UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

UNDERSTANDING FACTORS INFLUENCING LONELINESS AND CORRESPONDING QUALITY OF LIFE VIA INTERVIEWS, SMARTPHONE SENSING AND DATA MODELING

MASTER'S THESIS

Written by *Adam Honoré*
August 4, 2017

Supervised by
Katarzyna Wac

UNIVERSITY OF COPENHAGEN

UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

FACULTY:	Faculty of SCIENCE
INSTITUTE:	Department of Computer Science
AUTHOR(S):	Adam Honoré
EMAIL:	qgf142@alumni.ku.dk
TITLE AND SUBTITLE:	Understanding factors influencing loneliness and corresponding Quality of Life via interviews, smartphone sensing and data modeling -
SUPERVISOR(S):	Katarzyna Wac
HANDED IN:	07.08.2017
DEFENDED:	22.08.2017

NAME Adam Honoré

SIGNATURE Adam Honoré

DATE 04/08/2017

Resumé

Ensomhed er et udbredt fænomen, der påvirker mennesker verden over i alle lag af samfundet. Det har vist sig at øge sandsynligheden for tidlig død med op til 26%, hvilket understreger vigtigheden af at kunne måle og forudsige ensomhed. I dette speciale indsamler vi data om smartphonebrugsmønstre og måler ensomhed fra tre studerende over en periode på fire uger. Brugsmønstre bliver udregnet som funktioner af dataen fra én forsøgsperson og analyseres ved hjælp af datamodellering og maskinlæring for at afgrænse forholdet mellem disse brugsmønstre og den målte ensomhed. Flere prædiktive modeller er blevet evalueret, hvorefter vi har valgt den mest optimale model. Vores prædiktive model viser tegn på overtilpasning, men er stadig i stand til at bekræfte en sammenhæng mellem en stigning i mængden af fysisk aktivitet og nedsat ensomhed, som det også er set i lignende studier. Derudover fandt vi også en forbindelse mellem øgede mængder af tid tilbragt i nærheden af andre mennesker, målt som forsøgspersonens semantiske placering, og nedsat ensomhed. Vi fandt også en forbindelse mellem en stigning i antallet af gange, hvor forsøgspersonen brugte sin e-mail-app, og nedsat ensomhed. Baseret på vores arbejde i dette speciale validerer vi teorier fra relaterede studier via eksperimenter og lægger grundlaget for fremtidigt arbejde på dette område.

Abstract

Loneliness is a widespread phenomenon that affects people all over the world, in all layers of society. It has been shown to increase the likelihood of premature mortality by as much as 26%, which conveys the importance of being able to assess and predict loneliness. In this thesis, we gather data on smartphone usage patterns and assessed loneliness from three students over a period of four weeks. Smartphone usage features are extracted from the data of one subject and analyzed using data modeling and machine learning, to delineate the relationships between these features and the assessed loneliness. Predictive models have been evaluated, and the best performing model has been selected. Our predictive model shows signs of overfitting but is still able to confirm a link between an increase in physical activity and decreased loneliness, as is also found in related studies. Additionally, we also found links between increased amounts of time spent near other people, as measured by the subject's semantic location, and decreased loneliness. A link between an increase in the number of times that the subject used their Email app and decreased loneliness was also found. Based on our work in this thesis we experimentally validate theories from related studies and lay the foundation for future work in this area.

Contents

Chapter 1

Introduction

In this chapter, we give an introduction to the problem addressed in this thesis.

1.1 Problem Statement

Loneliness is a widespread phenomenon that affects people all over the world, in all layers of society. It affects people's lives and their health, and recent studies have shown that the amount of people who regularly feels lonely may be as high as 35% [35], which is predicted to increase to even higher numbers in the future [24].
The effects of loneliness [4] range from stress to depression and can have a significant impact on people's lives and health. Recent studies have even shown that loneliness might increase the likelihood of early mortality by as much as 26% [14], indicating that it could be a more significant factor of increased mortality than obesity [14]. This conveys the importance of being able to assess individuals' loneliness reliably, and potentially be able to help people avoid it.

In modern society, loneliness seems to be a sign of the times, but it may also be considered a taboo, making it hard for people to admit that they are lonely. This makes it difficult to help people, as they are less likely to seek help for something that they are not supposed to talk about. Saying *"I want a friend"* is not considered socially acceptable and highlights just why this problem is so hard to solve.

One might think that with the interconnectedness of the internet, and widespread use of "social" networks, people should not become more lonely. Several recent studies have shown quite the opposite [27, 36, 20, 19], indicating that extensive use of the internet and social media increased individuals' feelings of loneliness and perceived social isolation.

Our approach is that we can leverage the interconnectedness of today's world, and the accompanying ubiquity of smartphones in people's lives to solve this problem. We aim to collect extensive smartphone usage data and use this to computationally model the relationships between collected data and self-assessed loneliness. By modeling the relationships and analyzing the factors influencing them, we hope to gain insight into how and why people feel lonely, which might shed light on how we can help people reduce their feelings of loneliness.

1.2 Contributions

As a result of our work in this thesis, we have been able to make three main contributions to this specific area of research.

Firstly, we have explored a relatively new field of science (behavior modeling through smartphone sensing), and thereby helped increase the amount of work being done in this area. This is important, as can be seen from the small amount of related works that we were able to find, indicating a lack of research on the relationships between smartphone usage and psychological effects.

Secondly, by exploring this field, we have also collected empirical data between smartphone usage and psychological effects. By employing data modeling, we have constructed a classification model that can predict a specific subject's loneliness on a scale of "low," "medium," and "high" with a Cohen's Kappa performance of 0.87, indicating a substantial agreement between predicted loneliness and assessed loneliness.

Our empirical data could be of use in other studies, either as a part of any future work done to expand this field by building upon our findings in this thesis, or work done to confirm our findings, which would thereby help validate our results.

Finally, we were able to experimentally validate theories put forward by related studies. As reported by several of our related works [34, 31, 2], we also found links between an increase in physical activity and decreased loneliness, and by validating their theories, we help strengthen the arguments of their findings.

1.3 Thesis Overview

This chapter contains the introduction; chapter 2 discusses related work; chapter 3 includes the definitions; chapter 4 includes the methods used; chapter 5 analyzes and reports our findings; chapter 6 discusses our findings; chapter 7 concludes on our findings.

Chapter 2

Related Work

The prediction of certain psychological effects based on data models, created by analyzing data collected using smartphone sensing, is a relatively new field of research, with some of the earliest work on behaviour prediction using smartphone sensing data being done in 2014 by Wang et al. [34]. Since then, several studies [34, 31, 2, 11, 28] have been done to try and predict a range of psychological effects based on data collected from the now ubiquitous smartphone using smartphone sensing and various experience sampling methods. These psychological effects includes depression [2, 28], stress [34, 2], loneliness [31, 2, 11], among others. From the resulting predictive model, some studies [34, 31, 28] even developed a smartphone application to continuously predict the studied psychological effect.

Sanchez et al. [31] worked with a group of 100 seniors, who is commonly a part of the population which is not accustomed to using smartphones. This required them to collect smartphone usage data via interviews, which might have resulted in recall error as the seniors had to remember their behavior and how they used their smartphone during the last week. As the other studies [34, 2, 11, 28] worked with a student population in which smartphones are more common, they used various existing[1] or self-developed smartphone sensing applications to collect smartphone usage data from the subjects' phones, combined with experience sampling and psychological surveys. Farhan et al. [11] employed a clinician assessment to provide a ground-truth for their depression data. The widespread use of students in the related studies is likely because they were easily available, and usually have a high percentage of smartphone owners among them, making it easier to find viable subjects. The studies ranged from 4-10 weeks [34, 2, 28], up to seven months [11], and included from nine to 79 students.

For the data analysis, all the studies employed some method to extract meaningful features from the collected data, which were then combined with the measured labels of their respective psychological effect(s). Sanchez et al. [31] employed dataset balancing before doing their analysis, as they recognized the importance of a balanced dataset when doing classification using machine learning classifiers. Per-feature linear correlation between each study's respective psychological effect(s) and the extracted features was done, and some studies continued to do multi-feature correlation by using various machine learning classifiers [31, 11, 28]. Farhan et al. [11] employed clinical ground-truth for depression as

[1] AWARE (`http://www.awareframework.com/`)

part of their analysis to train their machine learning classifiers, and ensure a prediction quality close to that of a clinician assessment.

For the results, several of the related studies show meaningful correlations between specific features and their respective psychological effect(s). One study [34] found a link between activity duration, distance traveled, indoor mobility and loneliness. Another study [31] found a link between the average time spent outside of the home, total number of outings and loneliness, interpreting this as “*older adults who keeps busy doing different activities like going to elderly clubs, dancing, going to the church, going to supermarket, etc., tend to be less likely to suffer from loneliness*” [31]. For Ben-Zeev et al. [2], kinesthetic activity was associated with changes in loneliness, and they found “*associations between changes in depression and sensor-derived speech duration (p<.05), geospatial activity (p<.05), and sleep duration*” [2], where geospatial activity was calculated as the total distance covered daily. Farhan et al. [11] found “*significant correlation between depressive mood and social interaction (specifically, conversation duration and number of co-locations)*” [11], showing how important social interaction is for humans to avoid depression. Pulekar and Agu [28] “*synthesized machine learning classifiers that classified user interactions into ranges of loneliness with an accuracy of 98%, while factoring in user personality types*”, which shows how including extra features can provide a significant boost to classification accuracy. Pulekar and Agu [28] used three ranges to classify the loneliness scores into categories, and got an increased classifier accuracy of around 8%, for a total accuracy of 98%, by including the subjects’ personalities.

What is common between these related studies is that they all found some interesting link between the data collected from smartphones and their respective psychological effect(s), which goes to show how this type of study might be able to predict certain psychological effects, and help people before they realize themselves that they have a problem, or before it becomes serious.

Table 2.1 also provides a comparison with our study, as discussed and elaborated on in this thesis.

Year	Study	Goal	Methods	Participants	Analysis	Results
2014	Wang et al. [34]	Assess impact of student workload on Quality of Life	Smartphone sensing, ESM, health and behaviour surveys	48 students (30 undergraduates, 18 graduates) for 10 weeks	Correlation: Smartphone usage and psychological scales	Links: phone activity and loneliness score
2015	Sanchez et al. [31]	Predict loneliness from smartphone monitored activities	Loneliness and activity surveys	100 adults (ages 60-90)	Correlation: activity and loneliness; evaluate using smartphone app	Link: average time spent outside, total outings and loneliness score
2015	Ben-Zeev et al. [2]	Assess smartphone usage as behavioural marker for mental health	Smartphone sensing, experience sampling, mental health surveys	47 students (ages 19-30), 10 weeks	Mixed-effects linear modeling, penalized functional regression	Link: kinesthetic activity and loneliness score
2016	Pulekar and Agu [28]	Assess smartphone usage as behavioural marker for loneliness	Smartphone sensing, personality surveys, mental health surveys	9 students (ages 23-28), 4 weeks	Correlation analysis, machine learning classifiers	Subject's personality type as indicative variable
2016	Farhan et al. [11]	Assess feasibility of depression screening using smartphone	Smartphone sensing, mental health surveys, clinician assessment	79 college students (ages 18-25), 7 months	Machine learning (regression and SVM), clinical ground truth comparison	Neg. correlation: number of unique locations and depression scores
2017	**Honoré**	Assess smartphone usage as behavioural marker for loneliness	Smartphone sensing, ESM, DRM, loneliness surveys	3 students (ages 23-28), 4 weeks	Machine learning classification	Link: phone activity, time spent at school and loneliness

Table 2.1: Keyword comparison of related works

Chapter 3

Definitions

In this chapter we give an introduction to some important definitions used in this thesis.

3.1 Loneliness

A formal definition of loneliness can be found in Perlman and Peplau [26]:

> *"Loneliness is the unpleasant experience that occurs when a person's network of social relations is deficient in some important way, either quantitatively or qualitatively."*

To put this in a more informal way, we can describe loneliness as the subjective feeling a person has when their relationships either don't live up to the expected quality that a person desires, or if they do not have the number of social relations that they desire.

To further solidify the definition of loneliness it is important to note how it differs from *social isolation*, and how the two are related. Jong-Gierveld, Tilburg, and Dykstra [17] describe the feeling of social isolation as being related to the number of meaningful relationships a person has, as a socially isolated person will have a low amount of meaningful relationships. However, this does not mean that the individual is lonely, as Ang [1] shows that the number of meaningful relations needed to feel lonely differs from person to person, and across ages and genders. This subjective relation between loneliness and social isolation is used by Cacioppo, Fowler, and Christakis [3] to alternatively describe loneliness as *"perceived social isolation"*.

While there exist several subcategories of loneliness such as family loneliness, spousal loneliness, social loneliness and existential crisis loneliness as described in Sanchez et al. [31], we are focusing on the general feeling of loneliness, which is commonly measured using the UCLA Loneliness Scale (section 4.2)

3.2 Quality of Life (QoL)

An explicit definition of Quality of Life does not exist, but several proposals have been made, including one by the World Health Organization (WHO) as part of their development of a scale for measuring Quality of Life [12]. They clarify the concept of Quality of Life as:

> *"'individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns."*

Whenever the concept of Quality of Life is mentioned in this thesis, we will be referring to the definition above, as clarified by the WHOQOL group.

Chapter 4

Methods

In this chapter we introduce the methods used in this thesis, and motivate their usage.

4.1 Study Approach and User Recruitment

Research methods can be split into two categories; *qualitative* and *quantitative*.

Qualitative methods deal with unstructured data such as free-form text acquired through survey answers, interviews or observations of research subjects. While this type of data is hard to analyze due to its unstructured nature, it provides an in-depth understanding of individual research subjects, and of general human behavior.

Quantitative methods deal with structured data acquired through methods such as experiments or closed-question surveys. These empirically collected variables are usually numeric, which lends themselves to statistical analysis and inferences.

Creswell and Clark [6] describes and defines a method of research used since the 80's which combines both quantitative and qualitative data collection, called *mixed-methods research*. Their definition of mixed-methods research encompasses studies which are using both quantitative and qualitative data sources.

We employ the mixed-methods approach in our study, as quantitative data provides breadth, while qualitative data provides depth. Combining the two can give new perspectives by either complementing each other or contradicting each other, leading to different answers to research questions not possible by using only one method of data collection.

In their book, Creswell and Clark [6] describes six problems where using mixed-methods research is appropriate to provide better research results:

- A need exists because one data source may be insufficient
- A need exists to explain initial results
- A need exists to generalize exploratory findings
- A need exists to enhance a study with a second method
- A need exists to best employ a theoretical stance
- A need exists to understand a research objective through multiple research phases

We acknowledge the value of using mixed-methods research, not only to enhance our study but also because only using quantitative methods offers an insufficient explanation, which can be improved by combining it with qualitative methods. As such, our study consists of several qualitative [*entry- and exit survey, experience sampling method, day reconstruction method*] and one quantitative [*smartphone sensing*] method.

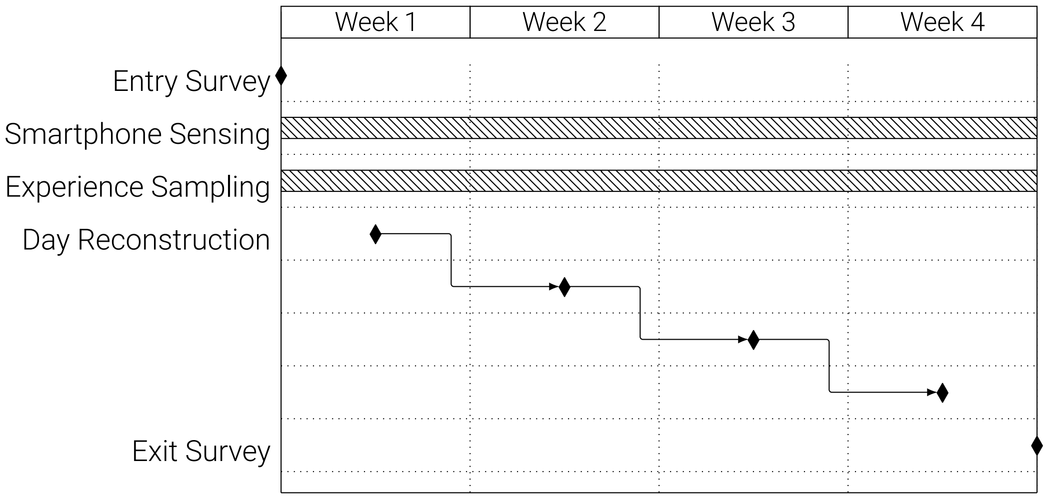


Figure 4.1: Timeline of methods used in our study

Our study has lasted for four consecutive weeks, which we assume to be representative of the subject's lifestyle and daily routines. Figure 4.1 illustrates the timeline of methods used in our study and shows both qualitative and quantitative methods. Quantitative methods include Smartphone Sensing (section 4.4) used to collect smartphone usage statistics. Qualitative methods include *entry- and exit surveys* used to assess socio-demographics, general lifestyle (e.g., if living alone) and measure loneliness based on the Danish UCLA scale (section 4.2.1), Experience Sampling Method (section 4.3) used to measure momentary loneliness using the Three-Item Loneliness Scale (section 4.2.2), and the Day Reconstruction Method (section 4.5) used to collect data about how people have spent their time in a period of the last 24 hours of their life. By combining the quantitative data collected, with the qualitative data, we can provide a more detailed explanation for the smartphone sensing data, which wouldn't have been possible otherwise.

After an initial survey round conducted in March 2017 with 102 responses, 11 people showed interest in participating in the study, where only five of those responded to our request to meet, and only three of those had a phone running Android, which is required for running our software to capture smartphone usage data, and for running the software for triggering the ESM surveys.

The processing of personal data in this project has been approved by the SCIENCE Faculty at the University of Copenhagen on March 28th. 2017.

4.2 Measuring Loneliness

In the field of psychology, the UCLA Loneliness Scale [30] is a commonly used tool to measure loneliness and is often quoted as being "the standard" way of measuring loneliness. The UCLA Loneliness Scale was originally published by Russell, Peplau, and Ferguson [29] in 1978 as a general measure of loneliness, consisting of a short 20-item survey with answers ranging from one to four. The scale enables an overall assessment of a person's subjective feeling of loneliness, calculated by summing the answers to all the questions, resulting in values of loneliness ranging from 20 (the lowest) to 80 (the highest).

The original scale consists of only negatively worded questions, for example *"there is no one I can turn to..."*, which could lead to response bias, as noted by Peplau and Cutrona [25], who then decided to develop a revised version of the scale called R-UCLA that included ten positively worded questions, for example *"there are people I can turn to..."* and ten negatively worded questions to prevent response bias, which was published in 1980 [25].

