TOWARD MENTAL HEALTH PREDICTION USING BROWSING HISTORY FOR PREDICTIVE AND SOFT LABELING

by

Sahar Nilipour

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Spring 2022

TOWARD MENTAL HEALTH PREDICTION USING BROWSING HISTORY FOR PREDICTIVE AND SOFT LABELING

by

Sahar Nilipour

Approved: _____
Matthew Louis Mauriello, PhD.
Co-Professor of the Department of Computer and Information Science

Approved: _____
Guangmo Tong, PhD.
Co-Professor of the Department of Computer and Information Science

Approved: _____
Rudolf Eigenmann, PhD.
Interim Chair for Computer and Information Sciences

Approved: _____
Levi T. Thompson, PhD.
Dean of the College of Engineering

Approved: _____
Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to acknowledge the contributions of the members of the Pervasive Wellbeing Technology Lab at Stanford University. Particularly, Marco Mora-Mendoza who provided us with the browsing data essential for this study, Deniz Akin and Medha Verma who designed the Qualtrics survey, and Dr. Pablo E. Paredes, the lab director. I would like to acknowledge Dr. Guangmo Tong, my committee member, who guided me towards the right direction with his technical expertise. Finally, I would like to acknowledge the relentless efforts of my advisor, Dr. Matthew Louis Mauriello, who was of great help in every step of the project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Web browsing data is increasingly being explored as a passive window into the daily lives of users to assess levels of internet addiction and other mental health markers. In this work, I determine whether such approaches can be applied to data generated by the HabitLab platform to assess the community of users for signs of anxiety, depression, and loneliness. HabitLab is a Chrome-based browser plugin that offers users tools to monitor and optimize the time they spend on various websites that negatively impact their productivity while passively logging their browsing sessions. As part of an initial explorative study, 66 HabitLab users completed a paid Qualtrics survey during the height of the COVID-19 pandemic in the US. I analyzed their response to several psychometric scales included on the survey and paired their results with their web browsing data. I then developed several features from this data to characterize their behaviors and used Machine Learning techniques (e.g., SVM, Random Forest) to attempt to learn relationships between their responses. I trained both Classification and Regression models, as I had access to both the real valued scores and their interpretations for most of the mental health scales. The results suggest that the models (specifically regression models), are capable of learning some of the scales; achieving over 80% accuracy for predicting Anxiety and Sleep Disturbance. As the long-term goal of the HabitLab team is to transform the plugin into a Digital Wellbeing and Occupational Health platform, these results can be used to inform: (i) future onboarding and demographic intake questionnaires, (ii) efforts to develop features based on web usage data and predictive models of mental health status, and (iii) facilitate anonymous community assessment through soft labeling approaches.

Chapter 1

INTRODUCTION

1.1 Background

The COVID-19 pandemic has had devastating consequences on global mental health [1] with a disproportionate burden on vulnerable populations [2]; requiring millions to undertake remote work or online learning with limited support is impacting mental health. While stress responses are personal and contextual, common patterns are emerging, including: anxious scrolling ("doomscrolling" [3]), overexposure to negative media [4], and lowered productivity due to overuse of social media [2]. To investigate the mental and social well-being of their users amidst the Coronavirus pandemic, the current HabitLab [5][6] and Home Sweet Office [7] team designed a Qualtrics Survey with questions around demographics, productivity, and mental health and incentivized their users to take part in it by Amazon Gift Cards. Consequently, they collected 66 responses that could be matched with the browsing history of the users. This enabled studying the relationship between browsing and mental health, with the ultimate goal of providing the users with just in time interventions in the event that their browsing history suggests any mental health difficulties.

1.2 Problem

The need for strategies to help people cope with stressors and manage time spent online, especially during remote education and work, is critical. To address this need, we explore the use of combining demographic and web-mining data to predict a variety of mental health markers. Similar to prior work that looked at Internet Addiction Disorder (IAD) [8], we propose exploring how similar processes might be used for Depression, Anxiety, Sleep Disturbance, Loneliness, Quality of Life, Professional

Fulfillment and Work Exhaustion markers. Such data could be collected from users web browsers and, after initial demographic collection, be used to passively monitor their emotional well-being and potentially provide mechanisms for early intervention.

1.3 Approach

In this work, I leverage data collected by HabitLab by creating both browsing behavior features such as "average seconds spent browsing each day", and browsing content features like "the most browsed web category by the user" (categories such as shopping, social media, business, etc). Then the generated web features will be used alongside the collected demographic features and all features will be fed into Machine Learning models such as SVM and Random Forest, with the goal of predicting the mental health markers. As it was possible to collect both raw scores and severity classes for most of the mental health scales, I explored both Regression and Classification tasks to determine the best approach for future related studies.

1.4 Contributions

The main contributions of this project are 1) Analyzing if and how the demographic and browsing features affect each of the seven mental health markers, and whether these features can be combined to develop robust predictive models. 2) Expanding upon the mental health scales previously studied in the context of web browsing; this project introduces Work Exhaustion, Professional Fulfillment, Sleep Disturbance, Loneliness, and Quality of Life in addition to two previously explored markers, i.e. Depression and Anxiety 3) While other related studies [8][13][14] rely on Cross Validation for assessing the performance of the models, I use Incremental Training/Testing in addition to Leave One Out Cross Validation. The former is better equipped to gauge the learning capabilities of a model trained on a small dataset, as it focuses not only on the final accuracy results, but also the trend in accuracy as the training set gets larger 4) A Github repository containing the feature generator functions written in

Python, which could be utilized and contributed to by other researchers. The link to the repository: https://github.com/Sensify-Lab/Web-Sense.git

Chapter 2

RELATED WORK

Several studies have been conducted to address the effects of the Covid-19 pandemic on the mental health of populations [9][10][11][12]. They point to increased levels of anxiety as depression among different populations, where children and young people seem to be impacted the most [10][11]. Covid has also led to increased reports of suicidal ideation and substance use [12].

There is ongoing interest on studying the correlation between browsing behavior and mental health, and several markers such as Anxiety, Depression, Obsessive Compulsive Disorder, Phobic Anxiety, Paranoid ideation, Interpersonal Sensitivity, Somatization, Hostility, Psychoticism, and Internet Addiction have been explored. [8][13][14]. Purwandari et al. were able to predict Internet Addiction and General Mental Health of 40 participants with around 65% accuracy using 10-fold cross validation [8]. Zaman et al. predicted anxiety levels of 104 participants using their Youtube and Google search histories with an average F1 score of 0.83 for binary classification, and an average mean squared error (MSE) of 1.87 for Regression [13]. Zhu et al. collected the browsing history of 47 participants over 4 weeks, and trained 9 SVM models for classifying 9 psychological dimensions according to SCL-90 [15], the cross validation accuracy results for the 9 dimensions range between 78%-100% [14].

I borrow from these works by using the questionnaire based methodology for collecting mental health labels (as used in all three papers), and building upon the approach discussed in [8] and [13] for creating content and behavior web features from the browsing data. I train SVM models as all three papers have done, while also using Support Vector Regression Machines[16] for predicting raw scores of the Mental Health scales. Similar to [8] and [13], I train Random Forest Classifiers, which tend

to work well with imbalanced data (since the survey was conducted at the height of the Coronavirus pandemic, it is expected for most of the participants to express higher than usual levels of Anxiety) [17]. Furthermore, as all of these studies have been (and perhaps continue to be) conducted with small datasets, I aim to improve upon the evaluation aspect of these works (Cross-Validation) by utilizing Incremental Training/Testing, which increases the training data gradually while monitoring the accuracy of the models in search of a trend.

Chapter 3

DATA COLLECTION AND PREPROCESSING

In this chapter, I discuss how the survey data was acquired with support from the current HabitLab team, how it was preprocessed from the self-reported mental health scales, and how It was aligned with the browsing history associated with each HabitLab user. I close by describing the criteria for selecting the window of time to use during the machine learning phase of this work.

3.1 Qualtrics Survey

Between November 2020 and January 2021, HabitLab users were asked to participate in a Qualtrics survey. Participation in the survey was solicited through a dynamic button that appeared in the user interface of the HabitLab dropdown window (i.e., to see the button the user would have to click on the Chrome extension button on top of their web browser interface). This method of solicitation was selected because prior communications with the HabitLab team suggested that the community was resistant, or would simply ignore, email and social media messages.

For those users who opted in to participation (n=66), the survey inquired about their demographic information as well as their lived experience around online work, stress, and productivity in general and in relation to the COVID-19 pandemic. Participants were allowed to skip any questions they did not wish to answer and were compensated with $20 dollar Amazon Gift Cards for their participation. Most relevant to this work, the survey contained several psychometric scales: DSM-5 Depression [18], DSM-5 Anxiety [18], UCLA Loneliness[19], Perceived Stress Scale [20], Work Exhaustion [21], Professional Fulfillment [21], WHOQOL-BREF (Quality of Life)[22], and

Sleep Disturbance [18]. Note that Quality of Life has four dimensions: Physical, Psychological, Social Relationships, and Environment. In this project, however, I only work with the average of the four dimensions, which I will refer to as "QOL mean".

Due to interactions with the HabitLab button to initiate the Qualtrics survey, the results also contain a column labelled "HabitLab user ID" which links the participants' responses to their HabitLab user identification number, making it possible to merge their survey responses with browsing data collected by the HabitLab Chrome Browser plugin. The initial dataset consisted of 66 records, out of which 36 were chosen for this project and aligned with their browsing data. These 36 participants were selected because they had completed the majority of the survey questions including all of the psychometric questions and all of the demographic questions.

3.1.1 Demographic Features

The Qualtrics Survey is very comprehensive in terms of the questions asked regarding the demographics. Features collected from the participants include but are not limited to: Age, Gender, Ethnicity, Education, Annual Household Income, Employment Status, Political Alignment, Country and state of Residence, Marital Status, etc. Five sample rows containing demographic features are depicted in table 3.1.

Among the 37 participants selected for analysis, 22 were male, 13 were female, and 1 person identified as non-binary. The age of our participants was acquired using age ranges as bins. Of these bins, 27 were between the ages of 18 and 30 while 9 were ages 30-55. Furthermore, 15 participants were Caucasian, 11 Asian, 4 Hispanic/Latino, and 3 were African American. The remaining three participants identified their race as mixed or preferred not to describe it. Finally, 13 participants have a Bachelor's degree,

Participant	Age	Gender	Ethnicity	Education	Household Income	Employment
P1	18-30	Male	Latino Or Hispanic	Postgraduate Degree	$25,000-$50,000	Employed
P2	18-30	Male	Caucasian	Some High School	$100,000-$200,000	Job Seeking
P3	18-30	Female	Asian	Bachelors Degree	< $25,000	Employed
P4	30-55	Female	Caucasian	Bachelors Degree	> 200,000	Employed

Table 3.1: Sample rows of the Qualtrics Survey Data - Demographic features

11 have a Postgraduate degree (Master's or PhD), 10 did not finish their undergraduate degree, and 2 have a high school diploma.

3.1.2 Psychometric Features

The most important information contained in the Qualtrics Survey is the Psychometric scales. Table 3.2 depicts five sample participants and the raw scores of some of their Psychometric measures. These scores were not calculated automatically by Qualtrics; instead, all of the mental health scale questionnaires were included in the survey, and the scores were manually calculated and added to the dataset as columns.

As an example, the DSM-5 questionnaire for Depression consists of 8 statements, one of them being "In the past 7 days, I felt that I had nothing to look forward to". The participants are then asked to indicate their agreement with that statement by choosing between Never, Rarely, Sometimes, Often, and Always. With that, each statement on the measure can be rated on a 5-point scale (1 = never; 2 = rarely; 3 = sometimes; 4 = often; and 5 = always) with a range in score from 8 to 40 with higher scores indicating greater severity of depression. For some items in some psychometric measures, the scores need to be inverted. For instance, in the UCLA loneliness scale, the statement "I feel in tune with the people around me" should be reverse scaled. Meaning that if the participant chooses "never", they should get 5 points for loneliness.

Once the raw scores are calculated, their interpretations [18][23] or lack thereof [24], were gathered from various sources. For example, Depression Scores range from 0-40 where scores 0-11 indicate No depression, 12-27 indicate Moderate Depression, and scores above 28 indicate Severe depression. Therefore a column containing these

Participant	Professional Fulfillment	Work Exhaustion	Depression	Anxiety	Sleep Disturbance	QOL mean
P1	2.36	0.875	9	23	23	75
P2	2.34	1.62	25	36	25	55
P3	1.75	0.875	17	23	19	60
P4	2.72	2.12	21	30	20	59

Table 3.2: Sample rows of the Qualtrics Survey Data - Psychometric features (Note that this is not a comprehensive table of all psychometric features. Also, these scales have different ranges, and can not be compared to one another without normalization)

Depression Score	Depression Level	Depression Num
15	None	1
26	Moderate	3
17	Mild	2
26	Moderate	3
25	Moderate	3
20	Mild	2
11	None	1

Figure 3.1: Three columns for the Depression Psychometric Scale

interpretations (None, Mild, Moderate, Severe) was added to the dataset as a new features. On top of that, the numerical representations of these interpretations were also added to the dataset, to facilitate the learning of the future Machine Learning models; 'None' is mapped to 1, 'Mild' is mapped to 2, 'Moderate' is mapped to 3, and 'Severe' is mapped to 4. Fig 3.1 demonstrates some sample rows of the data and how three columns were formed for Depression.

