"I'm so stressed, let's watch Youtube": Using sequential and non-sequential patterns in phone usage data to predict stress

Aaron Wijnker
Anr.: 187832
Snr.: 2030399

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Data Science Business & Governance
Department of Cognitive Science & Artificial Intelligence
School of Humanities and Digital Sciences
Tilburg University

Thesis committee:

Dr. A.T. Hendrickson
Dr. H. Brighton

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
December 2019

Preface

I would like to express my sincere appreciation to Dr. Drew Hendrickson and Dr. Giovanni Cassani, for providing the data, information and the help I needed to complete this thesis. I would also like to thank Dr. Henry Brighton, for providing a second opinion on the contents of this thesis.

Lastly, I would like to thank Vincent van Stiphout, Marijn Hollander, Pepijn van Teeffelen, Wietze Mulder, Sophie Vink, Jeroen Wijnker, Rejane Wijnker-Velten and Resa van Lith for the support, insightful discussions and feedback.

"I'm so stressed, let's watch Youtube": Using sequential and non-sequential patterns in phone usage data to predict stress

Aaron Wijnker

December 6, 2019

Abstract

Stress levels seem to have risen the past years. More people are claim to feel longer periods of stress. This can have negative health effects. Prediction of stress is important for stress detection, treatment and the prevention of chronic stress. Phone use has also increased worldwide. Phones play a big role in our everyday lives, which has led researchers to believe patterns in phone usage could identify people's emotions and personality. The present study uses these phone usage patterns to predict stress. Different models have already been built to obtain information from phone usage data, focusing on app frequency. Previous research suggested that the order of used apps could present additional information for stress prediction. The results of the present study showed that, while both non-sequential frequency and sequential patterns are able to predict stress better than the majority baseline, the non-sequential patterns were more useful for stress prediction. This suggests that sequential patterns might not provide additional information for stress prediction. The results also provide a new exciting stress prediction method, which could be combined with existing methods.

Keywords: stress prediction, mobile phone usage, bag-of-apps, sequential pattern mining, cSPADE

1 Introduction

Over the last years, perceived stress levels seem to have risen (Hagquist, 2010; Korn Ferry Institute, 2018; O'Malley, 2019). Higher perceived levels of stress can have serious ramifications, such as sleep difficulties, performance quality decrease and even depression. (AbuAlRub, 2004; Majeno et al., 2018; O'Malley, 2019). Prediction and detection of stress is important for health organisations and companies trying to identify risk groups or create a risk profile and has been done before (Maxhuni et al., 2016, 2017). It has been suggested that mobile phone usage data could be used to predict perceived stress, due to the

big role mobile phones play in our everyday lives (Lepp et al., 2014; Selkie, 2019; Weilenmann & Larsson, 2002). This study aims to create a general model for perceived stress prediction, by utilising mobile phone usage data.

Stress research has been done in many different ways. Some studies have used a more medical approach to perceived stress detection and prediction, using heart rate variability or galvanic skin response measures (Bakker et al., 2011; Pourbabaee et al., 2018). Because these data are costly and hard to obtain, other options have been analysed. One promising option was phone usage data. Approximately 5.1 billion unique mobile phone users were counted at the start of 2019, which is around 67% of earth's total population (Kemp, 2019). We use our phones continuously for a multitude of purposes. (Lepp et al., 2015; Selkie, 2019; Weilenmann & Larsson, 2002). With an average phone use of 3 hours per day, phones are used during a significant amount of our waking hours (Fennell et al., 2019). Do & Gatica-Perez (2010) created a framework, named bag-of-apps, to retrieve meaningful non-sequential patterns from phone usage data, mainly focused on frequency of app usage. It has since been suggested that sequential app usage patterns could provide additional information on different topics (Farrahi & Gatica-Perez, 2014; Yan et al., 2012). Alibasa et al. (2019) used sequential app usage patterns to predict mood. They suggested that this approach could be used to predict perceived stress levels. The current research aims to determine the stress prediction performance of models based on sequential patterns and models based on non-sequential app usage patterns. This led to the following research question:

To what extent can sequential or non-sequential app usage patterns predict daily stress levels?

This study hopes to determine the added value of sequential patterns for stress prediction. Stress 'levels' refers to a six-point Likert-type item, which is used in this research to measure stress. This method captures the different intensities of stress from low to high and is similar to the method used in Ferdous et al. (2015). To answer the research question, a model using sequential app usage patterns is created and tested. This model analyses sequential patterns between every individual app (category). A Bag-of-Apps type model is also created. This model uses the frequency of every individual app (category), without any sequential pattern between them. These models, using different types of patterns, are compared to analyse the importance of sequences in app usage for stress prediction. This is formulated in the following sub question:

How do sequential pattern mining models and bag-of-apps models differ on stress prediction performance?

Stress can be measured in levels (Ferdous et al., 2015) or binary, as proposed by Alibasa et al. (2019). It could be interesting to predict the severity of stress, as well as the presence of stress in general. Detection of general stress presence could be important for general prevention and treatment of symptoms, whereas categorical stress detection could be used to suggest specific intensity-related

treatments (Bakker et al., 2011; Glanz & Schwartz, 2008). The models in this study were tested on their stress prediction capabilities using stress measured on a six point Likert-type item and a binary scale, which measured the presence or absence of stress. This was summarised in the following subquestion:

What is the effect of binarised stress levels on the prediction performance of sequential pattern mining and bag-of-apps models?

Answers to these questions could provide a new method for stress prediction. Both the Bag-of-Apps (BoA) and the Sequential Pattern Mining (SPM) approach have not been used to predict stress, yet. The results showed that both the BoA and the SPM model could perform above baseline. Binarised stress levels improved the results of both models. Social media, messaging and lifestyle categories proved to be important for the prediction of the models. The BoA model outperformed the SPM model in all scenarios, which raised questions about the added value of sequential patterns for stress prediction.

2 Related Work

As mentioned, there seems to be a rise of perceived stress levels over the past years (Korn Ferry Institute, 2018; O'Malley, 2019; Hagquist, 2010). Pressure of constant change on the work floor, due to the fast-paced online world, is mentioned as a possible cause. Some studies mention mobile phones and the constant connection phones provide as an additional contributor to stress (Horwood & Anglim, 2018; Y.-K. Lee et al., 2014; Horwood & Anglim, 2019; Thomée et al., 2011; Kuss et al., 2018). More specifically, articles suggest that social media and work related apps, like email, could be nurturing constant feelings of stress (Kushlev & Dunn, 2015; K. B. Wright et al., 2014).

The consequences of stress have been known for decades. Quick et al. (1987) stated that badly managed stress can lead to a multitude of medical and psychological problems, ranging from alcohol abuse to cardiovascular diseases. In a more recent study on stress levels in the working age population, Wiegner et al. (2015) discovered that more than half of the participants reported to have stress to some extent. Among those who reported having stress to some extent, two thirds showed signs of exhaustion, anxiety and burnout. Other articles also reported sleep difficulties and depression as possible effects of stress (AbuAlRub, 2004; Majeno et al., 2018; O'Malley, 2019).

Acute stress is something we experience every day, whereas chronic stress can be a result of prolonged acute stress (Bakker et al., 2011). This chronic stress can cause the aforementioned health problems. (Bakker et al., 2011; Majeno et al., 2018). That is why this research focused on daily stress level prediction.

2.1 Stress prediction

The prediction of stress levels is important for stress detection and treatment, to prevent the health consequences described earlier. Stress prediction has

been done using a vast array of methods. Inventions have even been done to predict stress levels of drivers when driving in a motor vehicle (Woltermann & Schroedl, 2003). Bakker et al. (2011) predicted stress using galvanic skin response data. This data reflects sweat production, which was measured by sensors on a participant's body. Detection of arousals was used to determine when a person is stressed or not. The study had trouble identifying the different stressful events, mainly because the experiment was not done in a controlled environment. This made it difficult to interpret the different spikes in the data. Pourbabaee et al. (2018) researched stress based on heart rate variability. Stress was measured by adding the scores on rumination and locus of control. Both scores were measured on a five-point Likert-type item, which gave the total score a range between two and ten. Heart rate variability, measured by ECG scans, proved to be a good predictor for stress. Higher heart rates generally meant more stress. The problem with these methods lies in the obtainability of the data. ECG scans and galvanic skin response data are expensive and difficult to procure.

A study by Reddy et al. (2018) used surveys to predict stress. Other studies found different methods to predict stress. These methods focus on more accessible data, collected in an unconstrained environment. An example of such data was used by Soto et al. (2011). Phone usage data was used to determine a participant's socioeconomic status. The aggregated phone usage time was used as the primary parameter. As previously mentioned, smartphones are more widely used than ever, which is why it could provide researchers with an unconstrained look into the habits of a participant (Björkegren & Grissen, 2018; Fennell et al., 2019; Selkie, 2019).

Phone usage data has been used in numerous studies, with a variety of purposes. It has been used to predict loan repayment (Björkegren & Grissen, 2018), detect depressive and manic episodes (Osmani, 2015), predict the next app somebody is going to use (Baeza-Yates et al., 2015) and to predict Parkinson's disease progression (Anantharam et al., 2013). Ferdous et al. (2015) used smartphone app usage data to predict the participants' perceived level of stress. The apps were categorised to predict stress levels on a Likert-type item, ranging from zero to five. The prediction model used app category and time spent on the app as predictors. The results showed that the user-centric model was good for stress prediction, while the group behaviour model did not perform well. Stütz et al. (2015) also used mobile phone data to predict stress, collected by an app. This app sent out seven surveys per day to the participants. The results showed significant correlations between stress and smartphone data. Different basic usage features were created, such as mean app usage time and summarised session times. The prediction results of these features were not very accurate. This could be due to the small dataset, which used 15 participants and contained approximately 100,000 data points.

The present research aims at creating a general model with a bigger real-life dataset, not just suitable for user-centric stress prediction. Ferdous et al. (2015) showed that group behaviour models scored moderately or bad on stress prediction, when using duration of app use per category as the main variable.