A third version of the scale was developed by Russell [30] in 1996, as after having worked with the revised UCLA scale among elderly people [7], they encountered problems wherein some elderly people did not understand questions worded in a specific way. Especially questions with double negatives, like having to answer "never" to the question "I do not feel alone," were difficult for the elderly people to understand [30]. Their solution to this problem was to simplify and change the wording of the questions and the response scale. It was also adjusted to more easily be delivered using research methods such as telephone surveys, by prepending *"How often do you feel..."* to the questions. The third version of the scale was tested on a broad range of demographics by using data from previous studies of college students, nurses, teachers, and seniors. Results from their tests showed that the updated version of the scale had a high reliability [30].

4.2.1 Danish Loneliness Scale

The Danish version of the scale was developed in 2007 by translation of the original scale into Danish by Lasgaard [23], and by consulting with a person holding a Master of Arts for back-translation comparison, along with review by a professor of clinical psychology and four masters-level psychology students, resulting in several revisions to the translation.

The Danish translation consists of 20 questions, where 11 of those are positive worded, and nine are negative worded. This is a slight change from the UCLA Loneliness Scale which contains nine positive and 11 negative worded questions.

To test the validity of the Danish scale, the translation was administered to 379 8th grade students (ages 13-17) as part of a survey that included 22 randomly selected schools. The reliability of the test was found to be very high, comparable to that of the UCLA Loneliness Scale. The author does note that the age group used to test the scale is very narrow, but justifies this by referencing to the great success of using the UCLA Loneliness Scale in surveys with both young and adult people.

In our study, we apply the 20-item Danish loneliness scale in the entry- and exit survey to measure subjects loneliness before and after the study.

The Danish loneliness scale can be found in appendix A.1, and our translation of the T-ILS can be found as part of the ESM instrumentation in appendix B.1.

4.2.2 Three-Item Loneliness Scale (T-ILS)

To facilitate the measurement of loneliness in large-scale surveys, Hughes et al. [16] developed a shorter version of the UCLA Loneliness Scale. Called the Three-Item Loneliness Scale, this shorter version was developed to be used as part of large-scale telephone surveys where the 20 questions of the UCLA Loneliness Scale would not have been feasible.

By conducting an exploratory analysis of the UCLA Loneliness Scale, Hughes et al. [16] found statistical evidence to support a three-factor version of the scale and selected three items from the dominant factor to represent their new loneliness scale. The items were rephrased in second person, for example *"how often do you feel left out?..."*, as the questions are read to the subject and not by the subject themselves. The number of response options was also reduced, as initial tests showed that people had difficulties remembering all the possible response options.

The reliability of the Three-Item Loneliness Scale was tested by including it as part of a module on the American Health and Retirement Study of 2002. Analysis showed that the reliability of the scale was somewhat lower than usually reported for the UCLA Loneliness Scale, but internal consistency showed that the Three-Item Loneliness Scale reliably measured loneliness in large-scale telephone surveys [16].

In our study, we are using this shorter version of the UCLA Loneliness Scale to ensure that we can measure the subject's loneliness multiple times a day without it being too much of a burden for the subjects. By using the Danish translation of the UCLA Loneliness Scale [23] (section 4.2.1), we constructed a Danish version of the Three-Item Loneliness Scale by comparing the T-ILS questions taken from the UCLA scale, and taking the translations from the Danish translation of the UCLA scale as suggested in [23]. We changed the wording of the questions to be in first person *"I feel.."* as subjects are reading the questions themselves, and we also changed the questions to present tense *"...right now"* as we measure it several times a day:

- How often do you feel that you lack companionship?
  // Føler du, at du savner nogen at være sammen med lige nu?
- How often do you feel left out?
  // Føler du dig udenfor lige nu?
- How often do you feel isolated from others?
  // Føler du dig isoleret fra andre lige nu?

The Three-Item Loneliness Scale can be found in appendix A.2, and the Danish translation can be found as part of the ESM instrumentation in appendix B.1

4.2.3 Scale Conversion

When converting values between the two loneliness scales (sections 4.2.1 and 4.2.2) we use equation 4.1 to make comparisons more easy by having numbers on the same scale.

$$\text{NewValue} = \frac{(\text{OldValue} - \text{OldMin}) \cdot (\text{NewMax} - \text{NewMin})}{\text{OldMax} - \text{OldMin}} + \text{NewMin} \tag{4.1}$$

4.3 Experience Sampling Method (ESM)

The above Three-Item Loneliness assessment has been implemented via ESM, which was developed by Larson and Csikszentmihalyi [22], and described by them as:

> *"a research procedure for studying what people do, feel, and think during their daily lives, It consists in asking individuals to provide systematic self-reports at random occasions during the waking hours of a normal week."*

The strength of the ESM comes from it's ability to probe people for feedback during the actual experiences in their daily life, and therefore getting a better momentary assessment of the experiences people are having. By combining these self-reports, a diary of peoples everyday life and how they experience it can be created, giving great insight into peoples daily routines and their associated experiences.

In the original article for the ESM, electronic pagers were used to signal subjects that they were to fill out a report with a questionnaire about their objective situation and their subjective state at that moment. This is of course only a guideline by the authors, and they also note that the ESM provides great flexibility, as any question about a subjects current experience or situation can be added as part of the reporting questionnaire, and any device can be used.

In our study, we want to assess subjects' loneliness at various points throughout the day, and we find that this method is a viable solution. The loneliness level is assessed by giving people a questionnaire containing the Danish translation of the T-ILS (section 4.2.2), with the three questions delivered in a random order to prevent people from memorizing their response as a pattern.

ESM Implementation To facilitate the instrumentation of the ESM on the smartphones, we have employed a system called PACO[1]. PACO is described as a mobile platform for behavioral science and is used to run experiments on mobile platforms (Android, iOS) which can track subject behavior through custom surveys distributed using ESM.

[1] https://www.pacoapp.com/

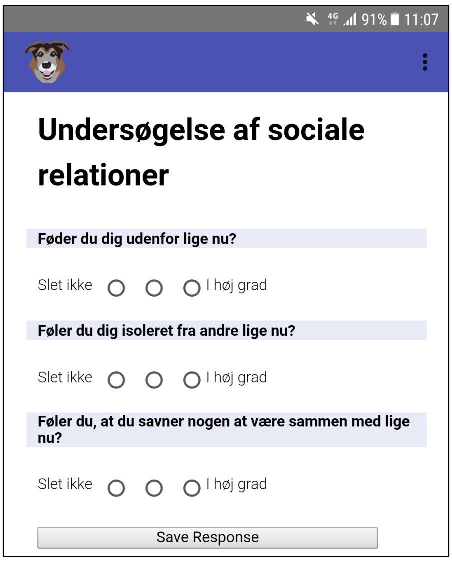


Figure 4.2: Screenshot of the ESM instrumentation running on a smartphone

By setting up an experiment on the PACO platform, we can create *triggers* by schedule or by phone events, and link them to *inputs* containing various survey style inputs (e.g. Likert scale, photo selection, free form text). In our case, we have set up a random sample (ESM) trigger which runs four times per day, with a minimum of 90 minutes in-between, every day from 09:00 to 21:00. This trigger triggers a notification to participate in the study, and by clicking the notification, the subject will be redirected to a survey inside the PACO application. This survey consists of three inputs, one for each item in the Danish translation of the T-ILS, shown in a random order to prevent response pattern memorization.

The setup for our experiment in PACO can be found in appendix B.1.

4.4 Smartphone Sensing

Since the advent of the smartphone, an ever increasing number of people are now carrying around small computers in their pockets during a large part of their daily life. In a report from 2013 by Smith [33], they estimate that 56% of American adults own a smartphone of some kind. This continuously increasing ubiquity of smartphones which contains an array of sensors such as accelerometer, microphone, GPS, WiFi, compass, and gyroscope has lead to the development of a category of methods called Smartphone Sensing.

When talking about Smartphone Sensing, we talk about methods leveraging the wide array of sensors in peoples smartphones and the close connection between peoples daily routines and their smartphones. These methods enable researchers to collect extensive data on peoples activities and their context without disrupting peoples lives.

mQoL-logger Developed by the mQoL Living Lab, mQoL-logger is a Smartphone Sensing tool that enables the collection of smartphone usage and sensor data to be used in various studies. The logger has been designed to easily be deployed as part of new studies, and transparently collects data about various aspects of phone usage, saved into .csv files on the mQoL server. Table 4.1 lists all data files and what kind of data they contain.

Filename	Collected Data
ActivityServService.csv	Information about currently running services
ApplicationsUsed.csv	Registers when a user starts "using" an app
CellIdsService.csv	Information about the cells the phone connectes to
PingService.csv	Measures the RTT (round trip time) to mQoL servers
TouchesBuffered.csv	Measures screen touches during a usage session
UserActivity.csv	Current user activity detected by the phone
UserPresenceEvents.csv	Screen and phone orientation events
UserPresenceLight.csv	Luminance measured by the phone's light sensor

Table 4.1: mQoL-logger Data Files

Data is collected by the logger either periodically (every minute), or on specific events. All collected data is saved on the mQoL Living Lab servers identified by the phones' IMEI.

A presentation of the mQoL platform has been made by De Masi et al. [8], which includes a discussion of the usages of the mQoL platform, and its role in research as part of machine learning and big data analytics.

4.5 Day Reconstruction Method (DRM)

The Day Reconstruction Method was proposed by Kahneman et al. [18] as a method for assessing cumulative daily experiences for the last 24 hours. The authors describe it as *"a method of measuring our daily affective experiences — our emotions at various moments throughout our day as we go about daily life"*. This means that by using the Day Reconstruction Method, we can assess how people spend their time, and how various experiences throughout the day are experienced, without interfering with the experience.

Before the Day Reconstruction Method, the standard way to measure daily experiences was the Experience Sampling Method [18]. This way of measuring experiences is very labor intensive and can interfere with the experience. Kahneman et al. [18] lists the following advantages of using the DRM method over the ESM method:

- Easier for the subjects
- Gives a complete picture of the day as opposed to random parts
- Shows how people use their time — how much time is spent on which activities

When employing the Day Reconstruction Method, people were asked to:

> "Think of your day as a continuous series of scenes or episodes in a film. Give each episode a brief name that will help you remember it (for example, 'commuting to work', or 'at lunch with B'...). Write down the approximate times at which each episode began and ended. The episodes people identify usually last between 15 minutes and 2 hours." - Kahneman et al. [18].

By asking people to recall their experiences throughout the day, we help them construct a short diary of the previous day. Moreover, by evoking the context of the actual experiences from the last 24 hours, we ensure a good recall of the actual experiences, and prevent memory bias by recalling their experiences shortly after they have been experienced.

As this method is used to collect data about daily experiences, it is very flexible and can be used in a wide variety of studies. The high flexibility in adapting the method to different studies makes it a good solution to provide qualitative data about a subject's smartphone usage and social interactions throughout the last day, which can refine our understanding of the data we collect using Smartphone Sensing and add context to the data we gather using the Experience Sampling Method.

The Day Reconstruction Method instrumentation can be found in appendix B.2.

4.6 Location Assessment Method

When working with cell tower location data, we have a desire to categorize the cell towers into commonly visited locations from a subject's everyday life, based on the geographical location of the towers. To do this, we employ a method for density-based spatial clustering.

The process of clustering a set of data can be defined as a method wherein a set of data points are grouped into clusters based on how similar the data points are. Density based clustering is a subset of data clustering wherein data points are grouped into clusters based on the local data point density. Areas with a high spatial density are grouped, whereas areas with a low spatial density are categorized as noise or cluster edges. This makes density based clustering a good fit for our problem of recognizing the subject's semantic locations, as often visited locations will be represented as spatially dense in the cell tower location data. One of the most popular density based clustering methods is DBSCAN.

DBSCAN Developed by Ester et al. [9] in 1996, DBSCAN (density-based spatial clustering of applications with noise) is a density based data clustering algorithm, used to identify clusters and noise in spatial data. The algorithm requires two parameters, *minPts* and ϵ. *minPts* defines the minimum amount of points required for a dense area to be classified as a cluster, and ϵ defines the maximum distance between points for them to be considered as part of the same cluster. By continuously selecting a random unvisited point in the data and calculating its ϵ-neighborhood, the point is either classified as part of a new cluster if the amount of points in the ϵ-neighborhood is larger than *minPts*, or classified as noise. If a point is classified as a new cluster, all the points in its ϵ-neighborhood are also added to that cluster, along with their own ϵ-neighborhoods. This process continues for each discovered cluster and stops when there are no unvisited points left.

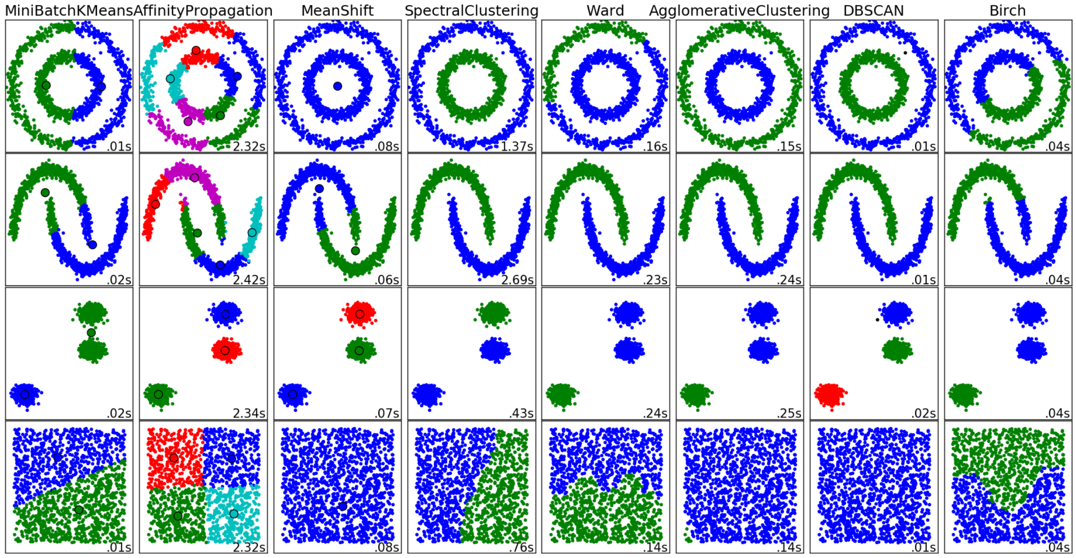


Figure 4.3: A comparison of the clustering algorithms in scikit-learn
source: `http://scikit-learn.org/stable/modules/clustering.html`

Figure 4.3 demonstrates eight commonly used clustering algorithms on four different datasets, which shows the strength in using DBSCAN for spatial clustering. Looking at the four datasets, it can be seen that DBSCAN is the only algorithm to correctly cluster the first three datasets, whereas the fourth dataset is harder, as there are no clear clusters.

4.7 Machine Learning Classification

When trying to estimate the variables/factors influencing the relationship between target variable, i.e., labels and features in a dataset, a common approach is to try and model the relationship using data modeling. In our approach, we leverage machine learning classifiers to create a model estimating this relationship, allowing us to analyze the model later and hypothesize about the factors and their importance. Classification through machine learning is the process of labeling new data based on the labels of existing data. By creating a model representing the relationships between data and labels, we can estimate the labels of new data.

Different types of machine learning methods exist. Unsupervised learning works on unlabelled data and is a common approach when trying to find structure or patterns. The clustering method described in section 4.6 is an example of unsupervised learning, where we try to find structure in the cell tower data without knowing what the output is supposed to be. Supervised learning is used for classification, and works on data where the labels for the target variables are known, and the classification method tries to model the relationship between the data/features and the labels.

As part of our analysis, we employ a set of machine learning classifiers from different categories. By picking a broad range of different classifiers, we increase our chances of finding a model that accurately, yet in a generalizable way, models our data, as some models might work better on our problem.

Logistic Regression Regression is the process of fitting a function to estimate the values of a dataset. In LR we try to fit a logistic function $f(x) = \frac{1}{1+e^{-k(x-x_0)}}$ to a dataset, to partition it into two labels, also known as binary classification. In our case, we work with multiple labels as output, which can be modeled by logistic regression using one-vs-rest, or multinomial logistic regression. One-vs-rest fits a binary logistic function for each label, whereas multinomial logistic regression uses the softmax function $\sigma(z) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} forj = 1, ..., K$ to translate the binary logistic function to multiary labels. This type of classifier works best on problems with linearly separable data.

Linear Discriminant Analysis LDA is a method wherein a set of linear decision boundaries are calculated as to separate the data into groups, where the inter-group variance is maximized, and the intra-group variance is minimized. This results in a projection of the dataset where data points are separated as much as possible using linear decision boundaries, which can then be used to predict the labels of future data as well. This type of classifier works best on problems with linearly separable data.

k-Nearest Neighbor KNN is an instance-based algorithm, which means that it keeps the training data internally and uses it to classify new data points. KNN is a simple method wherein a majority vote among the k-nearest neighbors of a data point is used to estimate the label of a new data point. This type of classifier works best on problems with low dimensionality.

Classification Tree CT, also known as decision trees, is an algorithm that uses a decision tree to model the data, where branches of the tree represent decisions based on the values of features, and leafs represent the label that is predicted by following the decisions down a set of branches to a specific leaf. To create the decision tree, a splitting strategy controlling how data should be split at each node is defined, and a criterion function to measure how good the split was, is required. Decision trees make it easy to reason about the features in a dataset, as the tree can be visualized and analyzed. This type of classifier tends to overfit and doesn't work well on datasets with unbalanced classes.

Random Forest and Extremely Randomized Trees RF and ET are ensemble methods, meaning that they create a model comprised of multiple less accurate methods, whose combined predictions is better than their individual predictions. The random forest algorithm works by constructing a "forest" of classification trees on random feature subsets and using their mean prediction to build the final prediction. The extremely randomized trees work like the random forest algorithm, but with the splitting strategy for each tree defined as the best among a set of randomly selected thresholds.

Gaussian Naive Bayes NB (GNB) is based on the Naive Bayes algorithm, wherein the features of the dataset are assumed to be conditionally independent. The probability of the labels of new data points can, therefore, be calculated using Bayes' theorem $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ by using the probabilities of the existing dataset. Gaussian Naive Bayes uses

Gaussian distributions to estimate probabilities for Bayes' theorem, by calculating the distributions using the mean and standard deviation of each feature for each class and therefore assumes the features of the dataset to be a Gaussian distribution. This type of classifier works best on problems where the dependence among features cancels each other out, and therefore support the conditional independence of Bayes' theorem.

Support Vector Machines SVM is a method wherein a hyperplane is calculated such that the classes of the dataset are separated as much as possible. The support vectors are the data points in the dataset, and the SVM method then finds the hyperplane that separates the data points into different classes, while maximizing the margin, defined as the distance from the hyperplane to the nearest data point. SVM can employ different kernel functions to transform the data points to a higher dimensional space, which might allow for a separation of data points into classes not possible at lower dimensions. This type of classifier does not work well on datasets with unbalanced classes.