3.2 HabitLab Browsing Data

The browsing data collected by HabitLab is discussed in this section, as well as the approach taken in processing it for the purposes of this work.

3.2.1 Synced Seconds per Domain

After extracting a collection called "Synced Seconds per Domain" from the MongoDB database where HabitLab Browsing data is stored, the HabitLab team was able to provide 35 data files. Each of the 35 files belongs to one of the participants, and it contains the number of seconds the user/participant spent in a browsing session without changing the domain. The most important columns in this dataset are shown in table 3.3. The key takeaways from this table are that the Time column is cumulative, and that a session is defined by the pair (Session ID, Domain).

Column Name	*Description*
Domain	The domain of the URL visited by the user
Session ID	The ID of the session; must be paired with Domain
Timestamp (Global)	The UTC time of the visit
Time	The cumulative number of seconds the user spent in that session

Table 3.3: The key columns of Synced Seconds per Domain

Time	Session ID	Habitlab ID	Domain	Timestamp	Local Timestamp
1	2913	3be71e7fb53ebd87c9da745c	twitter.com	2020-10-27 10:02:55.418000	2020-10-27 05:02:55.418
81	2913	3be71e7fb53ebd87c9da745c	twitter.com	2020-10-27 10:05:21.554000	2020-10-27 05:05:21.554

Figure 3.2: Sample rows from the Synced Seconds per Domain file of a random HabitLab user

Fig 3.2 shows a twitter session for a random HabitLab user with id = "3be7...". Both rows refer to the same session, as they share (Session ID = 2913, Domain = twitter.com). It is not clear why two separate rows are created for the same session; nevertheless, I assume that this is a result of the user taking an action (e.g., clicking on a tweet) inside a domain (i.e., twitter). For example, in fig 3.2 the user spent one second on 'twitter.com', and then took an action inside twitter which led to the creation of a new row. However, because the user did not create a new Web Page and type in twitter, the Session ID stayed the same; the user then spent an additional 80 seconds on twitter, resulting in a total time of 81 seconds in the second row. It follows from this example that the total time spent on a domain in one session can be obtained by looking at the 'Time' column in the last row with the aforementioned (session ID, Domain).

I computed the **"Local Timestamp"** column in Fig 3.2 by offsetting the global timestamp; the offset was calculated based on the country of residence for each participant which was self-reported in the survey. For participants residing in Canada, however, Eastern Time was used as their exact timezone could not be identified. Local Timestamp allows for a more accurate representation of a day for each participant.

3.2.2 Study Window

The browsing data collected after the participants took the survey is out of the scope of this study. Therefore, the date when the survey was taken by each participant was extracted (this is a column in the Qualtrics survey), and all of the browsing data collected after that date was removed in all of the 35 files. Then, It was observed that some users had years worth of browsing data while others only had a few weeks (Average Days before survey: 479.97, Standard Deviation of days before survey: 315.97). As a result, we calculated *days = (Survey taken date - oldest browsing record date)* for all of the users, and identified the minimum value for days, which is equal to 16.

In order to achieve comparable datasets, we calculated the start *date = (Survey taken date - 16)* for each participant and deleted all of their browsing history collected before "start date". This would ensure that all users have 16 days of browsing history before they took the Qualtrics Survey.

After this step, one participant had to be removed from the study as they had no rows in their "Seconds on Domain per Session" file; this is possibly due to being logged out of HabitLab for the duration of our study window. As a result, our final aligned data set contains 33 complete records to use for exploratory data analysis and subsequent modeling.

Appendix A provides more information about the 33 participants by presenting some interesting plots generated for the purpose of Exploratory Data Analysis. However, It is worth noting that no strong correlations were found between any pair of demographic and psychometric features.

Chapter 4

FEATURE ENGINEERING

In this chapter, I focus on my process for creating features that would be suitable for characterizing browsing behaviors for training the predictive Machine Learning models that follow. I discuss how I used the data contained in the "Synced Seconds per Domain" csv files to generate web-based features for all the participants, as well as the demographic features chosen from the Qualtrics Survey.

4.1 Web and Browsing Features

The exploratory hypothesis of this work is that browsing behavior could potentially be used to predict a user's score on some mental health scales. With that in mind, I looked into generating features that could characterize browsing behaviors. For example, some studies show that anxious people tend to spend more time online [25][26], which prompted me to generate the following features, all calculated based on the 'Synced Seconds per Domain' files:

Days Active: Represents the number of days the user was active during the 16 day study window. (Range: 3-16, Average: 12.51, Std: 4.16)

Average Seconds per Day: The average number of seconds the user spent browsing over the study window period. In the case that the user was not active during a given day, We assumed that they browsed for zero seconds. In order to compute this feature, we first had to compute the number of seconds the user spent on each domain. As mentioned before, this number is obtained by looking at the last row with a unique (session ID, Domain). We extracted all such rows, and added up all of their "Time" columns to get the total seconds spent each day, and then averaged those over all 16 days. (Range: 22-22323, Average: 10274 , Std: 6026.61)

Average Domains per Day: The average number of domains the user visits each day. Similar to "Average Seconds per Day", we extract all pairs of ¡session ID, Domain¿ and the number of such pairs determines the number of domains visited in each day. The values are then averaged over 16 days. (Range: 0-753, Average: 86.90, Std: 126.08)

Average Unique Domains per Day: Following the pattern of the previous features, this time we look at the unique pairs of (session ID, Domain) in each day and then average them over 16 days. (Range: 0-98,Average: 28.93, Std: 20.86)

Note that the minimum value for "Average Domains per Day" and "Average Unique Domains per Day" is zero. The reason is that if the participant was browsing for an average of 22 seconds (minimum value of "Average Seconds per Day") over the 16 day study window, then they probably did not browse at all for some of those days, leading to several zeros being included in the averages. On top of that, imagine the case where on the days that they did browse, they only visited a handful of websites. This could easily result in the average being zeroed out (perhaps due to rounding down).

4.2 Content Browsing Features

Apart from the frequency of visiting web pages and the time spent on each domain, the type of content surfed by the users could also play an important role in mental health prediction and/or into explaining certain behaviors [8][14][26]. In order to explore this hypothesis in the scope of this study, I identified the categories for all of the unique domains in the HabitLab browsing data using the "WebShrinker" API [28]. This API takes a URL as input, uses Machine Learning and Natural Language Processing to analyze it, and outputs one or more categories that best describe the URL. Some examples of these categories are: shopping, sports, social networking, travel, virtual reality, etc. The complete list of categories can be found in [31].

I created a table which maps all of the unique domains browsed by all participants to the output generated by WebShrinker. The following figure 4.1 depicts some of the rows of the aforementioned table:

wallacegarrisons	business
springplace.us12	business
www.mainewood	shopping,business
outlook.office.cor	chatandmessaging
www.bustle.com	entertainment,education
pt.stackoverflow.(	informationtech,messageboardsandforums
www.nike.com	shopping
email.ngpvan.cor	uncategorized
www.thosmoser.(	shopping,business
ymlp.com	business,informationtech
www.madewell.c(	shopping
tracking.lotusinte	uncategorized
ntg.omeclk.com	parked
www.nytimes.cor	newsandmedia

Figure 4.1: Sample rows of the csv file containing all the unique domains visited by all the participants and their content category; Parked category refers to websites that are no longer owned by the original user, or are being offered for sale.

The categories shown in Fig 4.1 were preprocessed so that each domain could be represented by only one category:

- If the domain had several categories associated with it, the first one was chosen (with the exception of the first category being equal to business). For instance, 'www.mainwoodworks.org' in Fig 4.1 is more closely related to shopping than it is to business. So shopping is chosen as its category.
- If there were several categories associated with a domain, and the first one was business, the second one was chosen. I made this decision based on the observation that many domains that represented businesses were categorized as business. However, what may be more important is the content and the focus area for that business. For example, even though 'ymlp.com' in Fig 4.1 is a business, the user who browsed it is more likely to be concerned with the information technology aspect of that domain.
- The domains labeled as uncategorized or parked were ignored in the feature generation phase.

After the preprocessing of the API results, the following groups of content-related features were created:

1. **Most popular categories**: The top three most popular categories of each participant were identified and were added to the dataset as three distinct features. Note that we regard each row in the 'Synced Seconds per Domain' data as one domain for the purposes of calculating these features. This is a reasonable approach as more activity during a session should in fact contribute to the domain category's popularity. Fig 4.2 exemplifies these three features. The three popularity features are categorical and nominal, therefore they were one-hot encoded before being fed into the Machine Learning models.

2. **Proportional categories**: for all participants, I calculated the percentage of their browsing history dedicated to each of the unique categories. For example, imagine that out of all the domains browsed by participant 1, 50% were entertainment, 30% were social networking, and 20% were education. Then participant 1 will have 0.5 in their entertainment column, 0.3 in social networking and 0.2 in education. The categories that were not visited at all by participant 1 will get the value 0.0 in the respective columns. Fig 4.3 exemplifies these features.

First Category	Second Category	Third Category
entertainment	informationtech	education
mediasharing	education	searchenginesandportals
education	informationtech	entertainment
education	business	searchenginesandportals
entertainment	socialnetworking	education
entertainment	newsandmedia	socialnetworking
entertainment	education	informationtech

Figure 4.2: Example of the top three popular categories for seven users; The first user has browsed domains with entertainment content the most, followed by information tech content and education content.

4.3 Demographic(Survey) Features

The survey columns that were used as features for the Machine Learning models are: Age, Education, Gender, Employment Status, Annual Household Income, Ethnicity, Political Alignment, and Marital Status. Out of these features, Age, Education and Income are ordinal features, while the others are nominal. The ordinal features were

business	shopping	chatandmessaging	entertainment	informationtech
0.010526	0.000000	0.052632	0.242105	0.231579
0.072275	0.106635	0.024408	0.044076	0.012559
0.049460	0.019948	0.004919	0.114223	0.122694
0.127044	0.053630	0.094506	0.038097	0.050523
0.006997	0.005831	0.022157	0.309621	0.006122

Figure 4.3: Example of proportional categories for five users; for the first user, business content took up 1% of the browsing traffic, while chats and messaging took up 5% and information technology took up 23%.

used in the models as they are, while the nominal features were converted to binary columns using one-hot encoding.

Chapter 5

MACHINE LEARNING SETUP

In this chapter, I discuss the labels and the features chosen for the Machine Learning models, since not all of them are equally suitable for the predictive Models (due to low variance among the data points, for example), and not all features are relevant to all labels. Then, I discuss the methods developed for training and testing the models for both regression and classification tasks.

5.1 Labels

It was mentioned earlier that I had access to several self-reported psychometric scales such as Loneliness, Anxiety, Sleep Disturbance, etc. Nevertheless, not all of these markers are equally valuable for our predictive models. This limitation is caused by the low variance among the majority of these scales; most notably Anxiety, Loneliness, and Sleep Disturbance. In other words, most of the participants reported similar levels of these scales, making it difficult to train a model that is accurate and unbiased at the same time. On the other hand, Depression and Work Exhaustion seem to have the most balanced responses. With that in mind, the results for all of the labels will be discussed.

In addition to the raw scores of the mental health markers, the interpretations of the scores and the numerical representation of the interpretations, a column consisting of binary values was added to the feature set for all of the psychometric labels, so that binary classification would be facilitated. In order to do this, the numerical interpretations were broken in the middle. For instance, None and Mild Anxiety will be labeled as negative, while moderate and High Anxiety will be labeled as positive. Authors of [8] and [13] adopted a similar approach for creating binary classes. For

	Positive	Negative
Work Exhaustion	12	21
Depression	14	19
Anxiety	29	4
Sleep Disturbance	23	10
Loneliness	17	16
Quality of Life	20	13
Professional Fulfillment	19	14

Table 5.1: The number of data points in the negative and positive classes for each of the seven psychometric scales. Note that as the classes are binary, the sum of the negative and positive classes will always be 33.

reference, the number of data points in both positive and negative classes for all of the potential labels are provided in table 5.1.

5.2 Feature Selection Method

Several feature selection methods such as Exhaustive Feature Selection and Backward Feature Elimination were considered for this study. However, due to the high number of features (over 140, mostly a result of one-hot encoding) and the utilization of incremental testing 5.3 which required training many models, I opted for Chi-square test and Variance Threshold.

The selector method takes the training set of features, the training set of labels, the variance threshold and k as input. Note that k indicates the number of features that perform the best on the Chi-square testing. The algorithm selects the best k features that highly correlate with the label we aim to predict. Of these features, only the ones with a greater variance than the variance threshold are chosen for the model. This quick algorithm will avoid low variance features which don't contribute much to the model while guaranteeing their helpfulness for predicting the label.