Do & Gatica-Perez (2010) created a framework, named bag-of-apps (BoA), to retrieve meaningful patterns from phone usage data, mainly focused on app usage of participants. The BoA model is based on the bag-of-words model. This model, frequently used in natural language processing, is used to quantify speech and text (Zhang et al., 2010). Sentences or parts of speech are quantified by measuring word frequencies in text, which results in a sparse matrix. The BoA model treats apps like the words in a bag-of-words model. It quantifies the app occurrence during a certain time period. This model proved to be capable of finding useful patterns in app usage data. It has since been suggested that sequential patterns in app usage could provide additional information on different topics (Farrahi & Gatica-Perez, 2014; Yan et al., 2012).

Alibasa et al. (2019) proposed that a model based on sequential patterns in phone use could predict perceived stress. They created a mood prediction model, which also used sequential patterns as features. Mood was measured as being positive or negative, and predicted by the presence or absence of a sequential pattern. Their sequential pattern model did perform better than the baseline, but only slightly. It was stated that a bigger dataset would likely increase the accuracy of the model. The current study recognises these statements and builds on previous research. A larger dataset is used, along with the frequency of patterns as features. A BoA model, a non-sequential pattern model, is built and trained to compare the results to a sequential pattern model.

2.2 Sequential pattern mining

The practice of obtaining sequential patterns is called Sequential Pattern Mining (SPM). This method originated from association rule mining, which was introduced by Agrawal et al. (1993). It uses baskets of items to find frequent itemsets. These itemsets occurred often together, but did not have any particular order.

An algorithm to mine such association rules was proposed by Agrawal & Srikant (1994). The algorithm uses support as a selection tool for selecting frequent itemsets. Support is defined as the proportion of the total number of baskets a rule appears in. If the support is higher than an arbitrary value, the rule is saved as a frequent association rule. Association rule mining is used to find unordered frequent itemsets.

To incorporate the extra information the order of items might contain, Agrawal & Srikant (1995) created the Sequential Pattern Mining (SPM) method. This method finds frequent itemsets in baskets of data. As a common example, market basket data is used. Baskets are created by grouping items per customer per transaction time (e.g. day). The SPM method finds frequent sequences of items in all baskets. Srikant & Agrawal (1996) proposed an improvement to their own algorithm, called the Generalised Sequential Pattern (GSP) algorithm. This algorithm focuses more on the transactions made per customer-id and less on the time constraint. However, it does allow users to set arbitrary time constraints for the itemsets.

Zaki (2000) introduced a different method, called cSPADE (constraint Sequential Pattern Discovery using Equivalence classes). This method uses a

vertical database layout, whereas the GSP method uses a horizontal layout. In the vertical layout, each event-id or transaction time per customer-id has its own basket, instead of a basket per customer-id. Furthermore, cSPADE provided the user with the ability to assign constraints to the mining process, such as the maximum size of itemsets and the maximum length of sequences. The two methods were compared on runtime by Zaki (2001) and Verma & Mehta (2014). Both studies found that cSPADE was more efficient than GSP.

Another method for finding sequential patterns is PrefixSpan, which operates using Pattern Growth algorithms (Pei et al., 2004). The algorithm avoids repeated scanning of the database for pattern growth, by recursively projecting a sequence database onto smaller partitioned pattern related datasets. PrefixSpan is likely to be slightly more efficient than cSPADE, when analysing long patterns (Verma & Mehta, 2014). The current study, however, used cSPADE to mine sequential patterns. cSPADE is better maintained and documented in python and R and provides better control over constraints, as mentioned earlier. Control over constraints is useful for monitoring the amount and length of sequences (Zaki, 2000).

The cSPADE algorithm has been used in many studies (Aseervatham & Osmani, 2005; De Smedt et al., 2019; Ibrahim & Shafiq, 2019; Wang et al., 2018; Julea et al., 2008). The most common implication of cSPADE is pattern recognition and pattern comparison between groups (Aseervatham & Osmani, 2005; Exarchos et al., 2008; Liu et al., 2017; Wang et al., 2018). The cSPADE method has also been used for a type of prediction, where the next item in a sequence is predicted. This is demonstrated by S. Lee et al. (2016), who predicted the next place a mobile phone user was going to be, based on their previous movements during a certain time period. Other studies have predicted the next prescribed medications (A. P. Wright et al., 2015)or the movement of taxis (Ibrahim & Shafiq, 2019). cSPADE has also been used to predict features not included in the sequences (Deeva et al., 2017; Smedley et al., 2018). It has, however, never been used to predict stress.

Alibasa et al. (2019) suggested that such a use of sequential pattern mining algorithms could be applied to app usage patterns, for a more accurate stress prediction. Consequently, the present study uses sequential patterns to create an SPM model and compare the results to a non-sequential BoA model. Both of these models use app usage patterns differently. This could improve stress detection and prediction, as well as provide new insights into the importance of sequentially ordered patterns in phone usage data for the prediction of daily stress levels.

As mentioned earlier, consistent perceived stress can cause multiple health problems (AbuAlRub, 2004; Majeno et al., 2018). The stress prediction methods used in this study could be employed to monitor stress for health organisations or provide personal warnings to users (Matic et al., 2014; Grünerbl et al., 2014). Different stress relieving methods could be suggested in these warnings.

3 Method

In this study, three different datasets were used. The first dataset contained the app usage of all participants. The second dataset contained all answers to a survey the participants had to fill in multiple times per day during the study, which is similar to the method used by Stütz et al. (2015). In the third dataset, the apps and their categories were reported. A full description of the datasets is provided in the experimental setup.

The method, used to build and evaluate the BoA and SPM models from these datasets, was divided into three different phases. The first phase was the cleaning and pre-processing phase, in which the datasets were cleaned and prepared for the feature extraction phase. In the feature extraction phase, two different sets of features were extracted. The model creation and evaluation phase was the third phase, where the models were finalised and evaluated. The methodology used in the different phases are described below. The procedure for performing the steps in the different phases is described in the experimental setup section.

3.1 Cleaning and pre-processing phase

The first phase consisted of merging and cleaning the datasets. After that, the apps were divided into categories. Ideally, every application would be used as a feature or in sequences under their own name. This would allow researchers to analyse the influence of every app. However, app usage is likely to be distributed according to Zipf's law (Zipf, 1932; Adamic & Huberman, 2002). This law describes how the most frequent word is used twice as much as the second most used word, three times as much as the third most used word, and so on. App usage in the dataset is approximately distributed according to Zipf's law, as is illustrated in figure 2 in the experimental setup. This meant that more than half of the used apps did not record a frequency rate higher than 50. It was likely that the apps would not be used by a classifier or for a split in a decision tree. To prevent information loss, the apps were bundled into categories.

The most used apps had a big impact on their categories. With the intention of attributing the information these apps provide to the apps themselves, these apps were not divided into an existing category. Instead, each of these apps was assigned to their own category. This could also spread out the information over the categories. A detailed description of this process can be found in the experimental setup.

3.2 Feature extraction phase

The second phase focused on feature extraction. Features for the BoA model were based on a bag-of-words representation of text data (Zhang et al., 2010). As mentioned earlier, it quantified the app occurrence during a certain time period, creating a sparse matrix. The features for the SPM model were based on the sequential representation of data (Zaki, 2000). Five steps were followed in the

mining process of the cSPADE algorithm (Agrawal & Srikant, 1995; Zaki, 2001). In the first step, the dataset is sorted into baskets, based on user-id and response date. The second step is the L-itemset step. In this step, all possible itemsets L are created. The itemsets contained all the items available in the baskets. The maximum length of itemsets is arbitrary. The itemsets are then mapped to create a single entity for each itemset. In the third step, all baskets are transformed. The items in the baskets are replaced by their respective mapped itemsets. This creates baskets of mapped itemsets, based on user-id and response date.

Sequences in these mapped frequent itemsets are then mined in step four, using the apriori algorithm. This algorithm mines a list of all possible itemsets that satisfy an arbitrary support threshold. Support is defined as the number of baskets an itemset appears in, divided by the total amount of baskets. In the fifth step, the maximal length of each sequence is found. Step four is repeated, adding a new itemset to a sequence every time. This is done, until no new sequences satisfy the support threshold. Every itemset that can be added to a sequence provides it with more specific information (Pei et al., 2004; Zaki, 2001).

The features for both models were based on the frequency of daily app usage patterns per user. This was done to ignore sudden short feelings of stress, which are not necessarily harmful (Quick et al., 1987; Bakker et al., 2011). As mentioned before, stress during longer periods of time (e.g. day) can create mental and physical health problems (Bakker et al., 2011; Majeno et al., 2018).

Stress itself was measured by the following survey statement: 'Since taking the last survey, I felt stressed (gestresst)'. Participants were asked to react, by choosing one answer from a six-point Likert-type item, ranging from zero ('not at all') to five ('extremely'). This was similar to the method used by Ferdous et al. (2015). To answer the third subquestion, stress levels were divided into two categories: (practically) no stress (0-1) and stress (2-5). These categories are based on Alibasa et al. (2019), who split mood into positive or negative. They suggested that this model could be used for stress prediction as well. This could also be useful for the early detection of chronic daily stress, which is important for prevention of chronic stress (Bakker et al., 2011).

3.3 Modelling phase

In the third phase, the BoA and SPM features were used to create the eponymous models. BoA and SPM models were constructed for each of the different research questions. To predict stress, the models were fed to the XGBoost classifier, which has won numerous Kaggle competitions (T. Chen & Guestrin, 2016). XGBoost uses eXtreme Gradient Boosting on decision trees (Friedman, 2001; T. Chen & He, 2015). It assigns a weight to every record in the dataset. The model runs a decision tree to classify the records. The weights of the incorrectly classified records are updated and a new decision tree is made, using the updated weights. It repeats this process sequentially, until an arbitrary maximum number of trees is reached. In the case of a classification problem, the class that is predicted by most trees is linked to a record (Z. Chen et al., 2018). To put the results of XGBoost into perspective, a Support Vector Machine (SVM) classifier was also

used to predict stress (Suykens & Vandewalle, 1999).