Multi-layer Perceptron MLP is a method wherein a set of input, output and hidden layers containing perceptron units are connected. Each perceptron calculates an output based on an activation function on its inputs using some weight. The weights of the network are optimized iteratively through the logarithmic loss function, using a solver such as stochastic gradient descent or limited-memory BFGS, etc. This type of classifier works well on data with many samples.

Each of the machine learning classifiers mentioned has a set of hyperparameters, which we try to tune in section 5.5.4 to find the best model for our dataset.

Chapter 5

Results

In this chapter we give an overview of the collected data, followed by a description of the location assessment process, the data merging process, and an analysis of the data.

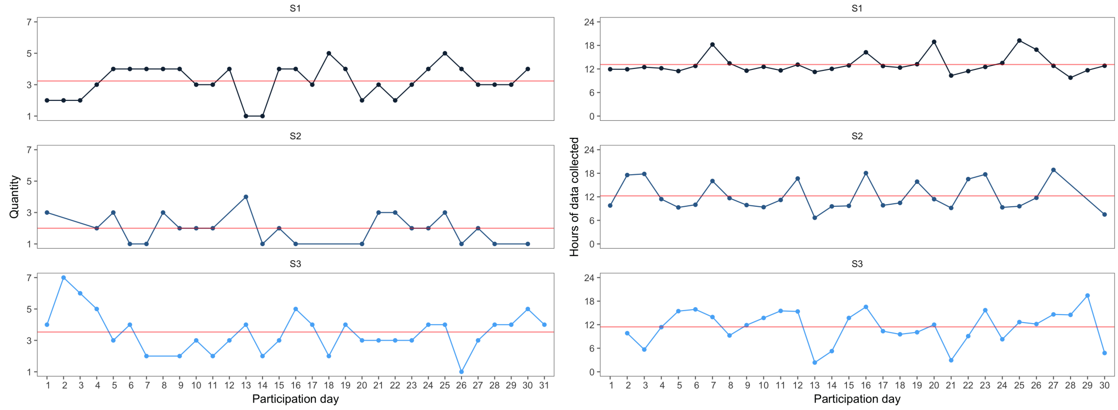
5.1 Collected Data Summary

The study was executed over a period of four weeks, with the first participant joining on April 2nd, 2017, and the last participant exiting on May 27th, 2017. Each participant was part of the study for 30 consecutive days and collected data during this period.

Subject	Gender	Age	Entry Loneliness	Mean ESM Loneliness	Exit Loneliness
S1	Male	27	27	20.31 $[n = 97, std = 1.74]$	26
S2	Male	23	49	34.35 $[n = 46, std = 15.73]$	40
S3	Female	28	35	20.38 $[n = 106, std = 1.91]$	36

Table 5.1: Subject Loneliness

Table 5.1 shows the gender, age, and loneliness assessed during the entry and exit survey, using the UCLA loneliness scale (min: 20; max: 80). It also includes the mean loneliness and standard deviation of all ESM entries, converted from the T-ILS scale (min: 3; max: 9) to the UCLA loneliness scale (section 4.2.3). As can be seen by the measured loneliness before and after the study, the UCLA loneliness is close to consistent in its measurement over four weeks, whereas the consistency between T-ILS and the UCLA loneliness scale is somehow further apart - most likely due to the less granular measurement of the T-ILS - although still close enough to be related.



(a) ESM Data Quantity

(b) mQoL Data Quantity

Figure 5.1: Quantity of data collected (mean = red line)

ESM The goal was to get four answers per subject per day, giving a maximum of 120 answers per subject for perfect data coverage. S1 and S3 had the highest data coverage at 80.83% and 88.33%, whereas S2 had a data coverage of 38.33%. Figure 5.1a shows the counts of how many responses were received per day, illustrating the differences in data quantity between S1 and S3 which both have a mean response rate of over three per day and S2 with a mean response rate around two per day. Table 5.2 shows the coverage of the data collected from all three participants.

It should be noted that S2 had technical difficulties with his phone force closing the ESM notification service, leading to the gaps in the collected ESM data (around days 16-20).

Subject	Expected Responses	Collected Responses	Coverage
S1	120	97	80.83%
S2	120	46	38.33%
S3	120	106	88.33%

Table 5.2: ESM Data Coverage

mQoL-logger In an ideal setting, each participant should have collected 24 hours of smartphone usage data, for 30 days. In reality, though, the amount of hours of data collected per 24 hours for all three subjects, is around 12 hours, which puts the total collected data coverage for all subjects at around 50%. Figure 5.1b shows the quantities of how many hours of data were collected per subject through all 30 days and illustrates the fact that the mean amount of hours of collected data for all three subjects is around 12 hours. Table 5.3 shows the coverage of the data collected from all three participants.

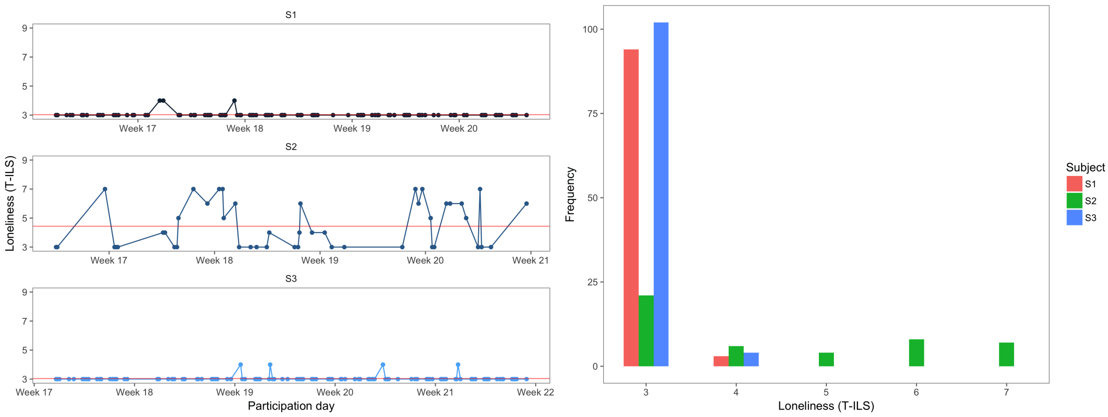
Subject	Expected Hours	Collected Hours	Coverage
1	720	393.95	54.72%
2	720	342.67	47.59%
3	720	331.90	46.10%

Table 5.3: mQoL-logger Data Coverage

5.2 Data Selection

By definition, a prediction is a way of estimating future behavior. However, if the behavior is constant (always the same), the prediction is obvious, meaning that the values that we are trying to predict need to have a certain amount of variation. Figure 5.2a shows all the measured loneliness values (min: 3; max: 9) for each subject in the study and illustrates the fact that there is close to no variation in the measured loneliness of S1 and S3.

This has lead us to only continue the analysis with the data from S2.



(a) Assessed Loneliness (mean = red line)

(b) Measured Loneliness Distribution

Figure 5.2: ESM Loneliness

S2 has a total of 46 assessed loneliness labels with a distribution as seen in fig. 5.2b.

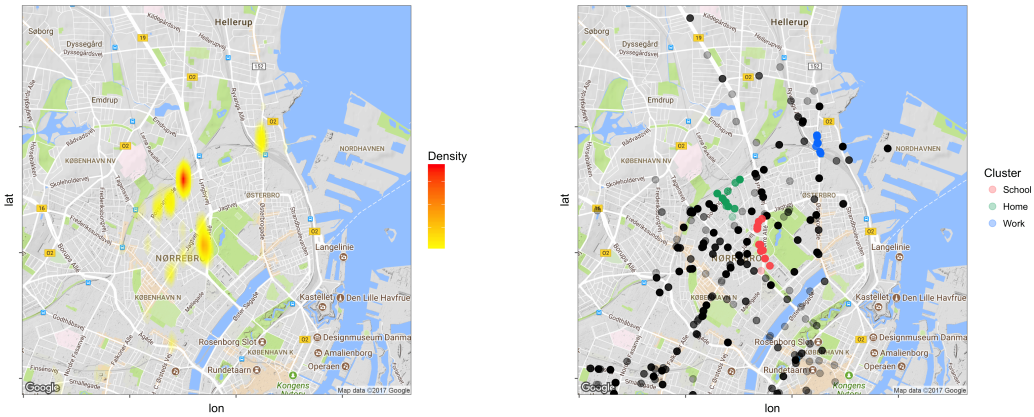
5.3 Location Assessment

To be able to identify the semantic location of S2 over the course of the study as precisely as possible, we analyzed the recorded cell tower IDs by correlating them with a cell tower location database, and applied a density-based spatial clustering technique to organize the cell towers into clusters of commonly visited locations.

The top visited locations were identified as *Home*, *School*, and *Work*. The *Home* location was manually identified by it being the most visited location, as people usually spend most of their hours in a day at home, which was corroborated by the locations recorded using the DRM. The *School* and *Work* location were identified through conversation with the subject during the DRM, where they mentioned their place of study and work.

Several methods exist for clustering cell towers based on their IDs, with some relying on a calculation of the estimated distance between them, based on the timestamps and relative time differences between recorded cell-IDs [10]. We decided to use the OpenCelliD[1] database for improved accuracy, and combined it with a spatial clustering algorithm.

OpenCelliD Maintained by Unwired Labs, OpenCelliD is a community built database of cell towers and their geographical coordinates, started in April 2008. It is promoted as *"The world's largest Open Database of Cell Towers"*. As the OpenCelliD database is community built, there is, of course, a certain amount of uncertainty in the data, which is mitigated by verifying the coordinates from multiple cell tower location samples. Samples are acquired by an application running on the contributors phones, which send in information on connected cell towers and the phones' geographical coordinates measured using GPS. The database covers 93.17% of the cell towers we have recorded for S2 in the 90-minute periods leading up to each recorded label, giving us a high location coverage.



(a) Cell-ID density

(b) Cell-ID clusters

Figure 5.3: Cell-ID clustering (S2)

As part of our location assessment analysis, we created a density plot to identify the locations where S2 were often connected to cell towers. Figure 5.3a shows a map of the city of Copenhagen, zoomed in on the area with a high density of connected cell towers. This density plot shows two main areas of cell tower activity, with one being the subject's home, and the other being S2's school. The third cluster we identified as the subject's work is not that noticeable on the map, which can be explained by looking at the DRM data, where the subject have not marked their location as being at work that often.

DBSCAN Using the density plot as a guideline, we applied the DBSCAN (section 4.6) algorithm to categorize the cell towers into clusters based on the three identified semantic locations. By manually tuning the variables of the algorithm, we ended up with a clustering that is a close match to the density plot, and which closely models the primary locations

[1] `https://opencellid.org/`

of S2 from the cell-IDs. Table 5.4 shows the distribution of cell towers into semantic locations based on not just data from periods labeled by a loneliness value, but all data collected from S2.

Cluster	Count	Percentage
Home	9826	52.83%
School	4695	25.24%
Work	844	4.54%
None	3233	17.38%

Table 5.4: Cluster Allocation (S2)

Figure 5.3b shows the resulting classification of S2's cell towers into clusters. The classifications mirrors the high-density areas identifiable in the density plot, and should therefore closely mirror the actual semantic locations where the subject spent their time. These cell tower location clusters will be used in our further analysis of S2's data, labeled with the semantic locations of the subject.

5.4 Data Merging

To be able to understand and analyze the data we have collected, we need to merge our quantitative and qualitative data sources together into a single dataset that can be used as input for the machine learning classifiers in our analysis. Figure 5.4 illustrates our process for merging together the data (highlighted in blue), wherein we assess the semantic locations of the subjects (labelled using DRM data), and then a set of feature vectors are extracted from the smartphone sensing data, and a set of accompanying labels for the specific time window are extracted from the experience sampling data.

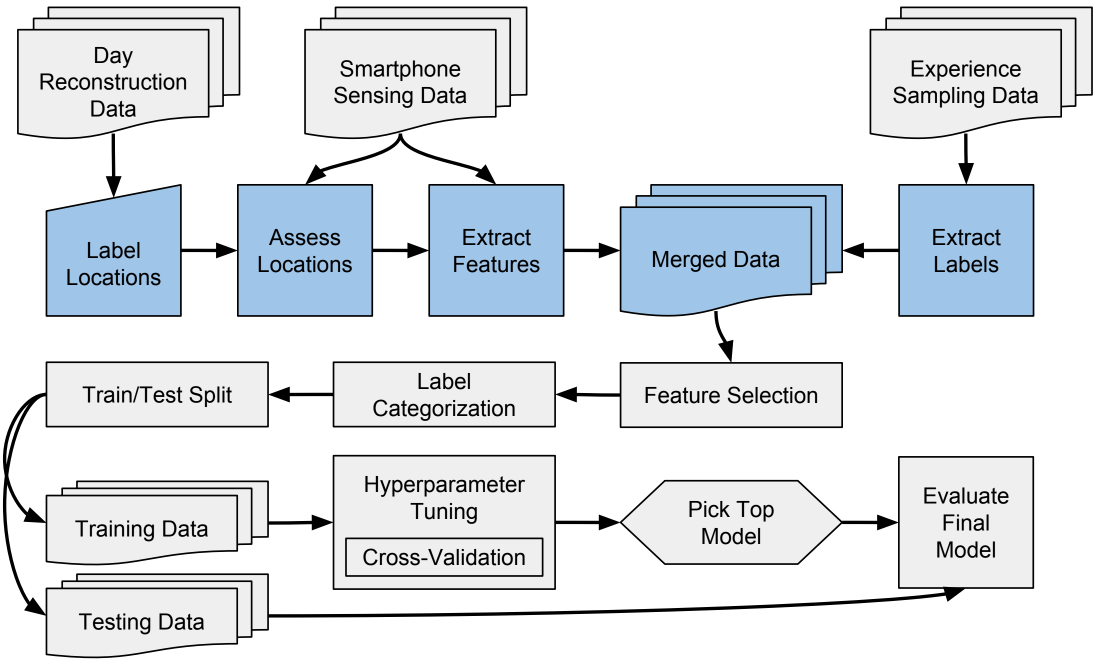


Figure 5.4: Data Merging Process (in blue)

Feature Extraction By summarizing the collected smartphone sensing data, we can deepen our understanding of the data in the form of extracted features. These features help us draw out meaning from the collected data, and can be useful when reducing multiple rows of data from a specific period, to a single row of data. This is particularly useful when merging two datasets of unequal length, like our smartphone sensing data, and our recorded loneliness labels (ESM). The data that we have collected using smartphone sensing is split into several files containing different information about how a subject have used their smartphone during the period of the study. From these files, as many features as we could come up with has been extracted:

- ApplicationsUsed.csv
  - Count of how often commonly used apps were used
- CellIdsService.csv
  - Percentage of time spent in semantic location
  - Count of total travelled distance (as distance between cell towers)

- PingService.csv
  - Percentage of time connected to cellular network
  - Percentage of time connected to WiFi network
  - Percentage of time connected to home WiFi network
- TouchesBuffered.csv
  - Count of phone screen touch sessions
- UserActivity.csv
  - Count of times registered doing specific physical activity
    [still, walking, running, on foot, on bicycle, in vehicle, tilting, unknown]
  - Percentage of time spent doing specific activity
- UserPresenceEvents.csv
  - Count of times phone was turned on
  - Count of times phone was turned off
  - Count of times phone was unlocked
  - Count of times phone was rotated

During our study, we have continuously collected Smartphone Sensing data, while the collection of Experience Sampling data have only been done up to four times per day. As the main goal of our analysis is to understand which factors derived from the Smartphone Sensing data affects the loneliness of our subjects, we have to extract the features in a time interval leading up to right before the time when the loneliness value was recorded. We define the size of this time interval as the *"window size"*.

Window Size We have defined our window size to be 90 minutes, as this ensures that the data used for feature extraction is both non-overlapping, momentary and of constant size. The non-overlapping property stems from the fact that we have configured the ESM with an interval of at minimum 90 minutes between each ESM pop-up. The momentary property stems from the fact that subjects are being asked via the ESM how they feel in the current moment, and we are only using the data collected in the 90 minutes leading up to each ESM pop-up.

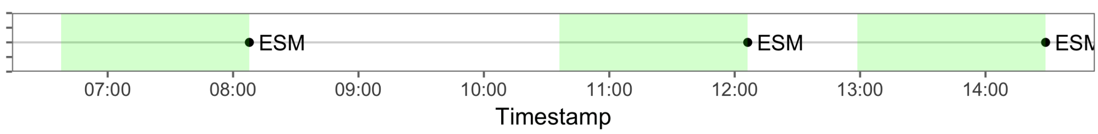


Figure 5.5: Illustration of window size (in green)

To merge the two data sources together, we go through each label in the experience sampling data. For each recorded label, we define a 90-minute window leading up to the time when the label was recorded. Figure 5.5 illustrates the collected labels (black dots) and the 90-minute windows (green rectangles) leading up to the registered label. A set of features is then extracted from the smartphone sensing data for these time periods and merged into a single data row consisting of the extracted features and the accompanying label from the ESM. Listing 5.1 outlines the data merging process in semi-pseudo-code.

Listing 5.1: Data Merging Pseudocode

```
windows = [(l.timestamp − 60 * 90, l.timestamp) for l in labels]
for i, (start, end) in enumerate(windows):
  row = [feature.extract(data, start, end) for feature in features]
  row += [labels[i]]
  data.append(row)
```

From the data merging process, 46 observations with 47 features were derived. Subject S2 had a total of 46 ESM loneliness measurements, resulting in a single feature row for each of the 46 measured loneliness labels, with the extracted features listed in appendix C. Moreover, with a total of 342.67 hours of smartphone sensing data collected from S2, and 46 loneliness measurements in periods of 1.5 hours (90 minutes), we end up excluding 79.86% of the smartphone sensing data collected from S2, for the analysis.