5.3 Incremental Testing

As only 33 records exist in the dataset, I choose an incremental approach to see whether the model is learning to predict the labels. First, I select 9 records randomly and set them aside as test data. Then, I train a model with 6 randomly selected

records, attempt to predict the test data and note the achieved accuracy. Next, I train a new model with 12 records (previous training data in addition to another 6 randomly selected records from what is left over), I attempt to predict the test data, note the accuracy again and move on to the final step. Finally, I add in the remaining 6 records and train and test the last model with 24 data points in the training set, and 9 data points in the test set. Note that for training any of the models, the feature selection algorithm is re-run on the randomly selected train data. So the features used in the model trained with a train set of size 6 might be different from the features used in the model trained with a train set of size 12.

5.3.1 Iterative Incremental Testing

Due to the limitation imposed by the size of the dataset, It is not reasonable to base one's judgment of the performance of the model on a single run of the incremental testing. With that consideration, I ran the training 100 times and calculated the average of the accuracy in each of the four rounds. In other words, I end up with a list of four numbers named *AccuracyAvg*, where *Accuracyavg[0]* represents the average of the accuracies of 100 models, where the training sets of the models contains 6 data points, *Accuracyavg[1]* represents the average of the accuracies of 100 models, where the training sets of the models contains 12 data points, *Accuracyavg[2]* represents average accuracies of models with training sets equal to 18, and finally, *Accuracyavg[3]* represents average accuracies of models with training sets equal to 24.

One can claim that if *AccuracyAvg* turns out to be strictly increasing, It can be deduced that the model is learning, as the larger train sets result in higher prediction accuracies. This hypothesis is put to test in the next chapter.

5.3.2 Evaluation Metrics

As previously discussed in the Qualtrics Survey section, the mental health scales provide both raw real-valued scores, and classes to signify the severity of the issue

(None, Mild, Moderate, High). Therefore, regression tasks are needed to predict the raw scores, and classification tasks are needed to predict the severity classes.

5.3.2.1 Classification

After running each model (with varying sizes of train data) and predicting the 9 data points in the test set, the following four values are calculated: True Positive (the model predicted the label of the test data point as 1, and the label of test data point is in fact equal to 1), True Negative (the model predicted the label of the test data point as 0, and the label of the test data point is in fact equal to 0), False Positive (the model predicted the label of the test data point as 1, however, the label of the test data point is 0), and False Negative (the model predicted the label of the test data point as 0, however, the label of the test data point is 1)

Note that even though there are four severity classes (i.e., None, Mild, Moderate, Severe), binary classification is preferred. This choice is driven by the fact that the test set is limited to 9 data points.

After evaluating all of the 9 predictions and calculating TP, TN, FP, and FN, the accuracy is calculated with the following formula:

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Note that in the case of this project, the denominator will always be equal to 9, as that is the fixed size of the test data.

5.3.2.2 Regression

After running each regression model, the errors of the predictions on the test data points (i.e., 9) data points are calculated using MAPE (Mean Absolute Percentage Error), and the accuracy is obtained by the following formula where n is equal to the size of the test data:

$$Accuracy = 100 \times (1 - \frac{\sum_{i=1}^{n} \frac{abs(predicted - real)}{real}}{n}) \tag{5.2}$$

5.4 Leave One Out

With this approach, I choose only one data point for testing, and use the remaining 32 data points as the train set. In each iteration, I select the best performing features according to the train data with the feature selection algorithm, then I train the model and use it to predict the single test data point. If it's a classification task, one of the four values TP, TN, FP, or FN will be incremented by one. For example, if the predicted label is 1 but the real label of the test data point is 0, then FP will be incremented by one. After running all of the 33 iterations, I look at the four values and compute the accuracy using formula 5.1. If it's a regression task, the MAPE error of the prediction is calculated. After training and testing all 33 models, accuracy is calculated by subtracting the average of the 33 MAPE errors from 100 5.2.

Chapter 6

MACHINE LEARNING RESULTS

In this chapter, I report the accuracy results of Incremental Testing and Leave One Out for both Random Forest and SVM models. The features that contribute the most to each model and the weight in which they do so will also be discussed.

For all models, the parameters k and *variance* are used for the feature selection algorithm discussed in 5.2; k denotes the number of features to be selected by the chi-square test and *variance* denotes the variance threshold that determines the minimum variance of the selected features. The values for k range from 10 to 60, and the values for *variance* range from 0.2 to 0.4. For each model and label, I report the k and *variance* values that resulted in the highest accuracy results. However, It is important to note that there were no stark differences in accuracy when changing these values. For example, if Incremental Testing showed an upward trend in accuracy for Loneliness, the trend did not change significantly as a result of changing k and *variance*. This was consistent with all the labels. When reporting accuracy, the accuracy of classification tasks is calculated using formula 5.1, and the accuracy of regression tasks is calculated using formula 5.2.

6.1 Incremental Testing - Classification

For incremental testing, the best results are reported in terms of 1) seeing an upwards trend in accuracy of predicting the psychometric label and 2) the highest prediction accuracy achieved.

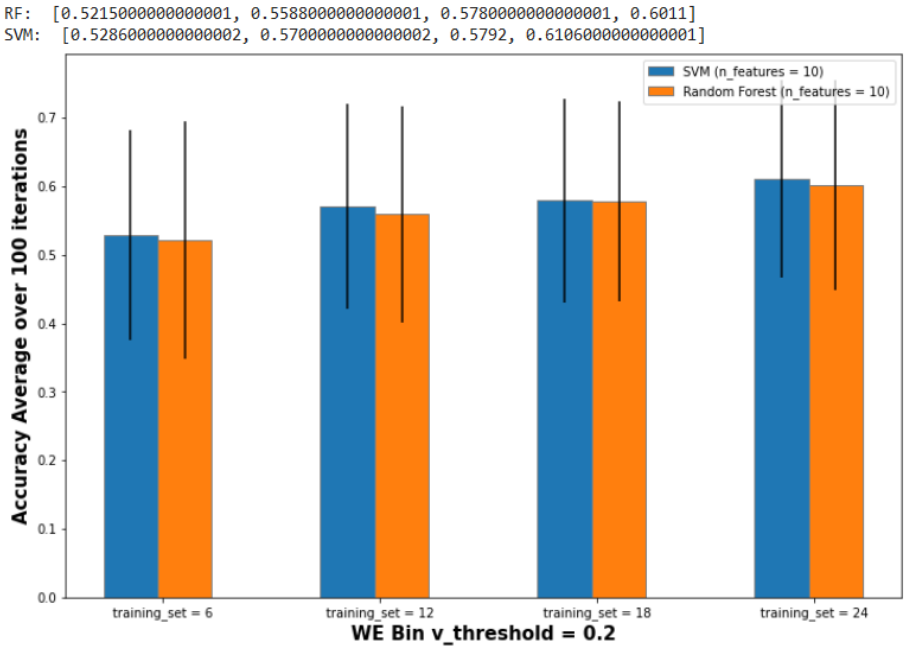
As previously discussed in 5.1, when reporting results for Incremental Testing on classification tasks, all of the labels are binary classes; meaning that the records with 'None' and 'Mild' interpretations will be labeled as negative, while the records

with 'Moderate' and 'High' interpretations will be labeled as positive. You can refer to the table 5.1 to see how the dataset is broken into two classes for each of the psychometric labels. As for model specifications, all the SVM models use Radial Basis Function kernels[32], but the number of estimators in different Random Forest Classifiers/Regressors might vary (the models were trained with various values for the number of estimators, but similar to k and *variance*, the number of estimators did not significantly change the trend of learning).

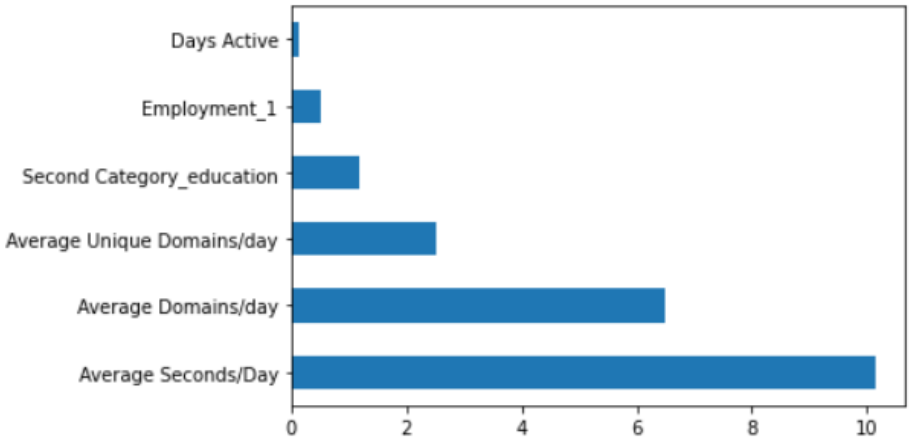
Work Exhaustion: An overview of the results of incremental training/testing for predicting Work Exhaustion is shown in Fig 6.1. As shown in Fig 6.1a, when the size of the training set increases, the accuracy of Work Exhaustion prediction also seems to improve for both Random Forest and SVM models.

Additionally, Fig 6.1b plots the most significant features for predicting binary Work Exhaustion over 100 iterations according to Random Forest. The feature importances are aggregates of the features of the best performing Random Forest models with the largest training sizes; that is, the models which use 24 training data points and predict Work Exhaustion binary class with an accuracy greater than 75%.

Other Psychometric Scales: After studying the plots of all other psychometric scales, it seems that the pattern of increasing accuracy was neither consistent nor strong enough to claim that the models were learning. You can refer to the plots in appendix B.



(a) Random Forest and SVM achieve higher accuracy for predicting binary Work Exhaustion labels as the size of the training set increases.



(b) The aggregate weight of the most important features according to several Random Forest classifiers that predict Work Exhaustion Binary class with accuracy $> 75\%$; Note that Employment1 refers to full-time employment

Figure 6.1: Random Forest and SVM accuracy results and feature importance for binary class Work Exhaustion

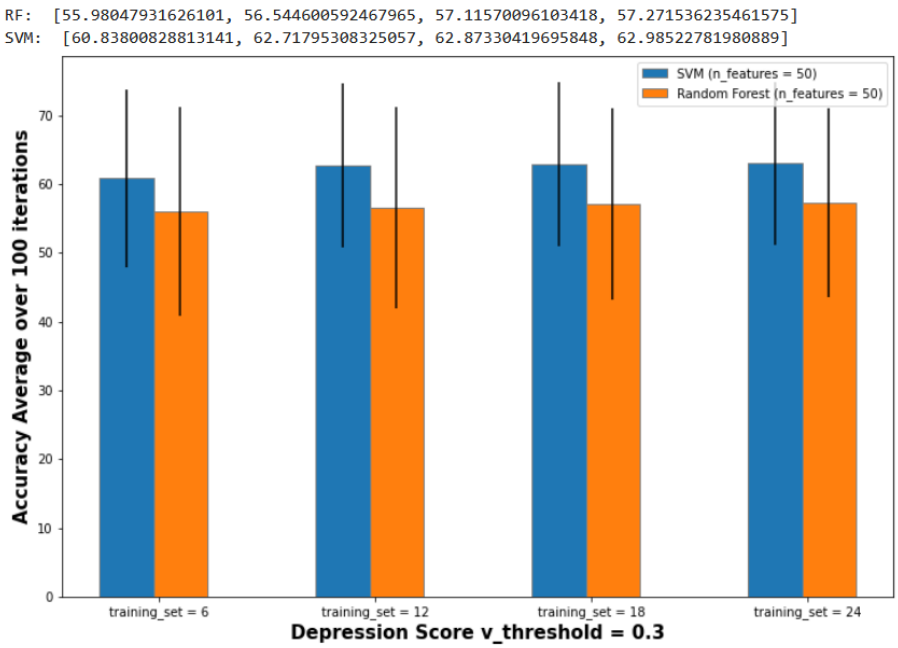
6.2 Incremental Testing - Regression

The goal of this section is to track the prediction accuracy of the Random Forest Regressor and SVR models on predicting the psychometric scores depicted in table 3.2

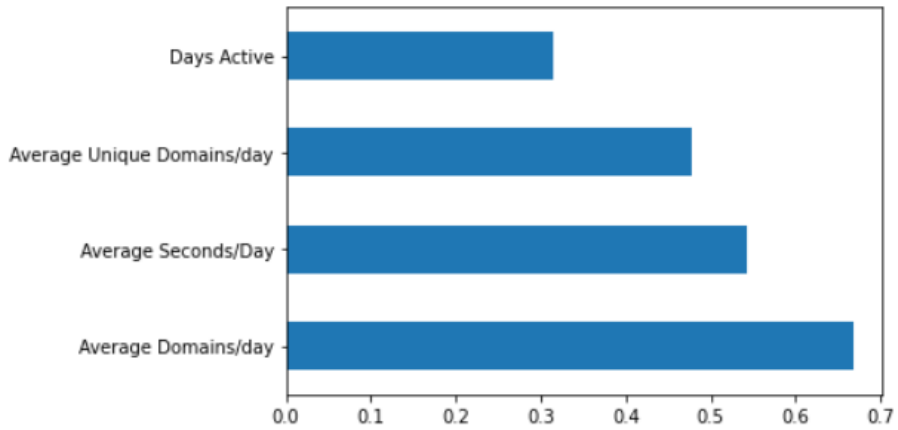
Depression Score: An overview of the results of incremental training/testing for Depression Score is shown in Fig 6.2. The bar plot in 6.2a demonstrates a slight increase in accuracy as the training set gets larger. Additionally, Fig 6.2b plots the most significant features for predicting Depression Score over 100 iterations according to Random Forest. The feature importances are aggregates of the features of the best performing Random Forest models with the largest training sizes; that is, the models which use 24 training data points and predict Depression Score with an accuracy greater than 80%.