Along with accuracy, recall was used to compare the performance of the models. Recall displays the amount of correctly labelled instances for one label, compared to the total amount of instances that should belong to that specific label. Because stress prediction is important for (early) detection and treatment of stress, identifying the correct stress level is more important than correctly identifying the instances that do not belong to a certain stress level. More specifically, it is important that people with stress are identified as such. It is less of a problem when less stressed people are labelled as stressed.

Feature importance for both models was also calculated, to determine which features has the most influence on the predictions. Features were compared by their F scores, which is a count of the number of times the feature was used to create a split in a decision tree (Z. Chen et al., 2018). The F score provides an indication of the relative value of features, when creating the decision trees in XGBoost. Comparing F scores between models is not done, since different models need a different total number of splits.

4 Experimental Setup

The following section describes the procedure for performing the steps in the different phases. The code can be found on GitHub (appendix A).

4.1 Datasets

As mentioned, three datasets were used to conduct this study. These datasets were collected by a third party, via the Ethica application. The first dataset contained all the phone usage information of 90 participants. Ethica tracked the apps a participant used, for an average of 23 days ($SD = 9.81$) per participant. The application also sent out a survey four times per day via a notification, at relatively random intervals between 9 am and 10:30 pm. Participants were required to complete a survey within two hours of it being sent or it would expire. Participants were rewarded with credit for completing as many surveys as possible. The complete survey and overview of apps per category could not be shared, due to ownership rights. The answers to the survey were combined in the second dataset. Participants filled in surveys for an average of 28 days ($SD = 9.60$). Due to an error, the application kept sending out surveys after the tracking process had finished. This data was excluded. The third dataset contained the 1,086 used apps and the corresponding 46 categories.

4.2 Cleaning and pre-processing

4.2.1 Apps without category

To provide the tracking data with app categories, the categories dataset was merged with the tracking dataset. This was done in R, using the sqldf package (Grothendieck & Grothendieck, 2017). This package allows for SQL-type merging

in R. The two datasets were merged on application name. Out of all 1,086 apps in the dataset, 642 did not have a category. This translated to 95,464 out of 464,286 tracking records. Figure 1 shows that only a few apps accounted for most of the tracking records without category. In fact, the top ten apps without a category made up for roughly 85% of all tracking records without category.

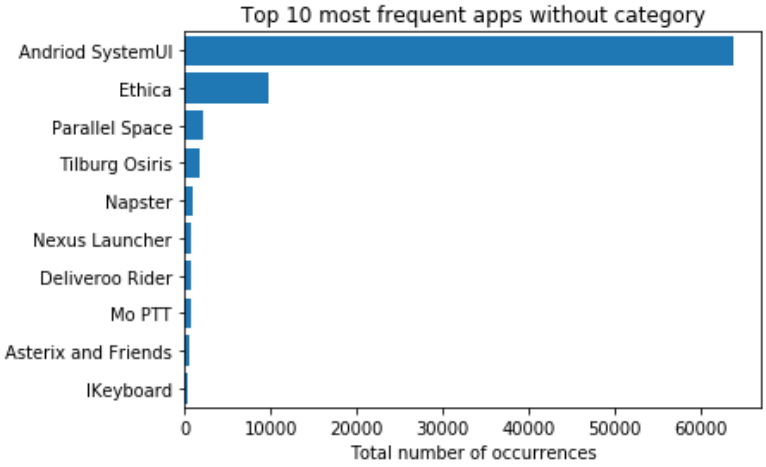


Figure 1: Frequency of ten most used apps without a category. IKeyboard Blue Love Heart theme has been shortened to IKeyboard, for fitting purposes.

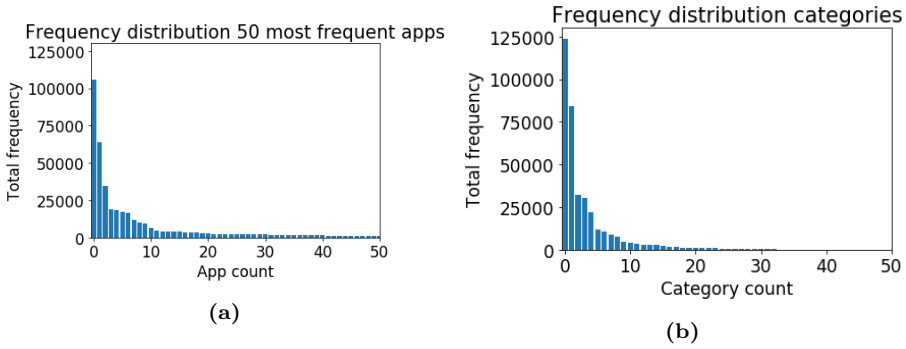
These apps have been categorised manually, by only using existing categories. This can be seen in the table in appendix C. It is important to mention that Ethica is assigned to its own category. Ethica is the application used to collect the data for this research. An interaction with the app was recorded when participants filled in a survey. Because this could have had an influence on stress levels, the Ethica data is not excluded. Besides that, this research aims at predicting stress with patterns in daily app usage of participants. Ethica was part of the daily app usage of the participants.

After these adjustments, 13,864 tracking records (2.98%) had no category. The average amount of records of the remaining categoryless apps in the dataset was approximately 22 records, which indicated that the remaining apps without a category are not used often. These records were not used for feature creation. Because the participants do not actively open or choose to open the apps categorised as 'Background_Process', these apps were also not used for feature creation. This category does not belong in the app usage of participants, because they did not use the app.

4.2.2 Hybrid app categorisation

As mentioned in the method section, app usage is distributed according to Zipf's Law. This can be seen in figure 2a. This figure only shows the 50 most frequent

apps, due to formatting reasons. Only a few apps accounted for most of the tracking records. This is also the case for the 46 app categories, as is depicted in figure 2b.



(a) (b)

Figure 2: Ordered frequency distribution of apps and app categories in the dataset. Figure 2a displays the frequency counts of the 50 most used apps. Figure 2b shows the frequency counts of all the categories.

Apps that were used more than 10,000 times were not divided into categories. This group consisted of: Whatsapp Messenger, Instagram, Snapchat, Google Chrome, Facebook, Spotify and Youtube. These seven apps are likely to provide considerable information, due to the amount of times they have been used. This would not be attributed to the apps if they are merged into a category. The hybrid categorisation slightly improved the spread of records across categories, as can be seen in the third graph in figure 3. Improvements are mainly visible outside of the top three most frequently used hybrid categories. After this procedure, 53 different categories were counted.

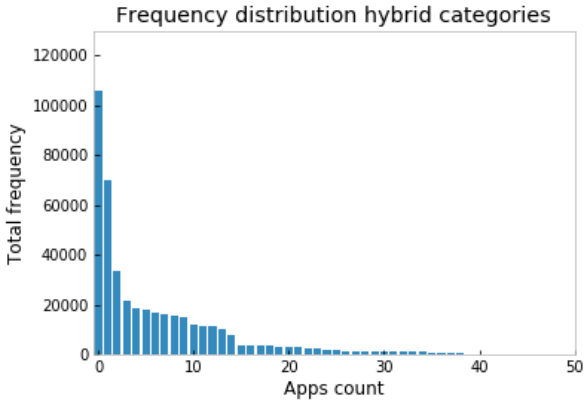


Figure 3: Ordered frequency distribution of hybrid app categories.

Next, some of the least used app categories were merged with comparable

categories. The category 'Job_Search' contained only one tracking record, which was a LinkedIn Jobseeker. This category was added to the 'Social_Networking' category, which is the same category as LinkedIn. The category 'Messages' contained only tracking records of the HTC messaging app, which was used two times by the same participant. This category was added to the 'Messaging' category. This category contained apps that are similar to the HTC messaging app, like the Android messaging app. The 'Music_&_Audio' category was added to the larger 'Music_Audio' category, because these categories contain the same type of apps. After this process, a total of 50 hybrid categories were counted.

The survey dataset was only used to obtain stress levels per day. Since stress was measured on a six-point Likert-type item from zero to five, every value above five can be labeled as an outlier. The stress data contained five outliers (10, 22, 55, 68 and 70). These values were removed, since they were not within the range of the six-point Likert-type item. As mentioned earlier, surveys were also sent to the participants after the tracking process was completed. Surveys could also expire after two hours. None of these surveys were used. After cleaning and pre-processing, the combined dataset contained roughly 320,000 tracking records, divided over 88 participants. A total of 1,713 unique user-id and response date combinations were found.

4.3 Features

4.3.1 stress level

The BoA and SPM models are built to predict daily stress levels. The distribution of these levels can be seen in table 1. To create the daily stress feature for both models, the response_time column in the survey dataset and the startTime column in the tracking dataset were modified. Both of these columns contained the date and time a participant had responded to a survey or started using an app. The *lubridate* package in R was used to split the date and time into two separate columns, titled 'response date' and 'response time' (Grolemund & Wickham, 2011). The stress level of participants was summarised per user-id and response date. These values were rounded to a whole number, to preserve the 0-5 Likert-type item classes. A separate stress dataset was created, containing user-id, response date and daily stress level. To create the binarised stress levels, a different dataset was created. Stress level 0 and 1 were labeled as 0 and the other labels were labeled as 1. The categorical stress levels seem to be imbalanced. More specifically, stress levels 3, 4 and 5 occurred less often than the other classes. The binarised stress levels were not heavily imbalanced. No stress (level 0) was the most frequent level.

Because imbalance in the categorical stress levels could be a problem for the model, the Adaptive Synthetic (ADASYN) sampling approach was used to perform oversampling on the training data. The test set was not used to perform oversampling. This was done to ensure no copies of the same point could end up in both the training and the test set, which would make them less independent of each other. ADASYN was proposed by He et al. (2008), who

Table 1: Frequency distribution of binary and categorical stress levels in the final dataset.

