BMI | 217 | numeric |
| | VisualAnalogueScale_nervous4a | 87 | numeric |
| | VisualAnalogueScale_hearttrate4b | 87 | numeric |
| | VisualAnalogueScale_sweating4c | 85 | numeric |
| | T3SASMean | 52 | numeric |
| | T3prePRPSAMean | 59 | numeric |
| | Amplituderust2 | 208 | numeric |
| | BPMrust2 | 208 | numeric |
| | HRVrust2 | 208 | numeric |
| | Amplitudespeech2 | 218 | numeric |
| | BPMspeech2 | 218 | numeric |
| | HRVspeech2 | 217 | numeric |
| | Amplitudebefore2 | 210 | numeric |
| | BPMbefore2 | 213 | numeric |
| | HRVbefore2 | 214 | numeric |
| medication2 | 2 | binary | |
| The delta attributes | BMI2 | 215 | numeric |
| | VisualAnalogueScale_nervous4a2 | 95 | numeric |
| | VisualAnalogueScale_hearttrate4b2 | 87 | numeric |
| | VisualAnalogueScale_sweating4c2 | 89 | numeric |
| | deltaHRVrust | 208 | numeric |
| | deltaBPMrust | 208 | numeric |
| | deltaAmplituderust | 208 | numeric |
| | deltaHRVbefore | 214 | numeric |
| | deltaBPMbefore | 214 | numeric |
| | deltaAmplitudebefore | 214 | numeric |
| | deltaHRVspeech | 219 | numeric |
| | deltaBPMspeech | 219 | numeric |
| | deltaAmplitudespeech | 219 | numeric |
| | deltaMedication | 3 | numeric |
| | deltaBMI | 233 | numeric |
| deltaVisualAnalogueScale_nervous4a | 100 | numeric | |
| deltaVisualAnalogueScale_hearttrate4b | 103 | numeric | |
| deltaVisualAnalogueScale_sweating4c | 96 | numeric | |
| deltaPreSASMean | 55 | numeric | |
| deltaPrePRPSAMean | 71 | numeric | |

Figure 12: This table shows all the features derived from the questionnaire or ECG data. In total there are 51 features available for the subgroups discovering task.

6 Results

This section will use experiments to answer the following questions:

1. Can we tune the alpha such that we obtain a desirable trade-off between subgroup size and the deviation of the target?
2. Is it better, in the purpose of this thesis, to use a tuned version than a default version of the quality measures?
3. Statistically, are the subgroups significant or just a coincidence, meaning is there an observed effect or correlation that indicates that the obtained subgroup is unlikely to be based on luck.
4. Which insights can be derived from the subgroups so that the psychological institute can get a better understanding of adolescents?

6.1 Experiment setup

During these experiments we will mainly be focused on the trade-off between the deviation of the target value and the subgroup size. By tuning the default WRAcc quality measure, we were able to get the right trade-off between these two values. To find the right α we will use the explanation of an expert. When the expert indicates that the coverage of the subgroups have to be at least 60, we have to tune the α such that the coverage of the subgroups are at least 60. In order to derive the right α , taking into account the argumentation of the expert, we measured two values:

- The average coverage of the top-5 subgroups for different values of alpha, in which α has a range of 0 to 2.
- The coverage of the top-1 subgroup for different values of alpha, in which α has the same range. We made this second measurement because we want to investigate what the effect of changing α is on the best subgroups.

In order to compare the tuned version of the CWRAcc to a default version of CWRAcc and the z-score, we had to measure the average target value and the average coverage per quality measure. With these measurements we can investigate whether a tuned version of the CWRAcc, the default version of the CWRAcc or the z-score is the best quality measure regarding to the explanation of the expert. The measurements are obtained using the social anxiety score at T1 as target value. Furthermore, the attributes containing data of T1, that is collected during the first of the two experiments as mentioned in section 5.1, are used as descriptors. Only the T1PrePRPSMean and VisualAnalogueScale_nervous has been excluded. These two attributes have a too high correlation to the target attribute.

For filtering the obtained subgroups we used the p-value for the statistical test. Unfortunately, this test can only be performed on the delta variables. In order to say something about the obtained subgroups with numeric values, we have calculate the p-value of the top-100 subgroups found with a nominal target. When the subgroups obtained using the delta variables are significant based on the p-value, we generalize it to the rest of the subgroups for which we cannot do this. For example, when we find that the top-46 subgroups are significant we can say that the top-5 subgroups having

a numeric target are also significant. For this experiment, the delta social anxiety score is used as target value. Furthermore, the delta attributes, which can be found in figure 12, are used as descriptors. Only the deltaPrePRPSMean and deltaVisualAnalogueScale_nervous has been excluded. These two attributes have a too high correlation to the target attribute.

After filtering the results, leaving only the subgroups who made it through the test, we will give a possible explanation and an interpretation of the subgroups. The top-3 will also be explained by an expert in this field.

6.2 Tuning alpha

This subsection gives answer to the following question:

Can we tune the alpha such that we obtain a desirable trade-off between subgroup size and the deviation of the target?

First of all, we implemented the tune-able CWRAcc formula in Cortana. This gives us the possibility to run the exact same subgroup discovery task but with a changing value of α . In order to know which α gives us the desirable trade-off, we have used the knowledge of an expert. The expert came up with the following explanation:

Social anxiety disorder occurs in around 3 percent of the adolescents. This means that a subgroup size of 10 can already be very interesting, but will lead to many subgroups which are not all interesting. Social anxiety is usually seen as a spectrum instead of something that you do or do not have. This means that it is possible that people without a disorder can behave like a person with a social anxiety disorder, but this is just a person who ends up high in the spectrum who does not have any issues in everyday life. With this in mind, the most liberal way to define a highly anxious group is to take the top 33 percent. But then there is a good chance that there are also people in the group who are not particularly socially anxious. Instead of 33 percent, we take 25 percent of the sample size as the minimal coverage of a subgroup. This will lead to a minimal coverage of 60, 25 percent of 237 adolescents.

We first set α on 0 and then we increase the value by 0.2 each time till the value reaches 2. Only between α is 1 and 1.2 we increased the value by 0.1, because the average coverage obtained using α is 1 is too low given the explanation of the expert. But using α is 1.2 the average coverage is too high. For each run, we calculated the average coverage of the top-5 subgroups and the coverage for the top-1 subgroup. This resulted in the visualization that can be seen in figures 13 and 14.

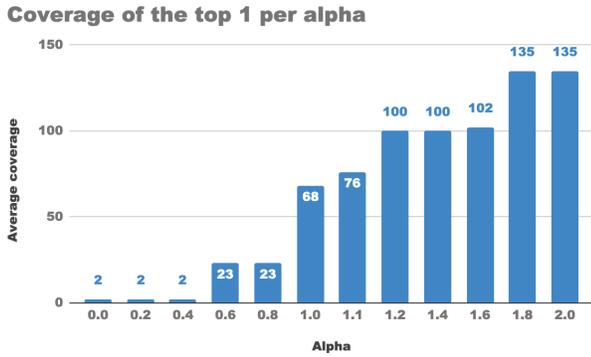


Figure 13: The coverage of the top-1 subgroups for different values of alpha, using the exact same attributes. Target attribute is the social anxiety score at T1 and the attributes containing data of T1 are the descriptors.