Anxiety: An overview of the results of incremental training/testing for Anxiety Score is shown in Fig 6.3. The barplot in 6.5a demonstrates a slight increase in accuracy as the training set gets larger. Additionally, Fig 6.4b plots the most significant features for predicting Anxiety Score over 100 iterations according to Random Forest. The feature importances are aggregates of the features of the best performing Random Forest models with the largest training sizes; that is, the models which use 24 training data points and predict Anxiety Score with an accuracy greater than 80%.

Other Psychometric Scales: Similar patterns to Depression and Anxiety were observed for Sleep Disturbance 6.4 and QOL mean 6.5. In contrast, Work Exhaustion and Professional Fulfillment could not be predicted by regression models at all (-∞ accuracy), and loneliness had fluctuating levels of accuracy among the four training sizes, while almost staying the same (no upwards or downwards trend) B.

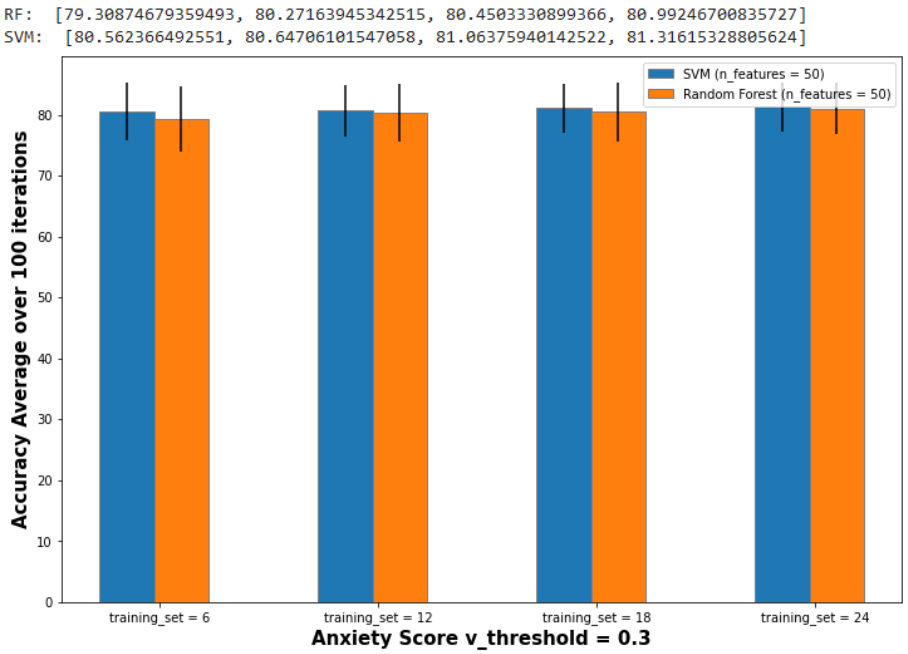


(a) SVR seems to perform better than Random Forest Regressor, while both models achieve slightly higher accuracy with more training data

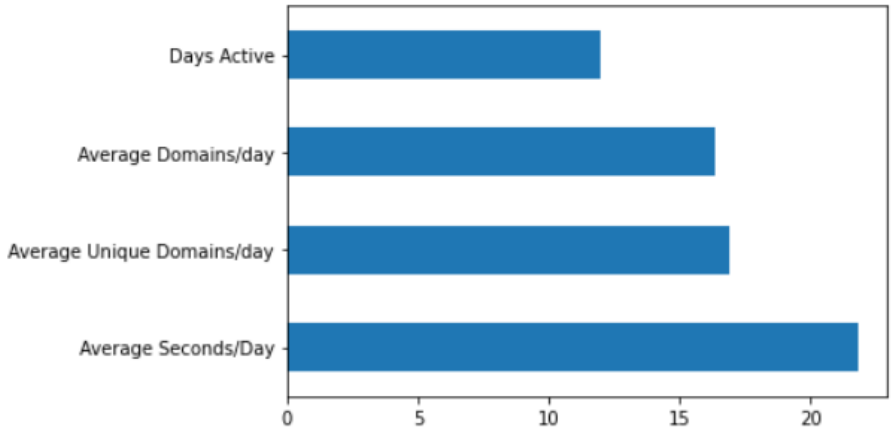


(b) The aggregate weight of the most important features according to several Random Forest Regressors that predict Depression score with accuracy $> 80\%$; all four are generated browsing behavior features.

Figure 6.2: Random Forest and SVR accuracy results and feature importance for Depression Score prediction

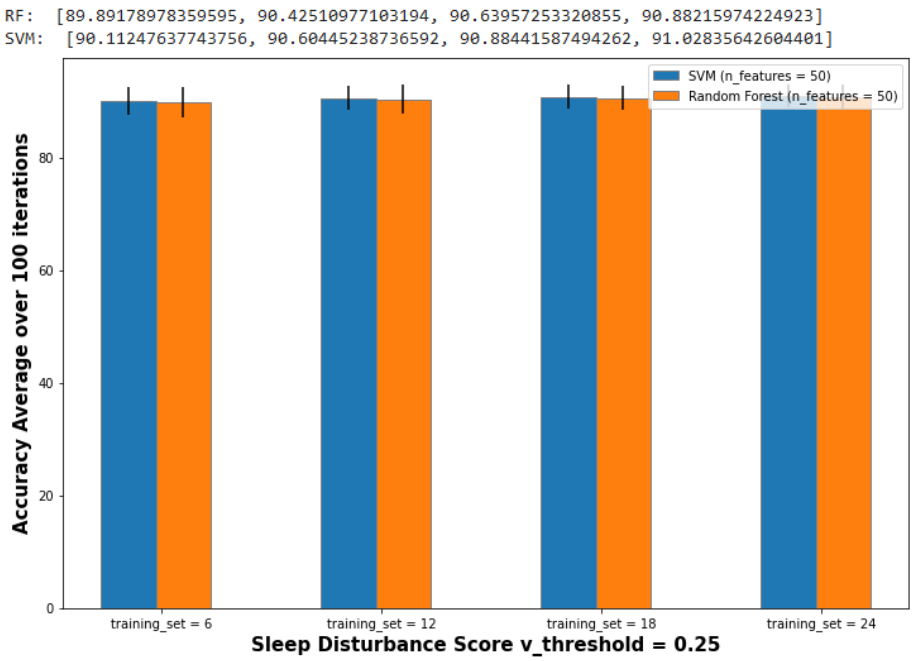


(a) Both SVR and Random Forest Regressor models achieve slightly higher accuracy with more training data

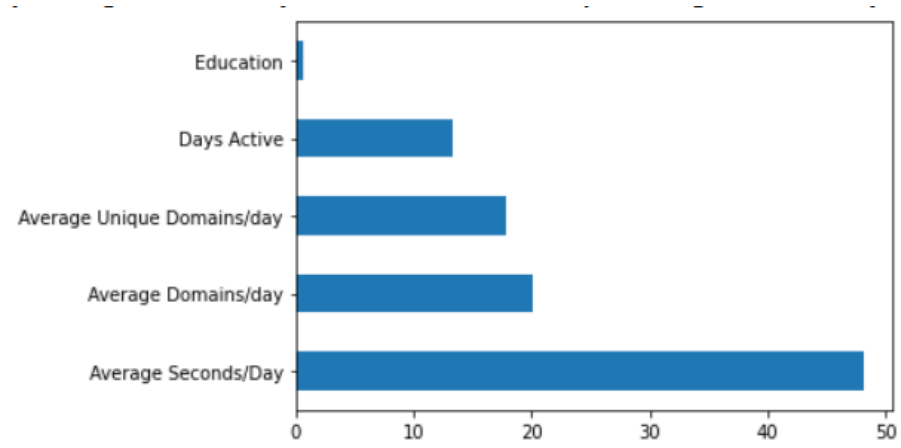


(b) The aggregate weight of the most important features according to several Random Forest Regressors that predict Anxiety score with accuracy $> 80\%$; all four are browsing behavior features.

Figure 6.3: Random Forest and SVR accuracy results and feature importance for Anxiety Score prediction

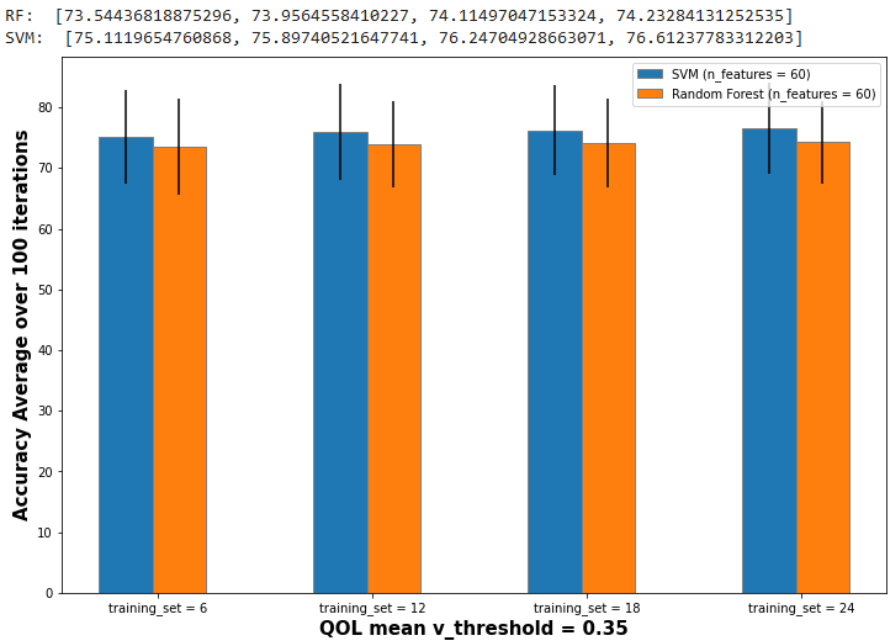


(a) Both SVR and Random Forest Regressor models achieve slightly higher accuracy with more training data

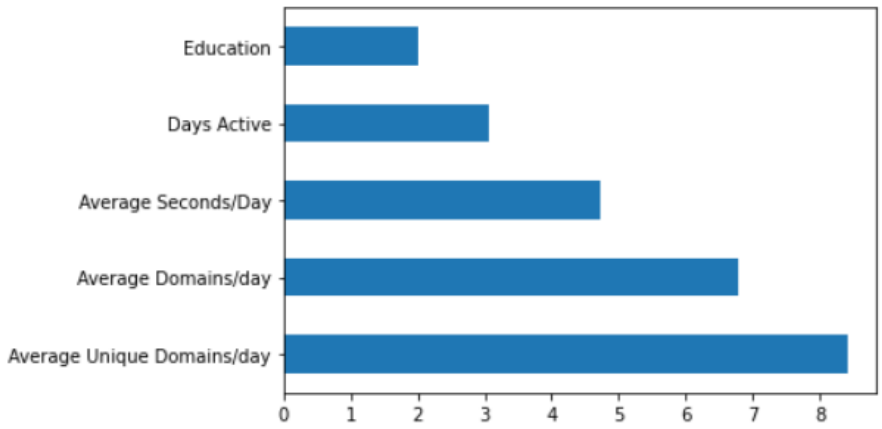


(b) The aggregate weight of the most important features according to several Random Forest Regressors that predict Sleep Disturbance score with accuracy $> 80\%$; Education is included alongside the four browsing behavior features.

Figure 6.4: Random Forest and SVR accuracy results and feature importance for Sleep Disturbance Score prediction



(a) Both SVR and Random Forest Regressor models achieve slightly higher accuracy with more training data



(b) The aggregate weight of the most important features according to several Random Forest Regressors that predict QOL mean score with accuracy $> 80\%$; Education is included alongside the four browsing behavior features.

Figure 6.5: Random Forest and SVR accuracy results and feature importance for QOL mean prediction

6.3 Leave One Out Results

Leave One Out Cross Validation allows uses small data sets to the fullest as it includes each data point in 32 training sets and exactly one test set. Comparing the results of Leave One Out with Incremental Testing might shed some light on why some labels are easier to predict than others, while providing guidance on the issues that need to be addressed before one could apply the approach in this project on a larger dataset.