5.5 Analysis

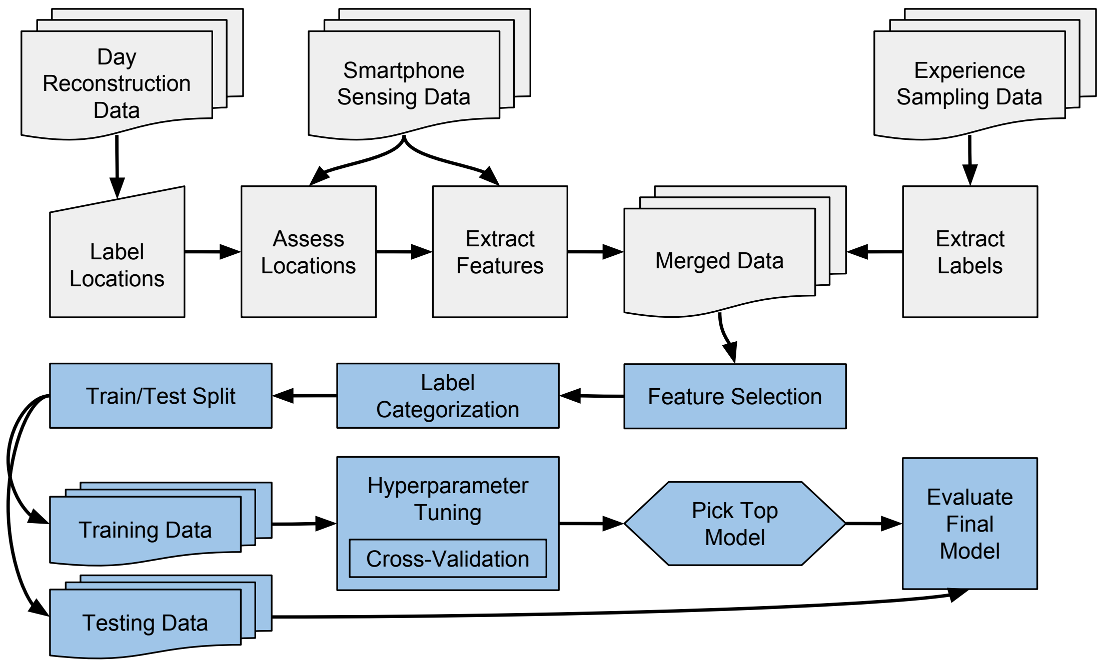


Figure 5.6: Classification Flowchart (in blue)

As for the analysis of our results, we are following the procedure outlined in fig. 5.6. A common approach to solving problems using machine learning[2] is to define the problem (section 1.1); prepare the data (sections 5.2 to 5.4); spot check algorithms; improve the results (section 5.5.4) and present the results (section 5.5.4). In our approach, we start by doing feature selection, skip the spot-checking of the algorithms, and instead proceed

[2] http://machinelearningmastery.com/process-for-working-through-machine-learning-problems/

directly to tuning the hyperparameters of a set of common machine learning classifiers, followed by reporting the performance of the top performing machine learning classifier.

5.5.1 Feature Selection

Several reasons exist as to explain why feature selection is an important part of solving machine learning problems. Guyon and Elisseeff [13] define the three most important reasons as:

- Improving the prediction performance of the predictors
- Providing faster and more cost-effective predictors
- Providing a better understanding of the underlying process that generated the data

By selecting a subset of the total features available, we also decrease the complexity of our final model, and at the same time we make it easier to interpret the results as we have fewer explaining features, helping us interpret which factors influence the outcome.

When doing feature selection, the problem of selecting the size of the subset arises. Although there is no precise guideline on picking this size, a paper by Hua et al. [15] on selecting the optimal number of features as a function of sample size N suggests a size between $N-1$ and $\sqrt{N}$ depending on the intercorrelation between the features, where in cases of highly correlated features one should pick a subset of $\sqrt{N}$ features. As we have a sample size of $N = 46$ for our merged data, we should ideally have a feature subset size between seven and 45, depending on the intercorrelation of the features.

In our feature selection process, we employ two methods of reducing the set of features, variance threshold and variable ranking, which results in a final feature subset size of 13.

Variance Threshold Is a feature selection method used to remove any features with a variance below a certain threshold. In our case, we set this to zero, such that any feature with a constant value for all samples will be removed as they will not have any impact on the model and only increase the computational power needed for the classifiers. Some features might be zero for all samples, as we only sample the 90 minutes leading up to each ESM pop-up, and the features might only be nonzero outside these time windows. The features removed by this method is as follows:

- applications_used.camera_count
- applications_used.contacts_count
- applications_used.gallery_count
- user_activity.running_count
- user_activity.running_percentage
- user_activity.walking_count
- user_activity.walking_percentage

Variable Ranking Is a feature selection method which takes a ranking measure, a count k, and keeps the top-k features ranked using the supplied ranking measure. In our case, we use the ANOVA F-test that calculates linear dependency between features and labels as the ranking measure, to only keep the features which have the highest linear dependence with the labels. We keep the top-$F/3$ features, where F is the total number of features in our dataset. The final feature set after using variable ranking is as follows:

Feature	Description	Range	S2
applications_used.calendar_count	Times the Calendar app was opened	$0 - \infty$	0-7
applications_used.email_count	Times the Email app was opened	$0 - \infty$	0-42
applications_used.messenger_count	Times the Messenger app was opened	$0 - \infty$	0-13
applications_used.mobilepay_count	Times the MobilePay app was opened	$0 - \infty$	0-7
applications_used.weshare_count	Times the WeShare app was opened	$0 - \infty$	0-7
cell_ids_service.cluster_school_percentage	Time spent at school location	0-100%	0-100%
cell_ids_service.cluster_unclassified_percentage	Time spent at unclassified location	0-100%	0-100%
ping_service.cellular_percentage	Time spent connected to cell towers	0-100%	0-100%
ping_service.wifi_home_percentage	Time spent connected to home WiFi	0-100%	0-100%
ping_service.wifi_percentage	Time spent connected to WiFi	0-100%	0-100%
user_activity.still_percentage	Time spent with phone being still	0-100%	0-100%
user_activity.tilting_count	Times phone has been tilting (moving)	0-∞	0-26
user_activity.tilting_percentage	Time spent with phone tiling (moving)	0-100%	0-50%

The linear correlation of all features and the ones we have selected (marked in blue) can be seen in fig. 5.7. The figure illustrates that most of the selected features are those with a high linear correlation with the assessed loneliness.

5.5.2 Feature Scaling

Several of the machine learning classifiers that we are going to use as part of our analysis will not function accurately on our dataset unless we apply a normalization to the features. This is because several of the classifiers are using the distance between features as part of their optimization process. An example is the κ-nearest neighbors algorithm, in which the classification of a point is the class that is most common among its κ-nearest neighbors, found using euclidean distance. Hence, the features should be normalized to the same range, so they are within the same proportions of each other, and no single feature dominates the others, thus potentially confusing the classification methods.

We standardized our dataset, scaling all 13 features to zero mean and unit variance.

5.5.3 Label Categorization

When working with classification, the complexity of the labels that we are trying to predict influences the complexity of the model needed for the prediction. The labels in our dataset are currently on an ordinal scale, with values ranging from three to nine (section 4.2). We have chosen to convert the ordinal labels to categorical labels, split into ranges of *low* (3-4), *medium* (5-6) and *high* (7-9) loneliness. This reduces the complexity of the classification task at hand and should make our final model more precise (due to a simpler task).

To illustrate the difference in classifier accuracy gained by converting the ordinal labels to categorical labels, a set of standard machine learning classifiers (section 4.7) were compared against each other. Each classifier was trained ten times using 3-fold cross-validation, and a boxplot of the accuracies can be seen in figs. 5.8a and 5.8b.

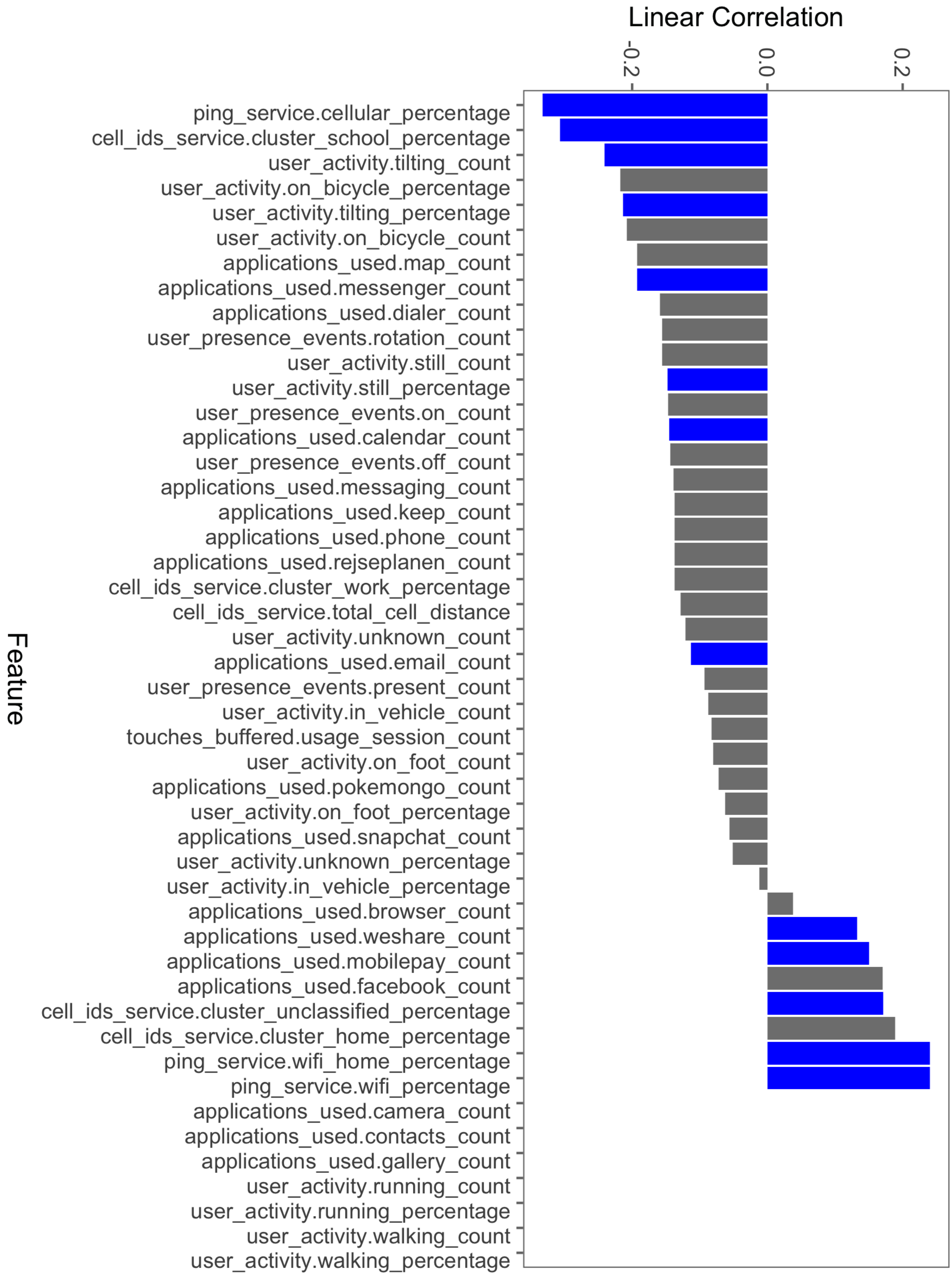
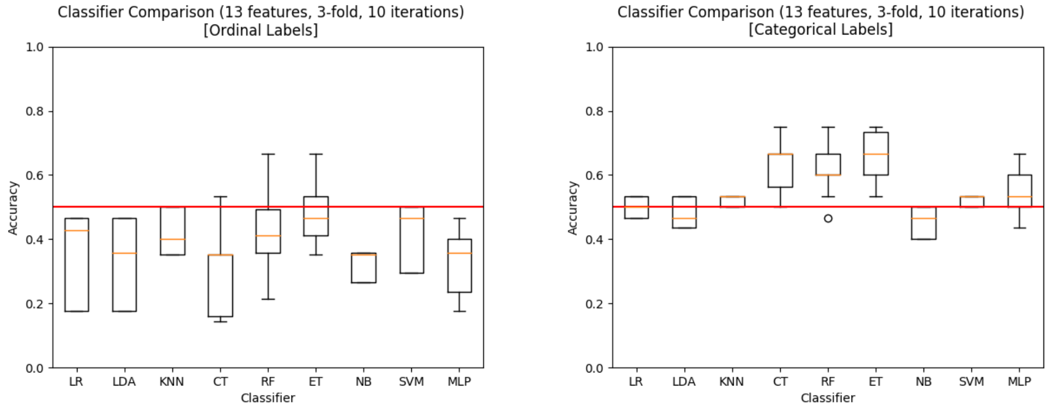


Figure 5.7: Linear correlation between features and target value



(a) Ordinal Labels (3-9)

(b) Categorical Labels (low, medium, high)

Figure 5.8: Classification accuracy for different labels

From figs. 5.8a and 5.8b it can be seen how there is a significant gain in general classification accuracy, which testaments our choice to simplify the labels. Figures 5.10a and 5.10b shows the distribution of labels before and after the categorization of the labels, with the guessing accuracy (50%) marked as a red line.

5.5.4 Classification

To analyze and model the relationship between the features representing the smartphone usage and the target variable, we apply a broad range of machine learning classification algorithms (section 4.7) to our data. These classifiers are trained, tuned and tested, leading us towards the model that closest fits the relationships in our dataset, measured by its performance.

Training and Testing

When modeling data using machine learning algorithms, there is always the risk of either underfitting or overfitting the data. Figure 5.9 illustrates underfitting and overfitting of data with a model. We will always aim for a model representing the middle graph, which is "just right." This means that our model should not be too general, or too specific to the data that it has been trained on.

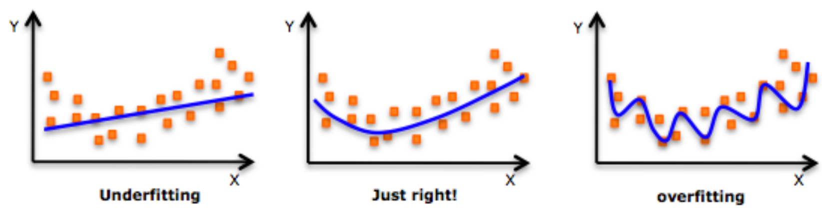


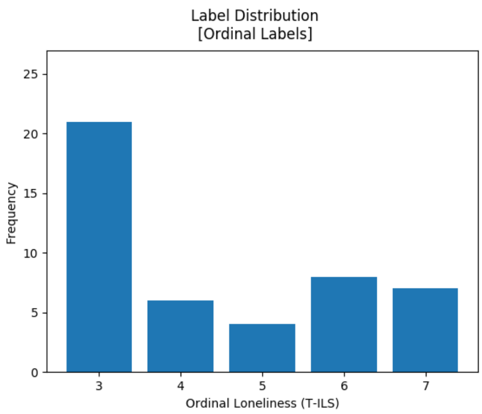
Figure 5.9: Overfitting vs. Underfitting
Source: `https://medium.com/towards-data-science`

To help reduce underfitting and overfitting, we are splitting up our dataset into two parts. One part to train our classifiers on, and one part to test the resulting model on, to ensure that our model is not under-fitted or overfitted to the problem at hand. For selecting the sizes of the training and testing subsets, we follow a common split of 80/20, where 80% of the data is kept for training, and 20% of the data is kept for testing. This follows the Pareto principle, which is a common split used in machine learning.

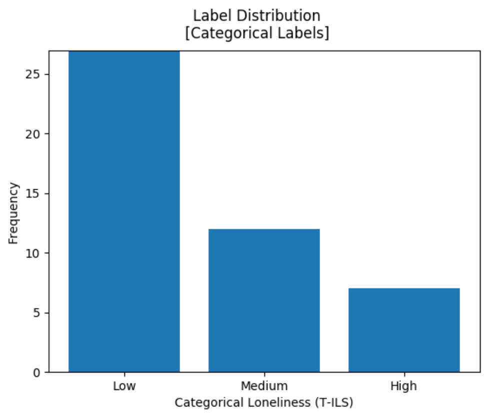
Cohen's Kappa

To ensure that we can reliably measure the performance of the different machine learning algorithms against each other, we have to take into account the fact that the classes we are trying to predict are unbalanced. As can be seen in figs. 5.10a and 5.10b, both the original (ordinal) and simplified (categorical) labels are unbalanced such that there are more than double the amount of "low" measures than of any of the other classes. This imbalance means that the random guess accuracy with the ordinal labels will be 46% when guessing majority class "3", and the random guess accuracy with the categorical labels will be 59% when guessing majority class "low."

Introduced in 1960 by Cohen [5], Cohen's Kappa is a performance measure that takes into account this class imbalance and can be used to measure the performance of classifiers in cases of unbalanced classes. The formula for Cohen's Kappa is defined as $\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$, where p_o is the observed accuracy between predicted and true labels, and p_e is the probability of each class being correctly classified by random choice. This means that a κ value of zero signifies that a classifier is as precise as a random guess, whereas a value of one means a perfect agreement between predicted and true labels.



(a) Ordinal Labels



(b) Categorical Labels

Figure 5.10: Distribution of labels (target variable)

To get a sense of the general performance of the nine machine learning classifiers that we are using for the analysis, an initial comparison was done to measure the κ value of each classifier without doing any hyperparameter tuning. Figure 5.11 shows a boxplot of the κ value of the classifiers, illustrating how the ensemble and tree classifiers are doing better than the other classifiers - that is, without any tuning.

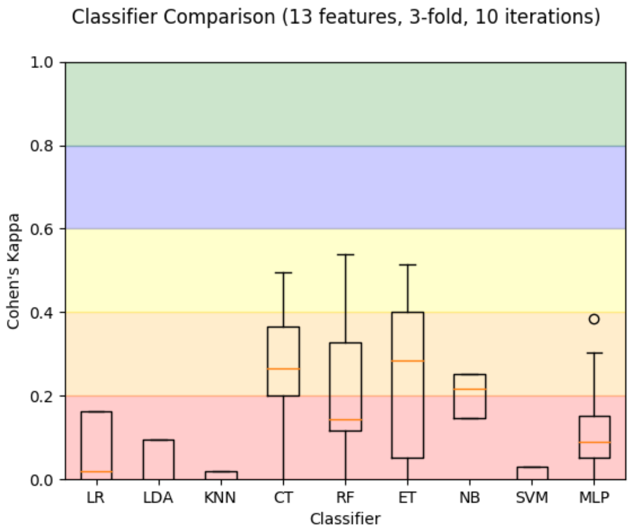


Figure 5.11: Classifier performance comparison

To interpret the reported κ values, we rely on the interpretation by Landis and Koch [21] who classify κ values in the ranges 0–0.2 as slight, 0.2–0.4 as fair, 0.4–0.6 as moderate, 0.6–0.8 as substantial, and 0.8–1 as almost perfect agreement between the predicted and the true labels.

Model Selection

After having defined how our dataset is split into training and testing subsets, and how we are going to measure the performance of the classifiers, we will now focus on tuning the machine learning classifiers and find the model that best fits our dataset. To tune the classifiers we apply cross-validated hyperparameter tuning, to find the best model.

Hyperparameter Tuning Each classifier that we are trying to optimize for our problem has a set of hyperparameters which can affect the performance of the classifier. By creating a grid of possible hyperparameter values for each hyperparameter in each classifier, we can iterate over any combination of these and find the combination of hyperparameters that gives the best performing model for each classifier in our set of classifiers. This is done through ten iterations of three-fold cross validation for each combination of hyperparameters, to ensure that the models are not overfitted to the training data. The hyperparameter grid with variables for each model can be seen in appendix D.