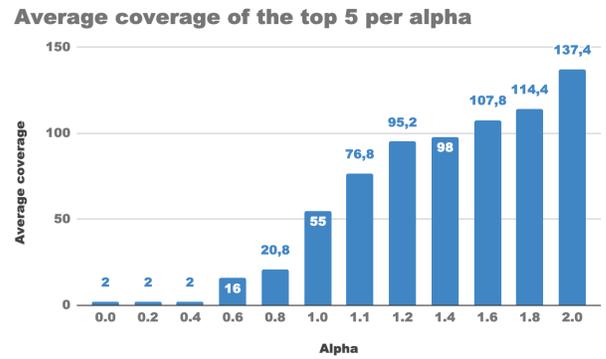


Figure 14: The average coverage of the top-5 subgroups for different values of alpha, using the exact same attributes. Target attribute is the social anxiety score at T1 and the attributes containing data of T1 are the descriptors.

We finally opted for an α value of 1.1 which will lead, using figures 13 and 14, to an average coverage of 76.8. This decision was made by an expert in the psychological domain. We will compare this value to the default version of the CWRAcc and the z-score in section 6.3.

6.3 Tuned vs default measures

This subsection gives answer to the following question:

Is it better, in the purpose of this thesis to use a tuned version than a default version of the quality measures?

In this experiment we compared three different quality measures: the default CWRAcc, Z-score, and the tuned CWRAcc. The tuned CWRAcc is the formula using α is 1.1 as mentioned in the previous section. The default version of CWRAcc uses α is 0.5. For each subgroup, we collected the target value. After that, we took the average of the 5 target values collected. We also collected the coverage of the 5 obtained subgroups and took the average of these five values. This has been performed for the three quality measures mentioned above. The results can be found in figures 15 and 16.

As we can see in figures 15 and 16, the tuned version of the CWRAcc is the only quality measure that returns subgroups with a size around 60. The Z-score and the default CWRAcc return subgroups with a very low subgroup coverage, but a large average of the target value. The average of the target "social anxiety score T1" for the entire data set is 2.26 which means that the Z-score and the default CWRAcc returns subgroups that have a large difference in average of the target. As mentioned in the previous section the expert explained that the minimal coverage have to be around 60 which lead to the use of the tuned CWRAcc but when one want subgroups with a large difference in the average of the target, the Z-score or default CWRAcc is the right quality measure. In conclusion, it is indeed better, for the purpose of this thesis, to use a tuned version instead of a default version of the quality measures. One can tune the quality measure so that the desired trade-off is made between the size of the subgroups and the average of the subgroups.

Average of the target value of the top-5 subgroups for different quality measures

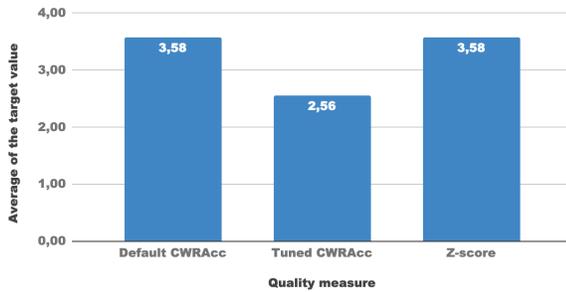


Figure 15: The average of the average target value of the top-5 subgroups for different quality measures. For each subgroup, we collected the average target value. After that, we took the average of the 5 average target values collected. Target attribute is the social anxiety score at T1 and the attributes containing data of T1 are the descriptors.

Average coverage of the top-5 subgroups for different quality measures

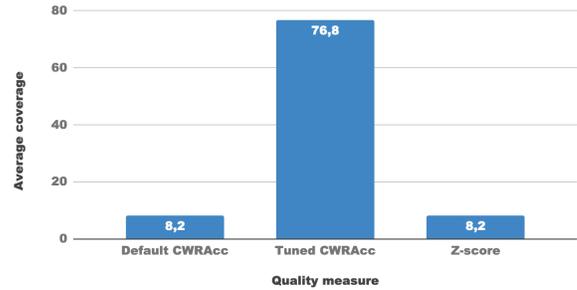


Figure 16: The average coverage of the top-5 subgroups per quality measure. Target attribute is the social anxiety score at T1 and the attributes containing data of T1 are the descriptors.

6.4 Filtering the results: Statistical tests

This subsection gives answer to the following question:

Statistically, are the subgroups significant or just a coincidence, meaning is there an observed effect or correlation that indicates that the obtained subgroup is unlikely to be based on luck.

In order to evaluate the subgroups whether they are significant or not, the p-value is calculated for each subgroup. The platform we are using, Cortana, can only calculate the p-value for nominal values. Because of this, we generalize the outcome of whether the subgroups are significant or not to the numeric values. In this experiment we will calculate the p-value for the top-100 subgroups found using the nominal social anxiety score as the target attribute which is explained in section 5.4. Based on the number of the top-100 subgroups having a p-value below 0.05, we will assume whether or not the top-5 subgroups found using a numeric target are significant as well. This means that if, for example, the top-46 subgroups with a nominal target are significant, we could say that the top-5 subgroups we have found using a numeric target are also significant as mentioned in section 5. We calculated the p-value for the nominal subgroups by using 200 random subsets of the entire data set in order to compute random quantities which is a setting in Cortana to calculate the p-value.

The result of this experiment was that at least the top-100 subgroups were significant, having a p-value below 0.05. As explained in section 2, a p-value below 0.05 means that there is enough evidence to reject the null-hypothesis. The result is generalised to the subgroups with numeric target, meaning that we treat the top-5 subgroups obtained using a numeric target as significant as well. During this project we have not studied the entire top-100 comprehensively because it did not fall within the time assigned to this project. We have added the full top-100 to the appendix. If we look at this very quick, we can see that there is a lot of overlap. But further research to the top-100 subgroups falls outside the scope of this project.

| Nr. | WRAcc | p-value |
|-------------|-------|---------|
| Subgroup 10 | 14,03 | 0.0052 |
| Subgroup 11 | 13,73 | 0.0061 |
| Subgroup 12 | 13,06 | 0.0086 |
| Subgroup 13 | 13,00 | 0.0091 |
| Subgroup 14 | 12,69 | 0.0100 |

Table 1: The p-values of the top-5 subgroups having a nominal target. The subgroup number as shown in the figure are related to the subgroup with the same number in the appendix.

6.5 Interpreting the results

This subsection gives answer to the following question:

Which insights can be derived from the subgroups so that psychologists can get a better understanding of adolescents?

During this experiment we took the top-5 subgroups of each of the following three setups resulting in a total of 15 subgroups:

1. Using the social anxiety score at T1 as target attribute and the attributes of T1 as descriptors, see figure 12 for the full list. The T1PrePRPSAMean and VisualAnalogueScale_nervous are excluded.
2. Using the social anxiety score at T2 as target attribute and the attributes of T2 as descriptors, see figure 12 for the full list. The T3PrePRPSAMean and VisualAnalogueScale_nervous are excluded.
3. Using the delta social anxiety score as target attribute and the delta attributes as descriptors, see figure 12 for the full list. The deltaT1PrePRPSAMean and deltaVisualAnalogueScale_nervous are excluded.