6.3.1 Leave One Out - Classification

The leave one out classification algorithm was run for all of the psychometric scales, where the binary psychometric classes are being predicted. I did not report the results of SVM to avoid redundancy, as they were almost identical to that of the Random Forest Classifier. The Random Forest models where trains with 350 estimators for all labels, and $k = 30$, $variance = 0.25$. The prediction accuracy of Random Forest Classifier for all binary psychometric labels is shown in table 6.1. The table also contains the number of instances of each of the four key metrics, namely, TP, TN, FP, and FN.

Psychometric Scale	Accuracy	TP	TN	FP	FN
Work Exhaustion	0.63	0	21	0	12
Professional Fulfillment	0.33	0	11	3	19
Depression	0.48	0	16	3	14
Anxiety	0.84	28	0	4	1
Sleep Disturbance	0.24	0	8	2	23
Loneliness	0.48	9	7	9	8
Quality of Life	0.39	0	13	0	20

Table 6.1: Random Forest Binary Classifier prediction evaluation for all psychometric labels

6.3.2 Leave One Out - Regression

For the regression task, Random Forest Regressor was trained with 500 estimators, with $k = 40$ and $variance = 0.3$. Similar to classification, the leave one out

Psychometric Scale	Accuracy	Errors Mean	Errors std
Depression	0.62	38.04	41.96
Anxiety	0.81	18.33	14.59
Sleep Disturbance	0.86	13.5	7.4
Loneliness	0.85	14.57	9.16
Quality of Life	0.76	23.95	18.94

Table 6.2: Random Forest Regressor prediction evaluation for all psychometric scores except Work Exhaustion and Professional Fulfillment

regression algorithm was run for all the psychometric scales excluding Work Exhaustion and Professional Fulfillment (because the regressors are not able to predict them at all). The accuracy results for each data point were calculated with MAPE and then averaged over the 33 data points. I also report the average of the error itself, as well as its standard deviation in table 6.2.

Chapter 7

DISCUSSION & CONCLUSION

In this chapter, I further discuss the results reported in chapter 6, and aim to explain my results with several hypothesis. I also talk about future steps to improve the immediate results, as well as how to modify and enhance the approach so that it can be utilized on a larger data set.

7.1 Promising Labels

Considering the great limitation imposed on the models by the small size of the data set, the accuracy results for the regression tasks seem generally promising; there is an upwards trend in accuracy for 4 out of the 7 mental health scales, with all of the accuracies being greater than 74%. Nevertheless, one should not be over-excited about some of the best results such as Anxiety Score and Sleep Disturbance Score, as the labels show very little variance in some of these cases. For example, 29 out of the 33 data points have "Positive" Anxiety, which means that we are dealing with a very anxious population. With such an imbalanced data set, not a great deal of credit can be granted to the learning abilities of the models.

7.2 Classification vs Regression

The classification and regression results are discussed in detail and compared for both Incremental Testing and Leave One Out testing.

7.2.1 Incremental Training/Testing

As presented in the previous chapter, binary classification only seems promising for Work Exhaustion, while regression accuracy seems to improve for Sleep Disturbance, Anxiety, Depression, and QOL mean. In the context of regression tasks, SVR

outperforms Random Forest for Depression and QOL prediction, but there is no significant difference in the two models' performance for predicting other scales. The superior results of the regression tasks could point to a flaw in the binary classification approach, where 'None' and 'Mild' were used as negative and 'Moderate' and 'High' were used as positive labels. However, there was no escape from compressing the four interpretive classes into binary classes, as the test data only consisted of 9 data points, and it would have been futile to test four classes on such a small test set. It is worth noting that a similar compression of multi-class labels into binary class was adopted in [8]. One immediate solution to this problem is discussed in the future work section 7.5.1. That being said, once the dataset becomes large enough, the four class, well-established interpretations can be used with no modifications, which will likely boost the prediction accuracy.

7.2.2 Leave One Out

One could argue that when using leave one out cross validation for regressing on such a small dataset, the results are sensitive to outliers and high variance labels. For example, if one data point is very different from the other 32 that are being used for training, the prediction is bound to be far from the real value of the test data point. Conversely, if the label is low variance (i.e., the data points are similar to one another), any set of 32 data points have a good chance of predicting the test data, resulting in a closer prediction. With that logic, and by studying table 6.2, it could be inferred that Anxiety and Sleep Disturbance have low variance, which perhaps is not surprising considering the survey took place in the midst of the Coronavirus pandemic. Nevertheless, it emphasises on the point made in promising labels 7.1.

As for Leave One Out with Binary Classification, it is clear from table 6.1 that the models mostly become very one-sided when predicting all of the labels. For example, Anxiety is mostly predicted as positive (TP or FP), while Sleep Disturbance is mostly predicted as negative (TN or FN). That could also point to the limitations of a small data set, as well as the prevalence of imbalance within the labels.

If we were to compare the results of Leave One Out Cross Validation with those of [8], they seem to outperform this work regarding the binary classification accuracies (they predict Internet Addiction Assessment and General Health Questionnaire results), with Anxiety being the only exception. Considering the range of precision, recall and F1-measures reported in [8], the IAD project seems to benefit from a more balanced dataset, which made it possible for non-accuracy metrics to be reported. This also further suggests that the highly imbalanced labels (i.e., Anxiety) are not a good measure of the performance of the model. Nevertheless, the authors also state that for the 10-fold cross validation, 70% of the data is used for training and the remaining 30% of the data is used for testing. This implies that the data points will be included in the test set more than once, further improving performance compared to a more common 90%-10% split.

7.3 Feature Importance

Looking at the important features reported for the Random Forest models in chapter 6, It is clear that all of the models heavily depend on Average Seconds per day, Average Domains per day, and Average Unique Domains per day to make their predictions. This means that the these three browsing behavior features passed the chi-square correlation test for all of the labels. Therefore, it is reasonable to claim that browsing behavior and mental health are correlated, as already noted in the literature [8][30], and it is recommended to continue engineering relevant and in depth features with browsing data.

On the other hand, the fact that the browsing behavior features dominated the models could also point to a problem with the other features. Since the variance threshold method is used as a filtering step in the feature selection method, one could argue that the web browsing features just have much higher variance than the others. This observation is, in fact, true; As shown in Fig 6.1b, when the variance threshold is lowered, other types of features are also picked up by the feature selector and ultimately Random Forest feature importance. Nevertheless, the browsing behavior

features still carry more weight than their demographic counterparts according to Fig 6.1b. Nonetheless, the web content feature seem to perform poorly compared to the browsing behavior features; this could be a result of using over 40 content categories, as apposed to 5 content categories used in [8]. Too many categories might have led to losing variance in the content features, preventing them from being picked up by the Feature Selector.

7.4 Limitations

In this section, I discuss why I chose to report certain evaluation metrics over others, and the reason behind the relatively low number of browsing behavior features.

7.4.1 Performance Metrics

For performance evaluation, Mean Absolute Percentage Error 5.2 was used for the Regression models, and the the number of true predictions over all predictions 5.1 was used for binary classification. However, in the case of binary classification, accuracy alone can not capture the level of bias and balance of the model; usually, precision (i.e., are the records predicted as true actually true?) and recall (i.e., did we miss a lot of records that should have been predicted as true?), and their aggregation, F1 measure are utilized for more rigorous testing. However, given that testing was done on only 9 data points, precision and recall resulted in zero denominator fractions in a lot of the cases, and hence, could not be reported.

7.4.2 Feature Engineering

Several useful features such as the maximum number of tabs open were being considered in the Feature Engineering stage, but were ultimately set aside. It was not possible to compute the maximum number of tabs accurately by solely relying on the 'Synced Seconds per Domain' data (which is the only data set provided), and any method for computing it would have been a guess. This could potentially be fixed with additional access to other collections in HabitLab's MongoDB database. Another example is the time in which each participant starts browsing every day, which was

another features that was not possible to compute with sufficient precision. That said, additional feature engineering would likely prove valuable.

7.5 Future Work

There are two approaches to continuing this work: 1) aiming to utilize the current data set to the fullest by generating more complex and relevant features, and 2) collecting more data, which would allow for creating more balanced training and test sets, ultimately leading to better accuracy results.

7.5.1 Modified Binary Classes

One approach that could be suggested for improving the accuracy of the binary classification models is to not set the separating threshold exactly in the middle of the four classes; meaning, one could aim to predict High Depression, and use 'High' as the positive class, while grouping the other three classes as negative. With the same pattern, it is easy to see that for each label with four interpretations, there are three possible separators which could be tested to potentially create more balanced datasets.

7.5.2 Advanced Feature Engineering

The current web browsing features discussed in 4.1 can be significantly improved upon by using Autoencoders [29]. In contrast with the current emphasis on "Averaging", an autoencoder will capture the nuances of the data and will more accurately distinguish data points from one another. On a different note, for each label, features can be generated that more closely correlate with that label. For example, given that we are able to create the browsing start time feature, that would be a great indicator of depression, as depressed individuals tend to have unusual sleeping habits [35].

7.5.3 Better Data

After observing the feature importance results, It is clear that the models rely heavily on the web browsing features 4.1. This could mean that the other features simply do not meet the variance threshold. In that case, collecting more data will

introduce more variance into several features such as the web content features 4.2, which could potentially make them much more useful for the models. For example, the proportion of browsing chat and messaging domains could potentially be very relevant to the loneliness scale. But at the moment the standard deviation for that feature is only 0.07. This could possibly change with more browsing data.

7.5.3.1 Maximize Study Window

In order to have a comparable set of browsing history for all participants, every user was limited to 15 days worth of browsing history, which was imposed by the participant with the lowest number of days in the system. Some users, however, had browsing data that went back over a year. Therefore, the study window can be maximized by looking at the second minimum, third minimum, and so on. This is an optimization problem where the trade-off is between the number of data points in the data set (i.e., participants), and the size of the study window. In this project, maximizing the number of participants was favored, but it was not necessarily the best possible approach.

7.5.3.2 Maximize the number of Participants

Even with the Qualtrics Survey data and Browsing history at hand, there is the possibility to free ourselves from the limitation of 33 data points.

Discard Labels: As discussed previously, the participants are chosen based on having complete demographics and psychometric records. However, when the model aims to predict Anxiety, for example, it does so without taking any other psychometric scales into account. So there is no need to have complete records for all of the psychometric scales when the model is predicting only one of them. With that, we can get varying sizes for our data set depending on which label is being predicted. For instance, by discarding Loneliness and Sleep Disturbance labels, 14 additional data points can be added to the data set to be used for predicting the other labels such as Depression, Anxiety, QOL mean, etc.

Request More Data: HabitLab recently ran another Qualtrics survey with more participants, possibly a result of better incentives and lower number of questions. The same methods in this project could be applied to the new, larger data set to investigate whether accuracy is actually increasing with more training data. That said, requesting new data potentially raises issues with temporality (similar to those described in [33][34]) which would also need to be explored.

7.6 Conclusion

Based on the results of the Machine Learning Models, both in terms of accuracy and feature importance, I can conclude that continuing to explore web feature engineering in the context of HabitLab browsing histories in an attempt to predict self-reported mental health markers is a worthwhile effort. Regression tasks seem to be more promising, as they provided positive results for Depression, Anxiety, Sleep Disturbance, and Quality of Life, whereas binary classification only seems fruitful for Work Exhaustion. This could point to the faulty method used for separating "Positive" and "Negative" classes. The more positive classification results discussed in [8] also point to the same conclusion. Nevertheless, increasing data points will inevitably provide better insights into the usefulness of this approach, be it more participants which would result in a larger data set, or a longer study window which would result in more reliable web browsing and content features. More data points will also mean larger test sets, which would eliminate the need for breaking the labels into two classes, so that multi-class classification could be performed for the four well-established categories in the mental health literature. Furthermore, the current regression models can be used to soft label the rest of the HabitLab population, accelerating the development and testing of just in time interventions, keeping in mind that the models can always benefit from more data (hard labels). However, the labeling process should account for temporal drift, as this study reflects the mental health status of HabitLab users in the midst of a pandemic, and the results will not necessarily translate to more normal times. [33][34]

REFERENCES

[1] Rajkumar, R.P. "COVID-19 and mental health: A review of the existing literature" *Asian Journal of Psychiatry*, 2020.

[2] Lavoie, J. A., and Pychyl, T. A. "Cyberslacking and the procrastination superhighway: A web-based survey of online procrastination, attitudes, and emotion" *Social Science Computer Review*, 2001.

[3] Garcia-Navarro, L. (2020, Your "'Doomscrolling' Breeds Anxiety. Here's How To Stop The Cycle" *NPR*, 2020.

[4] Twenge, J. M., Joiner, T. E., Rogers, M. L., and Martin, G. N. "Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased new media screen time" *Clinical Psychological Science*, 2018.

