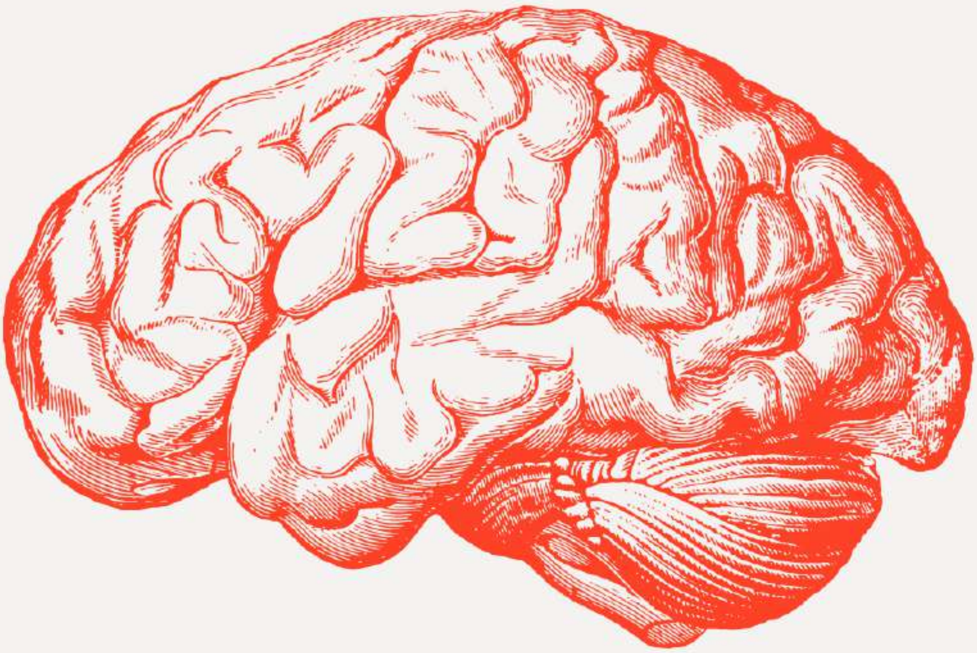
"'CRACKING THE PERSONALITY CODE' BRILLIANTLY BRIDGES TECHNOLOGY AND SELF-DISCOVERY, REVEALING HOW OUR SMARTPHONES ECHO OUR TRUE SELVES. A PIONEERING INSIGHT INTO THE DIGITAL REFLECTION OF OUR PERSONALITIES." **-THE YOUTH PERCEPTION**

Cracking the Personality Code:

A New Frontier In Personality Prediction



SWAVIMAN KUMAR,
IVANO BISON,
BRUNO LEPRI,
FAUSTO GIUNCHIGLIA

UNIVERSITÀ
DI TRENTO

Department of Sociology and Social Research

Master's Degree in
Data Science

FINAL DISSERTATION

CRACKING THE PERSONALITY CODE: A NEW FRONTIER IN PERSONALITY PREDICTION

Supervisors

Ivano Bison

Bruno Lepri

Fausto Giunchiglia

Student

Swaviman Kumar

Academic year 2022/2023

Permission To Use

In presenting this thesis as part of the requirements for obtaining a Master's degree from the University of Trento, I hereby grant permission for the University's Libraries to make it available for inspection. Furthermore, I consent to the potential copying of this thesis, either in its entirety or in part, for scholarly purposes. Such permission may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department in which my thesis research was conducted.

I understand and acknowledge that any reproduction or publication of this thesis or its components for commercial gain is strictly prohibited without my written consent. I also expect that proper acknowledgment and credit will be given to both me and the University of Trento in any scholarly use of the material from my thesis.

For inquiries or requests related to the reproduction or use of material from this thesis, whether in whole or in part, please direct your correspondence to the Head of the Department of Sociology and Social Research at the following address:

Head of the Department of Sociology and Social Research
via Verdi, 26
I-38122 Trento

Acknowledgments

I extend my gratitude to Dr. Ivano Bison, my supervisor, for giving me the opportunity to pursue this work, providing guidance along the way, and fostering an inspiring and supportive environment that made this journey both meaningful and enjoyable. I am grateful to Dr. Bruno Lepri and Dr. Fausto Giunchiglia, who agreed to co-supervise this work.
My heartfelt appreciation goes out to my parents for supporting my decision to pursue a master's degree. I am deeply grateful to my friends, whose companionship made this journey truly unforgettable.

I am grateful to all the benevolent individuals whose selfless acts of kindness, whether through their advice, shared wisdom, or a word of encouragement, have made a profound impact on my journey.

Lastly, I would like to express my gratitude to all the professors and colleagues who played a part in this academic endeavor. Thank you all for your contributions.

Contents

Abstract 3

1 Introduction 5

2 Literature Review 9
2.1 Evolution of Personality Concepts 9
2.2 Personality Traits and the Big Five Model 10
2.3 Smartphone data and Personality traits 10
2.4 Rationale for Our Study 12

3 Background 13
3.1 Supervised Machine Learning 13
3.1.1 Ordinary Least Square Regression 13
3.1.2 LASSO Regression 13
3.1.3 Ridge Regression 14
3.1.4 Random Forest 14
3.1.5 XGBoost 15
3.1.6 Support Vector Machine 16
3.1.7 K-Nearest Neighbour 17
3.1.8 Decision Tree 18
3.2 Baseline Models 19
3.2.1 Mean Model 19
3.2.2 ZeroR Model 19
3.3 Model Evaluation 19
3.3.1 Cross Validation 19
3.3.2 RMSE 20
3.3.3 Accuracy 20

4 Methods 21
4.1 Study Design and Data Collection 21
4.2 Ethics and privacy 22
4.3 The Sensor Data 23
4.4 Feature Extraction 24
4.4.1 Big Five Survey 24
4.4.2 App Usage Logs 25
4.4.3 Bluetooth 26
4.4.4 Ring Mode 28
4.4.5 Step Counter 30
4.4.6 Touch Event 30
4.4.7 Music Event 31
4.4.8 Screen Event 32
4.4.9 Battery Charge Event 33
4.4.10 Doze Event 34
4.5 Exploratory Data Analysis 36

4.5.1 Univariate Analysis . . . 36
4.5.2 Multivariate Analysis . . . 37
4.6 Prediction Analysis . . . 53
4.6.1 Regression Models . . . 54
4.6.2 Classification Models . . . 57

5 Results **61**
5.1 Regression Results . . . 61
5.1.1 Statistical Analysis (RMSE) . . . 61
5.1.2 Fit Line Plots Analysis . . . 61
5.1.3 Residual Plots Analysis . . . 61
5.2 Classification Results . . . 70
5.2.1 Statistical Analysis (Accuracy) . . . 70

6 Discussion and Future Work **79**
6.1 Summary . . . 81

Bibliography **83**

Abstract

Human behavior is inherently intricate, often challenging to explain through traditional mathematical models. To simplify this complexity, researchers frequently develop intermediate psychological models that capture specific facets of human behavior. These intermediate models, often derived from personality assessments, undergo validation using established survey instruments and tend to correlate with observable behaviors. Typically, these constructs are utilized to predict specific, standardized aspects of behavior.

The advent of novel sensing systems has ushered in an era of remarkably precise behavior tracking, raising the intriguing question of whether the reverse process is feasible: Can we deduce psychological constructs for individuals from their behavioral data? Modern smartphones are equipped with an array of sensors capable of capturing, filtering, combining, and analyzing data to generate abstract measures of human behavior. The ability to extract personal profiles or personality types directly from mobile phone data, without requiring participant interaction, holds potential applications in marketing, as well as in the initiation of social or health interventions.

In this study, our aim is to model a well-established personality inventory—the Big Five framework [17]. Activities of students were observed over a 2 months period using parameters readily available from the smartphone sensors of participants. Correlation analyses were performed and Supervised machine learning algorithms were implemented along with cross validation to make predictions about their personality traits using smartphone sensor data. The study illustrated that the root mean squared error was of a magnitude that allows for actionable predictions regarding an individual's personality based on smartphone data.

1 Introduction

In today's age of technology, smartphones have become an integral part of our daily lives. We use them not just for communication but also for entertainment, education, productivity, navigation and so on. What many don't realize however is that these devices are also capable of collecting vast amounts of data about us, even when we are not actively using them. With more than a dozen sensors housed within them, smartphones can track our movements, measure our physiological responses, and even monitor our environment.

This wealth of data presents a unique opportunity to gain insights into human behavior and personality traits. In particular, it offers the potential to measure and predict personality, a crucial aspect of human psychology that has long been studied only through self-reported questionnaires. Traditional ways of measuring personality through questionnaires have limitations [21]. They require a huge amount of time, resources and effort to administer the tests and there still remains the potential for bias. Such data collection can be biased due to social desirability bias or the individual's own lack of self-awareness. Hence, smartphones provide a new approach for researchers to measure personality in a more objective and passive manner leveraging vast amounts of sensor generated data [36].

Some of the conventional ways of understanding personality traits are using the Myers-Briggs Type Indicator (MBTI) [32], Big Five Personality Inventory [17] or NEO personality inventory [9]. The MBTI approach was based on Carl Jung's theory of personality types and includes 16 different personality types based on four dichotomies: Extraversion vs Introversion, Sensing vs. Intuition, Thinking vs. Feeling and Judging vs. Perceiving. This test was used to help individuals understand their own preferences and how they interacted with others, and could be used in personal development, team building, and leadership training. The NEO personality inventory measured an individual's personality across five dimensions such as openness, conscientiousness, extraversion, agreeableness, and neuroticism. The NEO-PI consists of 240 questions. Later a more compact approach gained popularity which was also based on the same five factor model (FFM) and was named The Big Five Inventory (BFI). It used only 44 questions instead of 240 questions in the NEO-PI approach. Another trade off between the two approaches is that the NEO-PI measures more specific facets within each dimension, whereas BFI assesses only the basic level of each dimension. Though the Big Five framework comes with its own flaws and has been subjected to criticisms on several instances for its inability to capture overall behavioral characteristics, this framework remains one of the most widely accepted inventory with consistent results across populations [45].

Based on the Big Five framework, every individual's personality consists of five latent dimensions, such as Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. The definitions and descriptions of these Big Five personality traits are as follows:

1. Openness - This trait refers to a person's openness to new experiences, ideas, and ways of thinking. Open individuals tend to be imaginative, curious, and creative. People who score high in Openness may be more willing to challenge traditional ways of doing things and to think outside the box. People who score low on this trait tend to be more practical and focused on the present.

2. Conscientiousness - This trait refers to a person's level of organization, responsibility, and self-discipline. Conscientious individuals tend to be reliable, hardworking, and detail-oriented. This is also called the orderliness dimension. People with high Conscientiousness happen to be very orderly and organized. People who score low on this trait tend to be more laid-back and less focused on achieving specific goals.

3. Extraversion - This trait refers to a person's level of sociability, outgoingness, and assertiveness.

They enjoy being around people and tend to be energized by social interactions. They may be seen as talkative and enthusiastic. People who score low on this trait tend to be more introverted and prefer quiet, solitary activities.

4. Agreeableness - This trait refers to a person's level of cooperativeness, empathy, and kindness. Agreeable individuals tend to be friendly, compassionate, and willing to compromise. People who score low on this trait tend to be more competitive and may prioritize their own needs over the needs of others.

5. Neuroticism - People who score high on this trait tend to experience more negative emotions, such as anxiety, stress, and sadness. They may be more sensitive to criticism and tend to worry more than others. People who score low on this trait tend to be more emotionally stable and less prone to experiencing intense negative emotions and more frequent mood swings.

Ideally, an assessment of personality traits should be done in an unobtrusive manner to ensure unbiasedness. An assessment is considered unobtrusive if it does not require any attention of the person being assessed [25, 52]. This not only makes the assessment more convenient for the person being assessed and can remove subjective bias, but also reduces the risks of measurements being affected by modified behavior due to the assessed person being consciously aware of the assessment [25]. This is where modern smartphones come to the rescue, since they come loaded with a host of sensors which can be employed to unobtrusively gather data about behavior of individuals [11]. Smartphones are a good option for this kind of study also because they are already widely used and are routinely carried around by people for most of their day [21]. Physical as well as logical sensors related to location, communication, phone state (e.g. screen lit, charging status), phone orientation, connections to other devices and to the internet can be put to use to understand activities like movements, interactions and daily habits [24].

The purpose of this study is to establish whether individual differences in personality traits can be detected through data collected from smartphones. For example, users who are actively using communication applications like whatsapp and telegram may score high on Extraversion. People who use productivity apps like calculator, calendar and to-do lists may score high on Conscientiousness. These hypothetical examples serve as guidelines to study the correlation of data with personality. With the use of a host of exploratory analysis, correlation analysis and machine learning algorithms plenty of existing research has already linked behavioral indicators derived from smartphone data to personality traits. However, as far as our knowledge goes, there is no publicly available dataset for investigating these connections. Therefore, we utilized data from the WeNet study, which was designed to address this and other research gaps. In our research, we performed an extensive exploratory analysis, considering self-assessed personality traits and indicators derived from smartphone data. Using feature selection, we determined indicators that were informative about the personality of people. We then adopted a predictive approach using linear as well as non-linear models, with a specific focus on whether combinations of features extracted from various smartphone sensors could assist in predicting individuals' personality traits.

My main hypothesis is that an individual's personality traits, specifically those related to the Big Five personality traits, can be predicted using real-world behavior data collected from smartphones.

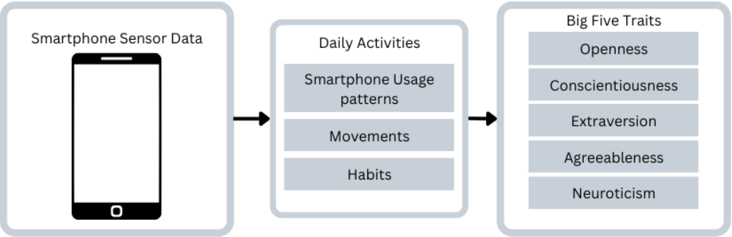


Figure 1.1: Graphical Representation of the Hypothesis

This thesis is composed of 6 chapters. Chapter 2 furnishes the foundational context necessary for this thesis. It covers previous research related to human behavior using diverse types of technology-mediated data and also delves into their constraints. It also addresses various types of personality traits. In Chapter 3, the research background is outlined which includes various machine learning models, baseline models and model evaluation techniques. In Chapter 4 the experimental configuration for gathering necessary data is outlined, along with thorough explanations of feature extraction techniques. This chapter also elucidates the methodologies and machine learning models utilized. Chapter 5 showcases the outcomes, while Chapter 6 offers a summary of the results and the significant contributions made in this research.

2 Literature Review

2.1 Evolution of Personality Concepts

Personality traits are patterns of thought, emotion, and behavior that are relatively consistent over time and across situations. They can be described with familiar words such as “reliable”, “sociable”, or “cheerful”, as well as more specialized terms such as “narcissistic”, “authoritarian”, or “conscientious”. Psychology has developed an impressive and useful technology for assessing personality traits, but personality assessment is not limited to psychologists: Everybody does it, every day. We all make judgments about our own personalities as well as of the personalities of people we meet, and these judgments are consequential [13].

Several non-human animal species also exhibit individual differences in behavioral patterns, indicating possible existence of personality traits that may even predate humanity [18]. It is likely that humans have long observed these differences among members of their community, even before they had a means to effectively record or communicate these ideas through writings. However, the specific ways in which prehistoric people conceptualized these differences, such as those between an individual who excelled at cave painting and one who was skilled at ensuring the safety of fellow tribe members during hunting, are likely lost to history.

The earliest known theory of personality can be traced back to the Greek physician Hippocrates at about 400 B.C. [30]. He suggested a classification of individual temperaments into four types: Sanguine (people who are optimistic and hopeful), Melancholic (people who are sad or depressed), Choleric (people who are irascible, i.e. easily angered), and Phlegmatic (people who are apathetic, i.e. indifferent and passionless).

This four temperament theory was further developed by Wilhelm Wundt, a German physiologist who is considered the “father of experimental psychology” [4]. According to Wundt’s model, the four temperaments represent the extreme ends of a two-dimensional space that is defined by the emotional vs. unemotional and changeable vs. unchangeable dimensions [30].

During the 1920s, Carl Jung, a Swiss psychiatrist, introduced the terms “extraversion” and “introversion” to describe different orientations of personality. Although initially ignored by academic psychologists, Jung’s work has endured despite being based on introspective and interpretive techniques within the psychoanalytic tradition established by Sigmund Freud. Extraversion and introversion are still considered important principles in modern models of personality [49].

Personality psychology as an academic field began to take shape in the 1930s, with the establishment of the first journal, “Journal of Personality,” in 1932. Hans Eysenck, a German-British psychologist, was an early influential figure who, in the late 1940s, introduced a three-dimensional model of personality that consisted of “extraversion”, “neuroticism”, and “psychoticism”. Initially personality psychologists struggled to establish personality as a relevant construct because behavior can change depending on context. But later they found out that a person’s behavior over a long period of time and in different situations remained consistent and were related to a person’s personality. The lexical approach proposed by Goldberg in 1982 involved creating a comprehensive list of adjectives to describe an individual’s character, which led to the discovery of the largely independent Big Five factors [16]. These factors are highly stable over time and are predictive of important life outcomes [29, 35]. As a result, the Big Five factors are now commonly employed both within and outside the domain of personality psychology.

2.2 Personality Traits and the Big Five Model

For this thesis, we are defining personality traits as patterns of behavior, thought, and feeling that remain consistent across various situations. We are using questionnaires to assess personality traits, which are based on item response theory [19, 41, 28]. In personality psychology, items usually consist of statements like "They like to take risks." and the test-taker responds on a Likert scale from "Very much like me" to "Not like me at all." This theory assumes that answers given in a test are informative about hypothetical latent variables that affect the answers. These latent variables cannot be measured directly, but can be inferred based on directly observed manifest variables. Factor analysis is used to discover the main dimensions along which people vary, and five factors consistently emerge in various populations. These are known as the Big Five and include extraversion, openness to experience, agreeableness, conscientiousness, and neuroticism [29].