The results of this experiment can be seen in the appendix. During this subsection we will use for example "Subgroup 1" which reference to the subgroup rule with subgroup number 1 in the appendix. In other words, subgroup 1 in the appendix is the same as subgroup 1 mentioned in this section. In this way you can look up the exact rule mentioned in the text.

General information

As we can see in figures 17 and 18, the average social anxiety score of the best subgroups do not have a large deviation from the average of the entire population. The reason for this minimal deviation is the tuned version of the CWRAcc, which says that the coverage of the subgroup is much more important than in the default version. As shown in the previous subsection, the default CWRAcc returns subgroups with an average social anxiety score of 3,58. In figure 19 we see that probabilityG is around 15% higher in comparison with probabilityD. This means that people satisfying one of the 5 subgroup rules, using the first setup, have on average around 15% more chance of a higher social anxiety score. As mentioned before, we do not take into account the size of the change which could be done in further research.

| Subgroup nr. | Coverage | CWRAcc | AverageD | AverageG | St. Dev |
|--------------|----------|--------|----------|----------|---------|
| 1 | 76 | 34,74 | 2,26 | 2,56 | 0.75 |
| 2 | 100 | 34,22 | 2,26 | 2,48 | 0.76 |
| 3 | 76 | 33,80 | 2,26 | 2,55 | 0.76 |
| 4 | 75 | 33,64 | 2,26 | 2,55 | 0.73 |
| 5 | 57 | 33,20 | 2,26 | 2,65 | 0.69 |

Figure 17: Statistics for the top-5 subgroups of the first experiment setup with a numeric target. The subgroup number corresponds to the subgroup in the appendix with the same subgroup number. The averageD is the average of the target in the entire data set. The averageG is the average of the target in the subgroup. The standard deviation is the variability of the target attribute values in the subgroups. Subgroup nr. 1 till 5 uses setup 1 mentioned above.

| Subgroup nr. | Coverage | CWRAcc | AverageD | AverageG | St. Dev |
|--------------|----------|--------|----------|----------|---------|
| 6 | 57 | 27,93 | 2,16 | 2,48 | 0.68 |
| 7 | 51 | 27,03 | 2,16 | 2,51 | 0.69 |
| 8 | 52 | 27,02 | 2,16 | 2,51 | 0.71 |
| 9 | 68 | 26,76 | 2,16 | 2,41 | 0.70 |

Figure 18: Statistics for the top-5 subgroups of the second experiment setup with a numeric target. The subgroup number corresponds to the subgroup in the appendix with the same subgroup number. The averageD is the average of the target in the entire data set. The averageG is the average of the target in the subgroup. Subgroup nr. 6 till 9 uses setup 2 mentioned above.

| Subgroup nr. | Coverage | WRAcc | ProbabilityD | ProbabilityG | P-value |
|--------------|----------|-------|--------------|--------------|---------|
| 10 | 77 | 14,03 | 45,3% | 57% | 0.0052 |
| 11 | 100 | 13,73 | 45,3% | 54% | 0.0061 |
| 12 | 58 | 13,06 | 45,3% | 60% | 0.0086 |
| 13 | 65 | 13,00 | 45,3% | 59% | 0.0091 |
| 14 | 54 | 12,69 | 45,3% | 61% | 0.010 |

Figure 19: Statistics for the top-5 subgroups of the third experiment setup with the delta variables. The subgroup number corresponds to the subgroup in the appendix with the same subgroup number. The first probability is the change of having an increasing target in the entire data set. The second probability is the change of having an increasing target in the subgroup. Subgroup nr. 10 till 14 uses setup 3 mentioned above.

In figure 20, the coverage with its corresponding average target value of the top-5 subgroups are displayed. We see that a higher value of the average target value results in a lower coverage. This means that subgroups with a lower coverage needs to have a higher average target value in order to end up high in the subgroup list. This is understandable since a lower coverage must be compensated by a larger deviation in the average target value. Reminding that the formula has two factors, the deviation of the average target value and the coverage. When we also look at figure 21, we see that the fluctuation of the quality score over the average social anxiety score is very low. This strengthens our statement mentioned above, because figure 20 and 21 together shows that a lower coverage must be compensated by a higher average target value in order to get an equal quality score.

Using figure 22, we can see that a higher standard deviation of the top-5 subgroups found using setup 1 leads to a higher coverage. This is quite understandable because with a higher coverage there is more chance of outliers that cause the standard deviation to increase. In figure 23, we can see that the standard deviation of the subgroups also increase in order to keep the quality score on the same level. The subgroups of T2 have a lot more fluctuation in the standard deviation, which may have been caused by environmental factors. Or because there are more older people in the data set, every participant has become 2 years older. The average age of the first experiment is 13.3 and of the second experiment 15.3, which is much higher.



Figure 20: Coverage over average target value for the top-5 subgroups using setup 1. The blue line corresponds to the subgroups 1 till 5. The red line corresponds to the subgroups 6 till 9. In this figure, the coverage and the average target value of the top-5 subgroups are displayed in a line-diagram.



Figure 21: CWRAcc over average target value for the top-5 subgroups using setup 1. The blue line corresponds to the subgroups 1 till 5. The red line corresponds to the subgroups 6 till 9. In this figure, the CWRAcc and the average target value of the top-5 subgroups are displayed in a line-diagram.

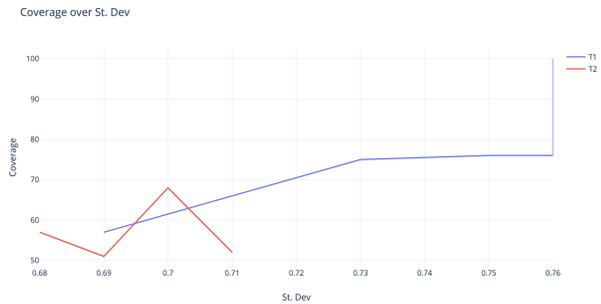


Figure 22: Coverage over standard deviation of the target attribute for the top-5 subgroups using setup 1. The blue line corresponds to the subgroups 1 till 5. The red line corresponds to the subgroups 6 till 9. In this figure, the coverage and the standard deviation of the top-5 subgroups are displayed in a line-diagram.



Figure 23: CWRAcc over standard deviation of the target attribute for the top-5 subgroups using setup 1. The blue line corresponds to the subgroups 1 till 5. The red line corresponds to the subgroups 6 till 9. In this figure, the CWRAcc and the standard deviation of the top-5 subgroups are displayed in a line-diagram.

Qualitative evaluation: Subgroups of T1

During this experiment setup we have used all parameters containing data of T1, except the visualAnalogueScale_nervous and T1PrePRPSMean because these parameters have a too high correlation with the target. The average social anxiety score of the entire population is 2,26 which means that the data is well distributed around the average of the range 1 to 5 which is 2.5.