	Stress levels					
Types of stress	level 0	level 1	level 2	level 3	level 4	level 5
Categorical stress	**498**	448	475	170	107	15
Binary stress	**946**	767				

stated that "the essential idea of ADASYN is to use a weighted distribution for different minority class examples according to their level of difficulty in learning, where more synthetic data is generated for minority class examples that are harder to learn compared to those minority examples that are easier to learn" (p. 1322). The *imblearn* package in python was used to perform ADASYN oversampling (Lemaître et al., 2017). Figure 4 shows the distribution of the ADASYN oversampled datasets, compared to the original dataset. The numbers this table is based on, can be found in appendix B.

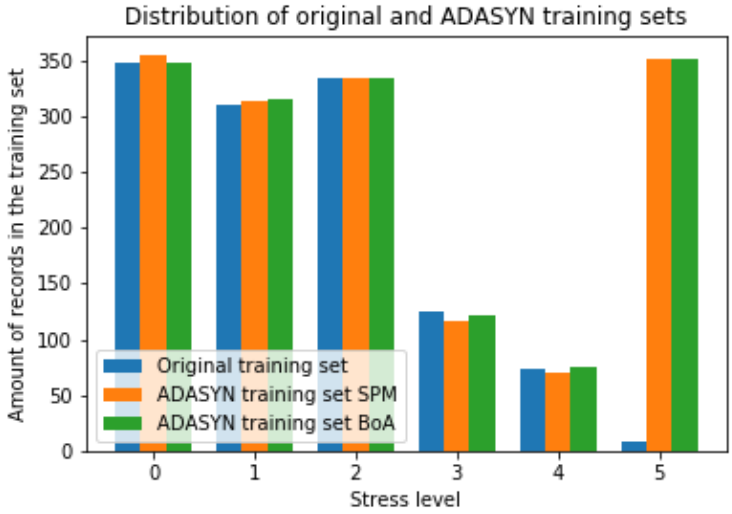


Figure 4: Stress level distribution of the original training set, compared to the ADASYN oversampled training sets for both the SPM and the BoA models.

Because ADASYN used the features to determine prediction difficulty, the SPM and BoA model have different numbers of synthetic samples. Judging from the high number of synthetic samples created for stress level 5, ADASYN predicted that this level was hard to learn. The amount of samples for the other classes remained approximately the same. The results of the models trained with the original dataset are compared to the results of the models trained with

the dataset altered by ADASYN in the results section.

4.3.2 Bag-of-Apps model

The features for the BoA model were created using the cleaned tracking dataset. The dataset was summarised using the *dplyr* package in R, creating a dataset with app categories organised per user-id and response date, as one long vector. A corpus of apps was created using the *tm* package in R (Feinerer, 2018). Using this corpus, the frequency of every app category per user-id and response date was calculated. This resulted in 50 features, which corresponds to the amount of hybrid categories. After that, the number of occurrences per user-id and response date were measured for each feature.

The 15 features with the highest frequency are shown in figure 5. This figure shows that 10 of the top 15 most frequent features are social media or internet related. Instant messaging is one of the most frequently used categories. Instant messaging offers real-time text transmission, where users are able to see who is online. This differs from the messaging category, which contains apps that do not facilitate real-time connection, but allow users to send text or items to unknown or previously known people. The difference between phone optimisation and phone tools is also important to explain. Phone tools are tools on your phone to help you (calculator, flashlight, etc.), whereas phone optimisation apps optimise phone usage or performance (cleaning apps, Parallel Space, etc.).

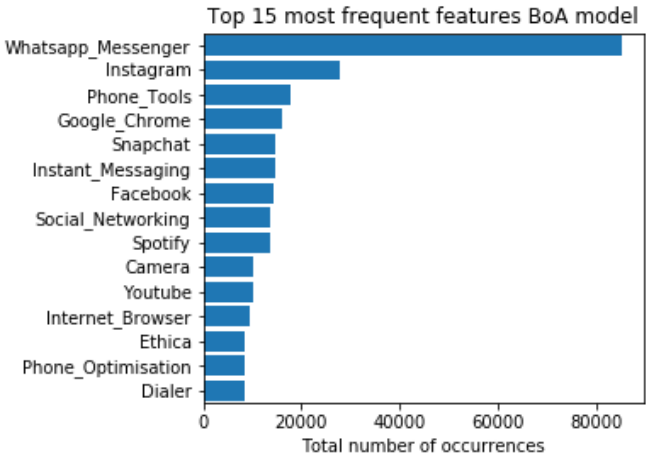


Figure 5: The 15 most frequent features in the dataset for the BoA model.

4.3.3 Sequential Pattern Mining model

The SPM model also used the cleaned tracking dataset to summarise the app categories per user-id and response date. The *arulesSequences* package provided

the C++ implementation of cSPADE for R (Buchta et al., 2007, 2019). The types and columns of the dataset were slightly modified or renamed, to match the requirements of the cSPADE function. The dataset was then transformed to the basket type, also provided by the *arulesSequences* package. The transformed dataset was fed to cSPADE. This algorithm used a support hyperparameter, which allowed users to select a certain support threshold, and the maxlength hyperparameter, which took an arbitrary number as the maximum length of a sequence. The sequences found by cSPADE were transformed to a dataframe, which allowed for easier manipulation. The dataframe was then transposed, creating a column for every sequence. The columns user-id, response date and app categories per day were added to the dataframe, to allow for frequency per user-id and day to be measured. The frequencies were calculated, partially using the *numpy* and the *re* package in Python (Friedl, 2006; Oliphant, 2006).

The final number of sequences, i.e. features, depended heavily on the hyperparameter settings of the cSPADE algorithm. Optimal features were essential for the final model, which is why different hyperparameter settings have been analysed. Due to the computational expensiveness of the algorithm, support could not be lower than 0.6 and the sequences could not be longer than 6. If no boundaries were set, too many sequences were found. This caused computing problems, when counting the frequencies of sequences per unique user-id and response day combination. The maximum lengths 4, 5 and 6 were tested, together with the support rates 0.6, 0.7, 0.8 and 0.9. The effect of different feature sets can only be properly measured when the feature sets are used for prediction. Therefore, the different feature sets were compared on their stress prediction accuracy, using the training set. Accuracy is not the only evaluation method used to evaluate models in this research, but it is used here because of its simple interpretation when comparing models.

The outcomes of the hyperparameter analysis can be seen in the results section, visualised in figure 7. The highest accuracy was achieved using a maximum sequence length of 6 and a support rate of 0.8. These parameter settings were used to create the SPM model. A total of 8,552 patterns were found and used as features. The 15 most frequent features are shown in figure 6. Due to fitting reasons, Whatsapp Messenger has been shortened to W-app. Noteworthy is the presence of Whatsapp Messenger in 14 of the 15 most frequent features, as well as the sequences that only exist of Whatsapp Messenger uses.

4.4 Model creation and evaluation

After the final feature set was chosen for the SPM model, the stress levels were added to the corresponding dates, using user-id and response date to merge the datasets. A dataset was created for both categorical and binarised stress levels. After the final datasets were created, the user-id and response date columns were removed. The datasets were split into a training set, which contained 70% of the data, and test set, which contained 30% of the data. This was done using python's *sklearn* package (Pedregosa et al., 2011). XGBoost classifier was used, by implementing the *XGBoost* package in python (T. Chen & Guestrin, 2016).

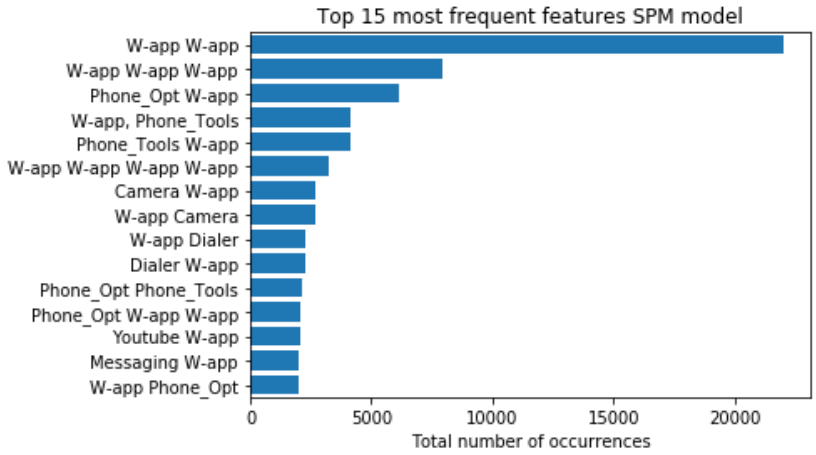


Figure 6: The 15 most frequent features in the dataset for the SPM model. Whatsapp Messenger has been shortened to W-app and Phone Optimisation to Phone_Opt, for to fitting purposes.

To put the XGBoost results into perspective, an SVM classifier has also been trained, using the previously mentioned *sklearn* package.

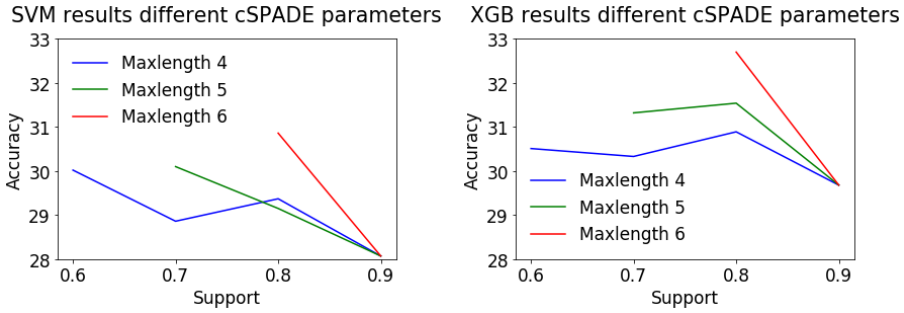
Parameter tuning was done, by using the Grid Search method from the *sklearn* package. Two parameters of XGBoost were tuned; depth of each decision tree (max_depth = 2, 4 or 6) and the number of decision trees (n_estimators = 50, 100 , 200). Two parameters of the SVM classifier were tuned. These parameters were the penalty parameter of the error term (C = 1, 10, 100 or 1,000) and the different kernels (kernel = linear, poly or rbf). The optimal settings for these hyperparameters are mentioned in the results section.