Figure 5.12 shows the performance distribution of all ten iterations of three-fold cross-validation of the nine classifiers and shows an improvement compared to the non-tuned models seen in fig. 5.11. The Random Forest classifier has the highest possible κ performing model, making it a good candidate for a final classification model.

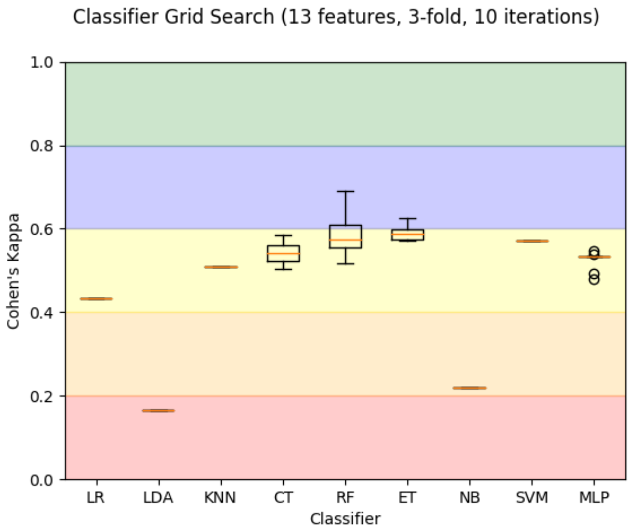


Figure 5.12: Tuned Classifier Comparison

Final Model To select the final model for our dataset, we looked at all models generated for each of the ten iterations of each classifier and picked the model with the highest performance. The performance measured is the median κ over the three folds of the cross validation done for each iteration of each classifier during the hyperparameter tuning in section 5.5.4. The final model is a Random Forest classifier with the hyperparameters configured as shown in table 5.5a. The performance of the final model on the training, testing and complete dataset, can be seen in table 5.5b

Hyperparameter	Value
Bootstrap	False
Estimators	10
Max Features	log2
Criterion	Entropy
Min Samples Split	2
Max Depth	4
Class Weight	Balanced

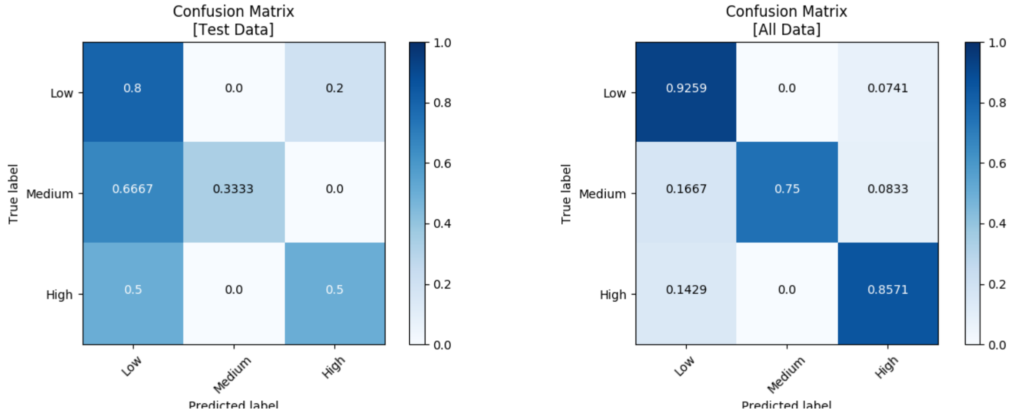
(a) Final Model Hyperparameters

Dataset	κ-performance
Training	0.69
Testing	0.60
Complete	0.87

(b) Final Model Performance

Table 5.5: Final Random Forest Model

To further evaluate the performance of our model, we employ confusion matrices to illustrate the relationship between the true labels and the labels predicted by the model. Figures 5.13a and 5.13b shows the classification accuracy by way of a confusion matrix, evaluated on both the testing data and on the complete dataset.



(a) Test Data (b) All Data

Figure 5.13: Confusion Matrices of Final Random Forest Model

Figure 5.13a shows how the model has a low precision when classifying anything but "low" on the testing data, where "medium" will be classified as "low" with a chance of 66.67%, and "high" will be classified as either "low" or "high" with a chance of 50%. When comparing this to fig. 5.13b which shows the classification precision on the complete dataset, we see a precision that is very acceptable, with a "low" precision of 92.59%, "medium" precision of 75%, and "high" precision of 85.71%. This could be caused by the hyperparameter tuning process having overfitted the model, as the difference in performance on the training data is much better than on the testing data.

Rank	Feature	Gini Importance
1	applications_used.email_count	0.23
2	user_activity.still_percentage	0.19
3	cell_ids_service.cluster_school_percentage	0.11
4	ping_service.cellular_percentage	0.09
5	ping_service.wifi_percentage	0.08
6	cell_ids_service.cluster_unclassified_percentage	0.07
7	applications_used.messenger_count	0.07
8	user_activity.tilting_percentage	0.06
9	user_activity.tilting_count	0.05
10	applications_used.calendar_count	0.03
11	ping_service.wifi_home_percentage	0.03
12	applications_used.weshare_count	0.00
13	applications_used.mobilepay_count	0.00

Table 5.6: Feature Ranking in Final Random Forest Model

In the last part of our analysis, we report the Gini importance of the features according to the model. Table 5.6 shows the ranks and Gini importance of all 13 features. Comparing this with the linear correlations reported in fig. 5.7, we can see that the model models some nonlinear relationships between features and assessed loneliness, as the ranks of the features in the model differ from the ranks according to the linear relationships.

Chapter 6

Discussion

In this chapter, we discuss the limitations of our study and the significance of our findings, followed by an evaluation of our problem statement.

6.1 Limitations of Study

We acknowledge the fact that there are certain limitations of our study that affects the generalization and implications of the results we have obtained.

Looking at the subject recruitment process, we focus on a student population right from the start. This narrowing of the subject pool means that we will not be able to generalize our results to a whole population, but only to the specific subject pool that we are focusing on. This is interesting in itself, but will not have as broad implications as results for a broader population would have. By further limiting the study to only include subjects who own an Android device (due to the mQoL-logger instrumentation limitation), we might eliminate certain socioeconomic variance from the subject pool, as iOS devices usually have a higher price than Android devices. iOS devices also have a higher market share in Denmark[1], meaning that we exclude a large part of the population from our subject pool by limiting our study to a specific platform.

Further limitations arise from the fact that we have only recruited three people for our study, whereas some of the related works recruited as much as 79 students [11]. This was likely caused by A relatively short recruitment period (2 weeks); A topic which might not be of interest to people; A small subject pool due to the limitation of the platform to Android-only. The low response rate could have been mitigated by providing the subjects with a reward for participating in the study, whereas the subject pool could have been increased by using a smartphone sensing framework such as AWARE[2], which works on both iOS and Android devices.

These limitations not only affect the number of participants in the study but also affect the amount of data that we can collect. Finally, as we only use the data from S2 in the analysis, any results obtained will not be generalizable among the subjects in the study.

[1] `http://gs.statcounter.com/os-market-share/mobile/denmark`
[2] `http://www.awareframework.com/`

However, we can still hypothesize based on our results, where any hypothesis can later be verified on a broader population, in a larger study.

6.2 Reliability of Methods

Considering the various quantitative and qualitative methods used in the study, some proved to be more valuable for our study than others.

ESM The experience sampling method was used to gather labels for our data and did so mostly successfully. As this method requires that the subject actively notice that they have a request on their phone for sampling their current experience, this requires the subject to be near the phone, and to be able to hear or see it, before we will be able to gather a response from them. This has led to some gaps in our data, as subjects either didn't see, hear or otherwise notice that they had an experience sampling request on their phone. Certain features are implemented in PACO to mitigate this, one being a snooze function where if a subject does not respond within a given period, the request will be sent again after a configurable time interval. From our data (fig. 5.1a), we can see that S1 and S3 both have a mean response rate of over three per day, whereas S2 has a mean response rate of around two per day. This illustrates that the subject type, and how they interact with their phone will also have a large effect on the reliability of the data collected using experience sampling.

Smartphone Sensing We picked the mQoL-logger framework for collecting smartphone sensing data, as it is a tried and tested method used by several studies in the mQoL Living Lab[3], and should, therefore, be reliable. However, this reliability is not reflected in the amount of data collected, as we only capture 50% of what is expected from each subject (fig. 5.1b). This could be explained by a phone running out of battery, a phone being turned off, or synchronization problems in the mQoL-logger - we have not been able to assess the cause of this, as battery information is not part of the logged data.

Although the mQoL-logger framework collects continuous data at a frequency of 60 seconds, giving a fine-grained view of the smartphone usage, it does not register as many facets of phone usage as other frameworks. The AWARE framework collects additional information such as data from the accelerometer, barometer, battery, and GPS. This data might have provided us with more features for the analysis, which could have shown how other aspects of phone usage might have influenced the assessed loneliness.

In hindsight, using the AWARE framework could have been a good choice, as it collects more data, and would enable us to include subjects running iOS on their phones.

DRM The day reconstruction method was arranged as meetings with each subject on a weekly basis, making the reliability of the method up to the flexibility of the subjects calendar, and ours. We did not have trouble scheduling the meetings, and each subject successfully participated in their weekly scheduled meeting. The qualitative data collected

[3] `http://www.qol.unige.ch/mQoL.html`

proved hard to quantify without any smartphone sensing features to link it to, except for the subjects locations which were used to estimate the names of oft-visited semantic locations throughout their day (section 5.3).

Having additional features from the smartphone sensing data in which the day reconstruction data could be used to enhance meaning, would have made the day reconstruction a much more useful resource in the study. For example, having access to call log data from the smartphone sensing, and information on whom people talked to from the day reconstruction method would enable us to gain valuable information on the social interactions of the subjects in the study. The value of this additional information is seen in the article by Sanchez et al. [31], who found links between different types of loneliness, and the amount of outgoing and incoming calls by the subject.

Mixed Methods In section 4.1 we listed several problems where Creswell and Clark [6] mentions how their mixed-methods strategy might provide better research results:

- A need exists because one data source may be insufficient
- A need exists to explain initial results
- A need exists to generalize exploratory findings
- A need exists to enhance a study with a second method
- A need exists to best employ a theoretical stance
- A need exists to understand a research objective through multiple research phases

After having employed their mixed-methods strategy and tested it in our study, we can see that especially one method benefitted from having the additional data of another method. When we did feature selection, having the semantic location available from the DRM enabled us to label the location clusters acquired from the clustering process (section 5.3). This goes to show how the smartphone sensing data in itself would have been insufficient, and that it can be enhanced with additional context through data acquired from a second method. It also helped us explain the initial results obtained, as we can hypothesize as to why S2 being at school is important for the assessed loneliness.

6.3 Reliability of Scales

In our study, we use two different scales for measuring the loneliness of subjects: The UCLA Loneliness Scale (section 4.2.1) and the Three-Item Loneliness Scale (section 4.2.2).

UCLA Loneliness Scale As mentioned in section 4.2, the UCLA Loneliness Scale is one of the most commonly used scales for measuring loneliness, with over 3000 citations (all three versions combined) on Google Scholar at the time of this writing. The Danish translation that we are using from Lasgaard [23], reports the reliability as *"highly comparable to the original version of the scale, indicate that the Danish version of UCLA is a reliable and valid measure of loneliness"*. When looking at the responses from our entry survey, fig. 6.1 shows a close relation between peoples' subjective quality of life and their assessed loneliness through the Danish UCLA Loneliness Scale, which indicates an excellent reliability of the loneliness measurements with the actual loneliness of the survey respondents.

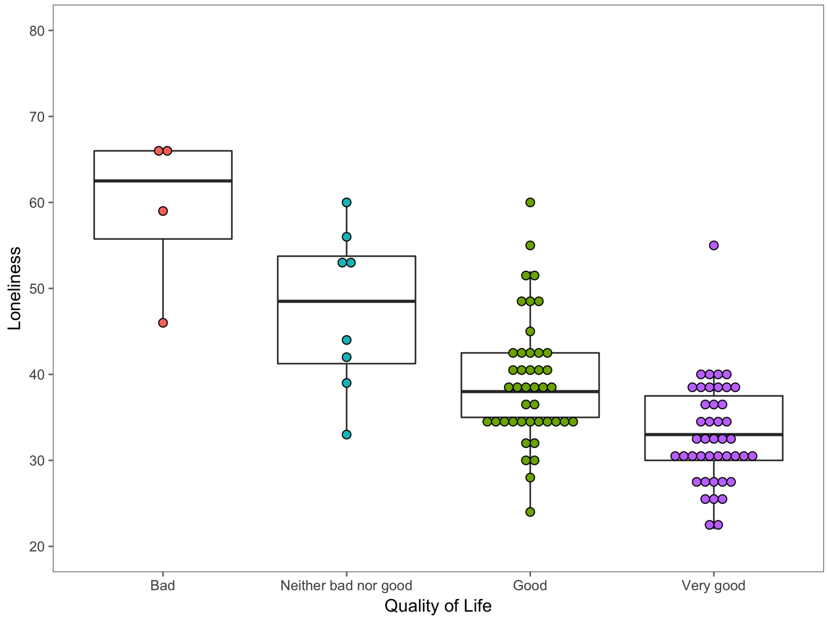


Figure 6.1: Relationship between quality of life and assessed loneliness

Three-Item Loneliness Scale As the T-ILS was developed as a response to the UCLA Loneliness Scale for large-scale surveys due to its simplicity, it was an optimal choice to be used as part of the ESM. Discussion with subjects during the DRM meetings made it clear that they did not feel annoyed by the popups, but that they definitely would have, had we used a scale with 20 questions. This was done by Pulekar and Agu [28] where subjects *"answer questions from the UCLA Loneliness Scale every 4 hours for the entire 2 weeks"*

Hughes et al. [16] says that the T-ILS *"appears to measure overall loneliness quite well"*, however, we found that the relation between UCLA measurements during the entry and exit survey does not relate as closely to the average T-ILS measurements. Table 5.1 shows the entry, exit and mean ESM loneliness, and illustrates how the measurements using the UCLA loneliness scale during the entry and exit survey remains stable across the four weeks, whereas the average T-ILS measurement is not close to the UCLA measurements.

We can see two reasons as to why this difference in measurements occurs. Firstly, Russell [30] notes about the first version of the UCLA loneliness scale that *"the fact that all items were worded in a negative or "lonely" direction created the possibility that loneliness scores would be affected by systematic biases in responding"*. This problem would also be present in the T-ILS, as all three questions are negatively worded. This could have skewed our assessed loneliness and might have had an influence on the low variance in ESM measurements. Secondly, as the T-ILS only contains three questions on a three-point Likert scale the amount of possible values is meager (3-9) versus the UCLA loneliness scale (20-80), making the scale less precise in its measurements.

We still believe that T-ILS is the right choice, as bothering people with 20 questions several times per day would not have been convenient for the subjects.

6.4 Reliability of Results

The reliability of the results obtained through our analysis is most likely to be affected by the low amount of data being used for the analysis. With only 46 samples and a reduced feature set of 13 features, we are very likely to have run into overfitting problems.

By reducing the labelspace for our problem, we greatly increase our chance of a successful model, as the relationships in the data we are trying to model will be simpler. We can see from figs. 5.8a and 5.8b how the performance even without tuning the models has already benefitted from this simplification.

Reducing the amount of features should, in theory, increase the performance of some of the classifiers in the tuning process, but it could also remove some features with a low linear dependency with the loneliness values, which could have had a substantial effect on the final model through some nonlinear relationship, or multi-feature relationship. Had time permitted, we would have liked to have tried a variation of feature subsets to see if any significant features were removed during our feature selection process, or if additional features might have resulted in a more accurate or more generalizable model.

When we trained our machine learning classifiers, we split the data into a training and a testing set. This was done to ensure that the trained model also performed well on the testing set and that it was not overfitted to the training data. If we look at the final model performance on the training and testing data in table 5.5b and the associated confusion matrices in figs. 5.13a and 5.13b, we can see that the performance on the testing data is not as good as on the training set, and the complete dataset. This is a sign that our model may be overfitted, which could have been mitigated if we had collected more data.

The final model, though, has a high κ-performance of 0.87 on the complete dataset, which Landis and Koch [21] would have interpreted as an almost perfect agreement between the predicted and the true labels. Moreover, the confusion matrix on the complete dataset underlines this, with a very high prediction accuracy of all three classes. This illustrates that our final model is a good fit for the data we have obtained.

6.5 Importance of Features

The feature importance reported by the final model as seen in table 5.6 is very interesting as it is both close to and far from what similar studies have shown.

The importance is measured across all ten classification trees of the final model, which can be seen in appendix E and the three most predictive features are:

- How many times the Email app was used
  $\rightarrow$ More usage leads to decrease in loneliness
- How high a percentage of the time the phone was still
  $\rightarrow$ Higher percentage leads to increase in loneliness
- How high a percentage of the time was spent at school
  $\rightarrow$ Higher percentage leads to decrease in loneliness

We can only theorize as to why the most important feature is how many times the Email app was used, but it will likely be related to the personality or behavior of the specific subject (S2), and therefore not be very generalizable. Having more subjects in the study, and more data, would have enabled us to analyze further why this feature is so important, or if it is also an important feature for other people and not just this specific subject.

The second most important feature though is very related to what is seen in other studies, where several of the related works also found links between physical activity - not being still - and loneliness [34, 31, 2]. This reaffirms their findings and underlines the importance of staying active as a way to reduce loneliness. We do not know if this link is due to the subject being active, or due to the subject being around people as a result of not being still.

The third most important feature is the time spent at school, which highlights the fact that being around other people is an effective way of reducing loneliness. Initial linear correlation of features with loneliness (fig. 5.7) also showed how being at home - being alone - had the opposite effect, thereby increasing loneliness. Again, this might be individual, as some people could, in theory, be lonely at school as well, if they were social outsiders.

Unlike reports from recent studies [20, 32] that showed increased use of social media leads to feelings of loneliness, and we have not been able to find any similar links in our analysis. This may be because of our analysis being conducted on a single subject, whom might not be affected by this, or might not represent the general pattern seen in the other studies.

What we see from the feature importance extracted from our model, is a mix of features. Features that could be explained as being personal to the specific subject in our study (S2), and features that could be more general, but of which we will not know for sure unless we have had more subjects and more data to generalize the results of our study.

Chapter 7

Conclusion

In this chapter, we will conclude on our results in relation to the problem statement, and talk about future work as a result of our work in this thesis.

7.1 Problem Evaluation

We described the importance of being able to assess loneliness, as a way to counteract the increasing spread of loneliness and its side effects in our society (section 1.1). Preventing loneliness will decrease the likelihood of mortality, and prevention would thereby indirectly increase the health of the population. Related works (chapter 2) showed varying success in predicting loneliness and associated psychological effects, through analysis, data modeling and machine learning classification.