The results of this experiment, shown in figure 32 in the appendix, shows that the heart rate is a very important parameter with regard to the social anxiety score. The visualAnalogueScale_hearttrate, BPMbefore and BPMspeech are present in almost all of the top-5 subgroups found. The top-1 subgroup tells us that people of an age above 10.7 with a visualAnalogueScale_hearttrate above 42, a heart rate before the presentation above 85 and a heart rate during the speech below 125 has a higher average social anxiety score, 2.56, relative to the average of the entire population. The coverage of this subgroup is 76, containing more participant with a high social anxiety score than with a low social anxiety score as shown in figure 26 below. The expert has also provided an interpretation of this subgroup. She came up with the following interpretation:

A heart rate before the presentation above 85 hardly excludes people. Figures 24 and 25 shows that there are almost no participants having a BPM below 85. A value of less than 25 occurs in BPM speech T2 which is not biologically plausible. This is probably

a measurement error. A heart rate during the speaking task below 125 indicates that the participants do not have an extreme heart rate. However, they do report a medium to high heart rate. This could be related to previous findings that people with social anxiety overestimate their physical stress responses. There are indications that they pay more attention to physical stress reactions, because they are afraid to show these symptoms during a speaking task and see this as failure, and therefore report that they experience physiologically more incentives than people who are less fixated on this. These criteria therefore do not seem to indicate a possible cause of social anxiety, but a phenomenon associated with it. The age criterion indicates that the participants satisfying this rule are not the youngest participants. This corresponds to an increase in social anxiety in the transition from child to adolescent.

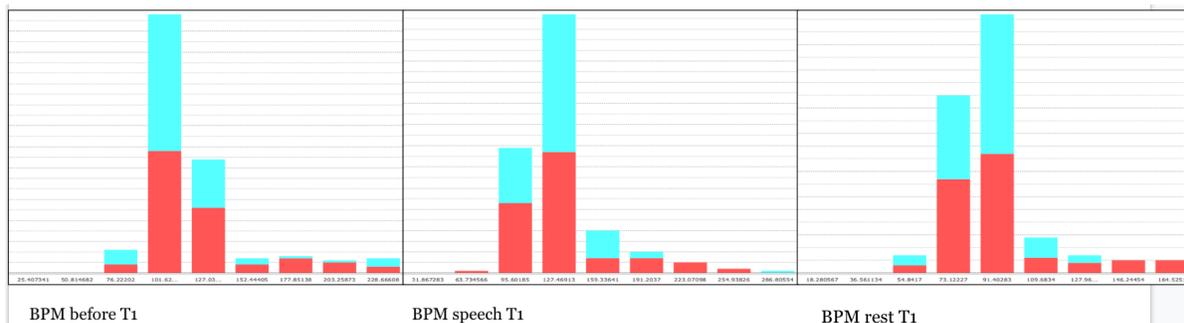


Figure 24: The distribution of the BPM feature for male and female at T1 for the entire data set. The red bar is male, and the blue bar is female.

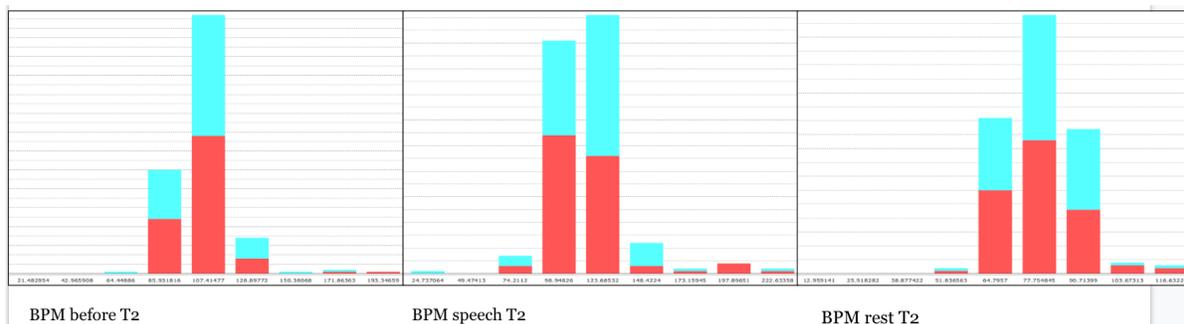


Figure 25: The distribution of the BPM feature for male and female at T2 for the entire data set. The red bar is male, and the blue bar is female. Note that the numbers on the x-axis are not the same as figure 24.

Furthermore, we see that HRV_{speech} is also an important parameter. The fifth subgroup corresponds to the fifth subgroup in the appendix, with a coverage of 57, tells us that people of an age above 10.7 with among other things an HRV_{speech} below 65 has a higher average social anxiety score, 2.65, relative to the average of the entire population. This observation confirms earlier research that a lower heart rate variability is associated with a higher social anxiety score [AQK+13]. The distribution of this subgroup relative to the distribution of the entire population can be seen in figure 27. The fourth subgroup corresponds to the fourth subgroup in the appendix, with a coverage of 75, also support this finding that a lower heart rate variability is related to a higher social anxiety score. This subgroup tells us that people of an age above 10.7 with among other things an HRV_{before} below 84 has a higher average social anxiety score, 2.55, relative to the average of the entire population.

Summarizing, the HRV and HR are two useful parameters in order to say something about the social anxiety score. A lower HRV is associated with a higher social anxiety score, and a higher HR is associated with a higher social anxiety score. But we see that participants with a higher social anxiety score do not have a large rise in the heart rate during the speaking task. The first subgroup shows this by saying that the BPMspeech is below 125. Furthermore, we see that the age in the top-5 subgroups is always above 10.7 years old which means that the social anxiety score is higher for adolescents above an age of 10.7. This corresponds to an increase in social anxiety in the transition from child to adolescent. A possible reason for this may be that a person becomes more aware of what is happening, which can result in that a person experience more anxiety. At the age of 11, children are often in the last year of primary school or in the first year of secondary school. If I may speak from my own experience, around this age I became more aware of what was happening. Because of this I got more nerves during presentations on secondary school than before. You were often afraid of being embarrassed, which caused you to experience more anxiety.

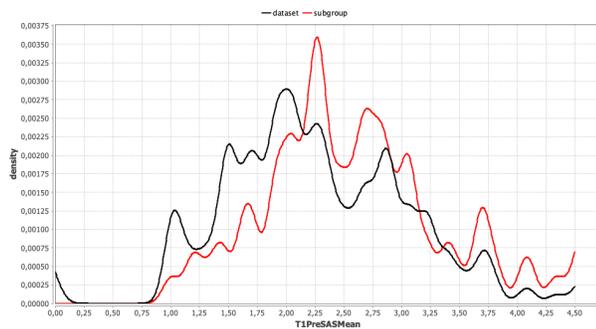


Figure 26: The distribution of subgroup 1, obtained using the first setup, relative to the distribution of the entire population using the "show model" option in Cortana.

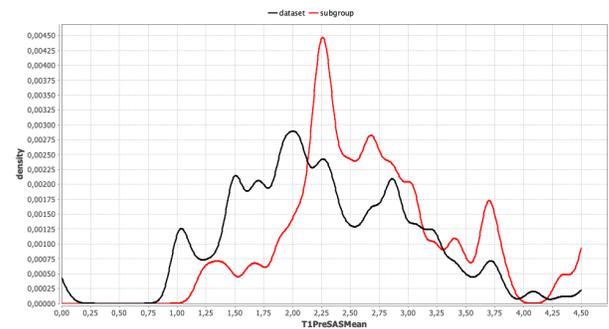


Figure 27: The distribution of subgroup 5, obtained using the first setup, relative to the distribution of the entire population using the "show model" option in Cortana.