The final models were tested, using the test set. The results of the BoA and SPM models were compared on their accuracy and their respective confusion matrices, as well as the recall score. A baseline was set for the accuracy and recall scores, using the majority baseline model. This model only predicts the majority class, which is stress level 0 for both the categorised and the binarised stress levels. The scores of the BoA and SPM models are compared to this baseline in the results section. Finally, feature importance for both models was calculated to determine which features were most important for stress prediction.

5 Results

First, the results of the cSPADE hyperparameter analysis are presented. These results can be seen in figure 7. The highest accuracies for the SVM and XGBoost algorithm were 31.86% and 32.70%, respectively. The overall difference between SVM and XGboost classifiers was only a few percent on every score. The accuracy scores for all maximum lengths using 0.9 support were the same. This is due to

the fact that no extra patterns were found with different support and maxlength settings. The highest accuracy on the training set was achieved, using the 0.8 support and 6 maxlength settings. These settings were used to create the final feature set for the SPM model.



(a) **(b)**

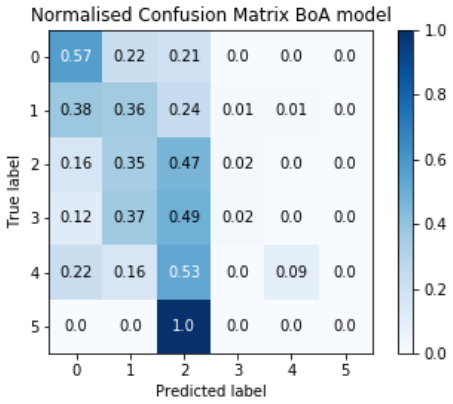
Figure 7: Accuracy results of different feature sets on the training set, generated by different cSPADE parameter settings. For both the SVM (7a) and the XGBoost (7b classifiers, a support rate of 0.8 with a maximum sequence length of 6 resulted in the highest accuracy.

In this paragraph, the categorical stress prediction results of the BoA and SPM model are presented. The results are listed in table 2. The accuracy for the majority baseline model is 29.07%, which is the relative frequency of the most frequent stress level (level 0), compared to the total amount of stress records. The recall for the majority baseline model is 16.67%, since level 0 has a recall of one and all other five levels have a recall of zero. Parameter tuning, using the Grid Search method, showed C = 10 and kernel = rbf as the optimal settings for the SVM. The XGBoost classifier scored the highest accuracy, using max_depth = 2 and n_estimators = 100. All models scored above the accuracy and recall baseline, although the SPM model was only slightly better. The BoA model, using XGBoost classifier, scored the highest accuracy (39.96%) and the highest recall (27.68%) on the test set. This suggests that the sequential patterns might not provide more information for stress prediction with phone usage data.

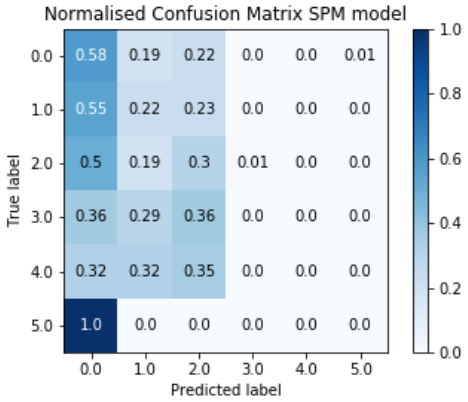
Table 2: Performance of the BoA and SPM models, using the SVM and XGBoost classifiers. Accuracy has been shortened to acc., for fitting purposes. The highest accuracy and recall scores on the test set for both models are depicted in bold font. Baselines for accuracy and recall were 29.07% and 16.67%, respectively.

	SVM classifier			XGBoost classifier		
	train acc.	test acc.	test recall	train acc.	test acc.	test recall
BoA model	39.41	39.18	25.27	39.68	**39.96**	**27.68**
SPM model	31.86	31.51	18.29	32.70	**31.71**	**18.77**

To add context to these scores, normalised confusion matrices were made for the XGBoost predictions of both models. These matrices are shown in figure 8. It seems that the BoA model is better at predicting stress level 1 and 2, while both models are approximately equally bad at predicting stress 3, 4 and 5. The BoA model seemed to have a strong tendency towards stress level 2, while the SPM model classifies most records as stress level 0. This could be explained by the fact that these stress levels are the most frequent stress levels, as was reported in figure 1 in the experimental setup section. This figure also showed that the classes are imbalanced, which could provide an explanation for the low scores on the less frequent stress levels.



(a)



(b)

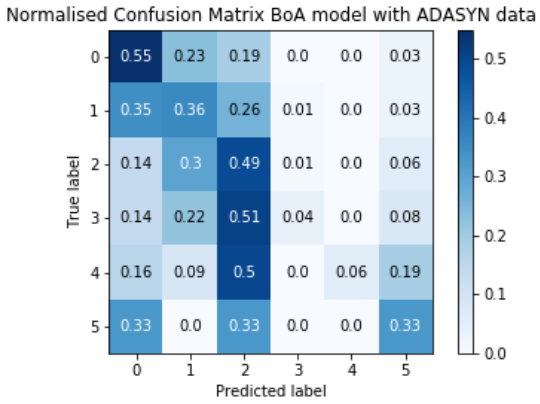
Figure 8: Normalised Confusion matrices of the XGBoost results on the test set. The results of the BoA model are depicted in figure 8a. The result of the SPM model are depicted in figure 8b. As can be seen, both models failed to predict stress level 3, 4 and 5.

To counter class imbalance, ADASYN was used to perform oversampling on the training data. The following paragraph describes the results of the ADASYN oversampling method and compares the results to the models trained with the original dataset. The accuracy and recall scores are listed in table 3. Because the test set was not changed by ADASYN, the baseline remains 29.07% for accuracy and 16.67% for recall. The BoA model, using the XGBoost classifier, scored the highest test accuracy (40.16%) and recall (30.55%). Both SPM results seem to indicate overfitting. The SPM model, using the SVM classifier, scored a lower accuracy score than the baseline (25.68%). The SPM model using the XGBoost classifier scored much better on the training set (43.50%) than on the test set (32.88%). Except for the SPM model with the SVM classifier, it seems that the addition of ADASYN oversampling only slightly increased accuracy and recall scores. This could be due to the fact that the test set is generated from the original dataset, which contained only a few level 5 records. This can be seen in table 1 in the experimental setup section.

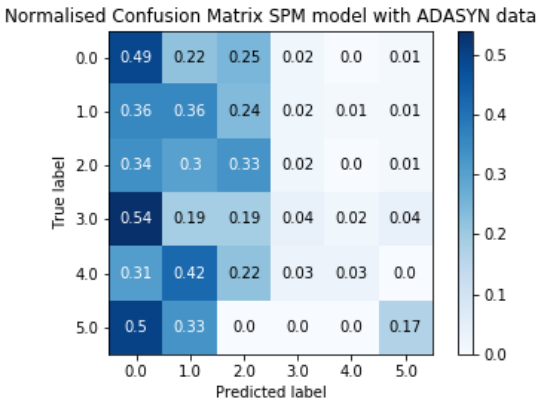
Table 3: Performance of the BoA and SPM models with ADASYN oversampled training sets, using the SVM and XGBoost classifiers. Accuracy has been shortened to acc., for fitting purposes. The highest accuracy and recall scores on the test set for both models are depicted in bold font. Baselines for accuracy and recall were 29.07% and 16.67%, respectively.

	SVM classifier			XGBoost classifier		
	train acc.	test acc.	test recall	train acc.	test acc.	test recall
BoA model	42.52	38.99	27.56	43.15	**40.16**	**30.55**
SPM model	31.20	25.68	21.07	43.50	**32.88**	**23.54**

For further analysis of the results, normalised confusion matrices were made for the XGBoost results of both models. These can be found in figure 9. The BoA model seems to be better at predicting stress levels 0, 2 and 5, compared to the SPM model. This contradicts expectations of prior research (Farrahi & Gatica-Perez, 2014; Yan et al., 2012). Compared to the results with the original dataset, the ADASYN trained SPM model is also able to better predict stress level 1 and 5. The prior tendency towards level 0 remained. The ADASYN trained BoA model is better than the BoA model trained on the original dataset, mainly because it was better at predicting stress level 5. This did not provide a high boost in test accuracy, since stress level 5 appears only scarcely in the test set. The BoA model also kept its prior tendency towards stress level 2.



(a)



(b)

Figure 9: Normalised Confusion matrices of the XGBoost results on the test set. The results of the BoA model are depicted in figure 9a. The result of the SPM model are depicted in figure 9b.

In this paragraph, the influence of different features on the BoA and SPM models is presented. The most important features of the BoA and SPM models, trained by the XGBoost classifier with the ADASYN training set, are compared. Reason for that is the higher score these models obtained, compared to the models trained with the original training set. The 15 most important features for the SPM model, based on F score, can be seen in appendix D. The sequence messaging - Whatsapp Messenger was the most important sequence for prediction in the SPM model. Whatsapp Messenger appears 14 times in the 15 most important sequences. Furthermore, almost all apps shown for the SPM model provide some form of connection with the outside world. Only phone tools, tools and Ethica do not provide social media services or a way to connect with people.