2.3 Smartphone data and Personality traits

There are several ways that researchers have attempted to measure personality traits using smartphone sensors. One approach is to use self-report questionnaires that are administered via smartphone apps. Another approach is to use data from smartphone sensors to infer information about behavior and activity patterns, which can then be used to make inferences about personality traits. Researchers gathered information on smartphone usage by relying on self-reports from participants. They then examined the relationships between personality traits and these self-reported patterns of smartphone usage. The dataset used for this analysis consisted of data from 112 participants who provided information about their smartphone usage, which was subsequently used to estimate their personality traits [5]. In the study, the researchers assessed participants' personality traits using two questionnaires: the Coopersmith Self-Esteem Inventory and the NEO-FFI (Neuroticism, Extraversion, Openness, Five-Factor Inventory) questionnaire. Additionally, participants completed an 8-item questionnaire related to smartphone usage. This questionnaire aimed to capture information about the amount of time participants spent on various smartphone activities, including making and receiving calls, sending and receiving SMS messages, playing games, changing ringtone and wallpaper, and other related activities.

The findings, based on regression analysis, indicated that individuals with extroverted personality traits tended to spend more time on making and receiving calls and changing wallpapers on their smartphones. Self-reported mobile gaming behaviors were used to identify the personality traits [38]. The results of the regression analysis showed that individuals who scored low on the Agreeableness personality trait were more likely to use mobile devices for playing games. Subsequently, a number of researchers delved into the potential for estimating human behavior through the automated extraction of smartphone sensor data, phone call information, and app usage data. An overview of various available smartphone sensors and specific areas of psychological research was presented [21]. App usage logs from smartphones were used to predict human personality [48]. In these studies, the app logs were collected and app usage categorized based on the type of usage. Usage was put in the following groups: communications, tools, productivity, games, media, and finance applications, etc. For instance, user traits were predicted using a snapshot of installed apps. App logs of 200 participants were collected, and the applications were grouped based on app purpose. An SVM classifier was employed to successfully infer an individual's religion, marital status, whether the user is a parent of small children, and their mother tongue. Personality traits were identified using individuals' app adoption [54]. In this study, app installation logs from a total of 2,043 Android users were analyzed to identify the Big Five personality traits. This identification was done based on the categories of apps available on the Google Play Store. The Big Five personality values were categorized into three classes: low, medium, and high. To model personality traits, a random forest classifier was employed. The results showed that the model was able to predict personality traits with a 50% success rate. It's important to note that this prediction was solely based on app adoption, meaning it considered which apps were installed by users but did not take into account how those apps were actually used.

Phone call behaviors have been employed as another method to estimate an individual's Big Five personality traits. In a specific study involving 39 participants, researchers extracted various call and SMS-related features from phone logs [34]. These features encompassed factors such as call duration, the timing of calls, and the quantity of text messages sent and received. These extracted features were then utilized to construct a social communication network. To predict the Big Five personality traits, a supervised learning approach based on Support Vector Machines (SVM) was employed. The results of this analysis yielded mean squared errors ranging from 0.73 to 0.86 on a 7-point scale. In another comprehensive approach, researchers harnessed both standard mobile phone information and GPS data to predict individuals' personality traits [31].

They collected conventional carrier logs, including phone calls and text messages, from a group of 69 participants. From this data, they calculated the entropy of calls and texts and also assessed the inter-event time between text messages and calls. In addition, GPS data was utilized to determine the radius of gyration, daily travel distances, and the number of distinct places visited by each participant. The participants' self-reported Big Five personality traits were categorized into three classes: low, average, and high. To build predictive models, they employed a Support Vector Machine (SVM) classifier. Using a ten-fold cross-validation approach, they were able to identify Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism traits with the following accuracies: 49% for Openness, 51% for Conscientiousness, 61% for Extraversion, 51% for Agreeableness, and 63% for Neuroticism. The researchers also noted correlations between specific personality traits and mobile phone behavior. For instance, they found that Extraversion and Agreeableness traits were associated with the entropy of participants' contacts, suggesting a connection between these personality traits and the diversity of their social interactions. Furthermore, the variance in the time intervals between phone calls was correlated with the Conscientiousness trait, indicating a potential link between conscientiousness and communication patterns.

A study of similar nature, but with a considerably larger sample size, incorporated Bluetooth sensor data to further explore the prediction of Big Five personality traits [33]. This extensive study involved the collection of mobile data, including telecommunication data (calls and texts), GPS data, and Bluetooth sensor logs, from a substantial group of 636 students over a period of 24 months. From this dataset, various features were extracted. These features encompassed aspects such as face-to-face contacts or physical proximity to others, as determined by Bluetooth signal strength, as well as geo-spatial mobility patterns, and the analysis of text messages, calls, and social network friends' contact lists. The Big Five trait values were categorized into three classes: low, medium, and high. Researchers utilized Support Vector Machines (SVM) to build models for predicting personality traits. However, the study reported successful identification only for the Extraversion trait.

Touch screen swipe behaviors have been employed as a means to identify personality traits. In a study involving 98 participants, researchers collected data on touch screen swipes and extracted various touch/swipe-related features [1]. These features included parameters such as average velocity, mean pressure, mean finger area, and others. To predict specific personality traits, the researchers utilized the self-reported Eysenck personality questionnaire in combination with the extracted touch screen swipe features. They employed machine learning classifiers, specifically K-Nearest Neighbors (KNN) and Random Forests, in their analysis. The outcome was the prediction of the Extraversion and Neuroticism traits with an average accuracy of 62.9%. In a related study, smartphone data, including call logs, SMS logs, Bluetooth scans, and app usage, was used to predict Big Five Personality traits in 83 participants over 8 months [7]. Features like Bluetooth IDs, call durations, unique contacts, SMS length, and app usage were extracted. Personality traits were categorized as low and high, and a Support Vector Machine binary classifier achieved accuracies of 69.3% for Openness, 74.4% for Conscientiousness, 75.9% for Extraversion, 69.6% for Agreeableness, and 71.5% for Neuroticism. The study also found correlations between personality traits and smartphone usage, e.g. Extraversion correlated with internet usage, while Conscientious individuals used media apps less, and Extroverts spent more time on calls.

Another study involved 32 participants who provided data from various sources, including application usage logs, phone calls, SMS messages, email messages, and self-reported mood states collected four times a week [27]. An application named Moodscope was developed as a tool for detecting and assessing mood based on smartphone usage, and it successfully demonstrated that mood can be inferred from sensor data. Researchers employed a multi-linear regression model to analyze this dataset and determine participants' mood. Remarkably, the study achieved a successful inference of participants' mood with an accuracy rate of 66%.

A study made use of app usage data, geospatial records (university arrival time and exit time) and behavioral parameters (such as charging time) collected from 80 students over 1461 days to estimate the personality inventory of a participants in an unobtrusive manner without the need of parsing the app-specific content and social media content [24]. This underscores the potential for utilizing smartphone data to assess and monitor individuals' emotional states and well-being.

2.4 Rationale for Our Study

There are several limitations to the previous studies.

- Numerous studies have relied on questionnaires and surveys to gather data, offering psychological insights that may not be directly accessible through observed behavior [12, 20, 22, 42].
- Certain approaches employ pervasive methods that involve analyzing personal email and social network activities, such as Facebook likes and the number of friends. However, these methods can potentially violate privacy laws or run counter to research ethics guidelines [2, 34, 15, 40].
- Certain implementations necessitated data such as the number of initiated or received phone calls, call response rates, phone contacts, SMS usage details (including the number of messages sent and message response rates), and social networking activities. However, this data is often proprietary and accessible only to service providers and social networking companies [2, 7, 31, 34, 15, 43].
- Several modern studies have modeled human personality through statistical analysis and classification models. They achieve this by transforming standard continuous personality trait values into discrete categories [1, 3, 7, 31, 23, 39, 43, 54].
- Numerous studies, with the exception of [24, 44], which employed machine learning models to predict human personality, often lack a thorough presentation of model performance. Instead, they tend to express results solely in terms of statistical metrics, without offering baseline models for comparison. This absence of baseline models raises questions about the reliability of their findings [7, 31, 33, 39].

In contrast to these, in this study,

- We utilized an unobtrusive sensing method to collect data, primarily focusing on easily accessible smartphone data. This included aspects like app usage (excluding internal app data), hardware sensors like Bluetooth, and software sensors such as screen events, charge events, and step counters. We specifically collected data on app usage, such as the number of times communication apps were used. Importantly, this data can be readily obtained by any app developer with the users' consent, eliminating the need to depend on social networking companies or network service providers.
- In contrast to previous studies that relied on a much smaller number of sensors for their analysis, we employed a total of nine different sensors, including both hard and soft sensors. In conjunction with the Big Five Inventory (BFI) data collected through questionnaires, this approach enabled us to create diverse combinations of features extracted from various sensors. This allowed us to conduct various correlation analyses and predictions using machine learning algorithms
- While the majority of contemporary studies treated the prediction of personality traits as classification problems, where continuous personality trait values were transformed into discrete classes based on value ranges, this study took a distinct approach. In our research, along with classification models we also employed regression models, which offered a continuous representation of personality within the Big Five model, presenting a different perspective on the subject.
- At the end, the results are presented with a comprehensive analysis of model performance, which includes an appropriate baseline model and an examination of fit behavior through residual analysis. The use of a baseline model for comparison is uncommon in extant research.

3 Background

3.1 Supervised Machine Learning

3.1.1 Ordinary Least Square Regression

Ordinary Least Squares (OLS) is a fundamental statistical method employed in various fields, including economics, social sciences, and data science, to explore and quantify the relationships between variables. OLS is particularly useful when investigating linear relationships, where a dependent variable is assumed to depend on one or more independent variables. The primary objective of OLS is to estimate the coefficients that define this linear relationship by minimizing the sum of the squared differences between observed data points and the predictions made by the linear model.

These coefficients reveal the strength and direction of the associations between variables, enabling researchers to make informed interpretations and predictions. Successful OLS application relies on meeting certain assumptions, such as linearity, independence of errors, homoscedasticity, normality of errors, and absence of multicollinearity among independent variables. Therefore, researchers must carefully assess these assumptions and tailor their OLS models accordingly. OLS not only serves as a foundational method but also paves the way for more advanced regression techniques, contributing to robust statistical analysis and informed decision-making across various domains. A linear regression model with one dependent and one independent variable when plotted along with the regression line would look like 3.1.

3.1.2 LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) regression is a statistical method used in linear regression analysis and machine learning for feature selection and regularization. It was introduced by Robert Tibshirani in 1996. LASSO is a variant of linear regression that adds a regularization term to the traditional least squares regression objective function. The primary purpose of LASSO is to prevent overfitting and to perform feature selection by shrinking the coefficients of less important predictor variables to exactly zero.

LASSO adds a regularization term to the linear regression objective function. This term is the absolute sum of the coefficients multiplied by a tuning parameter (λ):

$$\min \sum (y_i - \beta_0 - \sum \beta_j \cdot x_{ij})^2 + \lambda \sum |\beta_j|$$

λ (lambda) controls the amount of regularization. A higher λ results in stronger regularization, which in turn leads to more coefficients being shrunk towards zero.

One of the significant advantages of LASSO is its ability to perform feature selection. As λ increases, some of the coefficient estimates become exactly zero, effectively excluding the corresponding predictor variables from the model. This means that LASSO not only provides a predictive model but also identifies the most important features in the data.

The λ parameter introduces a bias-variance trade-off. A high λ shrinks more coefficients to zero, which increases bias but reduces variance, while a low λ allows the model to be more complex, reducing bias but increasing variance. To determine the optimal value of λ, cross-validation techniques are often used. By testing the model's performance with different values of λ, you can select the one that gives the best balance between predictive accuracy and feature selection.

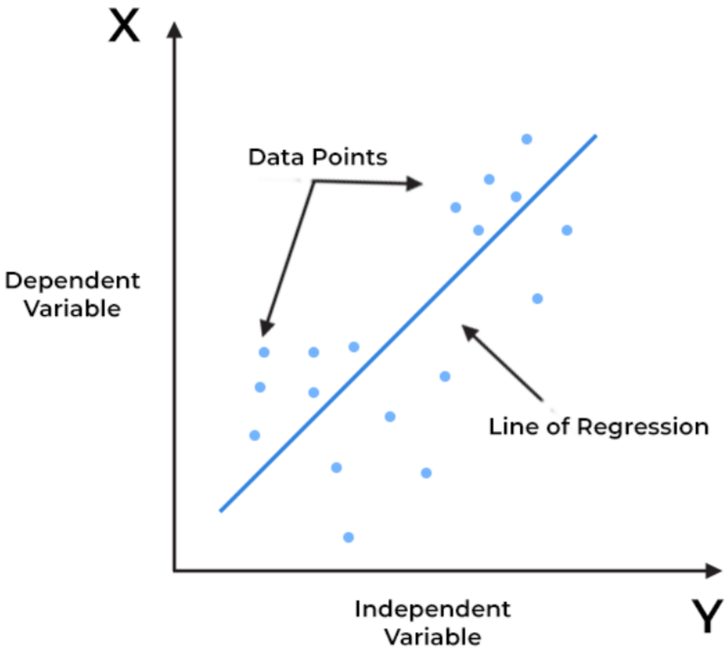


Figure 3.1: Example of a simple OLS Regression.

3.1.3 Ridge Regression

In simple linear regression, you have a dependent variable (Y) and one or more independent variables (X). The goal is to find the best-fitting linear equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

that minimizes the sum of squared differences between the predicted values and the actual values.

In OLS, the model can become overly complex when you have many features or when there is multicollinearity (high correlation between independent variables). This complexity can lead to overfitting, where the model fits the training data very closely but performs poorly on unseen data.

Ridge regression addresses overfitting by adding a regularization term to the OLS objective function. The objective function of Ridge regression can be defined as:

$$\text{Minimize: } \text{RSS} + \alpha \sum \beta_i^2$$

Where:

- RSS (Residual Sum of Squares) is the same as in OLS, measuring the sum of squared differences between predicted and actual values.
- α (alpha) is a hyperparameter that controls the strength of regularization.
- $\sum \beta_i^2$ is the sum of squared regression coefficients. The regularization term penalizes large coefficients.

The α parameter controls the trade-off between fitting the data well (minimizing RSS) and keeping the model simple (minimizing the sum of squared coefficients). A larger α leads to a more regularized model with smaller coefficient values, which is helpful in reducing the impact of multicollinearity and overfitting.

3.1.4 Random Forest

Random Forests are an ensemble learning method for decision trees. A decision tree is a graph structure created by splitting the data repeatedly into subsets, usually according to a single feature. A classic decision tree learning algorithm creates a model by splitting along the input dimension of greatest variance according to a heuristic or a cost function. Essentially, decision trees learn a hierarchy of (often true/false binary) decisions, leading to a classification of the data. A typical problem with decision tree learning is its tendency to overfit the data, that is to model the noise in the data as well as the underlying trend in the data itself, leading to poor model generalizability. To avoid these overfitting problems, random forests are employed. Random forests, as shown in figure 3.2 divide the whole data set into random small subsets (without replacement) and the decision tree is constructed for each subset. An aggregate statistic (usually the mean or mode) of the output of the ensemble (forest) of decision trees is taken as the actual answer. In regression tasks, Random Forest operates as a Random Forest Regressor. Instead of predicting discrete categories, it estimates continuous numerical values. Similar to the classification process, it constructs an ensemble of decision trees but employs a different aggregation method. Each tree in the ensemble predicts a numerical value, and the final prediction is obtained by averaging or taking the median of these individual tree predictions. Random Forest Regression is highly advantageous for capturing non-linear relationships and handling noisy data while avoiding the pitfalls of over-fitting. Additionally, it provides insights into feature importance, allowing researchers to identify the most influential variables in predicting the target variable. This flexibility and robustness make Random Forest a popular choice for both classification and regression tasks. For an overview of the various decision tree architectures, learning modes and applications see [26].

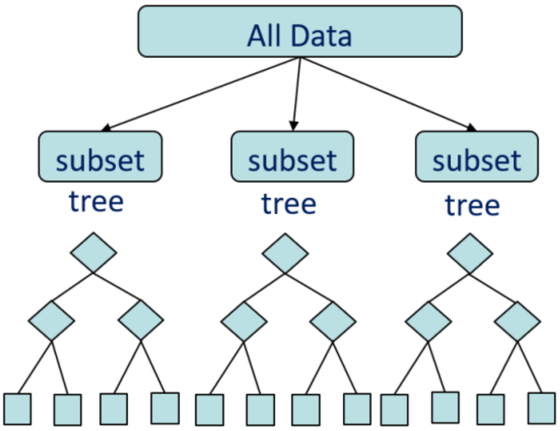


Figure 3.2: Sample Random Forests model created with three decision trees for the purpose of demonstration.