Qualitative evaluation: Subgroups of T2

During this experiment setup we have used all parameters containing data of T2, except the visualAnalogueScale_nervous and T2PrePRPSMean because these parameters have a too high correlation with the target. The average social anxiety score of the entire population is 2,16 which is a little bit lower than the average on T1. This means that in general the social anxiety score has been decreased in a time period of two years. The results of this experiment, shown in figure 33 in the appendix, shows a different result relative to the experiment using parameters containing data of T1. This means that older children have other parameters that are useful regarding to the social anxiety score.

The results of this experiment shows that the heart rate is an important parameter with regard to the social anxiety score. The visualAnalogueScale_hearttrate and BPMrust are present in almost all of the top-4 subgroups found. Furthermore, the amplitude is also a very important parameter. The top-1 subgroup found, corresponds to the sixth subgroup in the appendix, tells us that people of an age between 12.3 and 16.5 with a visualAnalogueScale_hearttrate above 33, an amplituderust above 1.12 and an amplitudebefore above 1.29 has a higher average social anxiety score, 2.48, relative to the entire population. The expert had the following possibility for this phenomenon:

This group is difficult to interpret. The amplitude exclude very few people shown in

figure 28 and 29. In figure 28 nobody is excluded by this rule. The lower-bound of the age criterion corresponds to the increase in social anxiety in the transition from child to adolescent. The upper-bound is difficult to interpret since it is unclear in the literature whether a decrease of social anxiety will follow later. Furthermore, the finding seems to suggest that social anxiety in this group is only increased among adolescents who do experience some increase in their heart rate during a speaking task. They say that young people, who experience a very low heart rate during a speaking task ($VAS < 33$), do not seem to suffer from social anxiety.

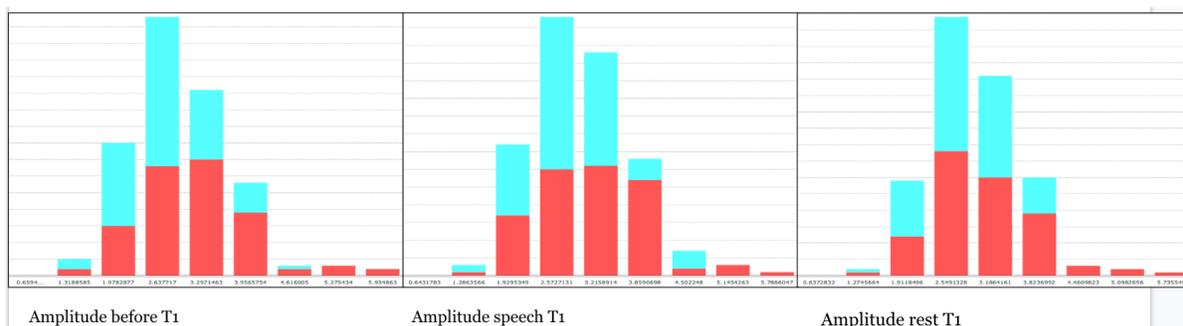


Figure 28: The distribution of the Amplitude feature for male and female at T1 for the entire data set. The red bar is male, and the blue bar is female.

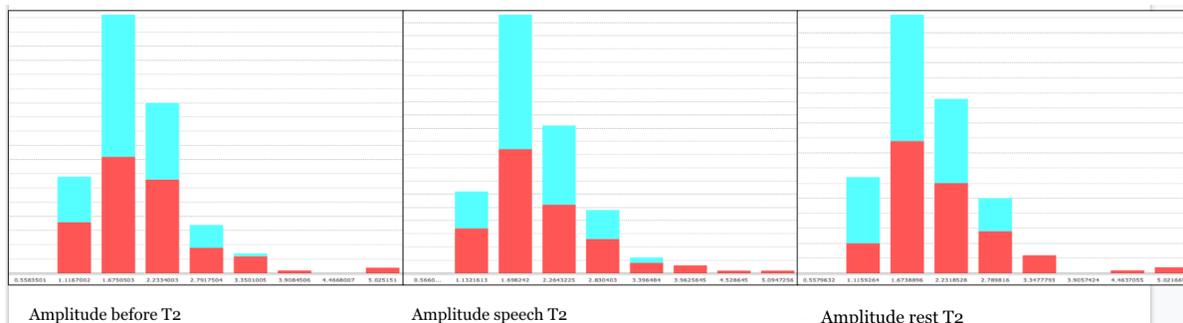


Figure 29: The distribution of the Amplitude feature for male and female at T2 for the entire data set. The red bar is male, and the blue bar is female. Note that the numbers on the x-axis are not the same as figure 24.

The second, third and fourth subgroup corresponds to the seventh, eighth and ninth subgroup in the appendix, shows that the medication parameter is also important, having a value of zero. This means that the participants do not take any medicines. Summarizing, the amplitude is not an useful parameter in order to tell something about the social anxiety score. A higher amplitude in rest and preparation phases is associated with a higher social anxiety score. But figures 28 and 29 shows that almost nobody is excluded by the amplitude rule. Also the expert found it difficult to explain this. Furthermore, we see that older children with an age between 12.3 en 16.5 have a higher social anxiety score. This may be the same reason as the reason we gave in the previous experiment setup, where the older children are more aware of what is happening making them more afraid. The lower-bound of the age criterion corresponds to the increase in social anxiety in the transition from child to adolescent. The upper-bound is difficult to interpret since it is unclear in the literature whether a decrease of social anxiety will follow later.

Qualitative evaluation: Subgroups using delta attributes

The results of the experiment using the delta variables, shown in figure 34 in the appendix, shows that the deltaBMI variable is important. The top-5 subgroups corresponds to the subgroup numbers 10 to 14 in the appendix, tells us that when the BMI of the participant has been increased since the first experiment, the social anxiety score has also been increased. The top-1 subgroup says that in 57% of the cases that the participant is a female below the age of 15,14 and has an increased BMI has also an increased social anxiety score. The reason for this can be that when a person has an increased BMI, he or she is less comfortable in his skin. This can cause that a person is more nervous because he or she thinks that other people think bad about her. The expert had the following possibility for this phenomenon:

The first two characteristics of this group fit well with what is known in the literature about social anxiety. First, women usually report a higher degree of social anxiety. Secondly, it is known that social anxiety generally increase during the transition from child to adolescent. This happens around the transition to high school, but what happens after this increase is not yet certain. Adolescence is also usually the period in which a social anxiety disorder develops. An increase in social anxiety by people with a risk for a disorder is certainly expected before they are 17 years old. The age at T1 is less than 15 which match this theory given above.

For the increased BMI there are two possible explanations that depends on the actual value of the BMI and the average age of the group.