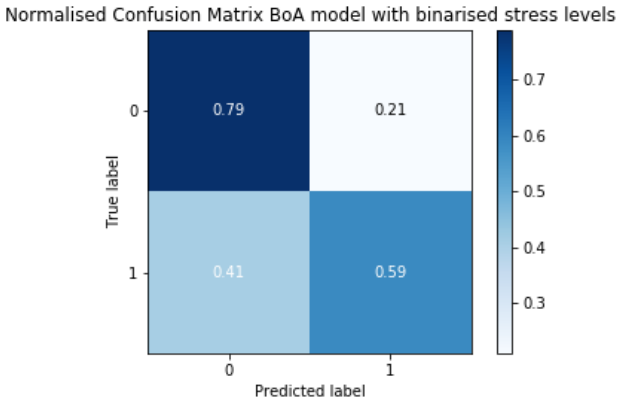
The same pattern can be seen in the most important features for the BoA model (appendix E). Only Ethica and phone tools do not provide some form of communication. Instant messaging category was the most important feature. It is interesting to see that the importance of Ethica, the app used to obtain the data for this research, has such an impact on the stress prediction in the BoA models. To a lesser degree, Ethica also had an impact on the prediction of the SPM model, appearing in 2 of the 15 most important sequences.

In this paragraph, the prediction results of binarised stress levels are presented. The majority baseline model scored an accuracy of 55.22% and a recall of 50.0%. For both the BoA and the SPM model, the results are shown in table 4. All models score above the baseline. The highest accuracy (69.79%) and recall (69.65%) scores were recorded by the BoA model, using the XGBoost classifier. Surprisingly, the accuracy test scores (SVM: 65.11%, XGB: 69.79%) of the BoA model were higher than the training scores (SVM: 62.23%, XGB: 68.32%) for the model. This could be coincidental, since the difference is modest. Another explanation could be that the test set contains more distinguishable cases, since both classifiers appear to underfit the BoA model. The highest score percentagewise, compared to the baseline, belonged to the BoA model with the XGBoost classifier, with a score of 14.57% above the accuracy baseline. Overall, the standard deviations of the binary stress prediction models were only slightly lower than the standard deviations of the models that predicted categorical stress. The BoA model scored higher on accuracy and recall with both classifiers, which seems to strengthen the previously mentioned suggestion that the sequential patterns might not provide more information than frequency patterns for stress prediction with phone usage data.

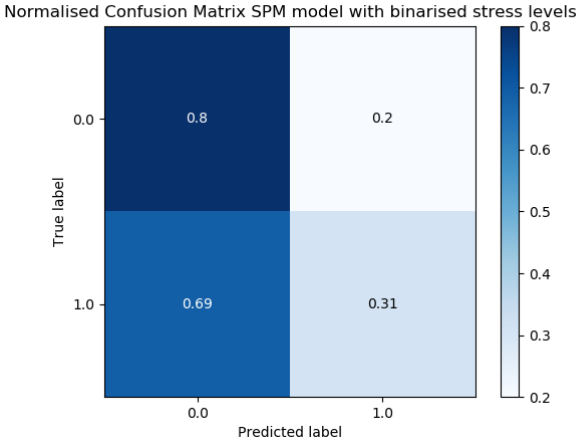
Table 4: Binary stress level performance of the BoA and SPM models, using the SVM and XGBoost classifiers. Accuracy has been shortened to acc., for fitting purposes. The highest accuracy and recall scores on the test set for both models is depicted in bold font. Baselines for accuracy and recall were 55.22% and 50.0%, respectively.

	SVM classifier			XGBoost classifier		
	train acc.	test acc.	test recall	train acc.	test acc.	test recall
BoA model	62.23	65.11	65.01	68.32	**69.79**	**69.65**
SPM model	59.41	56.42	52.87	62.69	**59.14**	**55.90**

To further analyse the results, normalised confusion matrices have been made. These are shown in figure 10. The matrices show that the difference in recall and accuracy scores is mainly due to the inability of the SPM model to correctly predict the presence of stress (stress level 1). The BoA model is able to somewhat accurately predict the presence of stress (59%), which was important for detection and treatment.



(a)



(b)

Figure 10: Normalised Confusion matrices of the XGBoost results on the test set. The results of the BoA model are depicted in figure 10a. The result of the SPM model are depicted in figure 10b.

The following section presents the most important features of the BoA and SPM model, when using the XGBoost classifier to predict binary stress levels. The 15 most important features of the SPM model are shown in appendix F. Appendix G shows most important features of the BoA model. The most important features for both models with binary stress levels differ from the most important features for categorised stress prediction. For the SPM model, Whatsapp Messenger and phone tools are still important. However, other apps appear to contribute to the model as well. Dialer, Ethica, camera and tools are new in the top four most important features. For the BoA model, Whatsapp Messenger is no longer in the top 15 most important apps. Ethica also dropped

in importance. The most consistent factor for the BoA model is the importance of the phone tools category. Where social media and communication categories dominated the BoA model trained with categorised stress levels, the BoA model with binarised stress levels is also influenced by lifestyle app categories, like personal finance, sports or phone personalisation.

6 Discussion

The aim of this study was to analyse the performance of two different models on stress prediction; the non-sequential BoA and sequential SPM model. The BoA model used daily frequency distributions of every app as features, whereas the SPM model used daily frequency distributions of sequentially ordered app usage patterns. The two models were compared to analyse the importance of the sequential orders for stress prediction. Importance of the dimension of the stress level has been tested using either binarised or categorical stress levels. The prediction performance of both settings was compared in this study. All suggestions for future research are combined in the eponymous subsection.

6.1 Hyperparameter testing cSPADE

The first results showed the outcomes of hyperparameter testing for the cSPADE algorithm on the training set. Optimal results for both the XGBoost and SVM classifier were found, using a support rate of 0.8 and a maximum sequence length of 6. Although these were the optimal settings, the accuracy scores did not differ greatly between settings. This could have been influenced by the fact that this study was only able to test a few hyperparameter settings, due to the computational expensiveness of the algorithm.

Previous studies did not set a maximum sequence length and used a lower support rate (Ibrahim & Shafiq, 2019; Smedley et al., 2018; A. P. Wright et al., 2015). A threshold was set in the present study, due to the computationally exhaustive algorithm. The fact that a maximum sequence length of six scored the highest accuracy is in line with Pei et al. (2004) and Zaki (2001). They suggested that a sequence gains more specific information for every item added to said sequence, provided that the support threshold is still met.

6.2 Categorical stress prediction

The BoA and SPM model were able to predict daily categorical stress levels. Both models outperformed the baseline. The BoA model recorded a high accuracy and recall, mainly due to the fact that the BoA model was better at predicting stress level 1 and 2. Both models were especially bad at predicting stress levels 3, 4 and 5. This could be due to class imbalance in the data.

To test this, the ADASYN oversampling method was used on the training set. The results showed that this only slightly improved accuracy and recall, compared to the models trained with the original data. After further analysis,

it became clear that the slight increase was mainly due to increased prediction performance on stress level 5. Although this seems positive, the result could be marginalised by the fact that the test set only contained six stress level 5 records. This could mean that sequential and non-sequential patterns are important for lower stress level prediction, but not for higher stress levels. However, this might be influenced by the aforementioned unbalanced classes in the training and test set, which could only partially be solved by the ADASYN oversampling method.

When comparing the models, it becomes clear that both the original and ADASYN trained BoA model outperformed the SPM model. This is not in line with prior research, which suggested that the sequential order of patterns provides additional information on the data (Farrahi & Gatica-Perez, 2014; Alibasa et al., 2019). This could mean that sequential patterns do not add as much information as thought, when it comes to categorical stress prediction. This could have been influenced by the fact that SPM was not fully used, due to a lack of computing power. As mentioned earlier, maximum sequence length is usually not limited and support rates are usually set lower than 0.8. Excluding these limitations could lead to more sequences, providing more information.

The relatively good performance of the BoA model, compared to the baseline, is in line with Do & Gatica-Perez (2010). The article presented the BoA model, because it could unravel meaningful patterns in phone usage data. Ferdous et al. (2015) showed that temporal app usage patterns are able to predict stress, on a user-specific level. The results of the BoA model provide support for the idea that the usage frequency of apps also contains information about daily categorical stress levels. A combination of the temporal and BoA model could be interesting. This is addressed in the suggestions for future research.

The results also displayed the most important patterns for both models, based on feature importance. Messaging and social media categories are dominant in both models. This adds to the conclusions made by Kushlev & Dunn (2015); K. B. Wright et al. (2014), who concluded that social media and work related apps, like email, could nurture constant feelings of stress. The findings in this study suggest that they are not only able to nurture stress, but might also have the ability to indicate the presence of stress. The Ethica app, used to obtain the data, seems to have had a significant impact on the prediction of the BoA model and a moderate impact on the prediction of the SPM model. It could be that the participants mostly answered surveys, when they were particularly stressed or relaxed. Another possibility could be that the apps might have influenced the stress level of participants, as is discussed in the suggestions for future research.

Lastly, it is interesting to note that no sequences containing five or six apps are in the top 15 most important features. This could be due to the way decision trees work and how the F score is calculated. A feature gains a higher F score when it is used for more splits in the decision trees. The more frequent features are in the dataset, the more likely it is they are used in multiple splits. As mentioned earlier, longer sequences have a tendency to contain more specific information (Pei et al., 2004; Zaki, 2001). One possibility is to focus on longer sequences by setting a minimum sequence length, as is also discussed in the suggestions for future research.

6.3 Binary stress prediction

The prediction of binary stress levels showed similar results to the daily categorical stress prediction. Again, the XGBoost classifier outperformed the SVM classifier and the BoA model outperformed the SPM model. The BoA model, using the XGBoost classifier, scored highest percentagewise compared to the baseline. This shows that the non-sequential BoA representation contains information about the distribution of stress levels. A BoA model could be used to detect stress among phone users. This could contribute to the identification of a risk group, which could result in better diffusion of anti-stress remedies, treatments or solutions.

The binary stress prediction results also show that the BoA model outperforms the SPM model. This could provide more evidence against claims about the added value of sequential patterns, as described in the previous subsection. This seems to indicate that app usage frequencies within a time frame offer more information about stress than the order in which apps are used. The value of sequential patterns for stress prediction seems questionable, although further research should investigate different hyperparameter settings to prove this claim.