3.1.5 XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is a powerful machine learning algorithm that falls under the category of ensemble learning. It was developed by Tianqi Chen and is known for its effectiveness in solving various machine learning problems, especially in structured data and tabular data scenarios. XGBoost is an implementation of the gradient boosting framework, a machine learning technique that builds predictive models by combining the predictions of multiple weaker models, typically decision trees. Gradient boosting works by sequentially training a series of weak learners and adjusting their predictions to minimize a specified loss function. It's an ensemble method, which means it combines multiple models to improve predictive accuracy. XGBoost primarily uses decision trees as base learners. Decision trees are simple, non-linear models that make predictions by partitioning the input data into subsets and assigning a constant value to each subset. The decision trees used in XGBoost are often shallow, with a limited number of nodes, which makes them weak learners. XGBoost incorporates several techniques to control overfitting and improve model generalization. Regularization is applied through L1 and L2 regularization terms added to the loss function, which penalize complex models. This helps prevent the model from fitting the training data

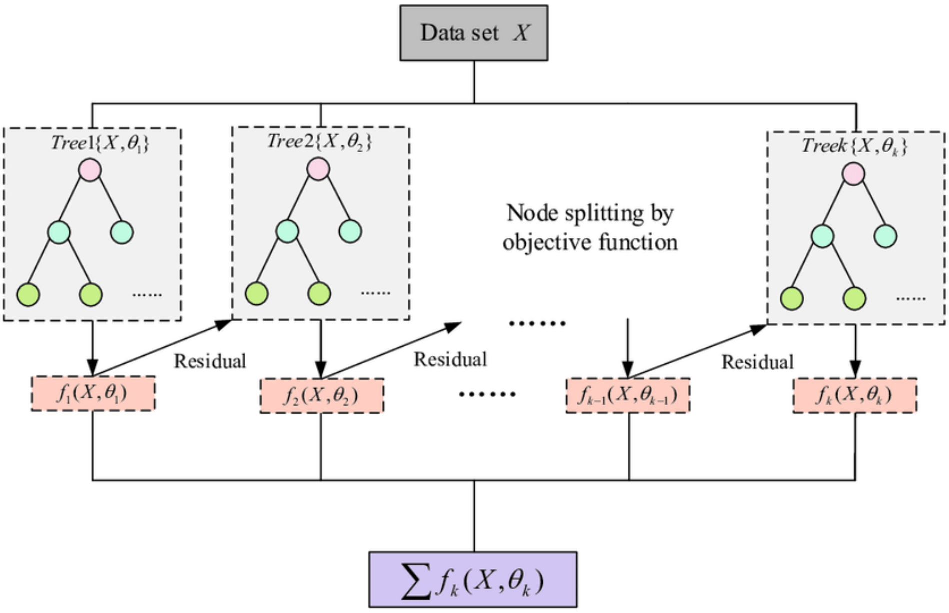


Figure 3.3: Sample XGBoost model created for demonstration

too closely.

While XGBoost uses popular loss functions like mean squared error for regression and log loss for classification by default, it also allows users to define custom loss functions, making it adaptable to a wide range of problems. The figure 3.3 shows a demonstration of how XGBoost works.

3.1.6 Support Vector Machine

Support Vector Machines (SVM) represent a powerful and versatile class of supervised machine learning algorithms used extensively in both classification and regression tasks. In classification, SVM aims to find a hyperplane that best separates different classes or categories within a dataset. It does so by maximizing the margin—the distance between the hyperplane and the nearest data points of each class. SVM is particularly effective in scenarios with complex decision boundaries or high-dimensional feature spaces, as it can employ various kernel functions (e.g., linear, polynomial, or radial basis function) to map data into higher-dimensional spaces (as shown in Fig. 3.4) where classes become more separable. This non-linear transformation enables SVM to handle intricate patterns and achieve high classification accuracy, making it a popular choice in image recognition, text classification, and bioinformatics.

In regression tasks, SVM transforms into a Support Vector Regressor, aiming to find a hyperplane that best fits the data while minimizing prediction errors. Unlike traditional regression techniques, SVM Regression can capture non-linear relationships by utilizing kernel functions to map input features into a higher-dimensional space. The objective is to find a hyperplane that maintains a specified margin around the predicted values, effectively balancing the trade-off between fitting the training data and generalizing to unseen data points. SVM Regression excels in scenarios where data exhibits nonlinear patterns, and it is robust to outliers due to its use of support vectors—data points that influence the position of the hyperplane. As a result, SVM is a valuable tool in both classification and regression domains, contributing to breakthroughs in fields such as finance, healthcare, and natural language processing.

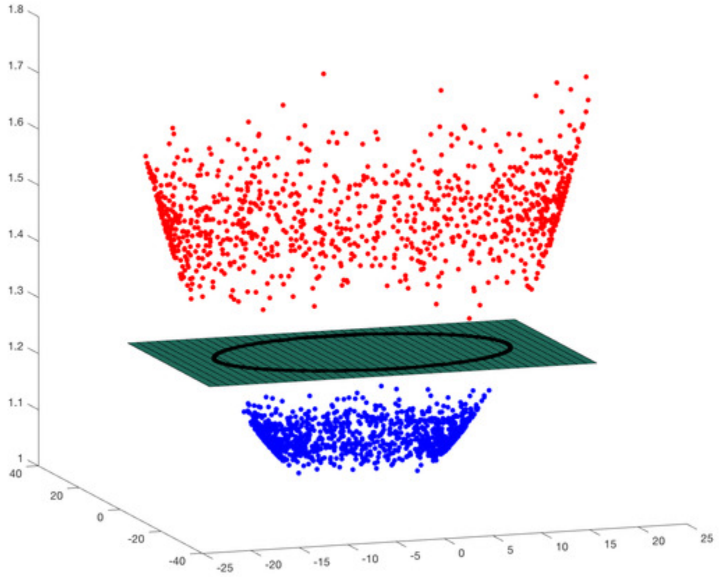


Figure 3.4: Visual representation of Support Vector Machine transforming the Non-liner separable data in to higher dimensional space

3.1.7 K-Nearest Neighbour

The K-Nearest Neighbors (KNN) algorithm is a fundamental and intuitive machine learning technique used for both classification and regression tasks. It is based on the principle of proximity, assuming that similar data points in a feature space tend to have similar target values or belong to the same class. KNN is considered a non-parametric and instance-based learning method because it doesn't make assumptions about the data distribution and makes predictions based on local information. To make a prediction for a new, unseen data point, KNN identifies the k-nearest neighbors from the training dataset. The distance metric (commonly Euclidean distance or Manhattan distance) is used to measure the proximity between the new data point and all other data points.

In classification tasks, KNN assigns the class label that is most frequently represented among the k-nearest neighbors to the new data point. This is often determined by a simple majority vote. In regression tasks, KNN calculates the average (or weighted average) of the target values of the k-nearest neighbors and assigns this value as the predicted target value for the new data point.The choice of the "k" parameter in KNN is crucial. A small "k" makes the model sensitive to noise and outliers, potentially leading to overfitting. A large "k" can over smooth decision boundaries, potentially leading to underfitting. KNN is often computationally expensive for large datasets, as it requires calculating distances between the new data point and all training data points. Various data structures (e.g., KD-trees) and optimizations can be used to speed up this process.Fig. 3.5) shows the Voronoi tessellation having 19 samples marked with a "+", and the Voronoi cell surrounding each sample. A Voronoi cell encapsulates all neighboring points that are nearest to each sample. For an overview refer to [37].

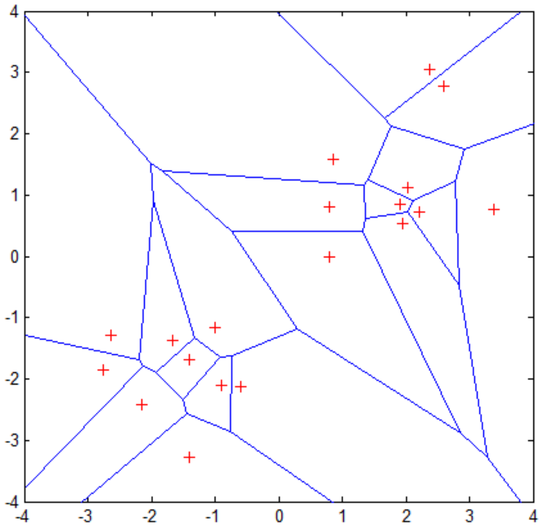


Figure 3.5: Voronoi tessellation showing a sample k-NN classifier

3.1.8 Decision Tree

Decision tree algorithm is a machine learning technique used for both classification and regression tasks. It is a non-linear and non-parametric model that builds a tree-like structure to make predictions by recursively splitting the data based on the most significant features. Each internal node in the tree represents a feature or attribute, and each branch represents a decision or rule based on that feature. The leaves of the tree contain the predicted class labels (for classification) or target values (for regression) for the corresponding subset of data.

The algorithm starts with the entire dataset at the root of the tree. It selects the feature that provides the best split, typically based on criteria like Gini impurity (for classification) or mean squared error (for regression). This split partitions the data into subsets, each sent down a branch of the tree. The process continues recursively for each subset, selecting the best feature for splitting at each internal node. The splitting stops when a predefined stopping criterion is met, such as reaching a maximum depth or having a minimum number of samples in a leaf node. Once the tree is built, the class label (for classification) or target value (for regression) assigned to each leaf node is used as the prediction for data points that reach that leaf during inference. Decision Trees serve as the building blocks for more advanced ensemble methods like Random Forests and Gradient Boosting, which combine multiple decision trees to improve predictive performance and robustness. We have made use of such ensemble techniques in our work too. Figure 3.6 refers to a sample decision tree based on binary target variable Y.

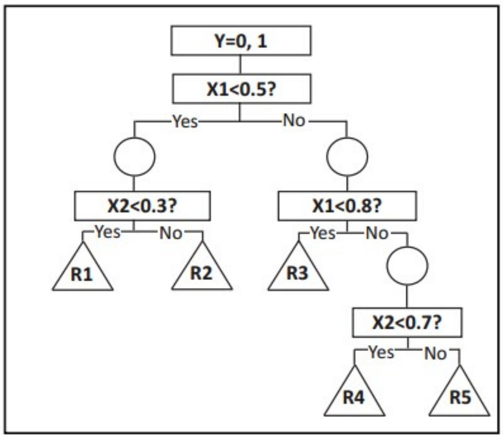


Figure 3.6: Sample decision tree based on binary target variable Y

3.2 Baseline Models

3.2.1 Mean Model

A Mean model serves as a fundamental baseline model for regression problems. In this simplistic model, the output value is forecasted solely as the population mean, disregarding any variations in the input values.

For instance, Table 3.1 includes data for five participants, encompassing three input features and an output feature known as the ”actual trait.” There is also a column labeled ”predicted trait,” which represents the values predicted by the mean model. As demonstrated, the mean model consistently returns a predicted trait value of 0.5 for all participants. This uniform prediction arises because the mean of the population, in this case, is calculated to be 0.5, and the mean model simply assigns this value as the prediction for each participant, irrespective of their individual input values.

Participants	Input 1	Input 2	Input 3	Actual Trait	Predicted Trait
Participant 1	10	5	15	0.8	0.5
Participant 2	7	11	3	0.9	0.5
Participant 3	11	13	17	0.2	0.5
Participant 4	15	8	6	0.5	0.5
Participant 5	3	9	9	0.3	0.5

Table 3.1: Sample Data for Mean Model

3.2.2 ZeroR Model

The ZeroR model, often referred to as the ’zero rules’ or ’most frequent class’ model, serves as a fundamental baseline model frequently used in classification problems. The ZeroR model adopts a uniform prediction strategy where the output class label remains constant and equal to the most prevalent class label observed in the dataset. So in this model, the prediction for the class label is always the most common class label regardless of the input features or their values.

For instance, consider Table 3.2, which contains data for five participants, including three input features and output class labels denoted in the ”Actual Trait” column. In this case, the possible output class labels are ’high’ (H) or ’low’ (L). Since the majority of the population in this dataset has ’H’ as the class label, the ZeroR model consistently predicts ’H’ as the class label for all participants, irrespective of their individual input feature values. This straightforward approach serves as a baseline for classification modeling, offering a straightforward way to gauge the performance of more intricate classification models, offering a clear and easily interpretable comparison in assessing the effectiveness of advanced techniques and algorithms.

Participants	Input 1	Input 2	Input 3	Actual Trait	Predicted Trait
Participant 1	10	5	15	H	H
Participant 2	7	11	3	H	H
Participant 3	11	13	17	H	H
Participant 4	15	8	6	L	H
Participant 5	3	9	9	L	H

Table 3.2: Sample Data for ZeroR Model

3.3 Model Evaluation

3.3.1 Cross Validation

Cross-validation is a technique used to assess the reliability and consistency of machine learning models. In the process of cross-validation, the entire dataset is divided into K folds, with one fold reserved as a validation set while the model is trained on the remaining K-1 folds. This approach allows for testing the model against unseen data without introducing potential bias that could arise from randomly selecting a particularly favorable or unfavorable test set.

For instance, consider a ten-fold cross-validation. The dataset is initially split into ten sets. During the first round, one set is designated as the validation set, and the model is trained on the remaining nine sets. This process is then iterated for the remaining sets. The accuracy or error for each round is computed separately, and the final accuracy or error for the model is determined by averaging the results from all ten rounds. This comprehensive evaluation provides a robust measure of the model's performance across different subsets of the data.

3.3.2 RMSE

The root mean squared error is a quality metric for regression models. It is computed by finding the square root of the mean of the squares of the difference between the actual values and predicted values. Since this is an error metric, a model with lower root mean squared error is considered as a better model. For the sample mean model data shown in Table 3.1 RMSE value is 0.28.

The Root Mean Square Error (RMSE) is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.1}$$

Where:

n : Number of data points

y_i : Actual value for data point i

$\hat{y}_i$: Predicted value for data point i

3.3.3 Accuracy

Accuracy serves as a crucial statistical metric for evaluating classification models. It is calculated by taking the ratio of the number of correct predictions made by the model to the total number of tests conducted. A model with a higher accuracy percentage is generally considered to be superior in its predictive capabilities.

For instance, let's consider a sample classification dataset as presented in Table 3.2. In this scenario, the model successfully made three correct predictions out of a total of five tests. Consequently, the accuracy of the model can be determined as 60% since it achieved a 60% accuracy rate by correctly classifying the majority of the test cases.

4 Methods

4.1 Study Design and Data Collection

The data for this study was collected under a project called WeNet, one of European Union's Horizon 2020 programs under the Grant Agreement number 823783 [51]. The data collection process spanned a six-week period and was organized into two stages:

1. The initial synchronic data collection involved the use of three standard close-ended questionnaires. This stage enabled the collection of self-reported general information regarding social practices.

2. The subsequent diachronic data collection took place through a smartphone app, facilitating the observation of the daily routines of the students.

As described in Figure. 4.1, the first two weeks were dedicated to the sample recruitment. This was performed by sending two initial questionnaires, i.e., invitation and assessment of habits. The remaining month was entirely dedicated to the data collection through the app installed on the students smartphone. During all the data collection a help-desk was active, and ready to support students in all the problems which were arising.

The questionnaires were managed with the LimeSurvey platform [47]. Invitations to participate in the online survey were sent out through LimeSurvey to the email addresses of students enrolled at various universities. This data collection method was based on the use of Time Diaries, which are a well-established tool in the social sciences. Time diaries ask respondents to record three key aspects of their daily lives: the activities they engage in, the locations they visit, and the people they interact with. Time diaries can be administered in two ways: as "leave-behind diaries," where respondents record their activities in real-time as the day progresses, or as "recall diaries," where respondents recall their activities from the previous day. In our study, we used the iLog app, which allowed students to provide real-time responses. The questions and answers were structured in accordance with the HETUS (Harmonised European Time Use Survey) standard [46, 55].

The sample was chosen from the entire student population of the University of Trento. An invitation to participate in the survey was extended to all students, with an initial exclusion criterion applied to those who did not possess a smartphone compatible with the study (specifically, only Android Operating System versions greater than 5.0) or those who did not regularly attend classes. Subsequent to the initial contact via email, the online questionnaire was dispatched to inquire about their habits and routines. The final stage involved the transmission of a password for downloading and installing the iLog application. This process resulted in 1042 responses. Among these responses, those from students born after 1993 (with the aim of restricting the involvement of latecomers to the university) and students who did not actively engage in university life were removed. From the pool of 860 eligible candidates, a total of 318 students were selected, with the sample size adjusted proportionally to each department's student population. This adjustment was made to prevent any misrepresentation of daily routines stemming from variations in schedules and university sub-communities. Data cleaning during the data preparation phase resulted in a final dataset that includes information from only 149 participants. This reduction in size, in comparison to the previously mentioned 318 students, is the outcome of excluding all participants with limited survey participation. Also due to computational limitations, for this particular work, the data collected between the time period November 2020 and December 2020 was used which contributed to the reduction in size.

	%
Gender	
Female	48.7
Male	51.3
Age	
<22	47.5
22-26	52.5
Department	
Stem	44.57
Non-Stem	55.42
Total	100 (N=149)

Table 4.1: Descriptive statistics of the participants

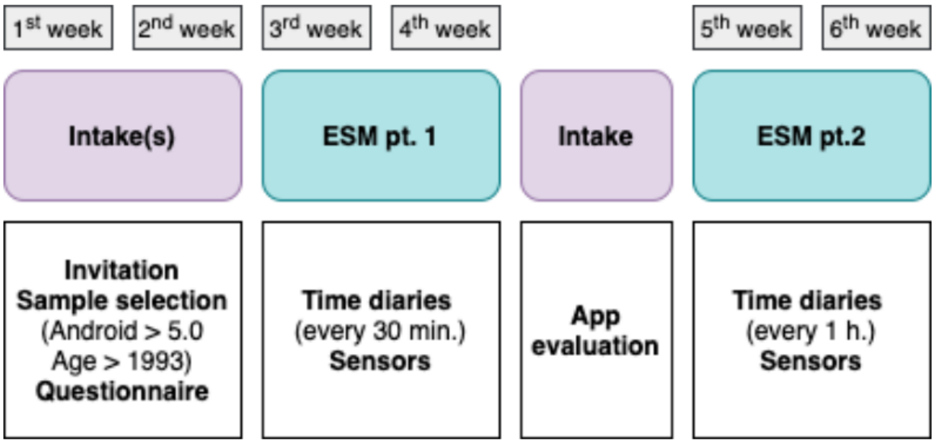


Figure 4.1: Schematic representation of the study protocol

Table 4.1 shows how the sample is balanced according to the main characteristics, namely gender, age and departments (whether stem or non-stem) in which the students were enrolled. Furthermore, it shows the range of annotations given from the participants. The psycho-social characteristics of the participants are described in Table 4.2. Concerning personality traits (BFI-10), the average of the scores is between 49.15 and 76.26, with a maximum standard deviation of 23.69 reached in the case of the Extraversion variable. The range of responses goes from 0 to 100. Students who were enrolled in the following departments: Engineering and Applied Sciences, Natural Sciences, Medicine and veterinary medicine, and Agricultural, were categorized as belonging to the STEM department. Conversely, students enrolled in departments such as Social Sciences, Business/Economics, Law, Humanities, and International Relations and Public Administration were categorized as non-STEM. Apart from psychosocial traits, data was also collected on various other aspects, including (i) daily and extraordinary journeys, which encompassed the times and means of transportation used; (ii) work routines; and (iii) study and class attendance routines. In total, 27 questions were posed, resulting in the collection of 78 variables. However, it's worth noting that for the purposes of this analysis, we did not utilize this additional data, and therefore, I won't be elaborating on it further.