[5] Kovacs, G., Wu, Z., and Bernstein., M. S. "Not Now, Ask Later: Users Weaken Their Behavior Change Regimen Over Time, But Expect To Re-Strengthen It Imminently". *In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery*, 2021

[6] Kovacs, G., Wu, Z., and Bernstein, M. S. "Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition"., *Proc. ACM Hum.-Comput. Interact*, 2018

[7] Paredes, P. E., Goel, Rahul and Mauriello, M. L., "SWEET - Towards a Digital Wellbeing and Occupational Health Platform in the Age of the COVID-19 Pandemic", 2020

[8] Purwandari, B., Wibawa, W. S., Fitriah, N., Christia, M., and Bintari, D. R. "Internet Addiction and Mental Health Prediction Using Ensemble Learning Based on Web Browsing History". *In Proceedings of the 3rd International Conference on Software Engineering and Information Management (ICSIM '20)*. Association for Computing Machinery, 2020.

[9] Deznabi, I., Motahar, T., Sarvghad, A. Fiterau, M., and Mahyar, N. "Impact of the COVID-19 Pandemic on the Academic Community Results from a survey conducted at University of Massachusetts Amherst"., *Digit. Gov.: Res. Pract.*, 2021

[10] Javed, B., Sarwer, A., Soto, E. B., and Mashwani, Z. U. "The coronavirus (COVID-19) pandemic's impact on mental health.", *The International journal of health planning and management*, 2020

[11] Manchia, M., Gathier, A. W., Yapici-Eser, H., Schmidt, M. V., de Quervain, D., van Amelsvoort, T., Bisson, J. I., Cryan, J. F., Howes, O. D., Pinto, L., van der Wee, N. J., Domschke, K., Branchi, I., and Vinkers, C. H. "The impact of the prolonged COVID-19 pandemic on stress resilience and mental health: A critical review across waves"., *Eur Neuropsychopharmacol*, 2022

[12] https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/

[13] Zaman, A., Zhang, B., Silenzio, V., Hoque, E. and Kautz, H. A. "Estimating Anxiety based on individual level engagements on YouTube and Google Search Engine"., *coRR*, 2020

[14] Nie, D., Ning, Y., and Zhu, T. "Predicting Mental Health Status in the Context of Web Browsing"., *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012

[15] R.Derogatis, L. and Unger, R., "Symptom Checklist-90-Revised", *The Corsini Encyclopedia of Psychology*, 2010

[16] Drucker, H., Chris, Kaufman, B. L., Smola, A., and Vapnik, V., "Support vector regression machines"., *Advances in Neural Information Processing Systems 9*, 1997.

[17] Chen, C, and Breiman, L., "Using Random Forest to Learn Imbalanced Data"., *University of California, Berkeley*, 2004

[18] American Psychiatric Association. *Desk reference to the diagnostic criteria from DSM-5 (R)*. American Psychiatric Association Publishing, 2013.

[19] Russell, D., Peplau, L.A., and Cutrona, C.E. "The revised UCLA Loneliness Scale: Concurrent and discriminant validity evidence". *Journal of Personality and Social Psychology*, 1980.

[20] Cohen, S., Kamarck, T., and Mermelstein, R. "A global measure of perceived stress". *Journal of Health and Social Behavior*, 1983.

[21] Trockel, M., Bohman, B., Lesure, E., Hamidi, M. S., Welle, D., Roberts, L., Shanafelt, T. "A Brief Instrument to Assess Both Burnout and Professional Fulfillment in Physicians: Reliability and Validity, Including Correlation with Self-Reported Medical Errors, in a Sample of Resident and Practicing Physicians". *Acad Psychiatry*, 2018.

[22] World Health Organization. *Programme on mental health : WHOQOL user manual*. World Health Organization ,2012.

[23] Mullins, L. C. *In Encyclopedia of Gerontology (Second Edition)*. 2007.

[24] Hawthorne, G., Herrman, H., and Murphy, B. "Interpreting the WHOQOL-Bref: Preliminary Population Norms and Effect Sizes" *Social Indicators Research*, 2007.

[25] Yayan, E. H., Arikan, D., Saban, F., Gürarslan Baş, N., Özel Özcan, Ö., "Examination of the Correlation Between Internet Addiction and Social Phobia in Adolescents". *Western Journal of Nursing Research*, 2017

[26] Shensa, A., Sidani, J. E., Dew, M. A., Escobar-Viera, C. G., and Primack, B. A., "Social Media Use and Depression and Anxiety Symptoms: A Cluster Analysis". *American journal of health behavior*, 2018

[27] Hohls, J. K., König, H. H., Quirke, E., and Hajek, A. "Anxiety, Depression and Quality of Life-A Systematic Review of Evidence from Longitudinal Observational Studies". *International journal of environmental research and public health*, 2021.

[28] 'AI Powered Website Categorization', https://www.dnsfilter.com/webshrinker

[29] Liou, C. Y., Cheng, W. C., Liou, J. W., and Liou, D. R. "Autoencoder for words". *Neurocomputing*, 2014.

[30] Ang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., and Campbell, A. T. "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones", *Association for Computing Machinery*, 2014

[31] https://docs.webshrinker.com/v3/web-shrinker-categories.html#categories

[32] Wang, J., Chen, Q., Chen, Y. "RBF Kernel Based Support Vector Machine with Universal Approximation and Its Application"., 2004

[33] Mauriello, M. L., Lincoln, T., Hon, G., Simon, D., Jurafsky, D., and Paredes, P., "SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems"., *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021

[34] Mauriello, M. L., Tantivasadakarn, N., Mora-Mendoza, M. A., Lincoln, E. T., Hon, G., Nowruzi, P., Simon, D., Hansen, L., Goenawan, N. H., Kim, J., and others., "A Suite of Mobile Conversational Agents for Daily Stress Management (Popbots): Mixed Methods Exploratory Study"., *JMIR formative research*, 2021

[35] Nutt, D., Wilson, S., Paterson, L. "Sleep disorders as core symptoms of depression", *Dialogues Clin Neuroscience*, 2008

Appendix A

EXPLORATORY DATA ANALYSIS OF QUALTRICS SURVEY

I conducted EDA on the Qualtrics data to explore the correlation between demographic features and the psychometric scales and I used scatter plots and barplots to visualize the results. Even though no significant correlations were found, the following plots demonstrate some of the interesting observations:

- According to Fig A.1, Hispanic and Asian participants seem to report higher levels of Depression than African-American and Caucasian respondents.
- According to Fig A.2, female respondents reported slightly higher anxiety levels than males.
- The survey responses demonstrate the established linear relationship between Depression and Anxiety, as is shown in the scatter plot in Fig A.3.
- The scatter plot in Fig A.4 depicts that the average QOL score decreases as the participants report higher levels of depression, which also matches the findings in the mental health literature [27].
- The scatter plot in Fig A.5 confirms the established positive correlation between Anxiety and Sleep Disturbance.
- The scatter plot in Fig A.6 depicts that higher intensity of Anxiety co-occurs with feeling increased levels of loneliness.

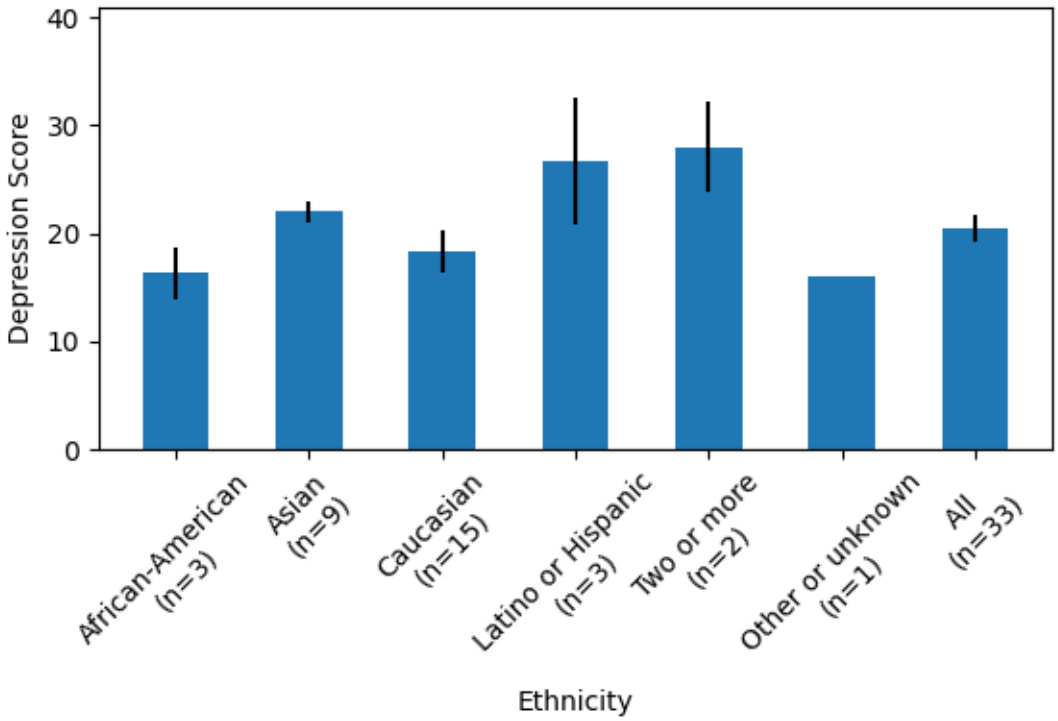


Figure A.1: Comparing Depression Scores among different Ethnicity groups; Hispanic/Latino and Asian participants seem to report higher levels of Depression than African-American and Caucasian participants

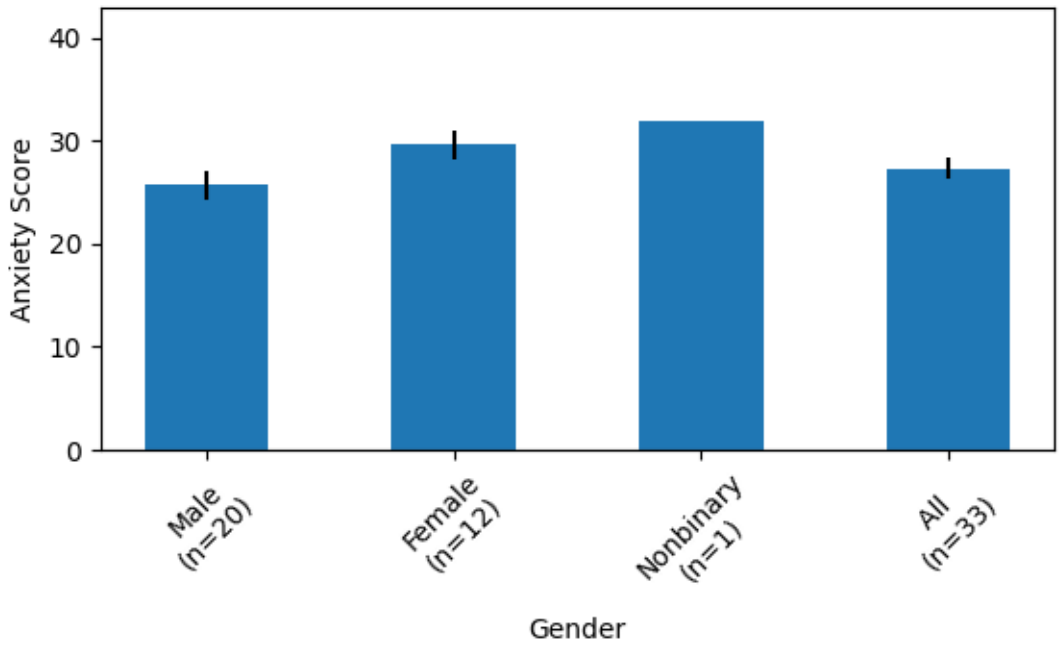


Figure A.2: Comparing Anxiety Scores across Genders; Female respondents report slightly higher anxiety levels than Males