It is also important to mention that the XGBoost classifier outperformed SVM classifier in all categorical and binary stress prediction problems. This in line with T. Chen & Guestrin (2016). They mentioned how XGBoost has won, and is winning, a great number of Kaggle competitions. Therefore, it has come as no surprise that XGBoost also outperforms SVM on this problem.

Looking at the most important features for binary stress prediction, it becomes clear that messaging and social media apps still have a significant impact on both models. The impact of the Ethica app increased marginally for the binary SPM model and dropped somewhat for the binary BoA model, compared to the categorical models. These findings support the claims made in the previous subsection, about the influence the data-collection app might have on the results. Future research should recognise this, as is further discussed in the suggestions for future research.

Another reoccurring result is the absence of sequences of length 5 or 6 in the top 15 most important features of the SPM model. This strengthens the claim made in the previous paragraph, about the exclusion of short sequences to improve importance of longer sequences. This is favored, because longer sequences are considered to contain more specific information than shorter sequences, as mentioned earlier (Pei et al., 2004; Zaki, 2001).

6.4 Suggestions for future research

The main limitation of the present study was the unavailability of the computing power necessary to analyse a wider variety of cSPADE hyperparameter settings. Future research should focus on testing a bigger range of hyperparameters for parameter optimisation. Most studies have used lower thresholds than the present study (Alibasa et al., 2019; Deeva et al., 2017; Ibrahim & Shafiq, 2019; A. P. Wright et al., 2015). Lower thresholds are likely to produce different results,

because the feature set changes with every setting.

The present study showed that non-sequential app usage patterns can be useful for binary stress prediction. As mentioned earlier, detecting the presence of stress can be valuable for providing general stress remedies and treatments. Ferdous et al. (2015) showed that temporal patterns can also be used for stress prediction. Future research could combine the non-sequential BoA approach with a temporal model. This could possibly lead to even better stress prediction results. It is advised that such research uses the XGBoost classifier as (one of) the classifier(s) to predict stress, because it outperformed the SVM classifier on every classification problem.

Future research could also decide to focus on longer sequences, using an arbitrary minimum sequence length. As mentioned earlier, these sequences are likely to contain more specific information. The more frequent shorter patterns dominated the classifiers. Setting a minimum sequence length negates this dominance and could reduce the amount of noise the classifier has to deal with, because less sequences are analysed. It could be especially useful for categorical stress levels, because distinguishing between multiple categories proved to be harder than distinguishing between two categories.

Finally, it is important that future research recognises the effect the data collection app might have on the study. It might be the case that the current way of measuring stress might not be as unconstrained as hoped. Although the Ethica app was only moderately important for the most successful models, it did appear in the top 15 most important features of every model. Because phone usage data is used in numerous studies and is likely to be used in the future, it might be interesting to further analyse the effect of the data collection app. The results could be used to design a method, aimed at minimising the effect of the app.

7 Conclusion

Two methods for stress prediction were presented. The BoA method focused on non-sequential frequencies of app categories, while the SPM method focused on the frequencies of sequential patterns between these app categories. The BoA method and SPM method both produced models that scored above the baseline. The BoA model outperformed the SPM model in every tested scenario. The best performing model was the BoA model for binary stress prediction, using the XGBoost classifier. This could indicate that app usage patterns are able to predict stress, especially on a binary level. The results also indicate that the added information that sequential patterns provide might not be able to produce better stress prediction results. Further research is needed to support this claim. An interesting extension of this study could be the exclusion of short sequences for the SPM model, or creating a combination of temporal and non-sequential app usage patterns to predict stress.

References

AbuAlRub, R. F. (2004). Job stress, job performance, and social support among hospital nurses. *Journal of nursing scholarship*, *36*(1), 73–78.

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the internet. *Glottometrics*, *3*(1), 143–150.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, pp. 207–216).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, vldb* (Vol. 1215, pp. 487–499).

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *icde* (Vol. 95, pp. 3–14).

Alibasa, M. J., Calvo, R. A., & Yacef, K. (2019). Sequential pattern mining suggests wellbeing supportive behaviors. *IEEE Access*, *7*, 130133–130143.

Anantharam, P., Thirunarayan, K., Taslimi, V., & Sheth, A. P. (2013). Predicting parkinson's disease progression with smartphone data.

Aseervatham, S., & Osmani, A. (2005). Mining short sequential patterns for hepatitis type detection. In *Ecml/pkdd 2005-discovery challenge workshop* (p. 6).

Baeza-Yates, R., Jiang, D., Silvestri, F., & Harrison, B. (2015). Predicting the next app that you are going to use. In *Proceedings of the eighth acm international conference on web search and data mining* (pp. 285–294).

Bakker, J., Pechenizkiy, M., & Sidorova, N. (2011). What's your current stress level? detection of stress patterns from gsr sensor data. In *2011 ieee 11th international conference on data mining workshops* (pp. 573–580).

Björkegren, D., & Grissen, D. (2018). Behavior revealed in mobile phone usage predicts loan repayment. *Available at SSRN 2611775*.

Buchta, C., Hahsler, M., Buchta, M. C., & Matrix, I. (2007). *The arulessequences package.*

Buchta, C., Hahsler, M., Diaz, D., Buchta, M. C., & Zaki, M. J. (2019). Package 'arulessequences'.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, T., & He, T. (2015). Higgs boson discovery with boosted trees. In *Nips 2014 workshop on high-energy physics and machine learning* (pp. 69–80).

Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., & Peng, J. (2018). Xgboost classifier for ddos attack detection and analysis in sdn-based cloud. In *2018 ieee international conference on big data and smart computing (bigcomp)* (pp. 251–256).

Deeva, G., De Smedt, J., De Koninck, P., & De Weerdt, J. (2017). Dropout prediction in moocs: a comparison between process and sequence mining. In *International conference on business process management* (pp. 243–255).

De Smedt, J., Deeva, G., & De Weerdt, J. (2019). Mining behavioral sequence constraints for classification. *IEEE Transactions on Knowledge and Data Engineering*.

Do, T.-M.-T., & Gatica-Perez, D. (2010). By their apps you shall understand them: mining large-scale patterns of mobile phone usage. In *Proceedings of the 9th international conference on mobile and ubiquitous multimedia* (p. 27).

Exarchos, T. P., Papaloukas, C., Lampros, C., & Fotiadis, D. I. (2008). Mining sequential patterns for protein fold recognition. *Journal of Biomedical Informatics*, *41*(1), 165–179.

Farrahi, K., & Gatica-Perez, D. (2014). A probabilistic approach to mining mobile phone data sequences. *Personal and ubiquitous computing*, *18*(1), 223–238.

Feinerer, I. (2018). Introduction to the tm package text mining in r. *Retrieved March*, *1*, 2019.

Fennell, C., Barkley, J. E., & Lepp, A. (2019). The relationship between cell phone use, physical activity, and sedentary behavior in adults aged 18–80. *Computers in Human Behavior*, *90*, 53–59.

Ferdous, R., Osmani, V., & Mayora, O. (2015). Smartphone app usage as a predictor of perceived stress levels at workplace. In *2015 9th international conference on pervasive computing technologies for healthcare (pervasivehealth)* (pp. 225–228).

Friedl, J. E. (2006). *Mastering regular expressions.* " O'Reilly Media, Inc.".

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Glanz, K., & Schwartz, M. D. (2008). Stress, coping, and health behavior. *Health behavior and health education: Theory, research, and practice*, *4*, 211–236.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25.

Grothendieck, G., & Grothendieck, M. G. (2017). *Package 'sqldf'.*

Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Oehler, S., Tröster, G., ... Lukowicz, P. (2014). Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics*, *19*(1), 140–148.

Hagquist, C. (2010). Discrepant trends in mental health complaints among younger and older adolescents in sweden: an analysis of who data 1985–2005. *Journal of Adolescent Health*, *46*(3), 258–264.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1322–1328).

Horwood, S., & Anglim, J. (2018). Personality and problematic smartphone use: A facet-level analysis using the five factor model and hexaco frameworks. *Computers in Human Behavior*, *85*, 349–359.

Horwood, S., & Anglim, J. (2019). Problematic smartphone usage and subjective and psychological well-being. *Computers in Human Behavior*, *97*, 44–50.

Ibrahim, R., & Shafiq, M. O. (2019, May 13). Detecting taxi movements using random swap clustering and sequential pattern mining. *Journal of Big Data*, *6*(1), 39. Retrieved from https://doi.org/10.1186/s40537-019-0203-6 doi: 10.1186/s40537-019-0203-6

Julea, A., Méger, N., Trouvé, E., & Bolon, P. (2008). On extracting evolutions from satellite image time series. In *Igarss 2008-2008 ieee international geoscience and remote sensing symposium* (Vol. 5, pp. V–228).

Kemp, S. (2019). *Digital 2019: Global internet use accelerates.* https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates. (Accessed: 24-09-2019)

Korn Ferry Institute. (2018). *Workplace stress continues to mount.* https://www.kornferry.com/institute/workplace-stress-motivation. (Accessed: 24-09-2019)

Kushlev, K., & Dunn, E. W. (2015). Checking email less frequently reduces stress. *Computers in Human Behavior*, *43*, 220–228.

Kuss, D. J., Kanjo, E., Crook-Rumsey, M., Kibowski, F., Wang, G. Y., & Sumich, A. (2018). Problematic mobile phone use and addiction across generations: The roles of psychopathological symptoms and smartphone use. *Journal of technology in behavioral science*, *3*(3), 141–149.

Lee, S., Lim, J., Park, J., & Kim, K. (2016). Next place prediction based on spatiotemporal pattern mining of mobile device logs. *Sensors*, *16*(2), 145.

Lee, Y.-K., Chang, C.-T., Lin, Y., & Cheng, Z.-H. (2014). The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress. *Computers in human behavior*, *31*, 373–383.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, *18*(1), 559–563.