The goal of this thesis was to be able to leverage the ubiquity of smartphones to assess the loneliness from phone usage patterns. We succeeded in gathering data from several subjects, and were able to analyze it and make assumptions about the relationships between usage patterns and assessed loneliness.

Based on our computational analysis and the resulting data model, we were able to assess the importance of the extracted features in relation to the assessed loneliness. The importance of the features as reported by the data model both confirm and refute links found in similar studies. We were not able to confirm any links between Facebook usage and loneliness, which was reported by several recent studies [27, 36, 20, 19]. This might very likely be due to our analysis only being conducted on one subject, which prevents us from generalizing our results. Although we cannot generalize from our results, we still found that increased physical activity, measured as the percentage of time the phone spent being still, was linked to a decrease in assessed loneliness. This confirms findings from related works [34, 31, 2], and helps support the arguments of these previous findings.

We also found links related to the assessed loneliness that was not identified in any related works, of which the cause may very well be our lack of generalization. For example, the number of times of which the subject opened their email app was reported as the most important feature in our final model, leading to a decrease in assessed loneliness. We can only speculate as to the meaning of this link, as it could be specific to the single subject that we included in our analysis. The second most important feature was the percentage of time spent at school, which leads to a decrease in assessed loneliness, which might very likely be caused by the social benefits of being around other people, decreasing the subjective feeling of loneliness.

The conclusions we can draw from our work is therefore that we can confirm the links between physical activity and assessed loneliness as seen in related works [34, 31, 2]. We can also conclude that for the specific subject in our analysis, leading a physically and socially active life (indirectly by use of email and by being at school) helps reduce their feelings of loneliness.

7.2 Future Work

The main goal of our future work would be to increase the generalization of the results, such that any resulting classification model could be instrumentalized. To increase the generalization of any future work, the study would have to include a much larger pool of subjects. By broadening the subject pool demographically by including not only students but people from different groups of the population, it would make it easier to find more subjects. Further increasing of the subject pool would be possible by using a smartphone sensing framework that supports iOS, as it is among the most used smartphone platforms. Being able to instrumentalize the resulting model and embedding it into smartphone applications would possibly allow for preventive measures reducing any impending feelings of loneliness.

References

[1] Chin-Siang Ang. "Types of Social Connectedness and Loneliness: the Joint Moderating Effects of Age and Gender". In: *Applied Research in Quality of Life* 11.4 (Sept. 2015), pp. 1173–1187. ISSN: 1871-2584. DOI: 10.1007/s11482-015-9428-5. URL: http://link.springer.com/10.1007/s11482-015-9428-5.

[2] Dror Ben-Zeev et al. "Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health HHS Public Access". In: *Psychiatr Rehabil J* 38.3 (2015), pp. 218–226. ISSN: 1559-3126. DOI: 10.1037/prj0000130.

[3] John T Cacioppo, James H Fowler, and Nicholas A Christakis. "Alone in the crowd: the structure and spread of loneliness in a large social network." In: *Journal of personality and social psychology* 97.6 (2009), p. 977. ISSN: 1939-1315.

[4] John T Cacioppo et al. "Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses." In: *Psychology and aging* 21.1 (2006), p. 140. ISSN: 1939-1498.

[5] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (Apr. 1960), pp. 37–46. ISSN: 0013-1644. DOI: 10.1177/001316446002000104. URL: http://dx.doi.org/10.1177/001316446002000104.

[6] John W Creswell and Vicki L Plano Clark. "Designing and conducting mixed methods research". In: (2007).

[7] Carolyn Cutrona, Dan Russell, and Jayne Rose. "Social support and adaptation to stress by the elderly." In: *Psychology and aging* 1.1 (1986), p. 47. ISSN: 1939-1498.

[8] Alexandre De Masi et al. "mQoL Smart Lab: Quality of Life Living Lab for Interdisciplinary Experiments". In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. UbiComp '16. New York, NY, USA: ACM, 2016, pp. 635–640. ISBN: 978-1-4503-4462-3. DOI: 10.1145/2968219.2971593. URL: http://doi.acm.org/10.1145/2968219.2971593.

[9] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[10] Marios Fanourakis. "Cell ID Similarities from Cell ID Traces for Clustering into Meaningful Places". In: ().

[11] Farhan et al. "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data". In: *2016 IEEE Wireless Health (WH)* (2016), pp. 1–8. DOI: 10.1109/WH.2016.7764553. URL: http://ieeexplore.ieee.org/document/7764553/.

[12] The Whoqol Group. “The World Health Organization quality of life assessment (WHO-QOL): development and general psychometric properties”. In: *Social science & medicine* 46.12 (1998), pp. 1569–1585. ISSN: 0277-9536.

[13] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182.

[14] J Holt-Lunstad et al. “Loneliness and Social Isolation as Risk Factors for Mortality: a MetaAnalysis”. In: *Perspectives on Psychological Science* 10.2 (2015), pp. 227–237. ISSN: 1745-6916. DOI: `10.1177/1745691614568352`.

[15] Jianping Hua et al. “Optimal number of features as a function of sample size for various classification rules”. In: *Bioinformatics* 21.8 (Apr. 2005), pp. 1509–1515. ISSN: 1367-4803. URL: `http://dx.doi.org/10.1093/bioinformatics/bti171`.

[16] Mary Elizabeth Hughes et al. “A short scale for measuring loneliness in large surveys results from two population-based studies”. In: *Research on aging* 26.6 (2004), pp. 655–672. ISSN: 1946-6242. DOI: `10.1124/dmd.107.016501.CYP3A4-Mediated`. arXiv: `NIHMS150003`.

[17] Jenny de Jong-Gierveld, Theo van Tilburg, and Pearl a Dykstra. “Loneliness and Social Isolation”. In: *Cambridge Handbook of Personal Relationships* (2006), pp. 485–500. ISSN: 0-521-82617-9 (Hardcover), 978-0-521-82617-4 (Hardcover), 0-521-53359-7 (Paperback), 9780521533591 (Paperback). DOI: `10.1192/bjp.bp.107.039859`.

[18] Daniel Kahneman et al. “A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method”. In: *Science* 306.5702 (2004), pp. 1776–1780. ISSN: 00368075. DOI: `10.1126/science.1103572`. URL: `http://www.jstor.org/stable/3839780%7B%5C%%7D5Cnhttp://about.jstor.org/terms%7B%5C%%7D0Ahttp://www.sciencemag.org/cgi/doi/10.1126/science.1103572%20http://science.sciencemag.org/content/306/5702/1776.full`.

[19] Junghyun Kim, Robert LaRose, and Wei Peng. “Loneliness as the cause and the effect of problematic Internet use: The relationship between Internet use and psychological well-being”. In: *CyberPsychology & Behavior* 12.4 (2009), pp. 451–455. ISSN: 1094-9313.

[20] Ethan Kross et al. “Facebook Use Predicts Declines in Subjective Well-Being in Young Adults”. In: *PLoS ONE* 8.8 (2013). ISSN: 19326203. DOI: `10.1371/journal.pone.0069841`.

[21] J Richard Landis and Gary G Koch. “The measurement of observer agreement for categorical data”. In: *biometrics* (1977), pp. 159–174. ISSN: 0006-341X.

[22] Reed Larson and Mihaly Csikszentmihalyi. “The Experience Sampling Method.” In: *New Directions for Methodology of Social & Behavioral Science* 15 (1983), pp. 41–56. ISSN: 0271-1249(Print).

[23] Mathias Lasgaard. “Reliability and validity of the Danish version of the UCLA Loneliness Scale”. In: *Personality and Individual Differences* 42.7 (2007), pp. 1359–1366. ISSN: 01918869. DOI: `10.1016/j.paid.2006.10.013`.

[24] T Linehan et al. “2030 vision: The best and worst futures for older people in the UK”. In: *London: Independent Age and International Longevity Centre-UK. Available online (www. ilcuk. org. uk/files/2030-vision-report. pdf), accessed on December* 21 (2014), p. 2016.

[25] Letitia A Peplau and Carolyn E Cutrona. “The revised UCLA Loneliness Scale: Concurrent and discriminant validity evidence”. In: *Journal of personality and social psychology* 39.3 (1980), pp. 472–480.

[26] Daniel Perlman and L Anne Peplau. “Toward a social psychology of loneliness”. In: *Personal relationships* 3 (1981), pp. 31–56.

[27] Brian A Primack et al. “Social Media Use and Perceived Social Isolation Among Young Adults in the U.S.” In: *American Journal of Preventive Medicine* 53.1 (July 2017), pp. 1–8. ISSN: 0749-3797. DOI: `10.1016/j.amepre.2017.01.010`. URL: `http://dx.doi.org/10.1016/j.amepre.2017.01.010`.

[28] Gauri Pulekar and Emmanuel Agu. “Autonomously Sensing Loneliness and Its Interactions with Personality Traits using Smartphones”. In: (2016), pp. 134–137.

[29] Dan Russell, Letitia Anne Peplau, and Mary Lund Ferguson. “Developing a measure of loneliness”. In: *Journal of personality assessment* 42.3 (1978), pp. 290–294. ISSN: 0022-3891.

[30] Daniel W. Russell. “UCLA Loneliness Scale (Version 3): Reliability, Validity, and Factor Structure”. In: *Journal of Personality Assessment* 66.1 (1996), pp. 20–40. ISSN: 0022-3891. DOI: `14.2327`.

[31] Wendy Sanchez et al. “Inferring loneliness levels in older adults from smartphones”. In: *Journal of Ambient Intelligence and Smart Environments* 7.1 (2015), pp. 85–98. ISSN: 18761364. DOI: `10.3233/AIS-140297`.

[32] Holly B Shakya and Nicholas A Christakis. “Association of Facebook use with compromised well-being: a longitudinal study”. In: *American journal of epidemiology* 185.3 (2017), pp. 203–211. ISSN: 0002-9262.

[33] Aaron Smith. “Smartphone ownership–2013 update”. In: *Pew Research Center: Washington DC* 12 (2013), p. 2013.

[34] Rui Wang et al. “StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 2014, pp. 3–14. ISBN: 9781450329682. DOI: `10.1145/2632048.2632054`.

[35] C Wilson and B Moulton. “Loneliness among older adults: A national survey of adults 45+”. In: *Prepared by Knowledge Networks and Insight Policy Research. Washington, DC: AAR Retrieved from: http://assets. aarp. org/rgcenter/general/loneliness_2010. pdf* (2010).

[36] Mike Z. Yao and Zhi Jin Zhong. “Loneliness, social contacts and Internet addiction: A cross-lagged panel study”. In: *Computers in Human Behavior* 30 (2014), pp. 164–170. ISSN: 07475632. DOI: `10.1016/j.chb.2013.08.007`. URL: `http://dx.doi.org/10.1016/j.chb.2013.08.007`.

Appendix A

Scales

A.1 Danish UCLA Loneliness Scale [23]

UCLA Loneliness Scale
(Russell, 1996)
(oversættelse Lasgaard, 2007)

De følgende sætninger beskriver, hvordan man nogle gange har det.
Angiv hvor ofte du har det som beskrevet.
Sæt kun ét kryds ved hvert sætning.

		Aldrig	Sjældent	Sommetider	Altid
1	Hvor ofte føler du, at du er på "bølgelænge" med dem, du er sammen med?	( )	( )	( )	( )
2	Hvor ofte føler du, at du savner nogen at være sammen med?	( )	( )	( )	( )
3	Hvor ofte føler du, at der ikke er nogen du kan henvende dig til?	( )	( )	( )	( )
4	Hvor ofte føler du dig alene?	( )	( )	( )	( )
5	Hvor ofte føler du dig som en del af en vennegruppe?	( )	( )	( )	( )
6	Hvor ofte føler du, at du har meget til fælles med dem, du er sammen med?	( )	( )	( )	( )
7	Hvor ofte føler du, at du ikke længere er tæt på nogen som helst?	( )	( )	( )	( )
8	Hvor ofte føler du, at dine interesser og ideer ikke deles af dem, du er sammen med?	( )	( )	( )	( )
9	Hvor ofte føler du dig udadvendt og venskabelig?	( )	( )	( )	( )
10	Hvor ofte føler du dig tæt på andre?	( )	( )	( )	( )
11	Hvor ofte føler du dig udenfor?	( )	( )	( )	( )
12	Hvor ofte føler du, at dine forhold til andre ikke er meningsfulde?	( )	( )	( )	( )
13	Hvor ofte føler du, at ingen rigtig kender dig godt?	( )	( )	( )	( )

14	Hvor ofte føler du dig isoleret fra andre?	( )	( )	( )	( )
15	Hvor ofte føler du, at du kan finde nogen at være sammen med, når du har lyst til det?	( )	( )	( )	( )
16	Hvor ofte føler du, at der er nogen, der virkelig forstår dig?	( )	( )	( )	( )
17	Hvor ofte føler du dig genert?	( )	( )	( )	( )
18	Hvor ofte føler du, at der er folk omkring dig, men ikke sammen med dig?	( )	( )	( )	( )
19	Hvor ofte føler du, at der er nogen, du kan tale med?	( )	( )	( )	( )
20	Hvor ofte føler du, at der er nogen, du kan henvende dig til?	( )	( )	( )	( )

A.2 Three-Item Loneliness Scale [16]

660 RESEARCH ON AGING

TABLE 1
Items in Revised UCLA Loneliness Scale (R-UCLA)[a] and Three-Item Loneliness Scale

R-UCLA Loneliness Scale

Directions: Indicate how often you feel the way described in each of the following statements. Circle one number for each.

Statement	*Never*	*Rarely*	*Sometimes*	*Often*
1. I feel in tune with the people around me.[b]	1	2	3	4
2. I lack companionship.	1	2	3	4
3. There is no one I can turn to.	1	2	3	4
4. I do not feel alone.[b]	1	2	3	4
5. I feel part of a group of friends.[b]	1	2	3	4
6. I have a lot in common with the people around me.[b]	1	2	3	4
7. I am no longer close to anyone.	1	2	3	4
8. My interests and ideas are not shared by those around me.	1	2	3	4
9. I am an outgoing person.[b]	1	2	3	4
10. There are people I feel close to.[b]	1	2	3	4
11. I feel left out.	1	2	3	4
12. My social relationships are superficial.	1	2	3	4
13. No one really knows me well.	1	2	3	4
14. I feel isolated from others.	1	2	3	4
15. I can find companionship when I want it.[b]	1	2	3	4
16. There are people who really understand me.[b]	1	2	3	4
17. I am unhappy being so withdrawn.	1	2	3	4
18. People are around me but not with me.	1	2	3	4
19. There are people I can talk to.[b]	1	2	3	4
20. There are people I can turn to.[b]	1	2	3	4

Three-Item Loneliness Scale

Lead-in and questions are read to respondent.

The next questions are about how you feel about different aspects of your life. For each one, tell me how often you feel that way.

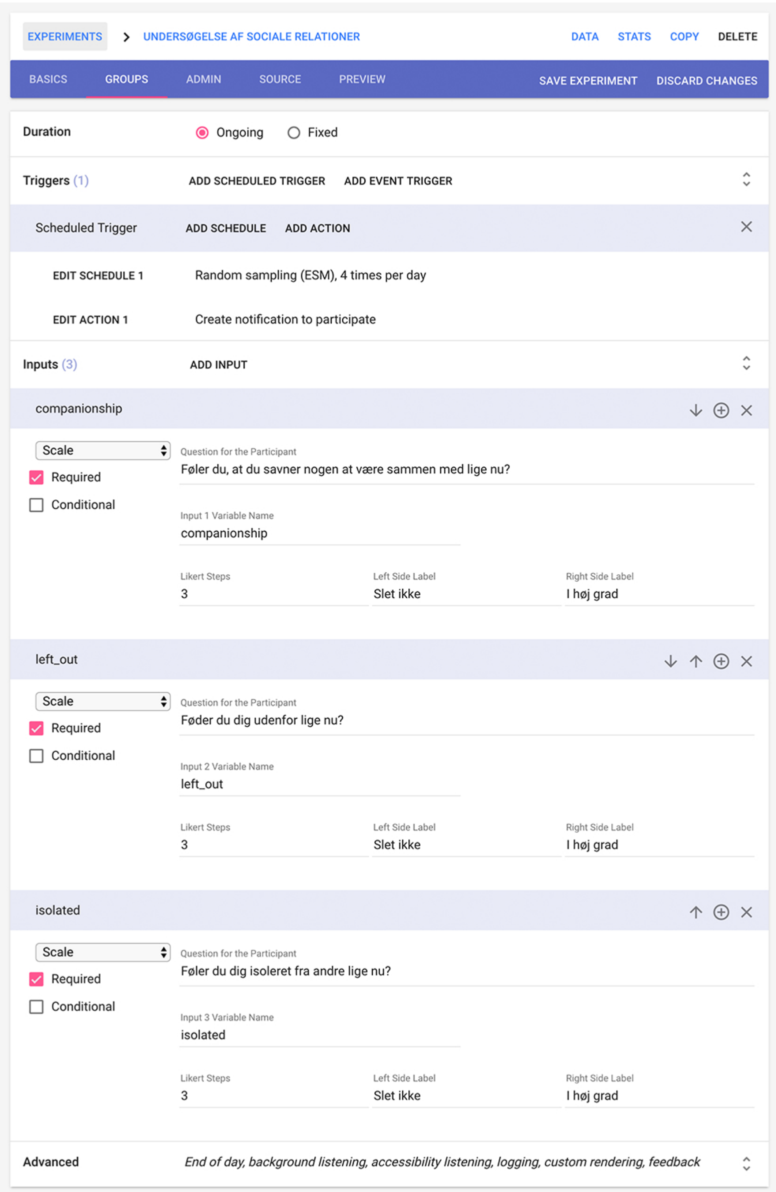
Question	*Hardly Ever*	*Some of the Time*	*Often*
First, how often do you feel that you lack companionship: Hardly ever, some of the time, or often?	1	2	3
How often do you feel left out: Hardly ever, some of the time, or often?	1	2	3
How often do you feel isolated from others? (Is it hardly ever, some of the time, or often?)	1	2	3

NOTE: For both scales, the score is the sum of all items.
a. Russell, Peplau, and Cutrona (1980).
b. Item should be reversed before scoring.

Appendix B

Study Instrumentation

B.1 Experience Sampling Method Instrumentation [1]



[1] `https://www.pacoapp.com/`

B.2 Day Reconstruction Method Instrumentation [18]

Subject ID ____________________________ Date _________________________

Studying the relationship between phone usage and social relations

- Day Reconstruction Method -

Vi vil gerne vide hvordan du oplevede brugen af din telefon og sociale relationer i de sidste 24 timer. Ikke alle dage er ens - nogle er kedelige, nogle er travle og andre er ret typiske. Vi spørger dig kun om de sidste 24 timer.