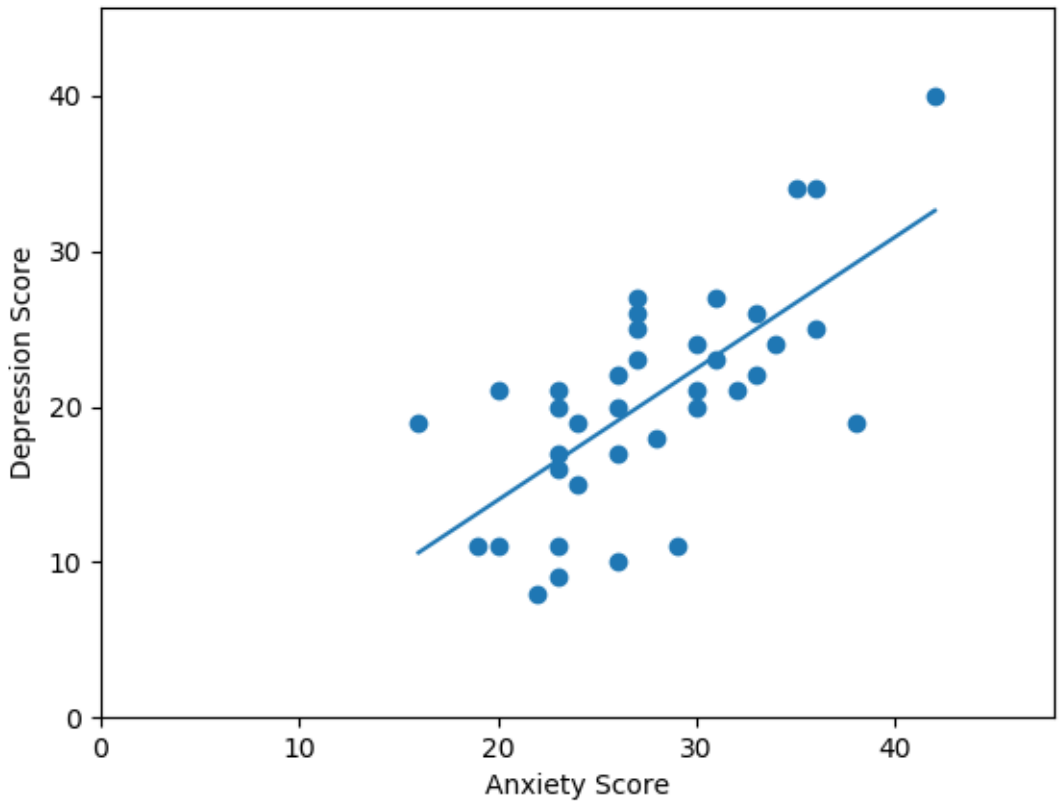


Figure A.3: Positive linear correlation between Depression and Anxiety scores

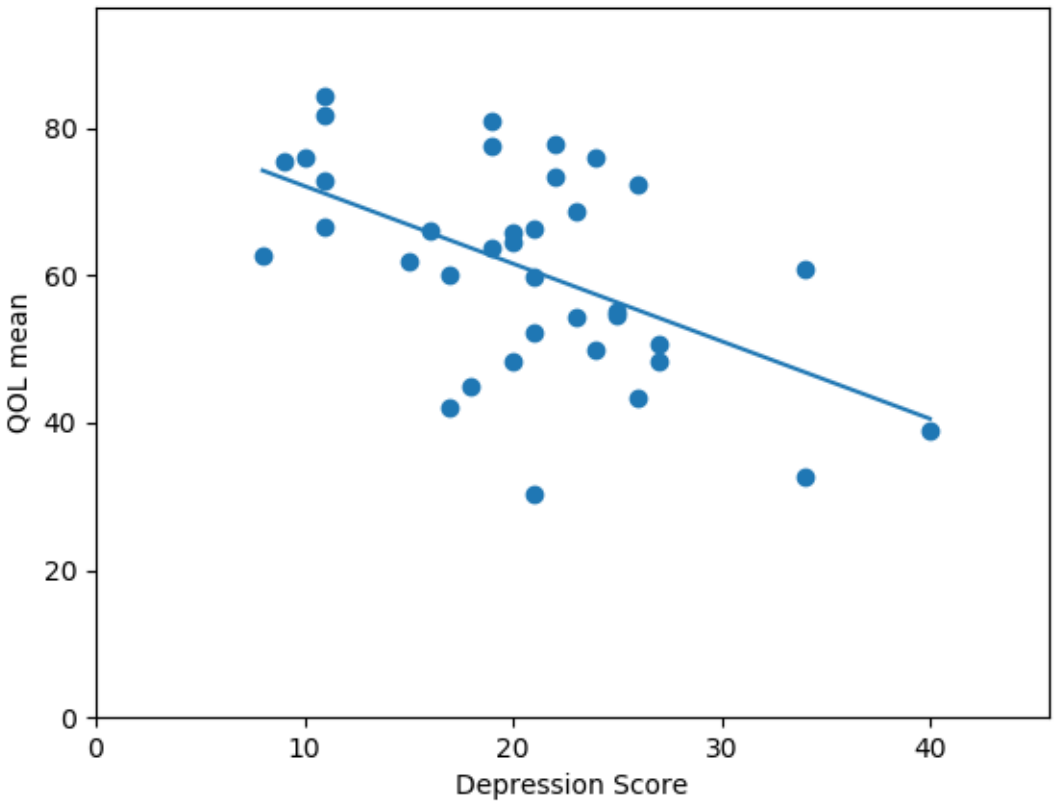


Figure A.4: Inverse linear correlation between Depression score and QOL mean

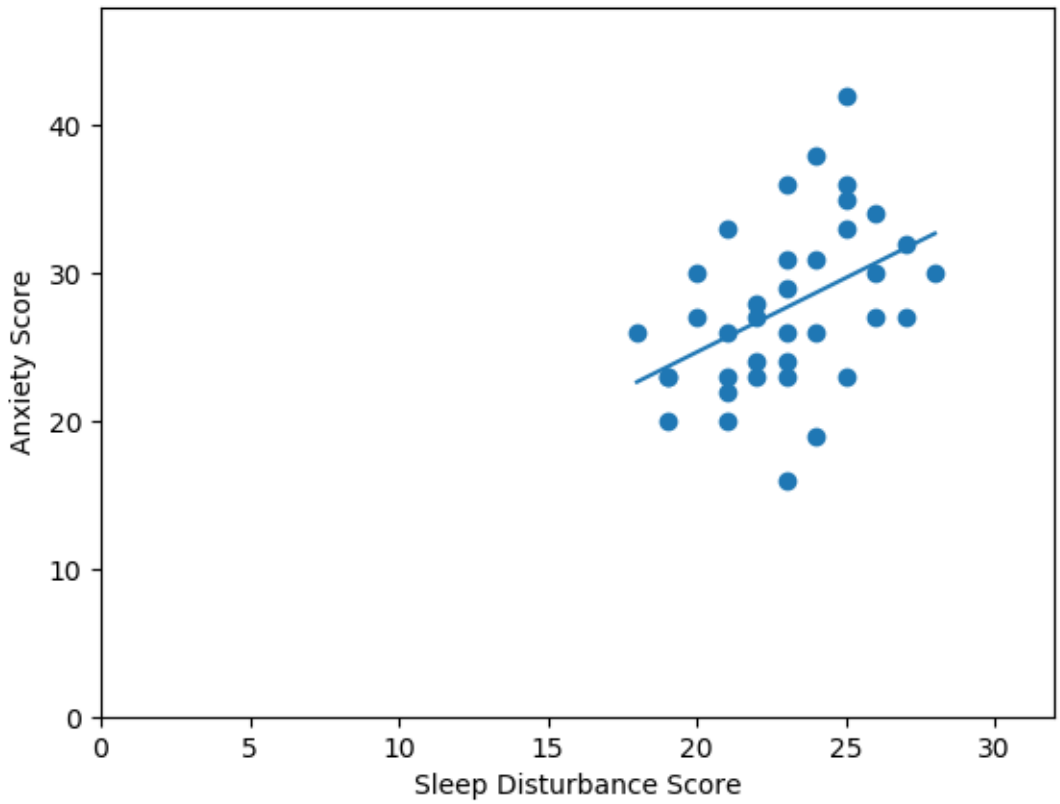


Figure A.5: Positive linear correlation between Anxiety score and Sleep Disturbance Score

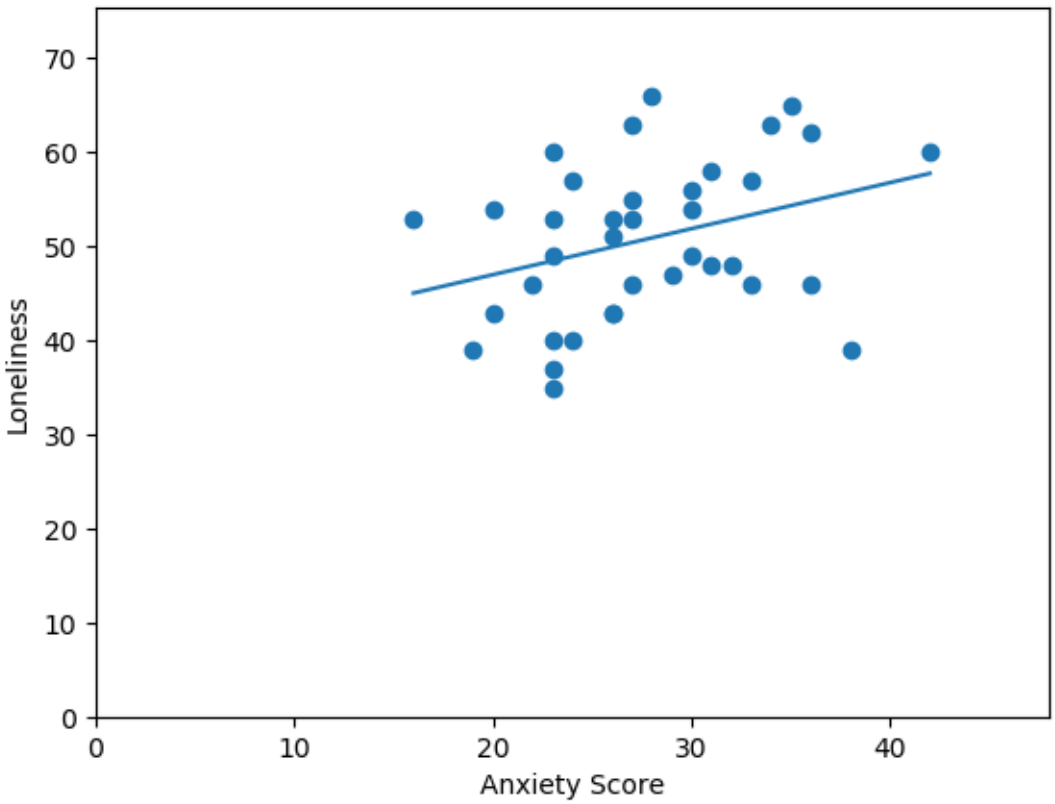


Figure A.6: Positive linear correlation between Anxiety score and Loneliness

Appendix B

INCREMENTAL TESTING NEGATIVE RESULTS

B.1 Binary Classification

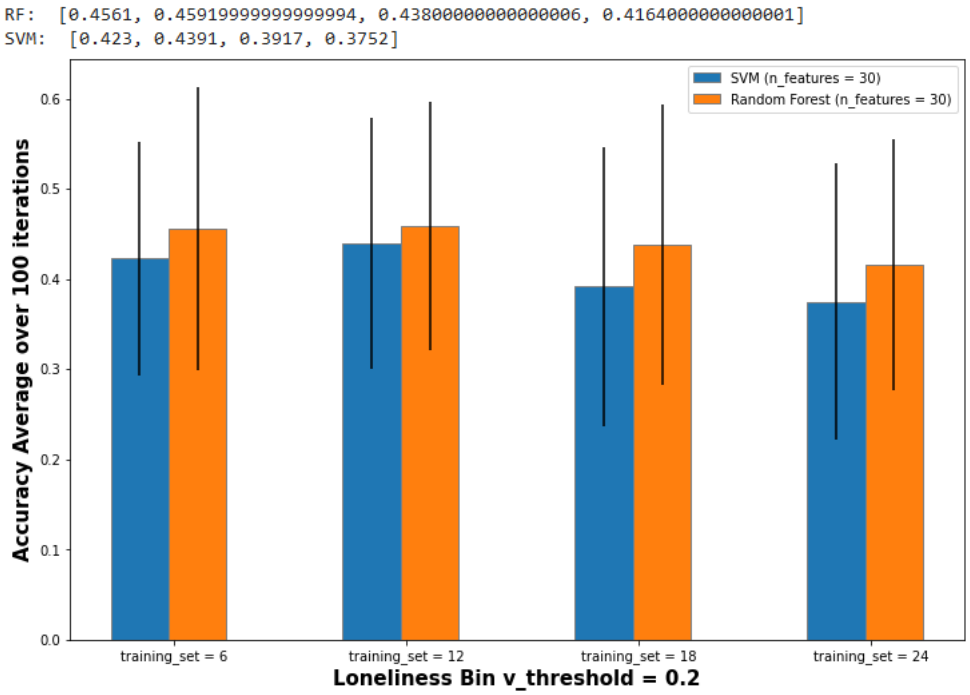


Figure B.1: The prediction accuracy for Loneliness Binary Class decreases as the training size gets larger