4.2 Ethics and privacy

All survey activities and the outcomes achieved at each site adhere to academic and national ethical standards, prioritizing privacy protection in accordance with applicable laws and guidelines. Further-

	mean	**std**	**median**	**min**	**max**
Total Population					
Extraversion	49.15	23.69	50.00	0.00	100
Agreeableness	76.26	15.56	75.00	25.00	100
Conscientiousness	64.88	19.27	62.50	12.50	100
Neuroticism	49.70	20.70	50.00	0.00	100
Openness	71.46	18.66	75.00	6.25	100
Female Population					
Extraversion	48.72	22.76	50.00	0.00	100
Agreeableness	80.15	14.27	81.25	25.00	100
Conscientiousness	63.69	20.40	62.50	12.50	100
Neuroticism	54.49	19.90	53.13	0.00	100
Openness	70.42	20.37	75.00	6.25	100
Male Population					
Extraversion	49.71	24.98	50.00	0.00	100
Agreeableness	71.09	15.77	75.00	31.25	100
Conscientiousness	66.47	17.62	68.75	18.75	100
Neuroticism	43.34	20.10	43.75	0.00	100
Openness	72.84	16.11	75.00	31.25	100

Table 4.2: Descriptive statistics of the psycho-social traits

more, for experiments conducted outside of Europe, the activities and results have been designed to align with the requirements of a specific European country, as stipulated by the European Commission. In this context, Italian legislation was chosen as the reference point. Additional details pertaining to these compliance measures are provided in [14].

4.3 The Sensor Data

The sensor data collected is characterized by its richness and diversity, and no other dataset with similar properties is currently known to us. Moreover, some of the selected sensors are rather unconventional. The sensor data can be categorized into two main groups:

1. Hardware (HW) sensors, which encompass sensors such as Bluetooth, Wi-Fi, GPS, and others. In this study, data from the Bluetooth and Accelerometer sensor were specifically utilized.

2. Software (SW) sensors, referring to all the software events that can be captured from the operating system and software applications. Examples include events related to Wi-Fi connectivity and more. A comprehensive list of the software sensors employed in our study is provided in Table 4.3.

Table 4.3 shows the list of sensors used in the study along with their measurement frequency and their respective number of observations. Among all the sensors listed in the table, the Step Counter is a software sensor whose data is derived from data recorded by a hardware sensor called the Accelerometer. The Step Counter data provides minute-by-minute records of the number of steps taken by users. Similarly, the Touch Event sensor contains analogous information, documenting the number of touch events occurring every minute. For the Screen Event, Battery Charge Event, Doze Event, and Music Event data, each row in the dataset corresponds to a change in the event, recorded as True/False or On/Off status. The Ring Event data comprises instances when changes in ring mode occurred, with three registered ring modes: Normal, Silent, and Vibrate. The Bluetooth sensor captures data regarding the device name, brand, and records of devices in close proximity to the user's smartphone. It includes a feature called "bond" to indicate whether the devices are paired with the host smartphone or not. This data offers insights into the level of populated areas where the user is situated at a given time. The data in its existing form could not be used for analysis and, therefore, needed to be cleaned, transformed, and preprocessed. Furthermore, meaningful features

Sensors	N. Obs.	Estimated Frequency
Screen Event Sensor (ON/OFF)	13,594,915	On Change
Battery Charge Event Sensor (True/False)	295,873	On Change
Doze Event Sensor (True/False)	583,665	On Change
Ring Mode Sensor (Normal/Silent/Vibrate)	102,540	On Change
Touch Event Sensor	1,235,536	On Change
Music Event Sensor (True/False)	374,898	On Change
Step Counter	10,227,444	Up to 20 times per second
Bluetooth Sensor	14,224,000	Every Second
Running Application	54,788,870	Once every 5 seconds

Table 4.3: Sensors Data

were extracted from this set of sensor data, and the process of feature extraction was discussed in detail in the chapter 4.4.

4.4 Feature Extraction

From the sensors that were available, the analysis was centered on data from Bluetooth, Doze Event, Ring Mode, Touch Event, Music Event, Step Counter, App Usage, Screen State, and Battery Charge Event. Each of these data streams contributed various features for the modeling process. The primary focus of the research was on the daily activities and behaviors of participants and their potential correlations with personality traits. The unit of observation for the extracted features was the person-day.

4.4.1 Big Five Survey

Each participant's Big Five personality traits were derived from a survey that they completed on their phones using the i-Log app. This survey employed a specific type of Five-Factor Model assessment called the Mini-IPIP [10]. The Mini-IPIP is a concise 20-item version of the 50-item International Personality Item Pool—Five-Factor Model measure. In this questionnaire, participants provided responses to 20 self-descriptive statements, rating their agreement on a scale of 1 to 5 (1 - strongly disagree, 2 - somewhat disagree, 3 - neither agree nor disagree, 4 - somewhat agree, 5 - strongly agree). For instance, questions such as "Don't talk a lot?" and "Likes Order?" were included. Using the responses collected from the survey, personality traits were calculated based on the scoring instructions provided in [10]. Prior to our analysis, the values representing these personality traits were normalized to fall within the range of 0 to 100. It's important to note that the Big Five questionnaire was administered only once at the beginning of the study. Here are the survey questions that were answered by the study participants.

1. Am the life of the party.
2. Talk to a lot of different people at parties.
3. Don't talk a lot
4. Keep in the background.
5. Sympathize with others' feelings.
6. Feel others' emotions.
7. Am not really interested in others.
8. Am not interested in other people's problems.
9. Get chores done right away.
10. Like order.
11. Often misplace things.
12. Make a mess of things.
13. Have frequent mood swings.
14. Get upset easily.
15. Am relaxed most of the time.
16. Seldom feel blue.
17. Have a vivid imagination.
18. Struggle with abstract ideas.
19. Am not interested in abstract ideas.
20. Do not have a good imagination.

Along with the Big-Five traits details such as participant's gender and department they are studying in are also collected. Table 4.4 shows a sample of the dataset that contained the BFI scores of all the users along with their gender and department information.

userid	gender	department	E	A	C	N	O
0	Female	Engineering and Applied Sciences	68.75	87.50	93.75	50.00	87.50
1	Female	Humanities	100.00	100.00	31.25	75.00	75.00
2	Male	Law	31.25	56.25	87.50	62.50	81.25
3	Female	Social Sciences	56.25	100.00	56.25	50.00	75.00

Table 4.4: Personality Scores, Gender and Department information on participants. O(Openness), C(Conscientiousness), E(Extraversion), A(Agreeableness), N(Neuroticism)

The department column here contains 9 department names which then was further classified into two main types named STEM and Non-STEM departments. Students who were enrolled in the following departments: Engineering and Applied Sciences, Natural Sciences, Medicine and veterinary medicine, and Agricultural, were categorized as belonging to the STEM department. Conversely, students enrolled in departments such as Social Sciences, Business/economics, Law, Humanities, and International Relations and Public Administration were categorized as non-STEM.

4.4.2 App Usage Logs

A total of 54,788,870.00 records were collected for application (app) usage data, including user IDs, timestamps, and app package names. Throughout the entire study period, 1970 unique applications were used by participants. To streamline the analysis, a decision was made to classify all these applications into app categories, utilizing the application classes from Google's Play Store [8]. Within the Play Store, there are 33 classes available for categorizing applications. These classes include "Game," which contains various subclasses like Arcade, Strategy, and Action. In our approach, all types of games were consolidated under the "Game" category. For example, applications such as Viber, WhatsApp, and Skype were grouped as communication applications. The list of categories used in this classification includes Art and Design, Auto and Vehicles, Beauty, Books and Reference, Business, Communication, Comics, Dating, Education, Entertainment, Events, Finance, Food and Drink, Game, Health and Fitness, House and Home, Libraries and Demo, Lifestyle, Maps and Navigation, Medical, Music and Audio, News and Magazines, Parenting, Personalization, Photography, Productivity, Shopping, Social, Sports, Tools, Travel and Local, Games, Video Players and Editors, and Weather.

Table 4.5 shows a sample of the original app usage dataset. The challenge at hand consisted of two major tasks. Firstly, determining the application names corresponding to the package names, such as identifying "com.miui.home" as the MIUI launcher and "org.telegram.messenger" as Telegram Messenger. Secondly, assigning the appropriate application class to each of these application names. Manually categorizing 1970 package names into application names and then into application categories would have been an overwhelming endeavor. Therefore, the utilization of large language models was employed to achieve this task. GPT-3.5 Turbo Large Language Model was used, with parameters set to a maximum token limit of 10 and a temperature setting of 0. This approach successfully classified all the package names into app names and app categories, resulting in the outcome resembling the sample presented in Table 4.6. Additionally, it should be noted that, all those entries registering usage of i-log application and usage of system apps such as Home Launchers were excluded from our analysis since the logs of these applications are not affected by users' personality traits. Package names that couldn't be associated with specific application names were designated as "Unknown apps" and were subsequently omitted from our analysis. This decision was made due to the inability to ascertain the corresponding applications, rendering the entries uninformative for our purposes.

The data was then grouped based on app categories to obtain the count of apps used by each user in each category daily. Upon averaging this data over the total number of days each user participated in the survey, we obtained the average app category count at the user level. A sample of this transformed data is presented in Table 4.7 and Table 4.8 shows all the extracted features from the original app usage data. So Table 4.7 shows that, on an average, user 0 has used Art & Design apps 0 number of times, Business apps 33.2 times and so on.

userid	timestamp	package_name
0	12/11/2020 12:38:49	it.unitn.disi.witmee.sensorlog
0	12/11/2020 12:38:55	com.miui.home
0	12/11/2020 12:39:00	com.miui.home
0	12/11/2020 12:39:05	com.miui.home
0	12/11/2020 14:42:46	com.google.android.gm

Table 4.5: App Logs

userid	timestamp	App_name	App_category
0	12/11/2020 14:42:46	Gmail	Communication
0	12/11/2020 14:43:01	WhatsApp Messenger	Communication
0	12/11/2020 14:43:06	WhatsApp Messenger	Communication
0	12/11/2020 14:43:11	WhatsApp Messenger	Communication
0	12/11/2020 14:43:16	Telegram	Communication

Table 4.6: Transformed App Log

4.4.3 Bluetooth

A total of 14,224,000 records were collected from the Bluetooth Sensor, and a sample of this data is presented in Table 4.9. In this dataset, the "User ID" and "MAC Address" columns played a crucial role in identifying the devices used by participants. It's important to note that this sensor logged information about all nearby Bluetooth devices, not just those with established connections. This included a wide range of devices such as printers, laptops, cars, and mobile phones. The "namecode" column provided the names of these nearby Bluetooth-enabled devices.

However, it's worth mentioning that each user may name their devices differently, and the names in the "namecode" column can vary depending on the manufacturer. In total, there were 12,461 unique name codes, and the goal was to classify them into more generic categories. For example, "ASUS_Z00AD" might be classified as a smartphone, and "VW BT 5689" could be categorized as a car, and so on. To accomplish this, we relied on the GPT-3.5 large language model. By utilizing a large language model classifier with parameters set to a maximum token limit of 10 and a temperature setting of 0, we labeled the namecode entries to indicate the type of device they represented. Once this classification was completed, we had all the device names labelled as shown in the sample Table 4.10.

Then we further segregated the nearby devices detected by the participants' Bluetooth sensors into two categories: smartphones and non-smartphone devices. Along with that, to account for the time of day element in machine learning predictions, a column named "Interval" was extracted from the timestamp data. This column was created by dividing the entire day into 4-hour intervals, with consideration given to the circadian rhythm of all the participants. It was observed that the majority of people's days typically commenced around 5 in the morning and concluded around 1 am the following day. Based on this pattern, another column was derived to specify whether each entry fell within one of the following intervals: 5 am to 9 am, 9 am to 1 pm, 1 pm to 5 pm, 5 pm to 9 pm, 9 pm to 1 am, or 1 am to 5 am. This additional column allowed for the capture of the time-of-day context

userid	Art and Design	Auto and Vehicles	Beauty	Business	Communication	
0	0.0	0.0	0.0	33.2	222.8	...
1	0.0	0.0	0.0	0.0	6061.6	...
3	16.5	0.0	0.0	0.0	1075.2	...
4	0.0	0.0	0.0	41.5	696.4	...
5	0.0	0.0	0.0	178.4	1348.2	...

Table 4.7: Sample of Transformed App Usage Data (showing only 6 of the 34 columns)

Sensor	Extracted Features
App Data	Art and Design
	Auto and Vehicles
	Beauty
	Books and Reference
	Business
	Communication
	Comics
	Dating
	Education
	Entertainment
	Events
	Finance
	Food and Drink
	Game
	Health and Fitness
	House and Home
	Libraries and Demo
	Lifestyle
	Maps and Navigation
	Medical
	Music and Audio
	News and Magazines
	Parenting
	Personalization
	Photography
	Productivity
	Shopping
	Social
	Sports
	Tools
	Travel and Local
	Games
	Video Players and Editors
	Weather

Table 4.8: Extracted Features from App Data, Shows Average number of times each app category was used

userid	MAC_Address	Timestamp	Contact Freq	namecode
1	24:4B:03:F7:75:41	14/11/2020 16:55	1	BCM20702B0 Generic USB Detuned
1	24:4B:03:F7:75:41	14/11/2020 16:55	2	BCM20702B0 Generic USB Detuned
1	1C:B7:2C:57:5B:12	03/12/2020 09:39	2	ASUS_Z00AD

Table 4.9: Sample Bluetooth Data

namecode	Bluetooth_Device
VW BT 5689	Car
HUAWEI Y5 2018	Smartphone
ID115Plus HR	Fitness Tracker
Samsung 7 Series (55)	TV
LG CJ45 (44)	Speaker

Table 4.10: Labelled Bluetooth Data Sample

userid	date	timestamp	namecode	Device_Type	Interval
1	03/12/2020	03/12/2020 09:39	ID115Plus HR	Non-Smartphone	9am-1pm
1	03/12/2020	03/12/2020 09:40	ID115Plus HR	Non-Smartphone	9am-1pm
1	03/12/2020	03/12/2020 09:41	ASUS_Z00AD	Smartphone	9am-1pm

Table 4.11: Transformed Bluetooth Data

userid	f1	f2	f3	f4	f5	f6
1	4.5	4.0	4.4	3.9	0.0	0.0
2	1.0	3.0	3.4	2.7	1.3	1.0
3	1.4	4.0	1.8	3.7	1.3	1.4

Table 4.12: Final Bluetooth Data Sample, f1 = Average No. of People nearby during 5am to 9am, f2 = Average No. of People nearby during 9am to 1pm, f3 = Average No. of People nearby during 1pm to 5pm, f4 = Average No. of People nearby during 5pm to 9pm, f5 = Average No. of People nearby during 9pm to 1am, f6 = Average No. of People nearby during 1am to 5am

for each data entry, thereby enhancing the precision of machine learning predictions. The resulting transformed dataset is structured as shown in Table 4.11.

To get to a dataset that contains only user level information in each row, we further transformed this dataset. Given the prevalent use of smartphones in today's society, it was reasonable to assume that the number of nearby smartphones detected by a user's smartphone Bluetooth is roughly equivalent to the number of people they are in proximity to. Consequently, we extracted features such as the average number of people nearby and the average number of non-smartphone devices nearby. The two final transformed datasets we ended up with, from the Bluetooth sensor ready for analysis looked like Table 4.12 and Table 4.13.

The extracted features from Bluetooth Sensor Data are listed in Table 4.14.