På de næste to sider vil vi gerne have dig til at beskrive din dag med hensyn til "telefonbrug", "sociale relationer" og "placering" (hvor du har været). Tænk på din dag som en serie af scener eller episoder i en film. Giv hver episode et kort navn, der hjælper dig med at huske det (for eksempel "pendling til arbejde" eller "til frokost"). Markér i tabellen de tidspunkter hvor hver episode begyndte og sluttede. De episoder, som folk normalt identificerer, varer mellem 15 minutter og 2 timer. Vi har lavet dem 30 minutter lange (dvs. hver række i tabellen svarer til 30 minutter), men du er velkommen til at gøre dem kortere eller længere. Angivelse af slutningen af en episode kan være ved ændring af lokation eller afslutning af en interaktion med din mobiltelefon.

Der er plads til at angive 48 episoder for hver af de sidste 24 timer, selvom du måske ikke har brug for dem alle, afhængigt af dine sidste 24 timer. Det er ikke nødvendigt at udfylde alle cellerne - brug den inddeling af dine sidste 24 timer der giver dig mest mening og bedst beskriver hvordan du oplevede din telefonbrug og sociale relationer. Prøv at huske hver episode i detaljer, og skriv et par ord i den tilsvarende række i tabellen. Prøv også at huske, hvordan din oplevelse var under hver episode.

Tag dig god tid til at udfylde tabellen.

Subject ID ______________________________ Date ___________________________

Klokkeslæt	Telefonbrug	Sociale Relationer	Din Placering
12:00 AM			
12:30 AM			
1:00 AM			
1:30 AM			
2:00 AM			
2:30 AM			
3:00 AM			
3:30 AM			
4:00 AM			
4:30 AM			
5:00 AM			
5:30 AM			
6:00 AM			
6:30 AM			
7:00 AM			
7:30 AM			
8:00 AM			
8:30 AM			
9:00 AM			
9:30 AM			
10:00 AM			
10:30 AM			
11:00 AM			

Subject ID ______________________________ Date ___________________________

Klokkeslæt	Telefonbrug	Sociale Relationer	Din Placering
11:30 AM			
12:00 PM			
12:30 PM			
1:00 PM			
1:30 PM			
2:00 PM			
2:30 PM			
3:00 PM			
3:30 PM			
4:00 PM			
4:30 PM			
5:00 PM			
5:30 PM			
6:00 PM			
6:30 PM			
7:00 PM			
7:30 PM			
8:00 PM			
8:30 PM			
9:00 PM			
9:30 PM			
10:00 PM			
10:30 PM			
11:00 PM			
11:30 PM			

Appendix C

Extracted Features

1. applications_used.browser_count
2. applications_used.calendar_count
3. applications_used.camera_count
4. applications_used.contacts_count
5. applications_used.dialer_count
6. applications_used.email_count
7. applications_used.facebook_count
8. applications_used.gallery_count
9. applications_used.keep_count
10. applications_used.map_count
11. applications_used.messaging_count
12. applications_used.messenger_count
13. applications_used.mobilepay_count
14. applications_used.phone_count
15. applications_used.pokemongo_count
16. applications_used.rejseplanen_count
17. applications_used.snapchat_count
18. applications_used.weshare_count
19. cell_ids_service.cluster_home_percentage
20. cell_ids_service.cluster_school_percentage
21. cell_ids_service.cluster_unclassified_percentage

22. cell_ids_service.cluster_work_percentage
23. cell_ids_service.total_cell_distance
24. ping_service.cellular_percentage
25. ping_service.wifi_home_percentage
26. ping_service.wifi_percentage
27. touches_buffered.usage_session_count
28. user_activity.in_vehicle_count
29. user_activity.in_vehicle_percentage
30. user_activity.on_bicycle_count
31. user_activity.on_bicycle_percentage
32. user_activity.on_foot_count
33. user_activity.on_foot_percentage
34. user_activity.running_count
35. user_activity.running_percentage
36. user_activity.still_count
37. user_activity.still_percentage
38. user_activity.tilting_count
39. user_activity.tilting_percentage
40. user_activity.unknown_count
41. user_activity.unknown_percentage
42. user_activity.walking_count
43. user_activity.walking_percentage
44. user_presence_events.off_count
45. user_presence_events.on_count
46. user_presence_events.present_count
47. user_presence_events.rotation_count

Appendix D

Hyperparameters

```
params = {
    'LR': [
        {
            'solver': ['newton-cg', 'lbfgs', 'sag'],
            'C': np.logspace(-4, 4, 5),
            'fit_intercept': [True, False],
            'class_weight': [None, 'balanced'],
            'multi_class': ['ovr', 'multinomial'],
            'max_iter': [10000],
        },
        {
            'solver': ['liblinear'],
            'C': np.logspace(-4, 4, 5),
            'fit_intercept': [True, False],
            'class_weight': [None, 'balanced'],
            'penalty': ['l1'],
        },
        {
            'solver': ['liblinear'],
            'C': np.logspace(-4, 4, 5),
            'fit_intercept': [True, False],
            'class_weight': [None, 'balanced'],
            'penalty': ['l2'],
            'dual': [True, False],
        }
    ],
    'LDA': {
        'solver': ['svd', 'lsqr'],
    },
    'KNN': {
        'n_neighbors': map(int, np.linspace(1, 15, 10)),
        'weights': ['uniform', 'distance'],
        'algorithm': ['ball_tree', 'kd_tree', 'brute'],
        'leaf_size': map(int, np.linspace(1, 50, 10)),
```

```
        'p': [1, 2],
    },
    'CT': {
        'criterion': ['gini', 'entropy'],
        'splitter': ['best', 'random'],
        'max_features': ['auto', 'sqrt', 'log2', None],
        'max_depth': [4, 8] + [None],
        'min_samples_split': range(2, 5),
        'class_weight': ['balanced', None],
        'presort': [True, False]
    },
    'RF': {
        'bootstrap': [True, False],
        'class_weight': ['balanced', None],
        'criterion': ['gini', 'entropy'],
        'max_depth': [4, 8] + [None],
        'max_features': ['auto', 'sqrt', 'log2', None],
        'min_samples_split': range(2, 5),
        'n_estimators': [1, 5, 10],
    },
    'ET': {
        'bootstrap': [True, False],
        'class_weight': ['balanced', None],
        'criterion': ['gini', 'entropy'],
        'max_depth': [4, 8] + [None],
        'max_features': ['auto', 'sqrt', 'log2', None],
        'min_samples_split': range(2, 5),
        'n_estimators': [1, 5, 10],
    },
    'NB': {},
    'SVM': [
        {
            'kernel': ['linear', 'rbf', 'sigmoid'],
            'C': [0.001, 0.01, 0.1, 1, 10],
            'probability': [True, False],
            'shrinking': [True, False],
            'class_weight': ['balanced', None],
            'decision_function_shape': ['ovo', 'ovr', None]
        },
        {
            'kernel': ['rbf', 'sigmoid'],
            'C': [0.001, 0.01, 0.1, 1, 10],
            'gamma': [0.001, 0.01, 0.1, 1],
            'probability': [True, False],
            'shrinking': [True, False],
            'class_weight': ['balanced', None],
            'decision_function_shape': ['ovo', 'ovr', None]
```

```
        },
        {
            'kernel': ['poly'],
            'C': [0.001, 0.01, 0.1, 1, 10],
            'degree': [1, 2, 3, 4],
            'gamma': [0.001, 0.01, 0.1, 1],
            'probability': [True, False],
            'shrinking': [True, False],
            'class_weight': ['balanced', None],
            'decision_function_shape': ['ovo', 'ovr', None]
        },
    ],
    'MLP': [
        {
            'solver': ['lbfgs', 'adam'],
            'activation': ['identity', 'logistic', 'tanh', 'relu'],
            'alpha': [0.0001, 0.001, 0.01, 0.1],
            'max_iter': [5000],
        },
        {
            'solver': ['sgd'],
            'activation': ['identity', 'logistic', 'tanh', 'relu'],
            'alpha': [0.0001, 0.001, 0.01, 0.1],
            'learning_rate': ['constant', 'invscaling', 'adaptive'],
            'max_iter': [5000],
        },
    ],
}
```