Lepp, A., Barkley, J. E., & Karpinski, A. C. (2014). The relationship between cell phone use, academic performance, anxiety, and satisfaction with life in college students. *Computers in Human Behavior*, *31*, 343–350.

Lepp, A., Li, J., Barkley, J. E., & Salehi-Esfahani, S. (2015). Exploring the relationships between college students' cell phone use, personality and leisure. *Computers in human behavior*, *43*, 210–219.

Liu, H., He, G., Jiao, W., Wang, G., Peng, Y., & Cheng, B. (2017). Sequential pattern mining of land cover dynamics based on time-series remote sensing images. *Multimedia Tools and Applications*, *76*(21), 22919–22942.

Majeno, A., Tsai, K. M., Huynh, V. W., McCreath, H., & Fuligni, A. J. (2018). Discrimination and sleep difficulties during adolescence: the mediating roles of loneliness and perceived stress. *Journal of youth and adolescence*, *47*(1), 135–147.

Matic, A., Osmani, V., & Mayora-Ibarra, O. (2014). Mobile monitoring of formal and informal social interactions at workplace. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 1035–1044).

Maxhuni, A., Hernandez-Leal, P., Morales, E. F., Sucar, L. E., Osmani, V., Muńoz-Meléndez, A., & Mayora, O. (2017). Using intermediate models and knowledge learning to improve stress prediction. In *Applications for future internet* (pp. 140–151). Springer.

Maxhuni, A., Hernandez-Leal, P., Sucar, L. E., Osmani, V., Morales, E. F., & Mayora, O. (2016). Stress modelling and prediction in presence of scarce data. *Journal of biomedical informatics*, *63*, 344–356.

Oliphant, T. E. (2006). *A guide to numpy* (Vol. 1). Trelgol Publishing USA.

O'Malley, S. (2019). *Fragile: Why we are feeling more stressed, anxious and overwhelmed than ever (and what we can do about it).* Gill & Macmillan Ltd.

Osmani, V. (2015). Smartphones in mental health: detecting depressive and manic episodes. *IEEE Pervasive Computing*, *14*(3), 10–13.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Dubourg, V. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., ... Hsu, M.-C. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on knowledge and data engineering*, *16*(11), 1424–1440.

Pourbabaee, B., Patterson, M., Brais, R., Reiher, E., & Benard, F. (2018). Daily mental stress prediction using heart rate variability. *CMBES Proceedings*, *41*.

Quick, J. D., Horn, R. S., & Quick, J. C. (1987). Health consequences of stress. *Journal of Organizational Behavior Management*, *8*(2), 19–36.

Reddy, U. S., Thota, A. V., & Dharun, A. (2018). Machine learning techniques for stress prediction in working employees. In *2018 ieee international conference on computational intelligence and computing research (iccic)* (pp. 1–4).

Selkie, E. (2019). Smartphone ownership as a developmental milestone. *Journal of Adolescent Health*, *64*(4), 419–420.

Smedley, B. M., Nova F.and Ellingson, Cloughesy, T. F., & Hsu, W. (2018, September 26). Longitudinal patterns in clinical and imaging measurements predict residual survival in glioblastoma patients. *Sci. Rep.*, *8*(14429). Retrieved from https://doi.org/10.1038/s41598-018-32397-z doi: 10.1038/s41598-018-32397-z

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011). Prediction of socioeconomic levels using cell phone records. In *International conference on user modeling, adaptation, and personalization* (pp. 377–388).

Srikant, R., & Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In *International conference on extending database technology* (pp. 1–17).

Stütz, T., Kowar, T., Kager, M., Tiefengrabner, M., Stuppner, M., Blechert, J., ... Ginzinger, S. (2015). Smartphone based stress prediction. In *International conference on user modeling, adaptation, and personalization* (pp. 240–251).

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293–300.

Thomée, S., Härenstam, A., & Hagberg, M. (2011). Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults-a prospective cohort study. *BMC public health*, *11*(1), 66.

Verma, M., & Mehta, D. (2014). Sequential pattern mining: A comparison between gsp, spade and prefix span 1.

Wang, Y., Hou, W., & Wang, F. (2018). Mining co-occurrence and sequence patterns from cancer diagnoses in new york state. *PloS one*, *13*(4), e0194407.

Weilenmann, A., & Larsson, C. (2002). Local use and sharing of mobile phones. In *Wireless world* (pp. 92–107). Springer.

Wiegner, L., Hange, D., Björkelund, C., & Ahlborg, G. (2015). Prevalence of perceived stress and associations to symptoms of exhaustion, depression and anxiety in a working age population seeking primary care-an observational study. *BMC family practice*, *16*(1), 38.

Woltermann, B., & Schroedl, S. (2003, July 29). *Personalized driver stress prediction using geographical databases.* Google Patents. (US Patent 6,599,243)

Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2015). The use of sequential pattern mining to predict next prescribed medications. *Journal of biomedical informatics*, *53*, 73–80.

Wright, K. B., Abendschein, B., Wombacher, K., O'Connor, M., Hoffman, M., Dempsey, M., ... Shelton, A. (2014). Work-related communication technology use outside of regular work hours and work life conflict: The influence of communication technologies on perceived work life conflict, burnout, job satisfaction, and turnover intentions. *Management Communication Quarterly*, *28*(4), 507–530.

Yan, T., Chu, D., Ganesan, D., Kansal, A., & Liu, J. (2012). Fast app launching for mobile devices using predictive user context. In *Proceedings of the 10th international conference on mobile systems, applications, and services* (pp. 113–126).

Zaki, M. J. (2000). Sequence mining in categorical domains: incorporating constraints. In *Proceedings of the ninth international conference on information and knowledge management* (pp. 422–429).

Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, *42*(1-2), 31–60.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1-4), 43–52.

Zipf, G. K. (1932). Selected studies of the principle of relative frequency in language.

Appendices

A

The following link leads to the GitHub repository, containing the code used to conduct this research:

Thesis code repository Aaron Wijnker

B

Table 5: Stress level distribution of the original training set, compared to the ADASYN oversampled training sets for both the SPM and the BoA models.

Stress level	Original training set	ADASYN training set SPM	ADASYN training set BoA
0	348	355	348
1	310	313	315
2	334	335	334
3	125	116	121
4	73	71	76
5	9	352	351

C

Table 6: Categorisation of the ten most used apps without a category. The 'source' column refers to a web page, where the app is explained and categorised by the developers. IKeyboard Blue Love Heart Theme has been shortened to IKeyboard, for fitting purposes.

App	Category	Source
Android SystemUI	Background_Process	Google Source (n.d.)
Ethica	Ethica	Ethica Data (2019)
Parallel Space	Phone_Optimisation	LBE Tech (2019)
Tilburg Osiris	Education	CACI bv (2019)
Napster	Streaming_Services	Rhapsody International, Inc (2019)
Nexus Launcher	Phone_Personalisation	xda-developers (2016)
Deliveroo Rider	Business_Management	Deliveroo (2019)
Mo PTT	Social_Networking	Mottx Co., Ltd. (2019)
Asterix and Friends	Game_Multiplayer	APKsHub (2019)
IKeyboard	Phone_Personalisation	Theme Design Apps for Android (2019)

D

Table 7: Top 15 most important features of the XGBoost BoA model, trained with the ADASYN oversampled training set. The features are ranked on F score. Whatsapp Messenger has been shortened to Whatsapp and Phone Optimisation has been shortened to Phone Opt, for fitting purposes.

Features	F score
Messaging, Whatsapp	47
Phone Tools, Phone Tools	42
Whatsapp, Youtube	35
Whatsapp, Whatsapp, Whatsapp	34
Whatsapp, Whatsapp	30
Youtube, Youtube	29
Ethica, Whatsapp	28
Whatsapp, Ethica	28
Phone Opt, Youtube	26
Phone Tools, Whatsapp	26
Google Chrome, Phone Tools	25
Phone Opt, Email	25
Phone Opt, Whatsapp, Whatsapp, Whatsapp	24
Youtube, Whatsapp	24
Email, Phone Tools	24

E

Table 8: Top 15 most important features of the XGBoost BoA model, trained with the ADASYN oversampled training set. The features are ranked on F score.

features	F score
Instant Messaging	104
Ethica	102
Phone Tools	99
Instagram	94
Whatsapp Messenger	77
Facebook	73
Google Chrome	65
News	65
Dialer	63
Streaming Services	61
Youtube	51
Snapchat	51
Email	50
Dating	45
Maps	44

F

Table 9: Top 15 most important features of the XGBoost SPM model, when predicting binary stress levels. The features are ranked on F score. Whatsapp Messenger has been shortened to Whatsapp and Phone Optimisation has been shortened to Phone Opt, for fitting purposes.

features	F score
Camera, Phone Tools, Dialer	10
Whatsapp, Dialer	10
Phone Opt, Whatsapp, Phone Tools, Ethica	8
Internet Browser, Ethica	7
Email, Phone Tools	7
Whatsapp, Whatsapp	7
Phone Opt, Whatsapp	7
Phone Opt, Phone Opt, Phone Opt, Ethica	6
Whatsapp, Whatsapp, Whatsapp	6
Whatsapp, Dialer, Whatsapp, Phone Tools	6
Whatsapp, Whatsapp, Youtube	6
Phone Tools, Whatsapp	6
Youtube, Whatsapp, Whatsapp	5
Phone Opt, Camera, Youtube	5
Phone Tools, Phone Tools, Camera, Whatsapp	5

G

Table 10: Top 15 most important features of the XGBoost BoA model, when predicting binary stress levels. The features are ranked on F score.

features	F score
Social Networking	25
Phone Tools	21
News	21
Google Chrome	15
Personal Finance	14
Internet Browser	13
Dialer	12
Phone Personalisation	12
Weather	11
Sports	11
Instagram	11
Snapchat	9
Streaming Services	9
Ethica	8
Education	7