1. It is known that girls with a BMI > 25 have an increased risk of depression. Social anxiety often occurs with depression and may even precede depression. Overweight could therefore also be a risk factor for social anxiety. In the relationship between overweight and depression, it is thought that changes in energy management play an important role. It is a little too simple to say that girls are unhappy because they are overweight. That is one of the reasons why we cannot conclude that girls with a rising BMI are more insecure about their appearance and are therefore more concerned about what others think of them. It would be a new finding that being overweight is a risk factor for social anxiety. But this is only relevant if the girls in this group also end up with a relatively high BMI. Further research into the mechanisms that underlie this relationship would be necessary.
2. Another possible explanation for the increase in social anxiety in girls under 15 with a rising BMI could be that BMI usually increases at the start of puberty, so there is an energy reserve before the chance of pregnancy arises. It would then be a "marker" for the start of puberty. The hormonal changes affect the stress system that will respond more strongly to (possible) rejection by peers. This in turn can lead to an increase in social anxiety. A comment on this statement is that 15 is quite late before the onset of puberty in girls.

BMI T1 over the deltaBMI

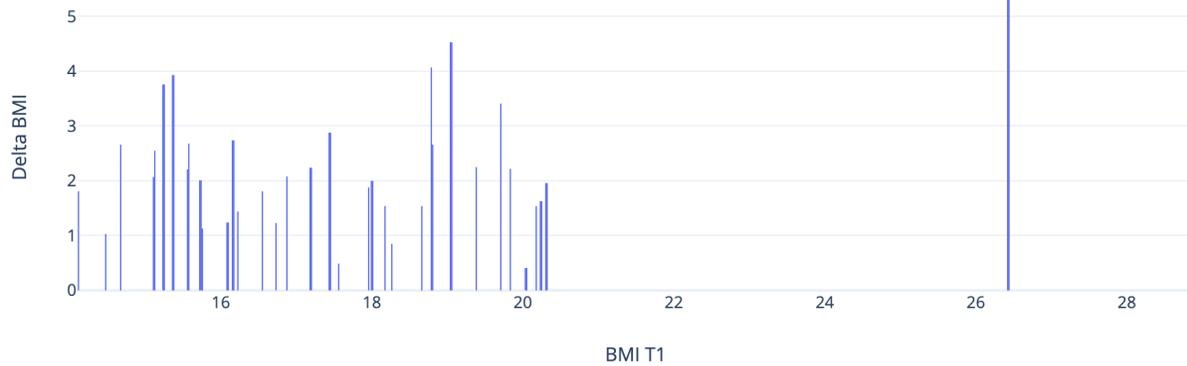


Figure 30: This figure shows the BMI at T1 and the deltaBMI of the participants satisfying subgroups number 10. It shows that not only girls with a relatively high BMI end up in this group. Especially girls who have an increase in BMI where the BMI is relatively low at T1 satisfy subgroup rule 10.

Age over the number of participants having that age

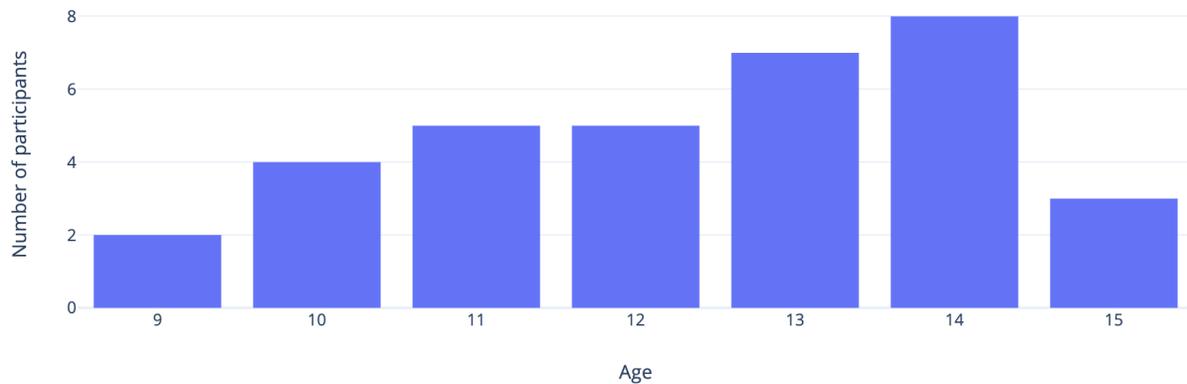


Figure 31: This figure shows the age over the number of participants having this age for the participants satisfying subgroups number 10. It shows that this subgroup has an average age of 12.38 but a high standard deviation. This matches the explanation of the expert that it could be a "marker" for the start of puberty.

Figure 30 shows that the explanation of the expert, that it would be a new finding that being overweight is a risk factor for social anxiety, does not apply here. But we can say, using the figure 30, that especially girls who have an increase in BMI where the BMI is relatively low at T1, have a higher change of a rising social anxiety score. This may suggest that girls with a relative low BMI respond worse to an increase in weight than the girls with a relative high BMI. Figure 31 shows that the explanation of the expert, that it could be a marker for the start of puberty, could apply here. The comment of the expert on this statement was that 15 is quite late before the onset of puberty in girls. But the average age of the participants satisfying subgroup number 10 is 12.38 which is relatively far below the age of 15.

We also see that in 61% of the cases the participant has a decreased HRV during rest, an increased BMI and experience an increased heart rate has also an increased social anxiety score. The increased heart rate causing a higher social anxiety score is quite obvious. Furthermore, this

subgroup confirms the result of another study saying that a reduction in the HRV leads to a higher social disorder [AQK⁺13].

Last but not least, we see that in 59% of the cases that the participant is a female below the age of 15,14 with an increased BMI and a decrease in the amplitude during the speech has also an increased social anxiety score. The decrease of the amplitude during the speech resulting in a higher social anxiety score is quite strange. Because research has shown that a higher amplitude is correlated to a higher effort, in terms of standing still versus start walking [HKYM95]. A possible explanation can be that mental effort has a different effect on your body than physical effort.

7 Discussion

In the following section we will discuss the results and method used in this project. We will discuss the small sample size, environmental factors, removal of outliers, missing important data properties and the reliability of Cortana.

7.1 Small sample size

In social experiments it often happens that the sample size is small since people must be willing to participate in an experiment. During this bachelor project we use data that consists of information about 236 participants. Because of this we experience the same problem of a small sample size. However, for the purpose of this bachelor project, providing new insights that can be further investigated, a sample size of 236 participants is large enough. The Developmental and Educational Psychology unit of Leiden University can use this information to conduct further research.

7.2 Data influenced by environment

During this Bachelor project, we used data that has been collected in 2009 [WBM⁺09]. For this reason, it is very difficult to find out whether the data obtained was affected by environmental factors. Environmental factors can influence the mood / behavior of a person. Examples of environmental factors can be:

- The way the participants were addressed
- The situation in the building in which the experiment was conducted. For example, was the temperature the same on day 1 and 2.
- The situation around the building. For example, there was a lawn mower working on day 1, and not on day 2.
- Period in the year, in the summer or in the winter.

The data used has been collected over a time period of more than 30 weeks. Furthermore, adolescents from different types of schools have been asked for this experiment. For this reason it is very difficult to keep the environmental factors the same. While collecting the data, they made sure that the start time remained constant and the presentations took place on school days. However, this information can no longer be retrieved and is not taken into account during this bachelor project.

7.3 Remove outliers

During the calculation of the HRV feature, a number of events were omitted from the data. The reason for this was that there were participants where the ECG data had an illogical shape, as shown in figure 9. This caused that no peaks could be detected, and therefore we could not derive the required features. For this reason we have chosen to exclude these participants from the data table. However, this illogical shape of the ECG data can contain relevant patterns, which say something about the social anxiety of a person. During this bachelor project it was agreed with the supervisor to filter this data from the data table.