4.4.4 Ring Mode

A total of 102,540 records were amassed for the ring mode event sensor. Ring mode could be set to Normal, Silent, or Vibrate. The data was found to be clean, devoid of any missing values or outliers. Table 4.15 displays a sample of the original dataset that was collected. Subsequently, the data was aggregated to transition from event level to user level, ensuring its convenience for machine learning algorithms. Initially, for each user, the number of times each ring mode was utilized was calculated for each day. This process yielded data indicating, for each user, the number of times they used the normal mode, silent mode, and vibrate mode day by day. Next, we computed the average of these mode counts across the number of days each user participated in the survey. This resulted in a dataset containing, for each user, the average number of times they employed the normal mode, the average number of times they utilized the silent mode, and the average number of times they utilized the vibrate mode. The resulting dataframe resembled the structure shown in Table 4.16. Table 4.17 gives an idea about the list of features we were able to extract from the original ring mode sensor data.

userid	f7	f8	f9	f10	f11	f12
0	1.0	2.3	3.2	6.5	7.0	2.3
1	2.0	2.2	3.1	6.6	6.6	2.3
2	3.0	2.7	6.1	5.0	8.3	3.5

Table 4.13: Sample of the final Bluetooth Dataset, f7 = Average Non-Smartphone Devices nearby during 5am to 9am, f8 = Average Non-Smartphone Devices nearby during 9am to 1pm, f9 = Average Non-Smartphone Devices nearby during 1pm to 5pm, f10 = Average Non-Smartphone Devices nearby during 5pm to 9pm, f11 = Average Non-Smartphone Devices nearby during 9pm to 1am, f12 = Average Non-Smartphone Devices nearby during 1am to 5am

Sensor	**Extracted Features**
Bluetooth	Average No. of People Nearby During 5AM to 9AM
	Average No. of People Nearby During 9AM to 1PM
	Average No. of People Nearby During 1PM to 5PM
	Average No. of People Nearby During 5PM to 9PM
	Average No. of People Nearby During 9PM to 1AM
	Average No. of People Nearby During 1AM to 5AM
	Average No. of Non-Smartphone Devices Nearby During 5AM to 9AM
	Average No. of Non-Smartphone Devices Nearby During 9AM to 1PM
	Average No. of Non-Smartphone Devices Nearby During 1PM to 5PM
	Average No. of Non-Smartphone Devices Nearby During 5PM to 9PM
	Average No. of Non-Smartphone Devices Nearby During 9PM to 1AM
	Average No. of Non-Smartphone Devices Nearby During 1AM to 5AM

Table 4.14: Extracted Features from Bluetooth Sensor

userid	**timestamp**	**status**
0	13/11/2020 23:00:01	mode_silent
0	14/11/2020 07:00:01	mode_silent
0	14/11/2020 08:13:32	mode_normal
0	14/11/2020 08:13:32	mode_normal
0	14/11/2020 08:13:32	mode_normal

Table 4.15: Sample of the Original Ring Mode Data

userid	**mode_normal**	**mode_silent**	**mode_vibrate**
0	4.44	5.22	2.25
1	10.00	3.60	7.83
3	12.80	15.37	0.00
4	3.33	3.58	3.65

Table 4.16: Transformed Ring Mode Data Sample, Each column represents Average number of times they utilized the corresponding ring mode

Sensor	**Extracted Features**
Ring Mode	Average Number of times Normal Mode used
	Average Number of times Silent Mode used
	Average Number of times Vibrate Mode used

Table 4.17: List of features extracted from the Ring Mode Data

userid	timestamp	value
144	20201117235956100	16716
144	20201117235955700	16716
144	20201117235945100	16716
144	20201117235938100	16716
144	20201117235933100	16716

Table 4.18: Original Step Counter Data Sample

userid	f1	f2	f3	f4	f5	f6
0	3.2	24.7	63.6	68.3	78.2	3.7
2	1500.1	1858.2	2770.7	3377.4	3042.0	6.0
3	10.8	46.1	65.5	142.5	152.2	6.2
4	242.1	717.3	1225.3	1893.4	1783.7	19.5
6	212.2	570.7	940.9	1774.6	2148.9	51.1

Table 4.19: Sample of the transformed data for Step Counter, f1 = Average Number of Steps Taken between 5am to 9am, f2 = Average Number of Steps Taken between 9am to 1pm, f3 = Average Number of Steps Taken between 1pm to 5pm, f4 = Number of Steps Taken between 5pm to 9pm, f5 = Average Number of Steps Taken between 9pm to 1am, f6 = Average Number of Steps Taken between 1am to 5am

4.4.5 Step Counter

A total of 10,227,444 records were gathered from the step counter. In Table 4.18, the top few rows of the original dataset collected are presented. To consider the time-of-day element in machine learning predictions, a column called ”Interval” was extracted from the timestamp data. This column was generated by dividing the entire day into 4-hour intervals, taking into account the circadian rhythm of all the participants. It was noted that the majority of people’s days typically commenced around 5 in the morning and concluded around 1 am the following day. Based on this pattern, another column was created to indicate whether each entry fell within one of the following intervals: 5 am to 9 am, 9 am to 1 pm, 1 pm to 5 pm, 5 pm to 9 pm, 9 pm to 1 am, or 1 am to 5 am. This additional column facilitated the capture of time-of-day context for each data entry.

Subsequently, a group-by operation was performed over the time intervals to determine the total steps taken each day during each interval. Then, another aggregation was conducted, considering the number of days each participant participated in the survey, in order to calculate the average number of steps taken by each user within a specific time interval. Once this transformation was completed, the resulting dataset exhibited a structure akin to the sample displayed in Table 4.19. Table 4.20 shows the list of features we were able to extract from the Step counter data.

4.4.6 Touch Event

A total of 1,235,536 records were collected from the touch event sensor. In Table 4.21, the top few rows of the original dataset are presented. In the touch event data, numerous outliers were identified. Some users exhibited exceptionally high numbers of touch events, with per-second clicks exceeding 100, a

Sensor	Extracted Features
Step Counter	Average Steps taken during 5AM to 9AM
	Average Steps taken during 9AM to 1PM
	Average Steps taken during 1PM to 5PM
	Average Steps taken during 5PM to 9PM
	Average Steps taken during 9PM to 1AM
	Average Steps taken during 1AM to 5AM

Table 4.20: List of Extracted Features from Step Counter

practically implausible scenario. To address this, a threshold of 6 clicks per second was established, as 6 touch events per second is a common occurrence for heavy users such as smartphone gamers. Any entry recorded above this threshold of 6 touch events per second was substituted with the median value from the entire dataset. To incorporate the time-of-day element into machine learning predictions, a column labeled "Interval" was derived from the timestamp data. This column was generated by dividing the entire day into 4-hour intervals, following the same methodology applied to previous sensors. This additional column allowed for the capture of time-of-day context for each data entry. Subsequently, an aggregation operation was conducted over the time intervals to determine the total number of touch events that occurred each day within each interval. Following that, another aggregation was performed, taking into account the number of days each participant participated in the survey, in order to calculate the average number of touch events recorded by each user during a specific time interval. Once this transformation was finalized, the resulting dataset exhibited a structure similar to the sample depicted in Table 4.22. Table 4.23 lists all the features which have been extracted successfully from the Touch Event data.

userid	**Month**	**Day**	**hour**	**min**	**touch**	**date**
1	11	12	11	17	23	12/11/2020
1	11	12	11	18	2	12/11/2020
1	11	12	11	21	5	12/11/2020
1	11	12	11	22	72	12/11/2020
1	11	12	11	23	53	12/11/2020

Table 4.21: Sample Touch Event Data

userid	**f1**	**f2**	**f3**	**f4**	**f5**	**f6**
1	861.3	6314	5557.2	5481	4797.5	859.8
2	3591.1	4700.4	7135.4	4254.1	5576.9	0
3	1015.9	7230.1	6582.1	4417.6	3762.6	1466
4	1007.4	1838.2	2052.6	1817.5	1721.6	199.7

Table 4.22: Sample of the transformed data for Touch Event Sensor, f1 = Average No. of Touch Events Registered between 5am to 9am, f2 = Average No. of Touch Events Registered between 9am to 1pm, f3 = Average No. of Touch Events Registered between 1pm to 5pm, f4 = Average No. of Touch Events Registered between 5pm to 9pm, f5 = Average No. of Touch Events Registered between 9pm to 1am, f6 = Average No. of Touch Events Registered between 1am to 5am

Sensor	**Extracted Features**
Touch Event	Average no of Touch Events Registered between 5AM to 9AM
	Average no of Touch Events Registered between 9AM to 1PM
	Average no of Touch Events Registered between 1PM to 5PM
	Average no of Touch Events Registered between 5PM to 9PM
	Average no of Touch Events Registered between 9PM to 1AM
	Average no of Touch Events Registered between 1AM to 5AM

Table 4.23: List of Features Extracted from Touch Event Sensor

4.4.7 Music Event

Table 4.24 displays a sample of the music event dataset. The data was devoid of any outliers or missing values, simplifying the task at hand. To incorporate the time of day into our analysis, the entire day was divided into 4-hour time intervals, and each entry in the data was labeled with its corresponding time interval. Utilizing a group-by operation, the total number of music events recorded for each interval was determined daily. Another group-by operation provided the average music event count recorded for each time interval, spanning from 5 AM to 9 AM and so forth. Table 4.25 exhibits a sample of the data extracted from the original dataset, while Table 4.26 enumerates the list of features derived from the music event sensor.

userid	timestamp	status
4	15/11/2020 12:22:43	TRUE
4	15/11/2020 12:22:43	TRUE
4	16/11/2020 07:25:18	TRUE

Table 4.24: Sample of the original Music Event Data

userid	f1	f2	f3	f4	f5	f6
4	2	5	2	2.3	4	0
13	27	39.1	17	68	43	0
15	61	217.6	123.6	270	196.7	112.5
19	130.6	560.9	788.7	638	46	0

Table 4.25: Sample of the Transformed Music Event Data, f1 = Average Number of Music Events Recorded During 5am to 9am, f2 = Average Number of Music Events Recorded During 9am to 1pm, f3 = Average Number of Music Events Recorded During 1pm to 5pm, f4 = Average Number of Music Events Recorded During 5pm to 9pm, f5 = Average Number of Music Events Recorded During 9pm to 1am, f6 = Average Number of Music Events Recorded During 1am to 5am

Sensor	Extracted Features
Music Event	Average Number of Music Events Recorded During 5am to 9am
	Average Number of Music Events Recorded During 9am to 1pm
	Average Number of Music Events Recorded During 1pm to 5pm
	Average Number of Music Events Recorded During 5pm to 9pm
	Average Number of Music Events Recorded During 9pm to 1am
	Average Number of Music Events Recorded During 1am to 5am

Table 4.26: List of features extracted from the Music Event Sensor

4.4.8 Screen Event

Table 4.27 presents a sample of the screen event dataset. The data was free of any outliers or missing values. To incorporate the time of day into our analysis, the entire day was segmented into 4-hour time intervals, and each entry in the data was tagged with its corresponding time interval. Through the utilization of a group-by operation, the total number of screen events recorded for each interval was computed on a daily basis. An additional group-by operation yielded the average screen event count registered for each time interval, encompassing intervals from 5 AM to 9 AM and so on. Table 4.28 showcases a sample of the data extracted from the original dataset, while Table 4.29 lists all the features derived from the screen event sensor.

userid	timestamp	status
0	12/11/2020 15:08:17	SCREEN_ON
0	12/11/2020 15:09:11	SCREEN_ON
0	12/11/2020 15:12:15	SCREEN_ON
0	12/11/2020 15:13:33	SCREEN_ON

Table 4.27: Sample of the original Screen Event Data

userid	f1	f2	f3	f4	f5	f6
0	8	36.7	23.1	29.8	17.3	3.8
1	94.5	371.8	308.3	261.8	129.5	19.3
2	229.7	363.5	433.9	295	154.1	113
3	56.9	216	205.6	157.8	118.6	20.3
4	44.1	82.8	118.5	108.6	42.5	4.2

Table 4.28: Sample of the Transformed Screen Event Data, f1 = Average Screen On Event During 5am to 9am, f2 = Average Screen On Event During 9am to 1pm, f3 = Average Screen On Event During 1pm to 5pm, f4 = Average Screen On Event During 5pm to 9pm, f5 = Average Screen On Event During 9pm to 1am, f6 = Average Screen On Event During 1am to 5am

Sensor	Extracted Features
Screen Event	Average Number of Screen Events Recorded During 5am to 9am
	Average Number of Screen Events Recorded During 9am to 1pm
	Average Number of Screen Events Recorded During 1pm to 5pm
	Average Number of Screen Events Recorded During 5pm to 9pm
	Average Number of Screen Events Recorded During 9pm to 1am
	Average Number of Screen Events Recorded During 1am to 5am

Table 4.29: List of all features extracted from the Screen Event Data

4.4.9 Battery Charge Event

Table 4.30 presents a sample of the battery charge event dataset, which was already clean. To incorporate the time of day into our analysis, the entire day was segmented into 4-hour time intervals, and each entry in the data was tagged with its corresponding time interval, mirroring the approach used for other sensors. Leveraging a group-by operation, the total number of charge events recorded for each interval was calculated daily. Another group-by operation furnished the average charge event count registered for each time interval, covering intervals from 5 AM to 9 AM and so on. Table 4.31 showcases a sample of the data extracted from the original dataset, while Table 4.32 enumerates the features derived from the charge event sensor.

userid	timestamp	source	status
0	15/11/2020 20:02:26	charging_ac	TRUE
0	17/11/2020 17:38:26	charging_unknown	TRUE
0	17/11/2020 20:12:19	charging_ac	TRUE
0	17/11/2020 20:12:19	charging_ac	TRUE

Table 4.30: Sample of the Charge Event Data

userid	f1	f2	f3	f4	f5	f6
0	0.0	1.0	4.0	2.0	1.5	0.0
1	38.0	16.2	6.6	12.5	10.2	0.0
2	38.0	218.1	162.8	174.6	130.3	113.0

Table 4.31: Sample of the Transformed Charge Event Data, f1 = Average Number of Charge Events Recorded During 5am to 9am, f2 = Average Number of Charge Events Recorded During 9am to 1pm, f3 = Average Number of Charge Events Recorded During 1pm to 5pm, f4 = Average Number of Charge Events Recorded During 5pm to 9pm, f5 = Average Number of Charge Events Recorded During 9pm to 1am, f6 = Average Number of Charge Events Recorded During 1am to 5am

Sensor	Extracted Features
Charge Event	Average Number of Charge Events Recorded During 5am to 9am
	Average Number of Charge Events Recorded During 9am to 1pm
	Average Number of Charge Events Recorded During 1pm to 5pm
	Average Number of Charge Events Recorded During 5pm to 9pm
	Average Number of Charge Events Recorded During 9pm to 1am
	Average Number of Charge Events Recorded During 1am to 5am

Table 4.32: List of Features Extracted from Battery Charge Event Data

4.4.10 Doze Event

Table 4.33 presents a sample from the doze event dataset, which was a clean and well-structured dataset. To integrate the time of day into our analysis, we divided the entire day into 4-hour time segments and assigned each data entry to the respective time segment, similar to our approach with other sensors. By using a group-by operation, we calculated the daily total count of doze events for each time segment. Additionally, another group-by operation allowed us to determine the average doze event count for specific time intervals, such as 5 AM to 9 AM, 9 AM to 1 PM, and so on. Table 4.34 showcases a sample of data extracted from the original dataset, while Table 4.35 lists the features derived from the doze event sensor.

userid	timestamp	status
0	12/11/2020 18:43:09	TRUE
0	13/11/2020 22:44:11	TRUE
0	14/11/2020 00:18:01	TRUE
0	14/11/2020 01:08:57	TRUE

Table 4.33: Sample of the Doze Event Data

userid	f1	f2	f3	f4	f5	f6
0	2.1	1.3	1.5	2	2.2	3.5
1	13.8	99.6	68.7	45.2	18.9	13.3
3	4.8	1.5	4	3.4	2.3	5.9

Table 4.34: Sample of the Transformed Doze Event Data, f1 = Average Number of Doze Events Recorded During 5am to 9am, f2 = Average Number of Doze Events Recorded During 9am to 1pm, f3 = Average Number of Doze Events Recorded During 1pm to 5pm, f4 = Average Number of Doze Events Recorded During 5pm to 9pm, f5 = Average Number of Doze Events Recorded During 9pm to 1am, f6 = Average Number of Doze Events Recorded During 1am to 5am

Sensor	Extracted Features
Doze Event	Average Number of Doze Events Recorded During 5am to 9am
	Average Number of Doze Events Recorded During 9am to 1pm
	Average Number of Doze Events Recorded During 1pm to 5pm
	Average Number of Doze Events Recorded During 5pm to 9pm
	Average Number of Doze Events Recorded During 9pm to 1am
	Average Number of Doze Events Recorded During 1am to 5am

Table 4.35: List of Features Extracted from Doze Event Data

Additionally, it should be noted that for all sensor data, feature extraction was performed after excluding users who participated in the survey for less than 7 days. This exclusion was made because such users were unlikely to provide valuable data that could demonstrate their regular patterns of smartphone usage. A consolidated list of all the features we extracted for our study from all nine sensors has been shared in Table 4.36

Sensor	Extracted Features
App Logs	Average No. of Times Art and Design app was used
	Average No. of Times Auto and Vehicles app was used
	Average No. of Times Beauty app was used
	Average No. of Times Books and Reference app was used
	Average No. of Times Business app was used
	Average No. of Times Communication app was used
	Average No. of Times Comics app was used
	Average No. of Times Dating app was used
	Average No. of Times Education app was used
	Average No. of Times Entertainment app was used
	Average No. of Times Events app was used
	Average No. of Times Finance app was used
	Average No. of Times Food and Drink app was used
	Average No. of Times Game app was used
	Average No. of Times Health and Fitness app was used
	Average No. of Times House and Home app was used
	Average No. of Times Libraries and Demo app was used
	Average No. of Times Lifestyle app was used
	Average No. of Times Maps and Navigation app was used
	Average No. of Times Medical app was used
	Average No. of Times Music and Audio app was used
	Average No. of Times News and Magazines app was used
	Average No. of Times Parenting app was used
	Average No. of Times Personalization app was used
	Average No. of Times Photography app was used
	Average No. of Times Productivity app was used
	Average No. of Times Shopping app was used
	Average No. of Times Social app was used
	Average No. of Times Sports app was used
	Average No. of Times Tools app was used
	Average No. of Times Travel and Local app was used
	Average No. of Times Games app was used
	Average No. of Times Video Players and Editors app was used
	Average No. of Times Weather app was used
Bluetooth	Average no of people nearby between 5AM to 9AM
	Average no of people nearby between 9AM to 1PM
	Average no of people nearby between 1PM to 5PM
	Average no of people nearby between 5PM to 9PM
	Average no of people nearby between 9PM to 1AM
	Average no of people nearby between 1AM to 5AM
	Average no of Non-Smartphone Devices nearby between 5AM to 9AM
	Average no of Non-Smartphone Devices nearby between 9AM to 1PM
	Average no of Non-Smartphone Devices nearby between 1PM to 5PM
	Average no of Non-Smartphone Devices nearby between 5PM to 9PM
	Average no of Non-Smartphone Devices nearby between 9PM to 1AM
	Average no of Non-Smartphone Devices nearby between 1AM to 5AM
Ring Mode	Average Number of time Normal Mode used

Continued on the next page

Sensor	Extracted Features
	Average Number of time Silent Mode used
	Average Number of time Vibrate Mode used
Step Counter	Average Steps taken during 5AM to 9AM
	Average Steps taken during 9AM to 1PM
	Average Steps taken during 1PM to 5PM
	Average Steps taken during 5PM to 9PM
	Average Steps taken during 9PM to 1AM
	Average Steps taken during 1AM to 5AM
Touch Event	Average no of Touch Events Registered between 5AM to 9AM
	Average no of Touch Events Registered between 9AM to 1PM
	Average no of Touch Events Registered between 1PM to 5PM
	Average no of Touch Events Registered between 5PM to 9PM
	Average no of Touch Events Registered between 9PM to 1AM
	Average no of Touch Events Registered between 1AM to 5AM
Music Event	Average Number of Music Events Recorded During 5am to 9am
	Average Number of Music Events Recorded During 9am to 1pm
	Average Number of Music Events Recorded During 1pm to 5pm
	Average Number of Music Events Recorded During 5pm to 9pm
	Average Number of Music Events Recorded During 9pm to 1am
	Average Number of Music Events Recorded During 1am to 5am
Screen Event	Average Number of Screen Events Recorded During 5am to 9am
	Average Number of Screen Events Recorded During 9am to 1pm
	Average Number of Screen Events Recorded During 1pm to 5pm
	Average Number of Screen Events Recorded During 5pm to 9pm
	Average Number of Screen Events Recorded During 9pm to 1am
	Average Number of Screen Events Recorded During 1am to 5am
Charge Event	Average Number of Charge Events Recorded During 5am to 9am
	Average Number of Charge Events Recorded During 9am to 1pm
	Average Number of Charge Events Recorded During 1pm to 5pm
	Average Number of Charge Events Recorded During 5pm to 9pm
	Average Number of Charge Events Recorded During 9pm to 1am
	Average Number of Charge Events Recorded During 1am to 5am
Doze Event	Average Number of Doze Events Recorded During 5am to 9am
	Average Number of Doze Events Recorded During 9am to 1pm
	Average Number of Doze Events Recorded During 1pm to 5pm
	Average Number of Doze Events Recorded During 5pm to 9pm
	Average Number of Doze Events Recorded During 9pm to 1am
	Average Number of Doze Events Recorded During 1am to 5am

Table 4.36: List of All Extracted Features along with their respective Sensors

4.5 Exploratory Data Analysis

In this section, uni-variate and multivariate analyses of all independent and dependent variables are presented to aid in establishing the dataset's structure and characteristics.