Appendix E

Random Forest Trees

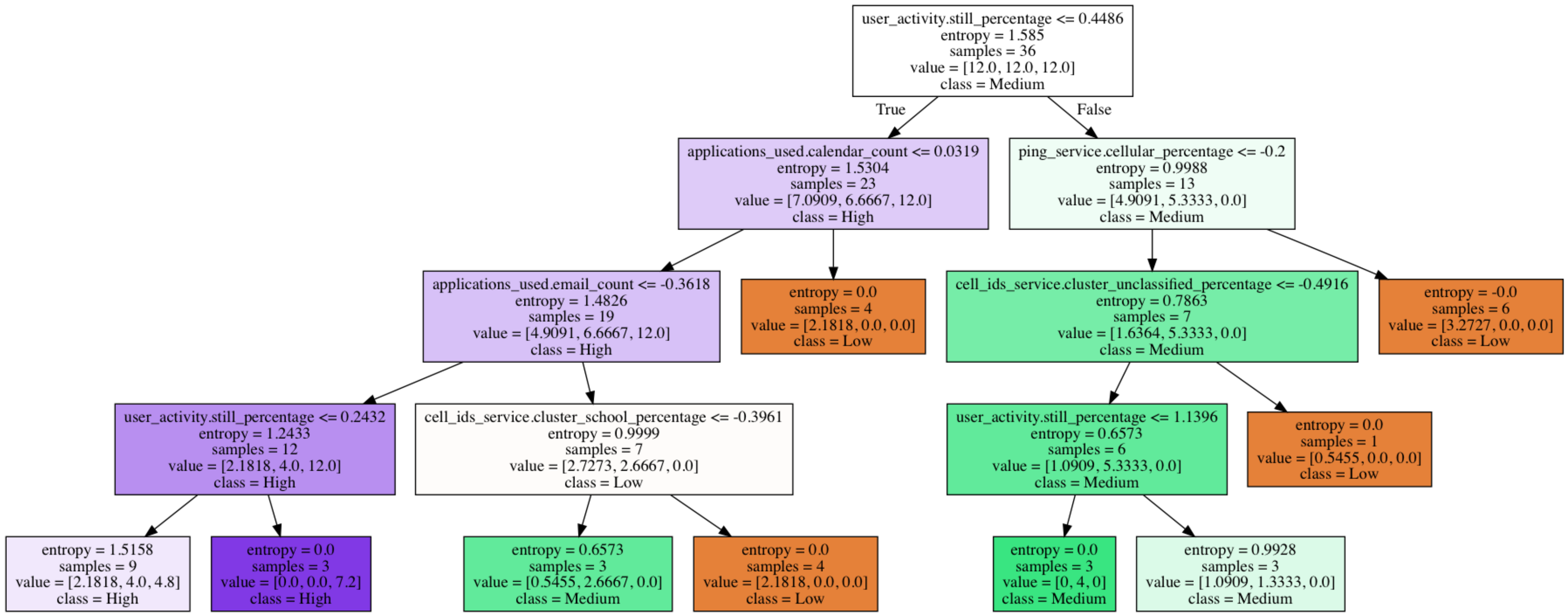


Figure E.1: Classification Tree #1

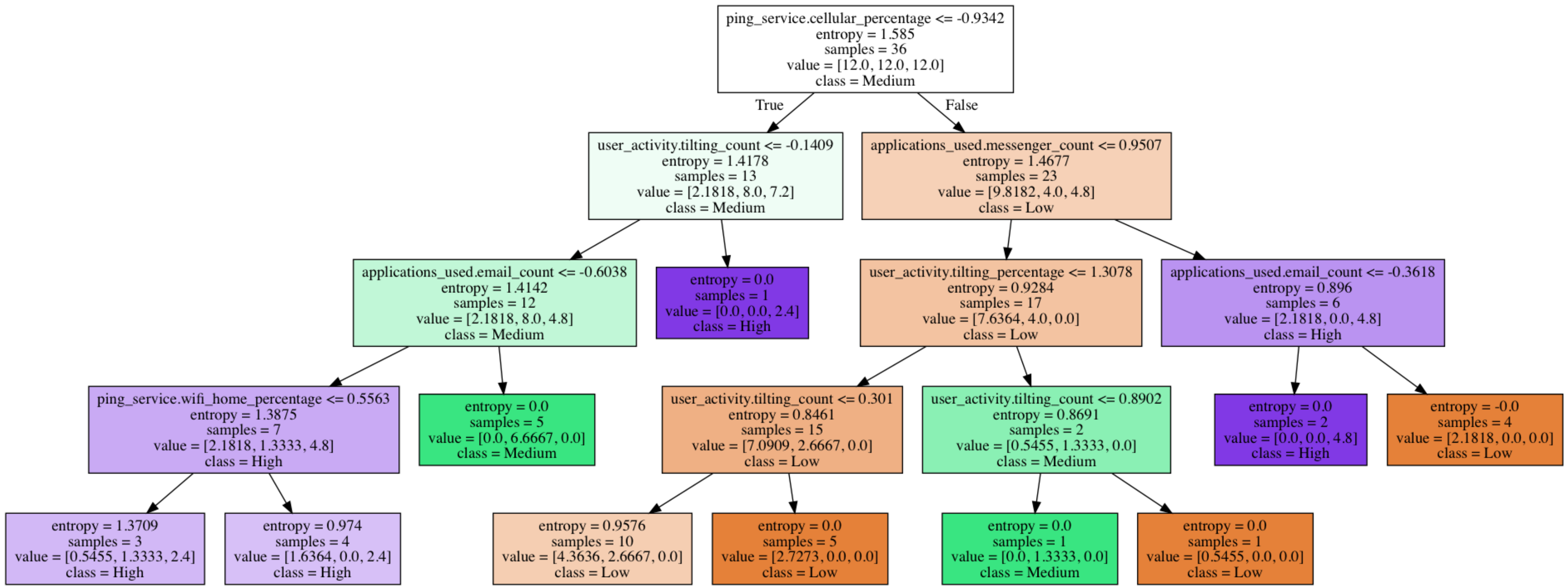


Figure E.2: Classification Tree #2

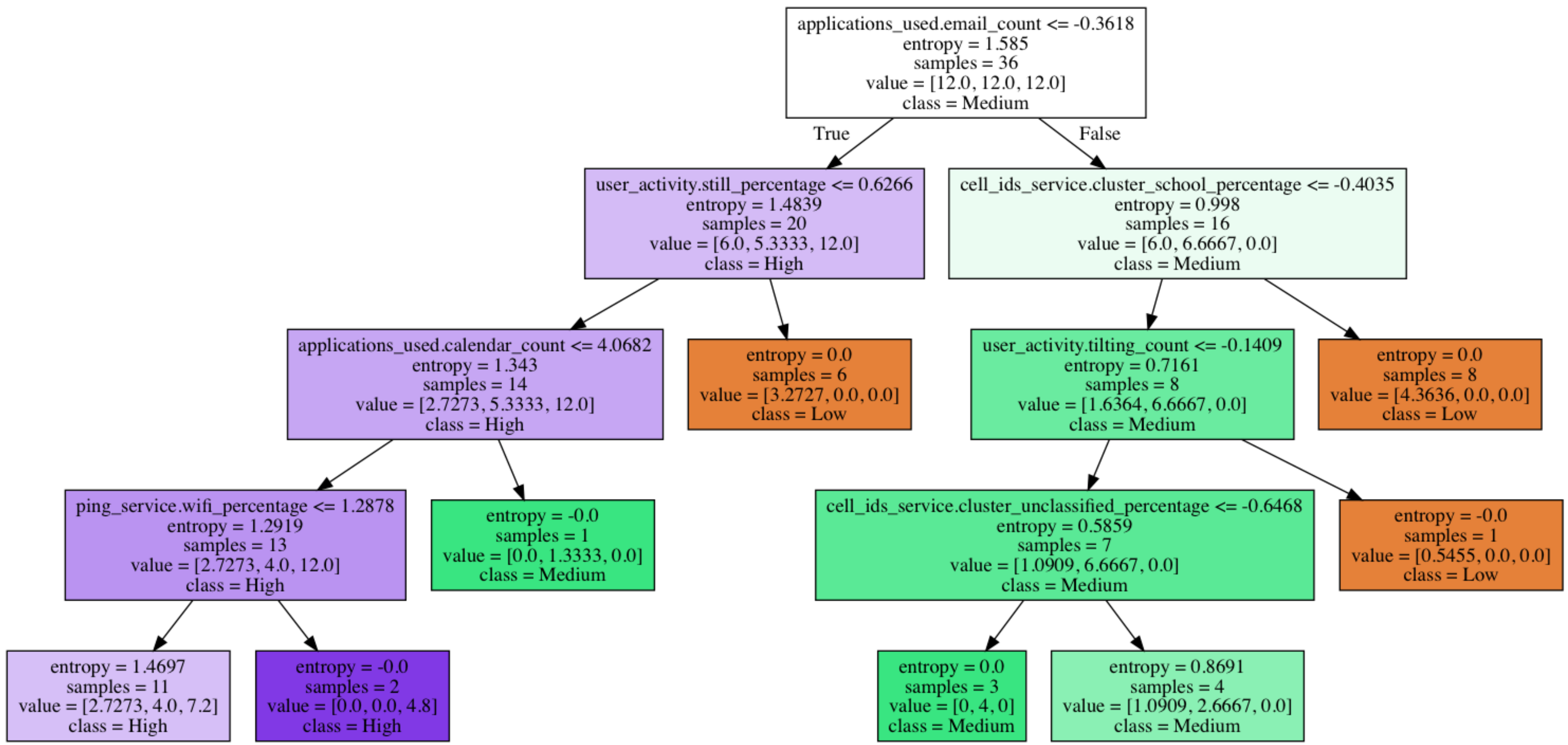


Figure E.3: Classification Tree #3

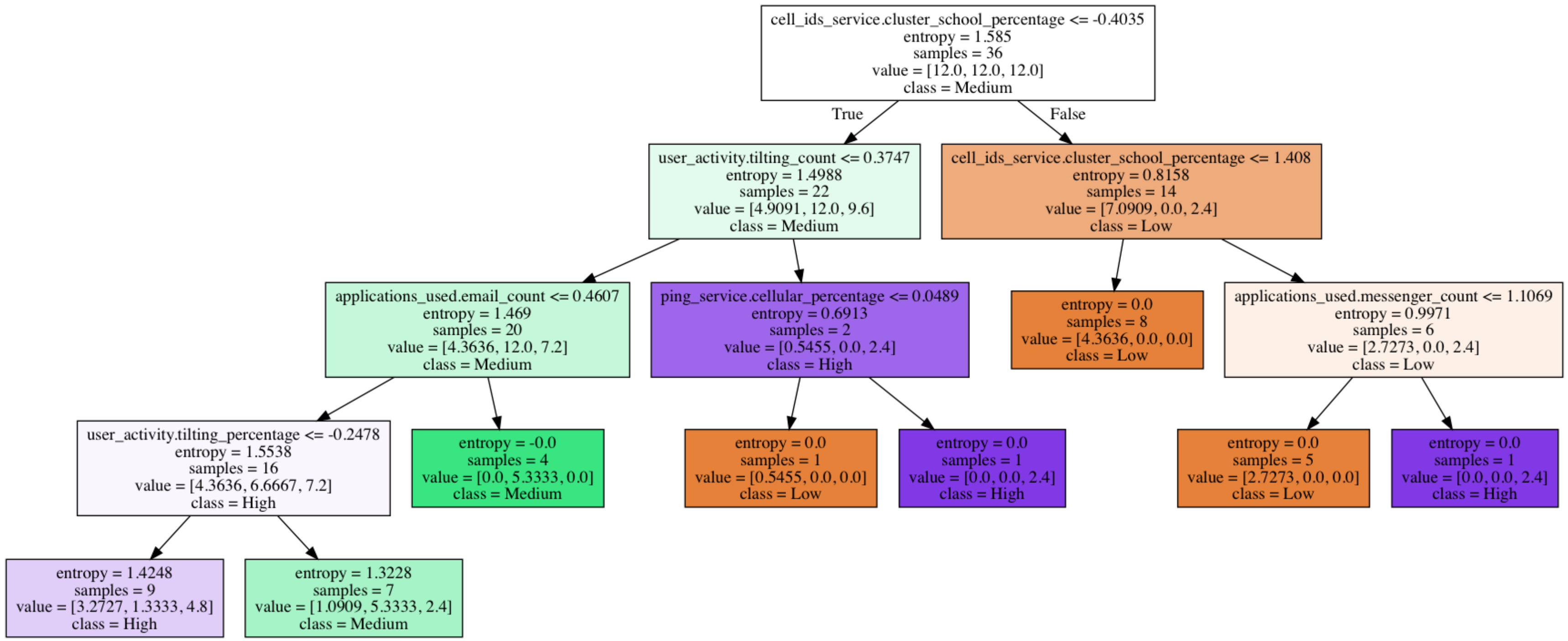


Figure E.4: Classification Tree #4

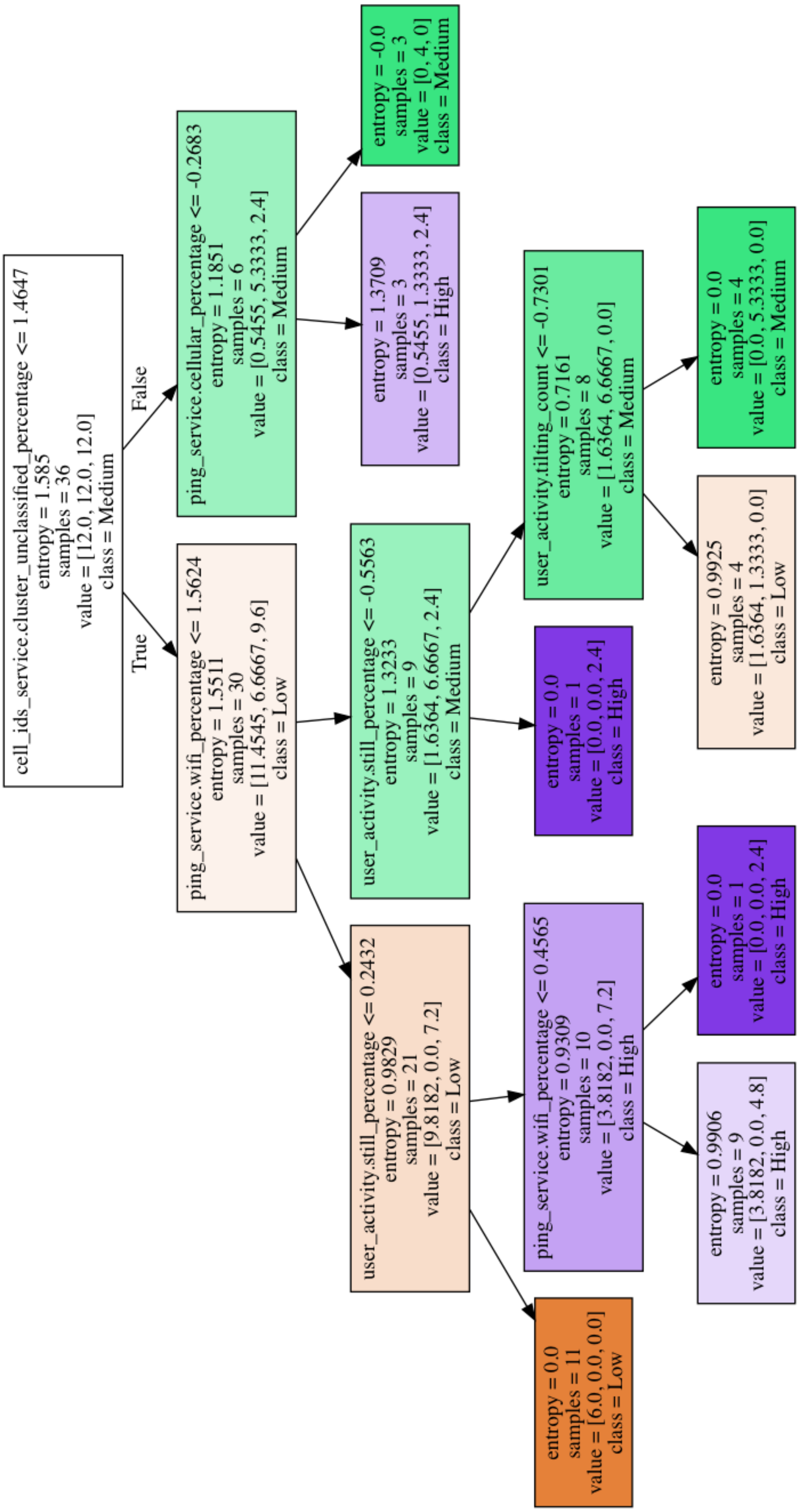


Figure E.5: Classification Tree #5

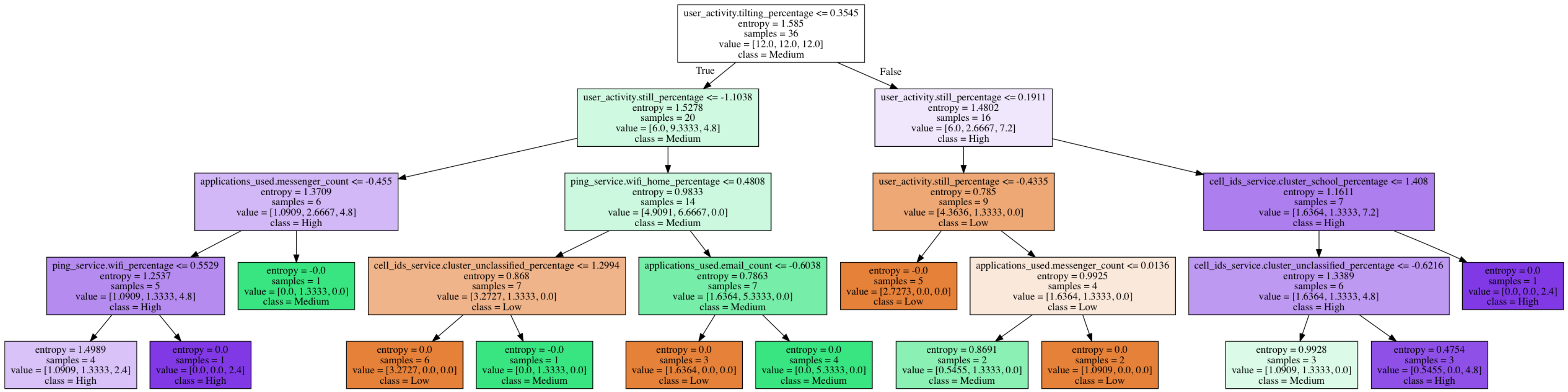


Figure E.6: Classification Tree #6

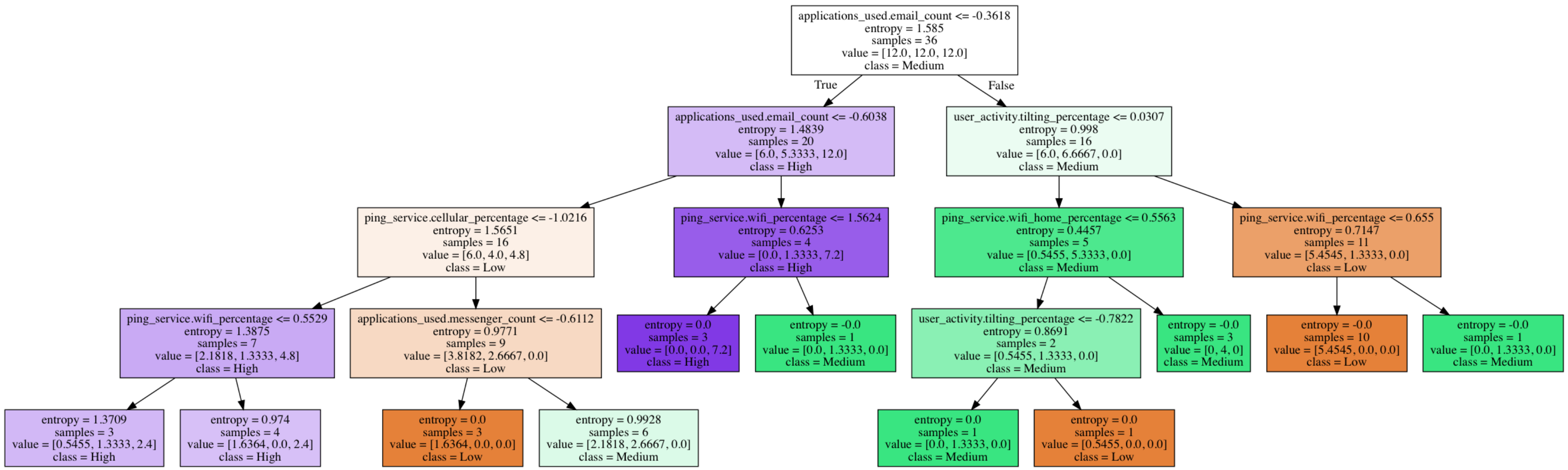


Figure E.7: Classification Tree #7

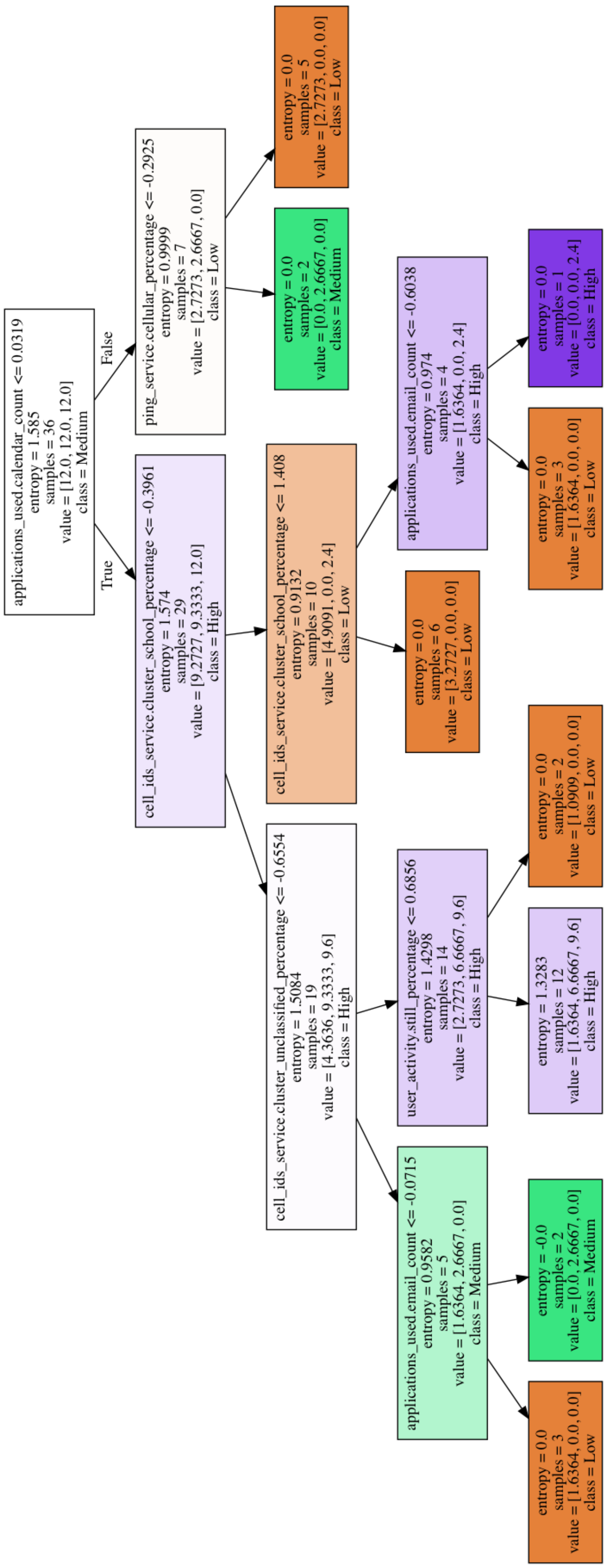


Figure E.8: Classification Tree #8

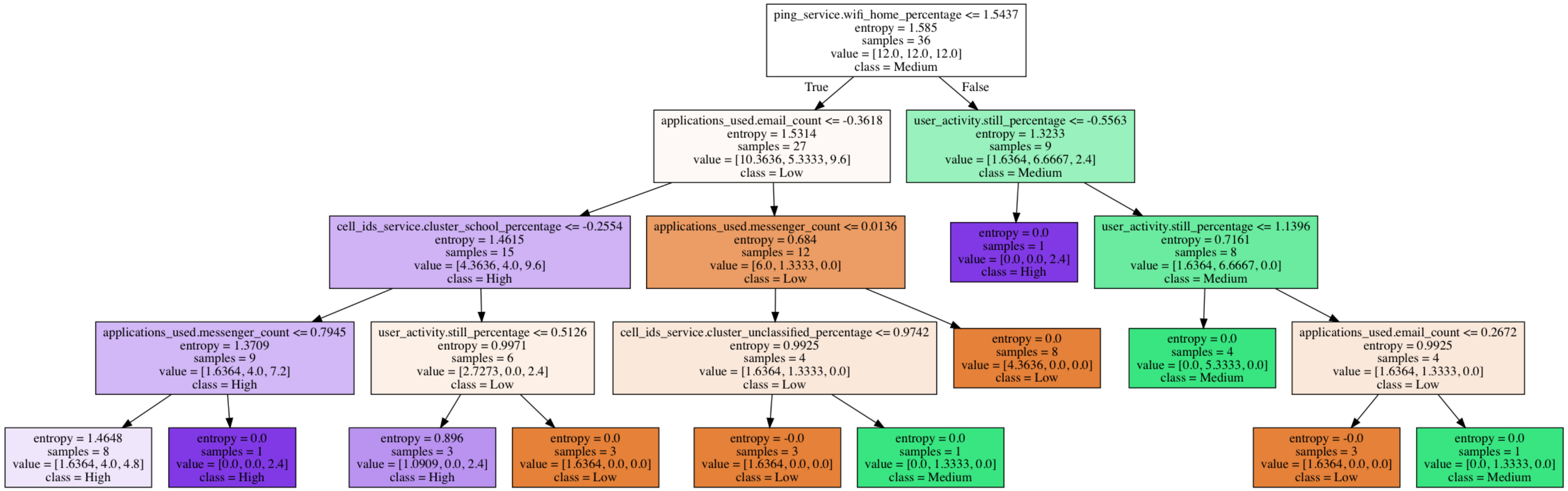


Figure E.9: Classification Tree #9

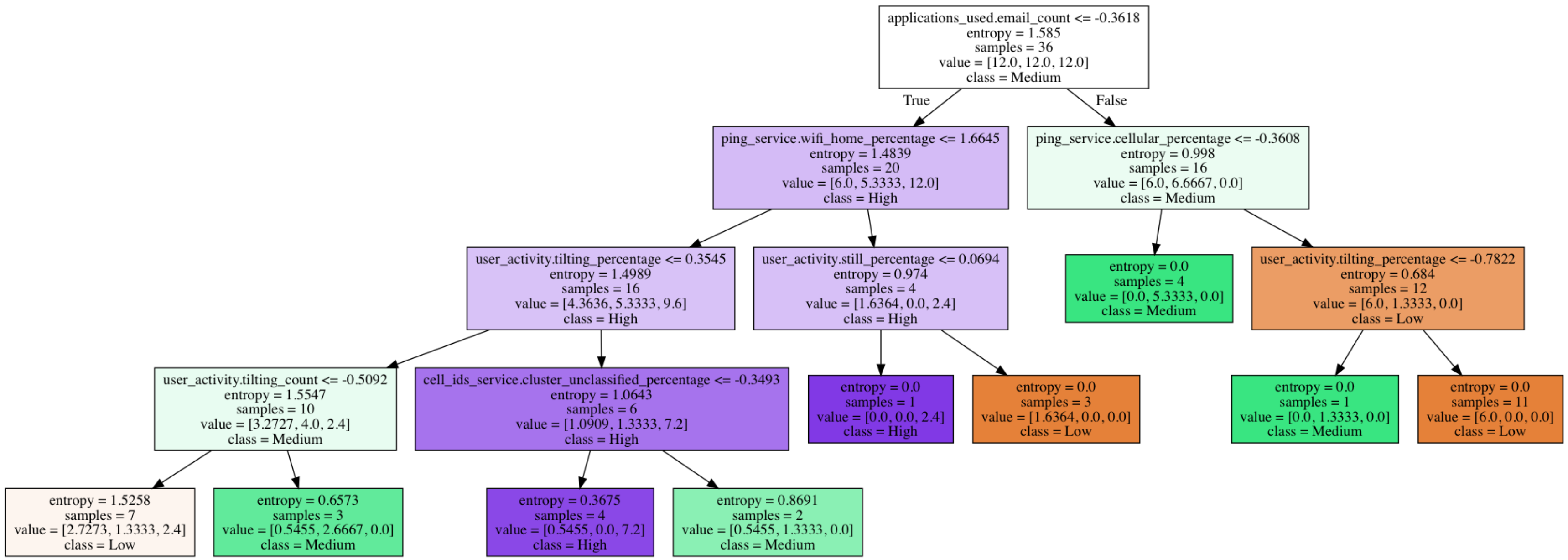


Figure E.10: Classification Tree #10