7.4 Missing important data properties

As explained in the "method" section, the participants of the experiment had to present for 5 minutes about the movie they like the most or about the movie they dislike the most. During this experiment they had not thought about writing down the choice of the participant. Because of this, we cannot retrieve this information and thereby we cannot use this data in this bachelor project. It would have been very useful if we had this data in order to tell something about the social anxiety score using the tone of the message(positive or negative).

7.5 Depending on the correctness of Cortana

This bachelor project has been done by using Cortana as the subgroup discovery tool. However, we are not totally sure if this tool is completely correct. It could be the case that certain functionalities in Cortana are not correct, which affect the data mining results negatively. In the data mining field they often say: "Garbage in is garbage out", which means that if the data quality is very low the quality of the output is also very low. So it is important to ensure that the data quality is high enough. But the phrase, garbage in is garbage out, also applies to the quality of data mining functions. Because if the process is not programmed correctly, the results will not be reliable.

8 Conclusion

In this thesis we investigated the following questions:

1. Can we tune the alpha in such that we obtain a desirable trade-off between subgroup size and the deviation of the target?
2. Is it better, in the purpose of this thesis, to use a tuned version than default version of the quality measures?
3. Statistically, are the subgroups significant or just a coincidence, meaning is there an observed effect or correlation that indicates that the obtained subgroup is unlikely to be based on luck.
4. Which insights can be derived from the subgroups so that psychologists can get a better understanding of adolescents?

It is possible to tune α in such a way that we can make a desirable trade-off between the subgroup size and the deviation of the target. The expert, with whom we are collaborating, indicated to take 25 percent of the sample size as the minimal coverage of a subgroup. This will lead to a minimal coverage of 60, 25 percent of 237 adolescents. By increasing α , we make the coverage of a subgroup more important. We finally opted for an α value of 1.1 which lead to an average coverage of 76.8.

Comparing to the Z-score and the default CWRAcc, the tuned version is the only quality measure that return subgroups with a size around 60. The Z-score and the default CWRAcc return subgroups with a very low subgroup coverage, around 8, but a large average of the target value. In this bachelor project, as mentioned in the previous section the expert explained that the minimal coverage have to be around 60 which lead to the use of the tuned CWRAcc but when one wants subgroups with a large difference in the average of the target, the Z-score or default CWRAcc is the right quality measure. So it is indeed better, in the purpose of this thesis, to use a tuned version instead of a default version of the quality measures. One can tune the quality measure so that the desirable trade-off is made between the size of the subgroups and the average of the subgroups.

The result of these experiments also shows that at least the top-100 subgroups with nominal targets were significant, having a p-value below 0.05. This result is generalised to the subgroups with numeric target, meaning that we treat the top-5 subgroups obtained using a numeric target as significant as well. We used this generalization because Cortana, the platform we are using, does not support p-values for numeric targets.

During the qualitative evaluation we found that the HR and HRV are two very important features. A lower HRV for children with an age above 10.7 is associated with a higher social anxiety score regarding the subgroups found. Also a higher HR for children with an age above 10.7 during the preparation and after the speaking task is associated with a higher social anxiety score, and having no large rise in the heart rate during the speaking task.

Furthermore, the amplitude and age are also two important features. We see that the subgroups consist most of the time of older children between an age of 12.3 and 16.5 or above 10.7 which can be caused by being more aware of what is happening. A higher amplitude is also associated with a higher social anxiety score. This result fit well with previously conducted research, as mentioned in section 2, in which was indicated that the amplitude mainly increases during exertion. This would make it understandable that a higher amplitude during the speech was associated with a higher social anxiety score. But figures 28 and 29 shows that only a few people are excluded by the amplitude. This indicates that the amplitude used in this project has little to say about a person's social anxiety.

Last but not least, the deltaBMI is also a very useful feature. The top-1 subgroup says that in 57% of the cases that the participant is a female below the age of 15,14 and has an increased BMI has also an increased social anxiety score. The reason for this can be that when a person has an increased BMI, he or she is less comfortable in their skin. This can cause that a person is more nervous because he or she thinks that other people think bad about them.

9 Future research

9.1 Make use of the tone of the message

The data we are using for this bachelor project has been collected in 2009. For this reason, it is very difficult to retrieve information about the tone of the message during the presentation. The tone of the message can be positive, when the participants present about the movie they like the most, or negative, when the participants present about the movie they dislike the most. It should be interesting to perform the same experiment but with the information about the tone of the message. With this information we can maybe say something about how the tone of the message influence the anxiety of a participant. This can result in particular deviations in the data when a participant presents about a movie they dislike the most.

9.2 Develop a data mining program in Python

This bachelor project has been performed using Cortana as subgroup discovery platform. This platform was already developed and because of this we depend on the accuracy of the program. For future research, it seems interesting to develop a Python program which can perform the same experiment. Using this program we can validate the results which gives a confirmation that the results obtained using Cortana are correct.

9.3 Make predictions about the psychological condition of a person

This bachelor project has used the descriptive type of data mining. It gives insights in the collected data by discovering patterns using subgroup discovery. In future research it is interesting to make use of the prediction type of data mining. Using this type of data mining, we can create a model that is able to predict the psychological condition, with regard to social anxiety, of a person. In order to do this, it is important to have a huge amount of data points of good quality in order to learn a model. The data we have used consists of 237 data points which is to little for creating a well-learned prediction model.

9.4 Collecting data about the environmental condition in which the adolescents are presenting

As we have mentioned before, this data has been collected in 2009. For this reason, it is very difficult to find out whether the data obtained was affected by environmental factors. Because of this, it seems effective to collect data about the environmental factors during the presentation. Maybe, certain deviations in the data simply occurs because there was a lawn mower working which lead to distraction of the presenter. By collecting the data of environmental factors, we can ensure a better quality of the results by taking this data into account.