4.5.1 Univariate Analysis

Independent Variables - Input Features/Extracted Features

Figure 4.2 to Figure 4.5 depict the distributions of all extracted input features across six time intervals, based on our participants' circadian clock. In subplots, of Figure 4.2 to Figure 4.5, the x-axis is used to represent the time intervals, while the y-axis displays the aggregate value. Several observations can be made based on our data for all the users:

The interval between 5 PM and 9 PM shows the highest number of steps taken by participants indicated by Figure 4.2a. The time interval between 1 AM and 5 AM exhibits the lowest step count, which is reasonable given that this is when most users are asleep. Figure 4.2b indicates that, except during late night and early morning hours, the average touch event count remains relatively consistent throughout the day for all users. A similar observation can be made from Figure 4.2c, which indicates that all participants charge their smartphones consistently across all time intervals. Figure 4.2d suggests that the highest number of doze events are registered during the afternoon, potentially due to increased smartphone usage during daytime.

Figure 4.3 reveals that afternoon is the time when most users listen to music. In contrast, Figure 4.4 shows that the average screen event count increases as the day progresses, peaking in the evening before gradually decreasing at night. This pattern might be associated with students returning home after attending university, providing them with more free time to interact with their smartphones.

Figure 4.5a demonstrates the distribution of the average smartphone count in close proximity. The highest number of smartphones in close proximity is observed between 9 AM and 5 PM, which is understandable, as students are usually surrounded by other students while at the university or during their commute. Figure 4.5b is dedicated to the average count of non-smartphone devices in close proximity, which peaks in the evening between 5 PM and 9 PM. This behavior may be attributed to students returning to their residences in the evening and using various Bluetooth devices like speakers, headphones, fitness trackers, household appliances such as TV, printer and other Home Automation Devices.

Figure 4.6a displays categories of applications with their respective usage percentages. Communication apps and social apps are the most frequently used categories, followed by tools, video players, and games. Figure 4.6b exhibits the percentage distribution of the population concerning the frequency of ring mode use. Normal Mode is the most commonly used, while Vibrate Mode is the least utilized.

Please note that these distributions encompass all users, and behaviors may vary within user subgroups, such as those related to specific personalities. To comprehend personality-specific behaviors, we explored multivariate analysis and is discussed in Chapter 4.5.2.

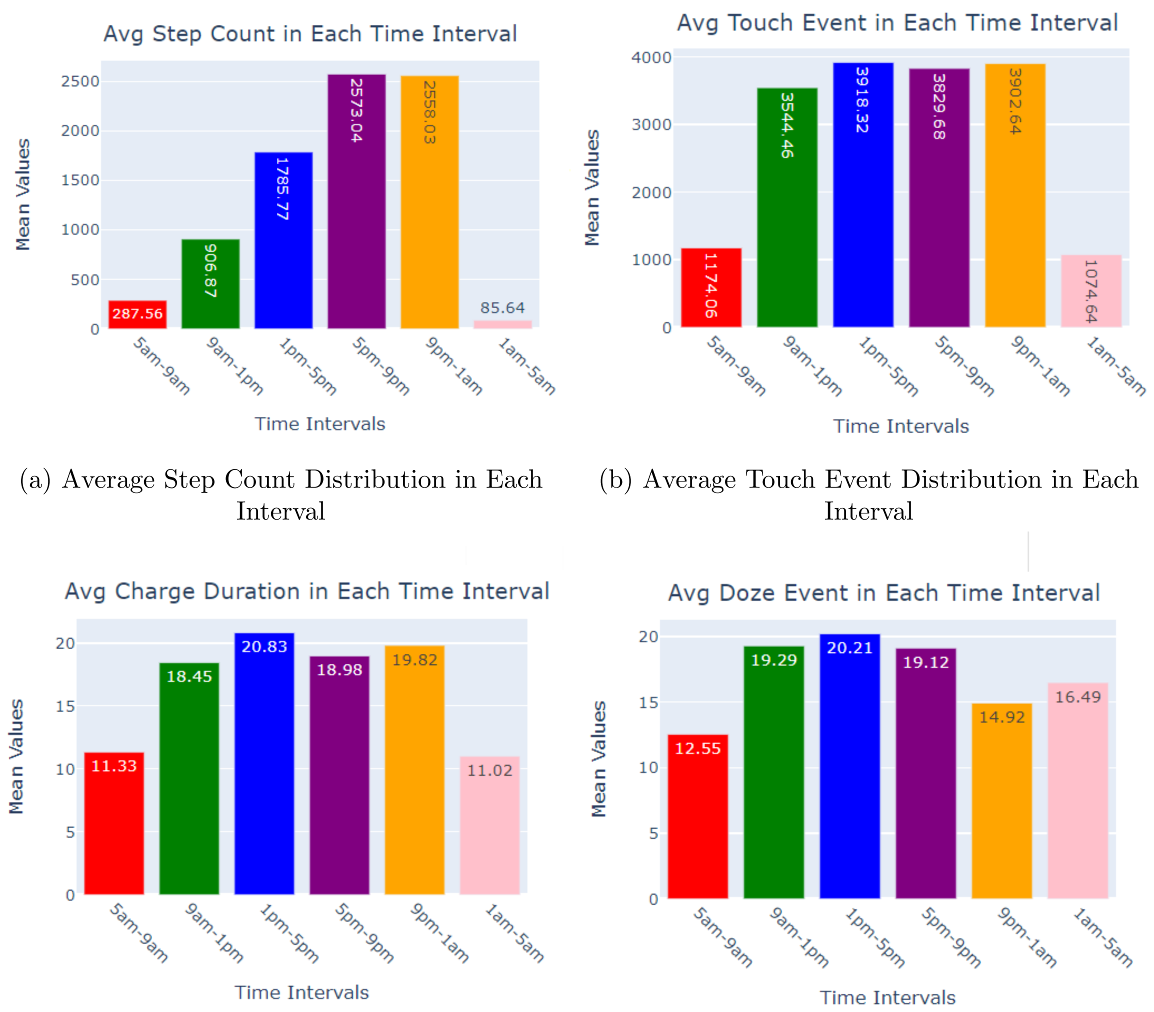
Dependent Variables - Big Five Traits

Figure 4.7 and Table 4.37 depict the histogram distributions and statistics for Big Five personality traits. As observed, Openness, Conscientiousness, and Agreeableness exhibit higher values, which aligns with expectations since the traits were gathered from university students known for their receptiveness to new learning experiences, social interactions, and adherence to class schedules. Extraversion, on the other hand, demonstrates a wide range of values, reflecting the presence of participants with varying levels of extraversion. Notably, Neuroticism and Extraversion exhibit notably high Standard Deviations compared to the other three traits, signifying a broader dispersion of data values across a wide range. It's worth noting that the statistics for Openness, Extraversion, and Neuroticism resemble those reported in previous personality studies conducted on diverse populations [15]. From Figure 4.7, it becomes evident that none of the distributions exhibit a Gaussian, or normal, nature. Instead, there is a noticeable presence of slight skewness in all of the distributions.

4.5.2 Multivariate Analysis

Pearson's correlations, scatter plots, 3-dimensional plots, and ANOVA tests were employed to ascertain the existence of linear or ordinal relationships between the measured features and Big Five personality traits. Hereby, the correlation results are expounded upon.

In the course of various sensor analyses, distinct groups were established to examine whether personality played a role in distinguishing between these groups. As an illustration, when considering step counter data, users were categorized as either fast walkers or slow walkers based on their average steps taken per hour. Consequently, an individual who took, for instance, 10,000 steps within an hour was designated as a fast walker, while another individual who accumulated 11,000 steps throughout



(a) Average Step Count Distribution in Each Interval

(b) Average Touch Event Distribution in Each Interval

(c) Average Charge Event Distribution in Each Interval

(d) Average Doze Event Distribution in Each Interval

Figure 4.2: Histograms of Extracted Features

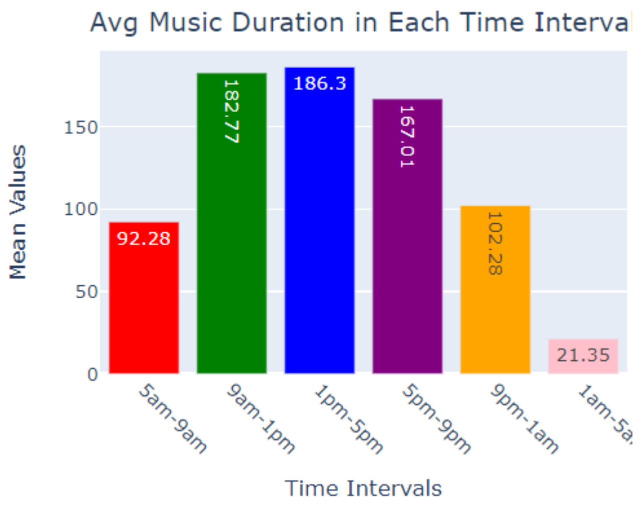


Figure 4.3: Average Music Event in Each Interval

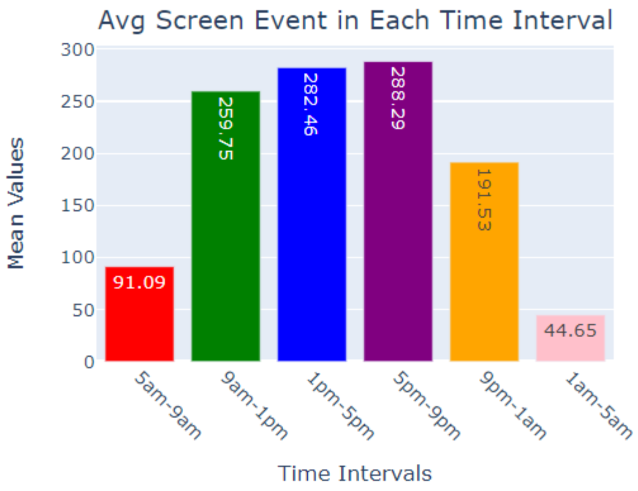
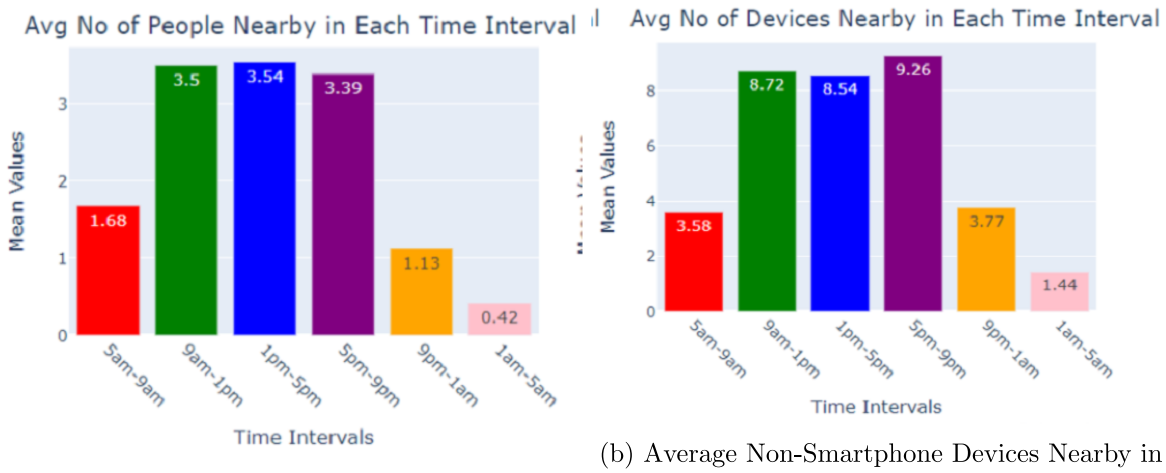


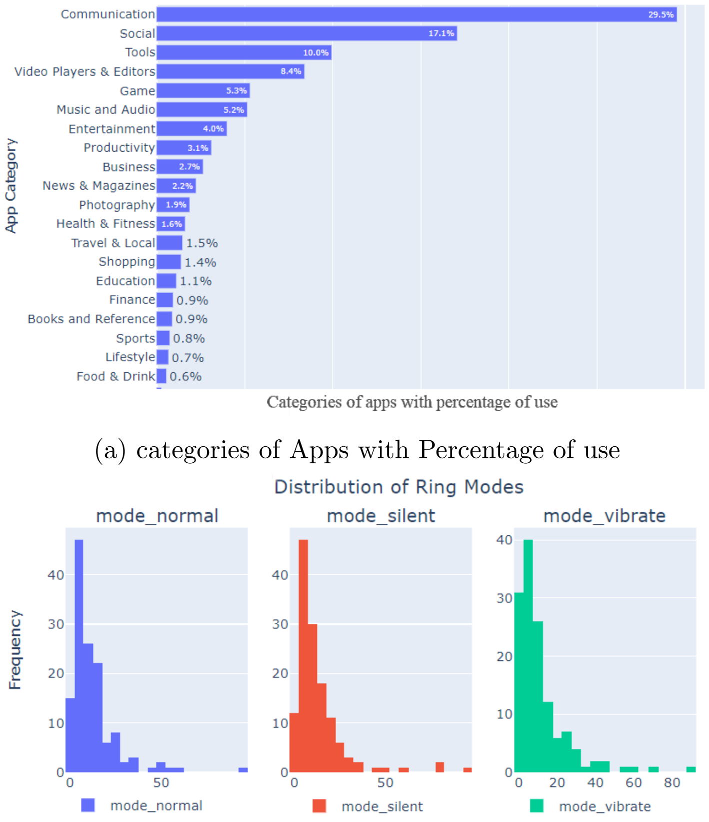
Figure 4.4: Average Screen Event Count in Each Interval



(a) Average No. of people nearby in Each interval

(b) Average Non-Smartphone Devices Nearby in Each Interval

Figure 4.5: Histograms of Extracted Features



(a) categories of Apps with Percentage of use

(b) Distribution of different ring modes by frequency

Figure 4.6: Histograms of Extracted Features

Predictors	mean	std	median	min	max
Total Population					
Extraversion	49.15	23.69	50.00	0.00	100
Agreeableness	76.26	15.56	75.00	25.00	100
Conscientiousness	64.88	19.27	62.50	12.50	100
Neuroticism	49.70	20.70	50.00	0.00	100
Openness	71.46	18.66	75.00	6.25	100
Female Population					
Extraversion	48.72	22.76	50.00	0.00	100
Agreeableness	80.15	14.27	81.25	25.00	100
Conscientiousness	63.69	20.40	62.50	12.50	100
Neuroticism	54.49	19.90	53.13	0.00	100
Openness	70.42	20.37	75.00	6.25	100
Male Population					
Extraversion	49.71	24.98	50.00	0.00	100
Agreeableness	71.09	15.77	75.00	31.25	100
Conscientiousness	66.47	17.62	68.75	18.75	100
Neuroticism	43.34	20.10	43.75	0.00	100
Openness	72.84	16.11	75.00	31.25	100