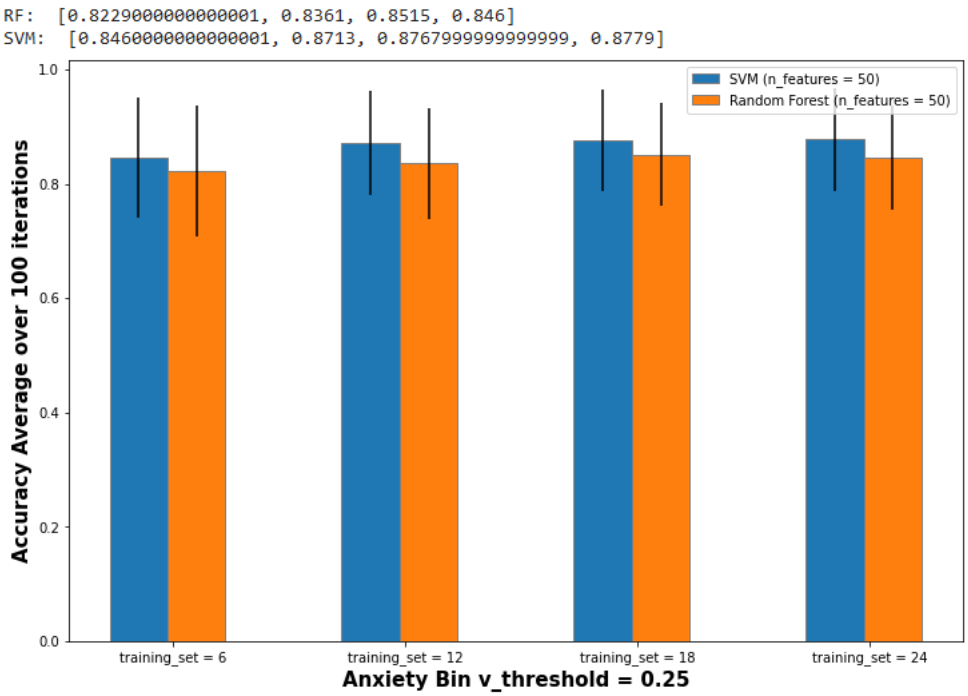


Figure B.2: There is a slight increase in the prediction accuracy for Anxiety Binary class as the training set gets larger. However, the binary class is extremely imbalanced, as shown in table 5.1

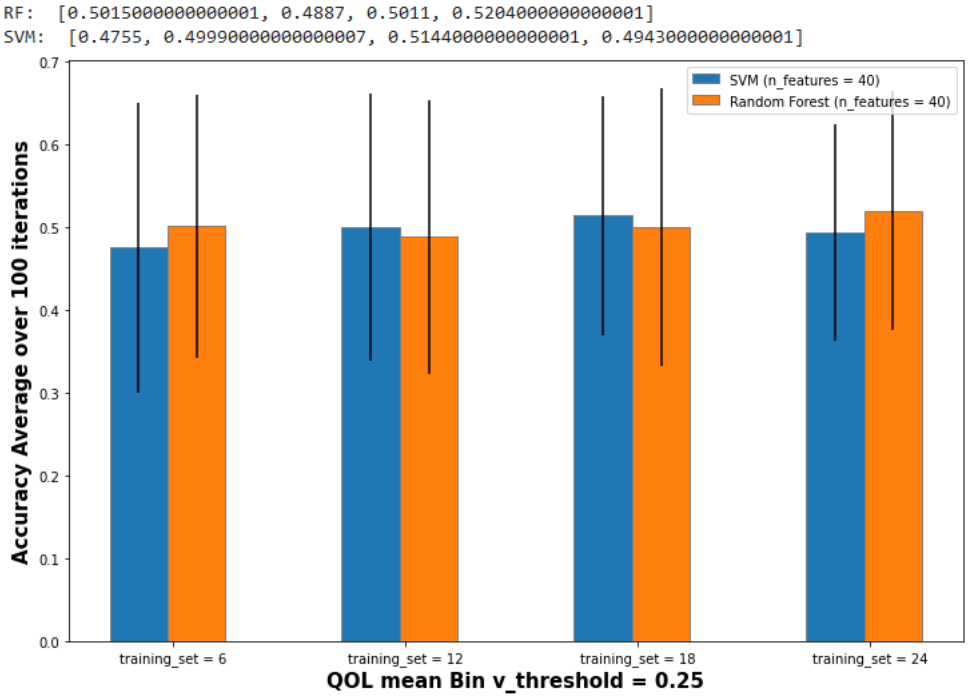


Figure B.3: The increase in accuracy is not consistent among the four training sets for Quality of Life mean Binary Class (observe the difference between train set = 18 and train set = 24 for RF)

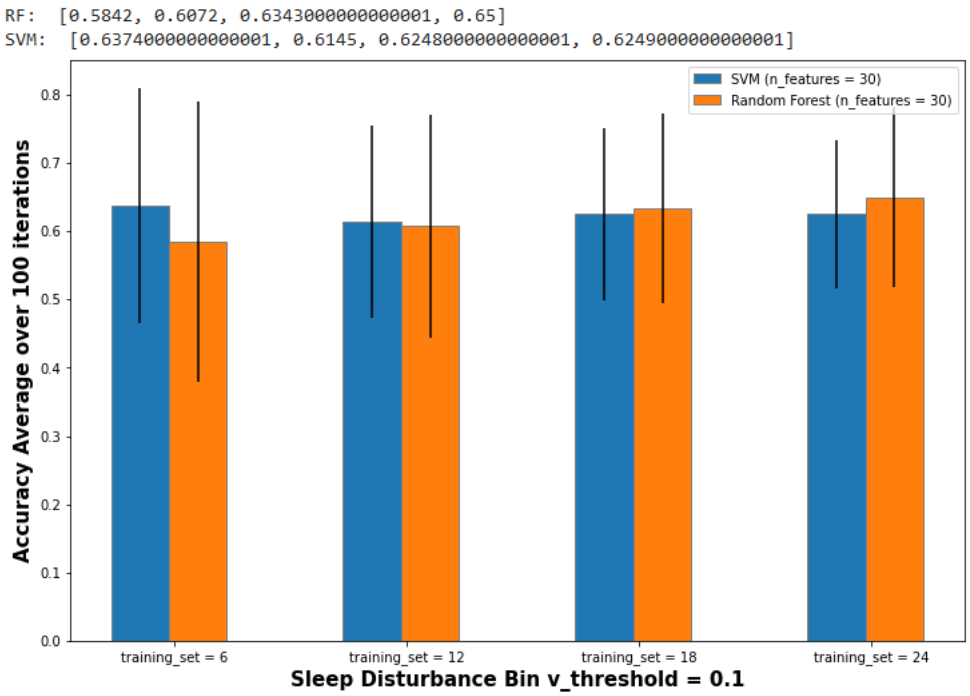


Figure B.4: The increase in accuracy is not consistent among the four training sets for Sleep Disturbance Binary Class (observe the difference between train set = 6 and train set = 12 for RF)

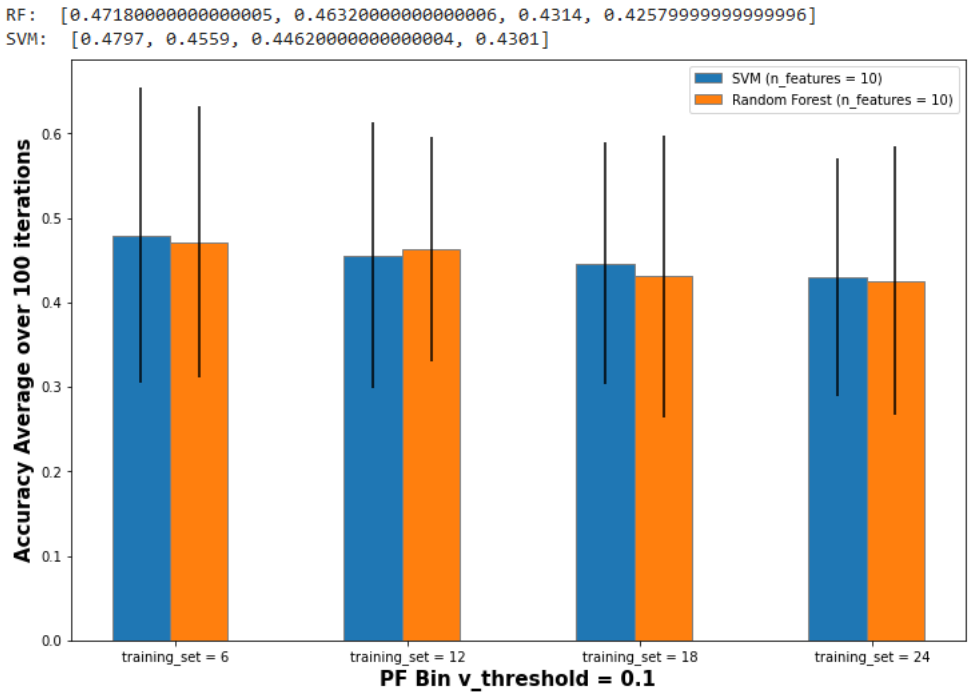


Figure B.5: The prediction accuracy for Professional Fulfillment Binary Class decreases as the training size gets larger

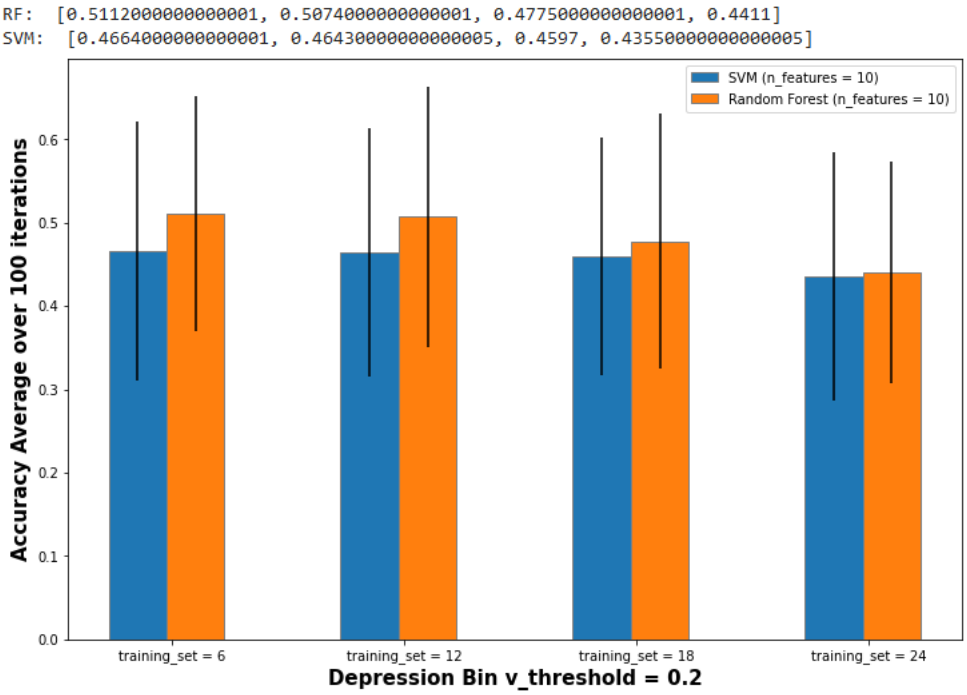


Figure B.6: The prediction accuracy for Depression Binary Class decreases as the training size gets larger

B.2 Regression

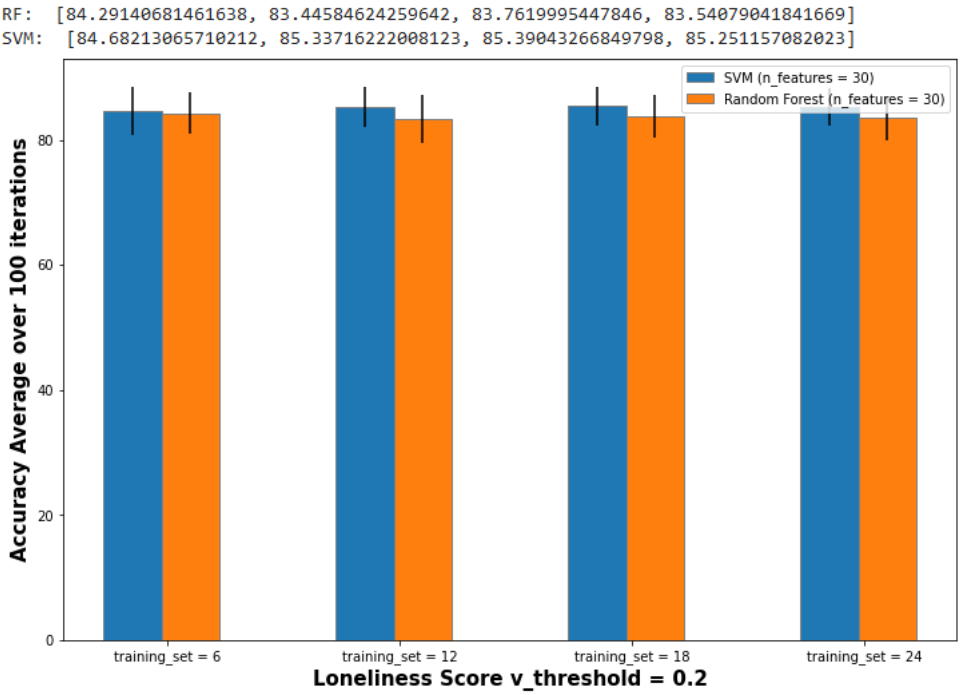


Figure B.7: The prediction accuracy for Loneliness Score does not seem to change in a meaningful way when the training data gets larger