References

- [ACT⁺16] Lauri Ahonen, Benjamin Cowley, Jari Torniainen, Antti Ukkonen, Arto Vihavainen, and Kai Puolamki. Cognitive collaboration found in cardiac physiology: Study in classroom environment, Jul 2016.
- [AQK⁺13] Gail A Alvares, Daniel S Quintana, Andrew H Kemp, Anita Van Zwieten, Bernard W Balleine, Ian B Hickie, and Adam J Guastella. Reduced heart rate variability in social anxiety disorder: associations with gender and symptom severity, Jul 2013.
- [com19] The SciPy community. `scipy.signal.find_peaks`, May 2019.
- [DK11] Wouter Duivesteijn and Arno Knobbe. Exploiting false discoveries – statistical validation of patterns and quality measures in subgroup discovery. *2011 IEEE 11th International Conference on Data Mining*, 2011.
- [Far17] Bryn Farnsworth. Heart rate variability - how to analyze ecg data, Jul 2017.
- [Fay96] Usama M. Fayyad. *Advances in knowledge discovery and data mining*. AAA/MIT Press, 1996.
- [HGCJ11] F Herrera, P Gonzalez, C J Carmona, and M J D Jesus. An overview on subgroup discovery: Foundations and applications, Dec 2011.
- [HKYM95] J He, Y Kinouchi, H Yamaguchi, and H Miyamoto. Exercise-induced changes in r wave amplitude and heart rate in normal subjects, Apr 1995.
- [Inf19] InformedHealth. What is an electrocardiogram (ecg)?, Jan 2019.
- [K97] Jan Komorowski and Jan M. ytkow. *Principles of data mining and knowledge discovery First European Symposium, PKDD 97, Trondheim, Norway, June 24-27, 1997: proceedings*. Springer, 1997.
- [Lam15] Rachel Lampert. Ecg signatures of psychological stress, Aug 2015.
- [LCGF04] Nada Lavra, Bojan Cestnik, Dragan Gamberger, and Peter Flach. Decision support through subgroup discovery: Three case studies and the lessons learned. *Machine Learning*, 57(1/2):124127, 2004.
- [LK12] Matthijs Van Leeuwen and Arno Knobbe. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 25(2):208242, 2012.
- [MK11] M Meeng and A J Knobbe. Flexible enrichment with cortana - software demo, 2011.
- [nat] Social anxiety disorder: More than just shyness.
- [PCRB88] A C Petersen, L Crockett, M Richards, and A Boxer. A self-report measure of pubertal status: Reliability, validity, and initial norms, Apr 1988.
- [Sau17] Asena Atilla Saunders. The history of data mining, Jan 2017.

- [Tak19] Frank Takes. Data science and process modelling, Feb 2019.
- [Tet18] Salomon Tetelepta. Exploring heart rate variability using python, Aug 2018.
- [Tho15] Jamie Thomas. Where is the world supposed to put all of its data?, Feb 2015.
- [vdBW15] Esther van den Bos and P Michiel Westenberg. Two-year stability of individual differences in (para)sympathetic and hpa-axis responses to public speaking in childhood and adolescence, Mar 2015.
- [WBM⁺09] P Michiel Westenberg, Caroline L Bokhorst, Anne C Miers, Sindy R Sumter, Victor L Kallen, Johannes van Pelt, and Anke W Blte. A prepared speech in front of a pre-recorded audience: subjective, physiological, and neuroendocrine responses to the leiden public speaking task, Oct 2009.

10 Appendix

10.1 Peak detection function

```
prominenceFactorPeaks= 1; averagePeaks= 0; foundPeaks= True
#We set the prominence to one, because in most cases this was the
#correct value
while foundPeaks == True:
    foundPeaks= False
    peaks, _ = find_peaks(y, prominence=prominenceFactorPeaks)
    for n, item in enumerate(peaks):

        #Calculate the average distance between two peaks
        if n < len(peaks)-1:
            averagePeaks+= peaks[n+1]-peaks[n]
        if len(peaks)>0:
            averagePeaks= averagePeaks/len(peaks)
    for n, item in enumerate(peaks):
        #Iterate through peaks to find a distance between two peaks which
        #is greater than 1.5 times the average calculated distance
        if foundPeaks == False:
            #If there is already found a distance between two peaks which
            #is greater than 1.5 times the average calculated distance,
            #we don't have to look further
            if prominenceFactorPeaks > 0.6:
                #If the prominence is already 0.6 we stop because a
                #lower prominence we think it is an outlier, where
                #the data contains too many errors
                if n < len(peaks)-1:
                    if peaks[n+1]-peaks[n] > (1.5*averagePeaks):
                        #If the distance between two peaks is greater
                        #than 1.5 times the average peak, we decrease
                        #the prominence with 0.1
                        prominenceFactorPeaks= prominenceFactorPeaks
                                                -0.1

                foundPeaks= True
```

10.2 Obtained subgroups in T1

| Subgroup nr. | Rule |
|--------------|--|
| 1 | VisualAnalogueScale_heartrate4b >= 42.0 AND BPMbefore >= 85.135796 AND AgeT1 >= 10.735113 AND BPMspeech <= 125.40409 |
| 2 | VisualAnalogueScale_heartrate4b >= 42.0 AND BPMbefore >= 85.135796 AND AgeT1 >= 10.735113 |
| 3 | VisualAnalogueScale_heartrate4b >= 42.0 AND BPMbefore >= 85.135796 AND AgeT1 >= 10.735113 AND BPMbefore <= 112.54425 |
| 4 | VisualAnalogueScale_heartrate4b >= 42.0 AND BPMbefore >= 85.135796 AND HRVbefore <= 83.841896 AND AgeT1 >= 10.735113 |
| 5 | VisualAnalogueScale_heartrate4b >= 42.0 AND BPMbefore >= 85.135796 AND HRVspeech <= 65.08277 AND AgeT1 >= 10.735113 AND BPMbefore <= 111.54309 |

Figure 32: Subgroups obtained with all T1 columns enabled, only visualAnalogueScale_nervous and T1PrePRPSMean is disabled because these attributes have a too high correlation to the target, and the social anxiety score as the target attribute.

10.3 Obtained subgroups in T2

| Subgroup nr. | Rule |
|--------------|---|
| 6 | VisualAnalogueScale_heartrate4b2 >= 33.0 AND Amplituderust2 >= 1.1148207 AND Amplitudebefore2 >= 1.2865313 AND AgeT1 <= 14.483231 AND AgeT1 >= 10.335386 |
| 7 | VisualAnalogueScale_heartrate4b2 >= 33.0 AND Amplituderust2 >= 1.1148207 AND Amplitudebefore2 >= 1.2865313 AND medication2 = '0' AND AgeT1 <= 14.75154 AND AgeT1 >= 10.302532 |
| 8 | VisualAnalogueScale_heartrate4b2 >= 33.0 AND Amplituderust2 >= 1.1148207 AND medication2 = '0' AND BPMrust2 >= 63.59406 AND AgeT1 <= 14.387406 AND Amplitudespeech2 <= 2.2704577 |
| 9 | VisualAnalogueScale_heartrate4b2 >= 33.0 AND Amplituderust2 >= 1.1148207 AND medication2 = '0' AND BPMrust2 >= 63.59406 AND AgeT1 <= 14.387406 |

Figure 33: Subgroups obtained with all T2 columns enabled, only visualAnalogueScale_nervous and T2PrePRPSMean is disabled because these attributes have a too high correlation to the target, and the social anxiety score as the target attribute.

10.4 Obtained subgroups in delta variable

| Subgroup nr. | Rule |
|--------------|---|
| 10 | geslacht = 2.0 AND AgeT1 <= 15.140315 AND deltaBMI = '1' |
| 11 | geslacht = 2.0 AND deltaBMI = '1' |
| 12 | deltaHRVrust = '0' AND deltaBMI = '1' AND AgeT1 <= 13.136209 |
| 13 | geslacht = 2.0 AND AgeT1 <= 15.140315 AND deltaAmplitudespeech = '0' AND deltaBMI = '1' |
| 14 | deltaHRVrust = '0' AND deltaBMI = '1' AND deltaVisualAnalogueScale_heartrate4b = '1' |

Figure 34: Subgroups obtained with all delta columns enabled, only deltaVisualAnalogueScale_nervous and deltaT2PrePRPSMean is disabled because these attributes have a too high correlation to the target, and the delta social anxiety score as the target attribute.