Table 4.37: Statistics for Dependent variables

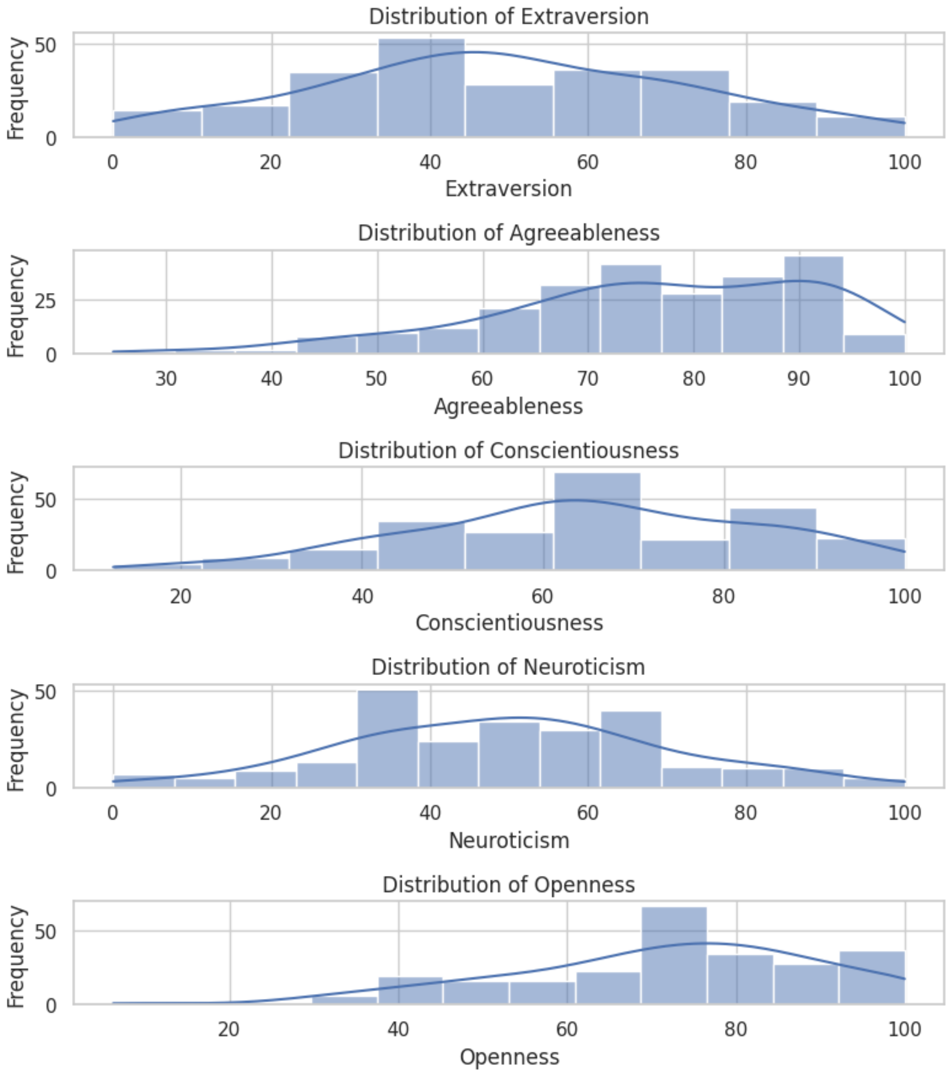


Figure 4.7: Distributions of Dependent Variables

ANOVA Test	Fast Walker		Slow Walker		**f-statistics**	**p-value**
Trait	Mean	std	Mean	std		
Extraversion	46	26.1	31.2	21.5	421953	0
Agreeableness	72.4	19.4	50.7	23.8	805800	0
Conscientiousness	70.9	13.1	74.6	15	58822	0
Neuroticism	46.9	22.2	62.3	19.8	551402	0
Openness	64.6	16.1	86.1	19.3	1201401	0

Table 4.38: ANOVA Test between fast and slow walkers

Personality	Average Ring Mode Count Per Day	r	p-value
Conscientiousness	normal	0.25	0.00
Conscientiousness	silent	0.20	0.02
Conscientiousness	vibrate	0.21	0.02
Neuroticism	vibrate	-0.18	0.03

Table 4.39: Pearson Correlation between Ring Mode and Personality

a day was labeled as a slow walker, even if the latter recorded more steps in a day. Subsequently, when comparing the mean personality scores, statistically significant differences were discerned across all personality traits. An ANOVA test was conducted for this purpose, yielding the most substantial difference in the case of openness, characterized by an F-statistic of 1,201,401 and a p-value of 0.00. These outcomes are documented in Table 4.38.

For ring mode data, a Pearson's correlation analysis was executed, correlating the Average value of each ring mode event per day for all users with their respective personality scores. This analysis revealed significant correlations between all three ring modes and conscientiousness, as well as a correlation between neuroticism and the "ring mode vibrate". The result is denoted in Table 4.39. These findings align with the results reported by [6].

Table 4.40 displays the outcomes of several Pearson's correlation tests conducted between all the extracted features measured throughout the time intervals of the day and the users' personality scores. Given the small sample size of 149 users in our dataset, a decision was made to employ a significance threshold of 0.1, as opposed to the conventional 0.05.

Results marked with a "*" indicate statistically significant correlations and only those features are mentioned in the table for which statistically significant correlation was found with at least one personality trait. Notably, Average step count appears to be higher for introverts consistently throughout the day, while disagreeable individuals tend to walk more during the latter part of the day. Additionally, conscientious individuals demonstrate a preference for listening to music during the morning hours, whereas neurotic individuals tend to do so in the evening, particularly between 5 pm and 9 pm. This observation aligns with findings from prior research, such as [6]. Another noteworthy observation is that individuals scoring high in conscientiousness exhibit increased smartphone activity throughout the day, as evidenced by the strong positive correlation between conscientiousness and average screen event data. This finding corroborates the results of earlier research conducted by [24]. Regarding Bluetooth sensor data, the dataset was initially divided into users with nearby smartphones and those with nearby non-smartphone devices. Given the prevalent use of smartphones in today's society, it is reasonable to assume that the number of nearby smartphones detected by a user's smartphone Bluetooth is roughly equivalent to the number of people they are in proximity to. Consequently, we extracted features such as the average number of people nearby and the average number of non-smartphone devices nearby. Subsequently, a correlation test was conducted between these features and the BFI scores, revealing an intriguing pattern: neurotic individuals tend to be in the vicinity of more people during the morning hours. This observation also aligns with one of the findings from [7, 33].

Furthermore, Figure 4.8 through Figure 4.15 illustrate scatter plots depicting the relationships between the extracted features measured at regular intervals throughout the day and personality

Pearson's Correlation	E	A	C	N	O
Avg Step Count in 5am-9am	-0.15*	-0.04	-0.04	-0.03	-0.10
Avg Step Count in 9am-1pm	-0.18*	-0.06	0.07	-0.04	-0.08
Avg Step Count in 1pm-5pm	-0.18*	-0.11	0.09	-0.06	-0.09
Avg Step Count in 5pm-9pm	-0.19*	-0.13*	0.06	-0.03	-0.10
Avg Step Count in 9pm-1am	-0.17*	-0.15*	0.09	-0.03	-0.10
Avg Step Count in 1am-5am	-0.10	-0.15*	-0.07	-0.06	0.06
Avg Music Event in 5am-9am	-0.26*	-0.08	0.10	0.06	-0.01
Avg Music Event in 9am-1pm	0.11	0.10	0.21*	-0.14	-0.02
Avg Music Event in 5pm-9pm	-0.14	-0.07	-0.17	0.27*	-0.18
Avg Screen Event in 5am-9am	-0.03	0.06	0.18*	-0.01	-0.05
Avg Screen Event in 9am-1pm	-0.03	0.02	0.15*	0.02	-0.07
Avg Screen Event in 1pm-5pm	-0.02	0.02	0.16*	-0.01	-0.06
Avg Screen Event in 5pm-9pm	-0.02	0.04	0.17*	-0.02	-0.08
Avg Screen Event in 9pm-1am	-0.06	-0.01	0.12*	-0.04	-0.07
Avg Doze Event in 5am-9am	0.06	0.05	0.07	-0.16*	-0.03
Avg Doze Event in 9am-1pm	0.14*	0.12	0.12	-0.16*	-0.02
Avg Doze Event in 1pm-5pm	0.12	0.11	0.11	-0.17*	-0.01
Avg Doze Event in 5pm-9pm	0.10	0.09	0.07	-0.16*	-0.01
Avg Charge Event in 9pm-1am	-0.01	0.02	0.17*	0.09	0.00
Avg Charge Event in 1am-5pm	-0.10	-0.16*	0.06	0.12*	-0.05
Avg Touch Event in 5am-9am	-0.08	-0.05	0.09	0.02	-0.20*
Avg Touch Event in 1am-5pm	0.04	0.06	-0.02	-0.12*	0.11
Avg No of People Nearby in 5am-9am	-0.10	0.06	0.02	0.22*	-0.11
Avg No of People Nearby in 1am-5am	0.02	0.17*	-0.17*	0.03	0.02

Table 4.40: Pearson's Correlation between Extracted Features and Big Five traits. A * here represents p value < 0.1.

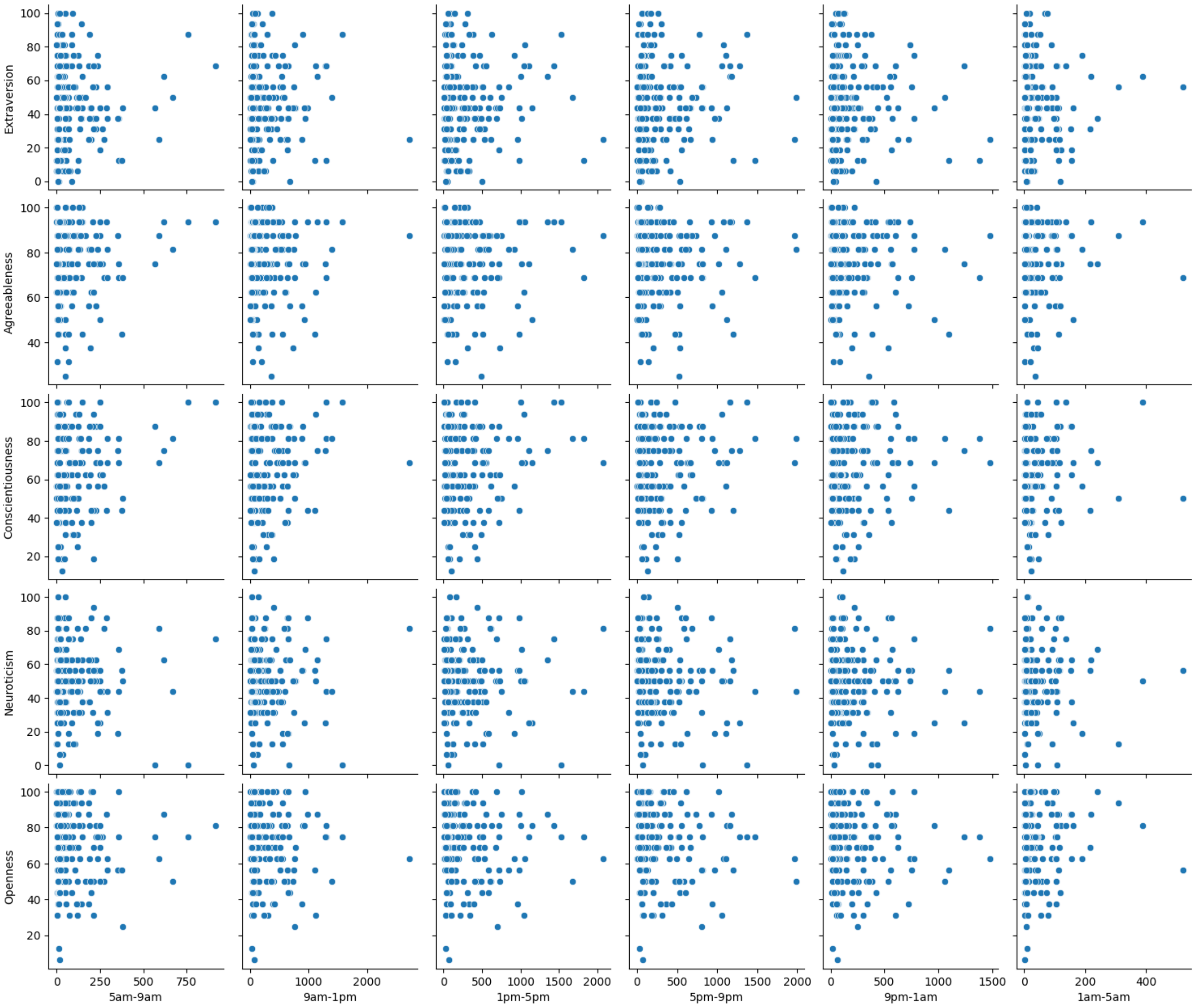


Figure 4.8: Scatter Plot between Average Screen Event in each time interval and Big-Five traits.

traits.

Table 4.39 shows the strong correlation that exists between normal ring mode and conscientiousness. This result aligns with the findings of [6].

When we conducted a Pearson's correlation test to examine the relationship between the average daily use of various application categories and users' personality trait scores, we uncovered intriguing findings, as detailed in Table 4.41. Our analysis revealed that extroverted individuals tend to utilize communication, finance, maps and navigation, and personalization apps more frequently. In contrast, introverts display a greater inclination towards using game apps, reflecting their preference for solitary activities. This finding is aligned with previous literature [38]. Highly agreeable individuals are more likely to engage with travel and local apps, while exhibiting a reduced preference for news and video player applications. For conscientious individuals, there is a proclivity to use events and sports apps, possibly influenced by their orderly disposition, and a decreased preference for maps, navigation, and health and fitness apps. Emotionally stable individuals tend to use beauty apps less frequently but display a greater affinity for finance and video apps. Open-minded individuals are inclined to employ art and design apps more frequently, in line with their openness to new experiences and learning. Moreover, open individuals are also more likely to engage with personalization and social apps, suggesting their propensity for creative customization of smartphone interfaces, indicative of their creative dimension.

Additionally, Table 4.42 presents Pearson's correlation coefficients among the Big Five traits themselves.

Figure 4.16 presents a 3D plot illustrating the relationship between Average Music Event Count, Hour of the Day, and Personality Score. Meanwhile, Figure 4.17 displays a 3D plot representing the

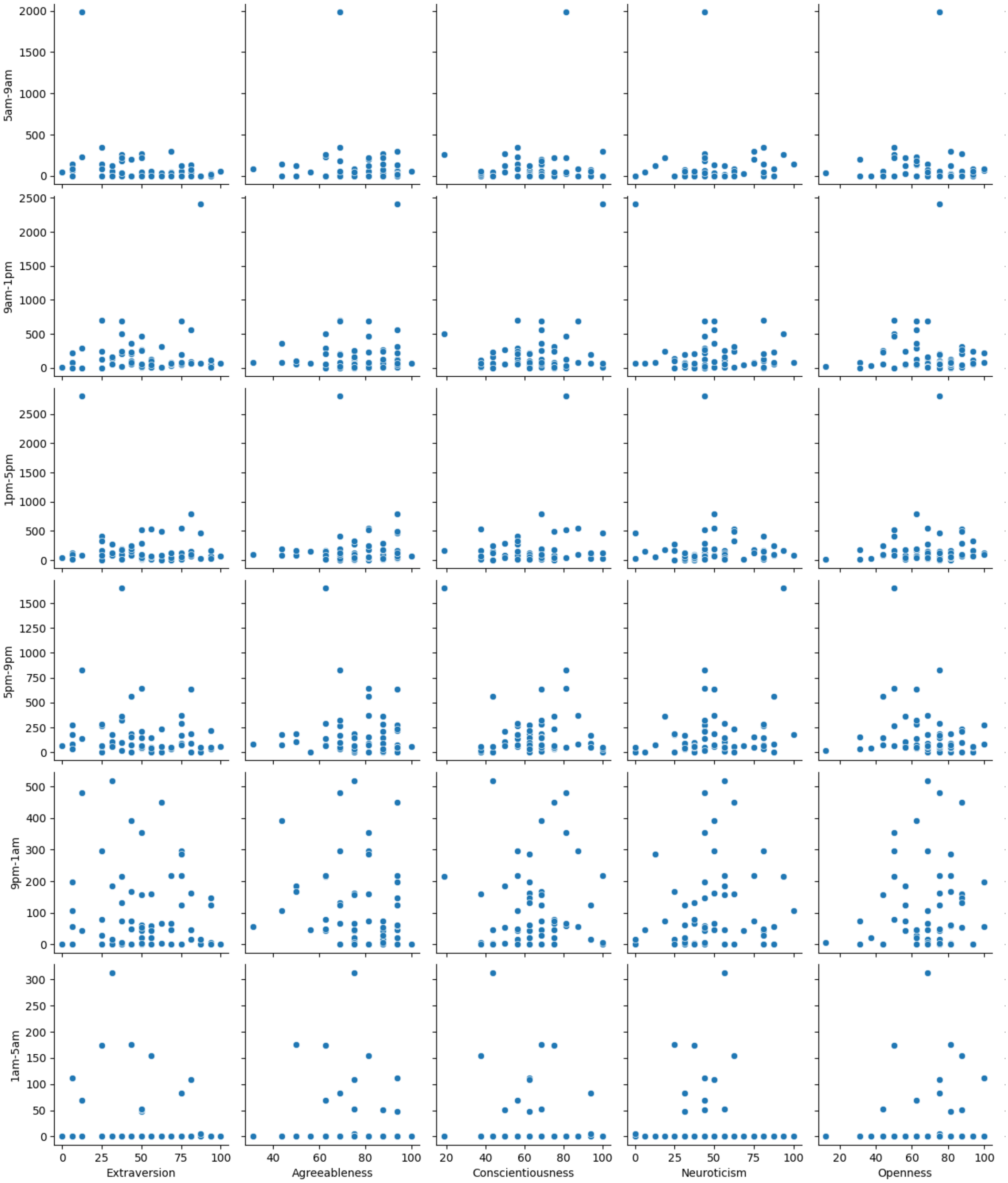


Figure 4.9: Scatter Plot between Average Music Event in each time interval and Big-Five traits.

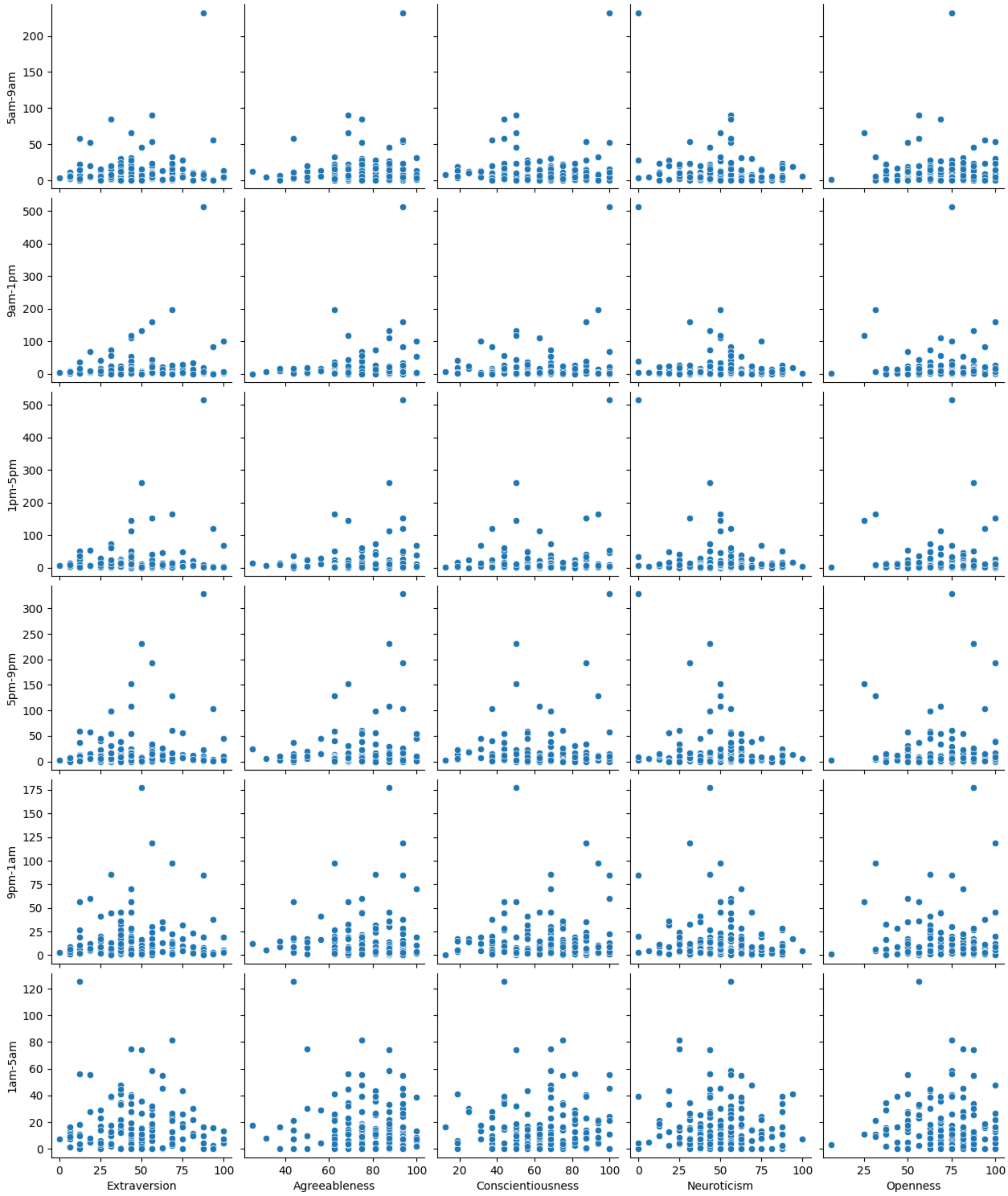


Figure 4.10: Scatter Plot between Average Doze Event in each time interval and Big-Five traits.

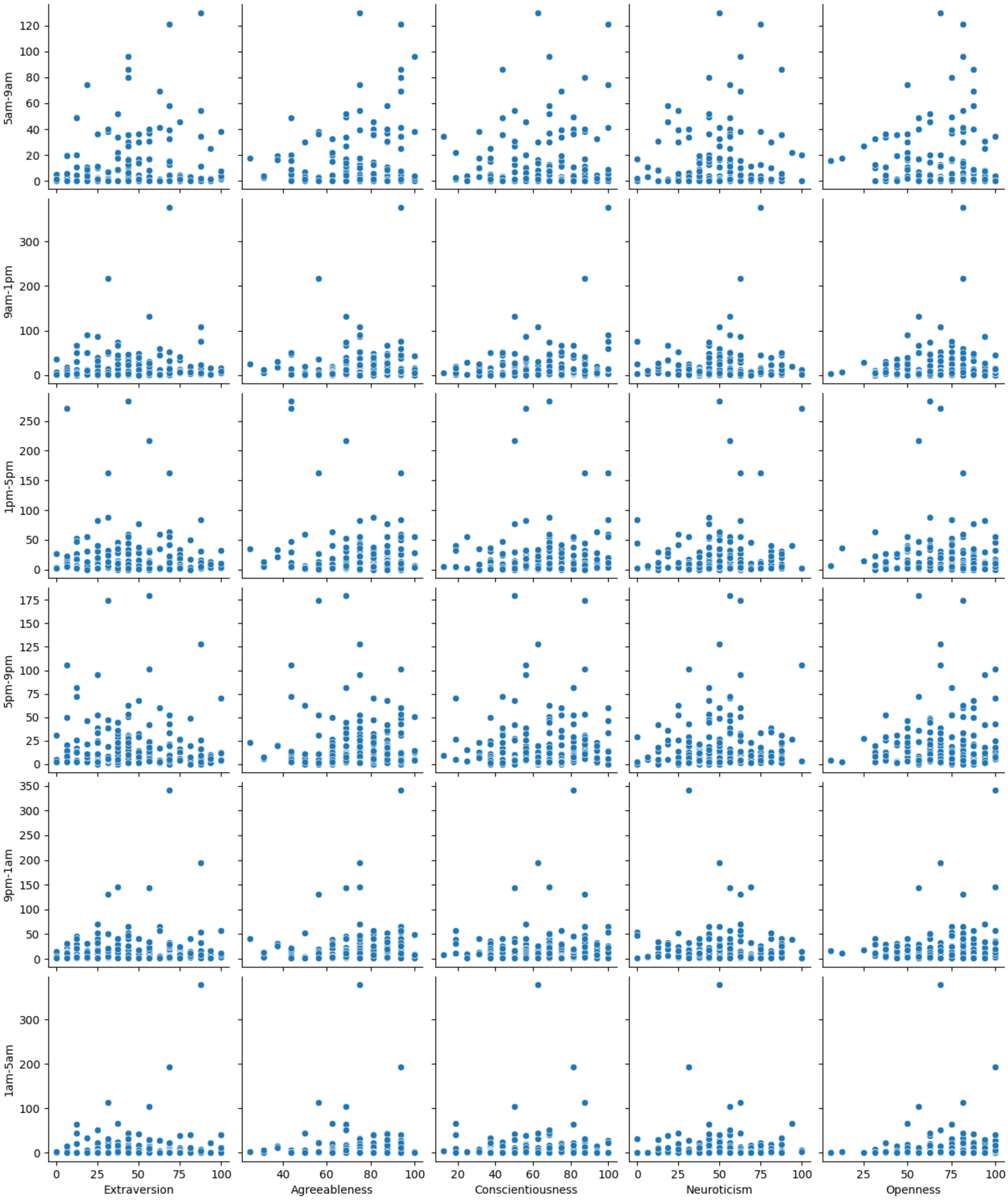


Figure 4.11: Scatter Plot between Average Battery Charge Event in each time interval and Big-Five traits.

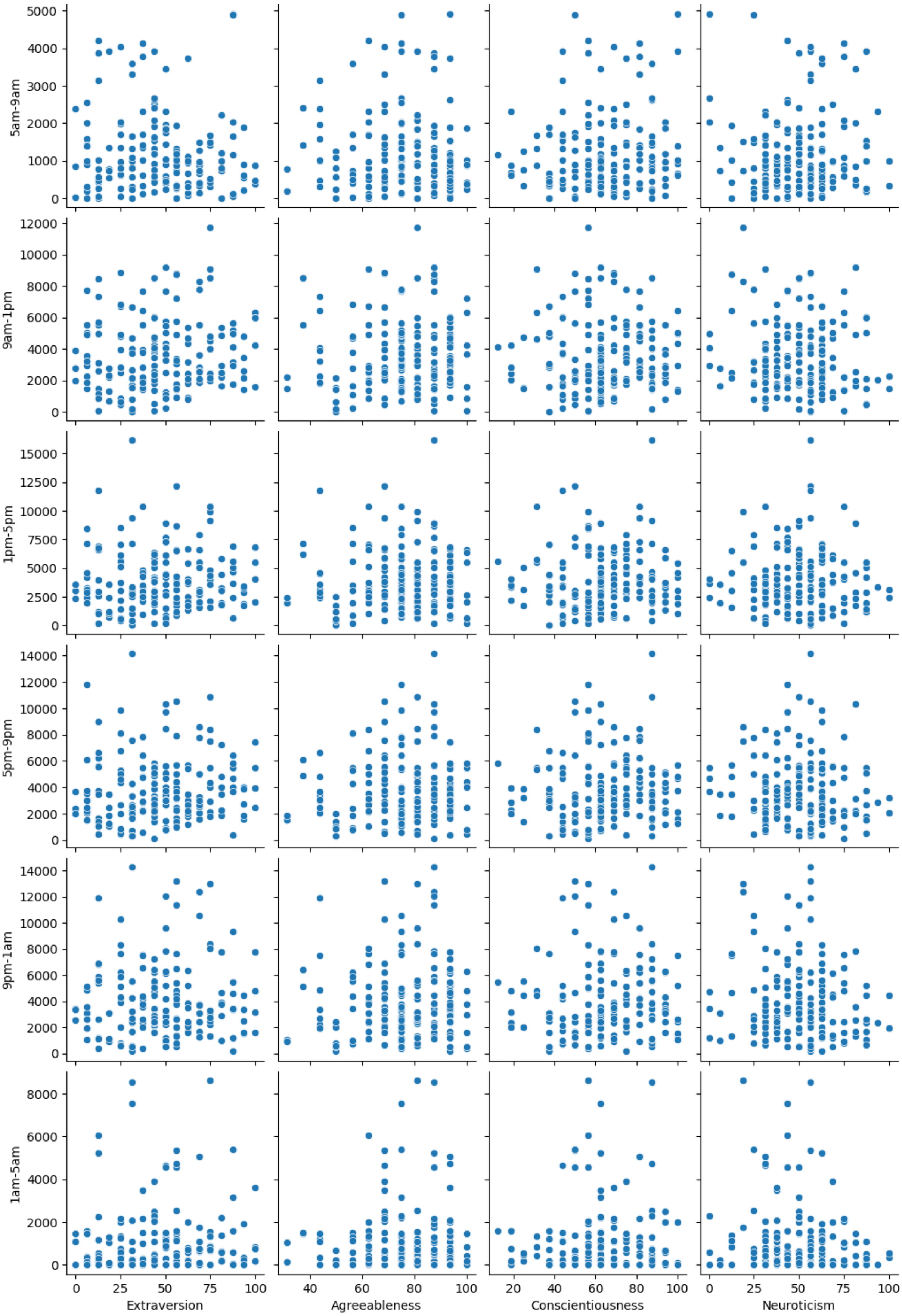


Figure 4.12: Scatter Plot between Average Touch Event in each time interval and Big-Five traits.

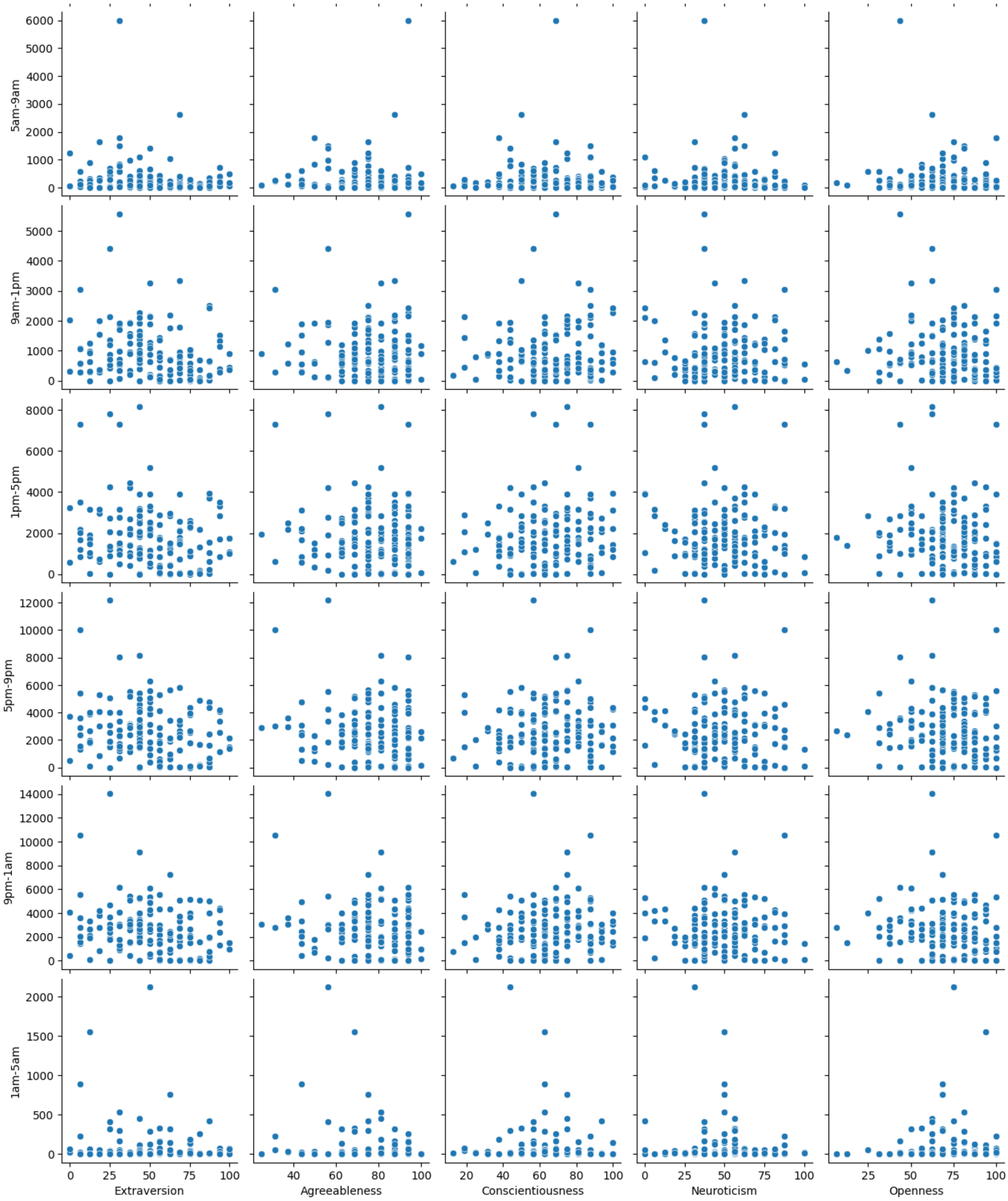


Figure 4.13: Scatter Plot between Average Step Count in each time interval and Big-Five traits.

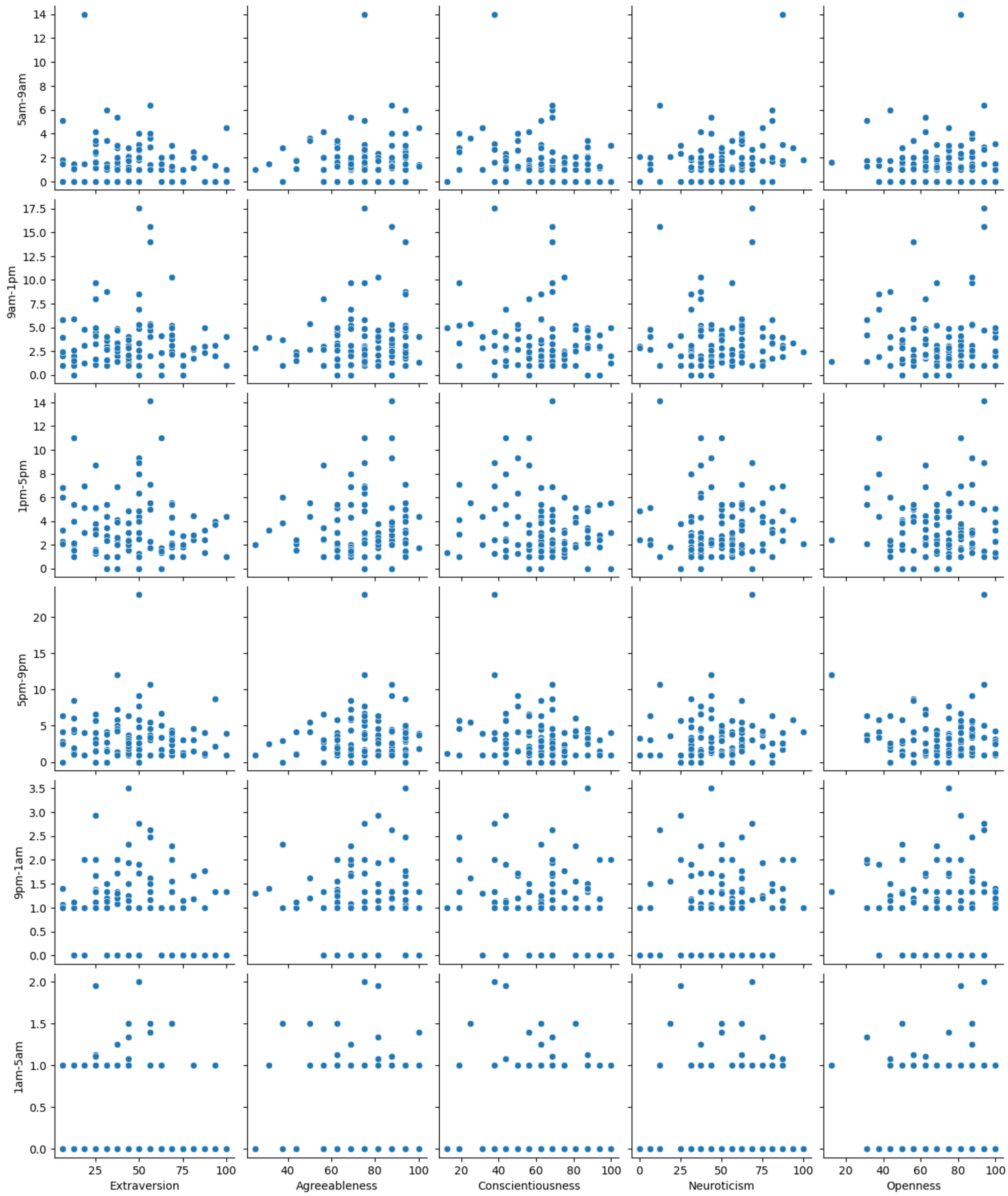


Figure 4.14: Scatter Plot between Average Number of People Nearby in each time interval and Big-Five traits.

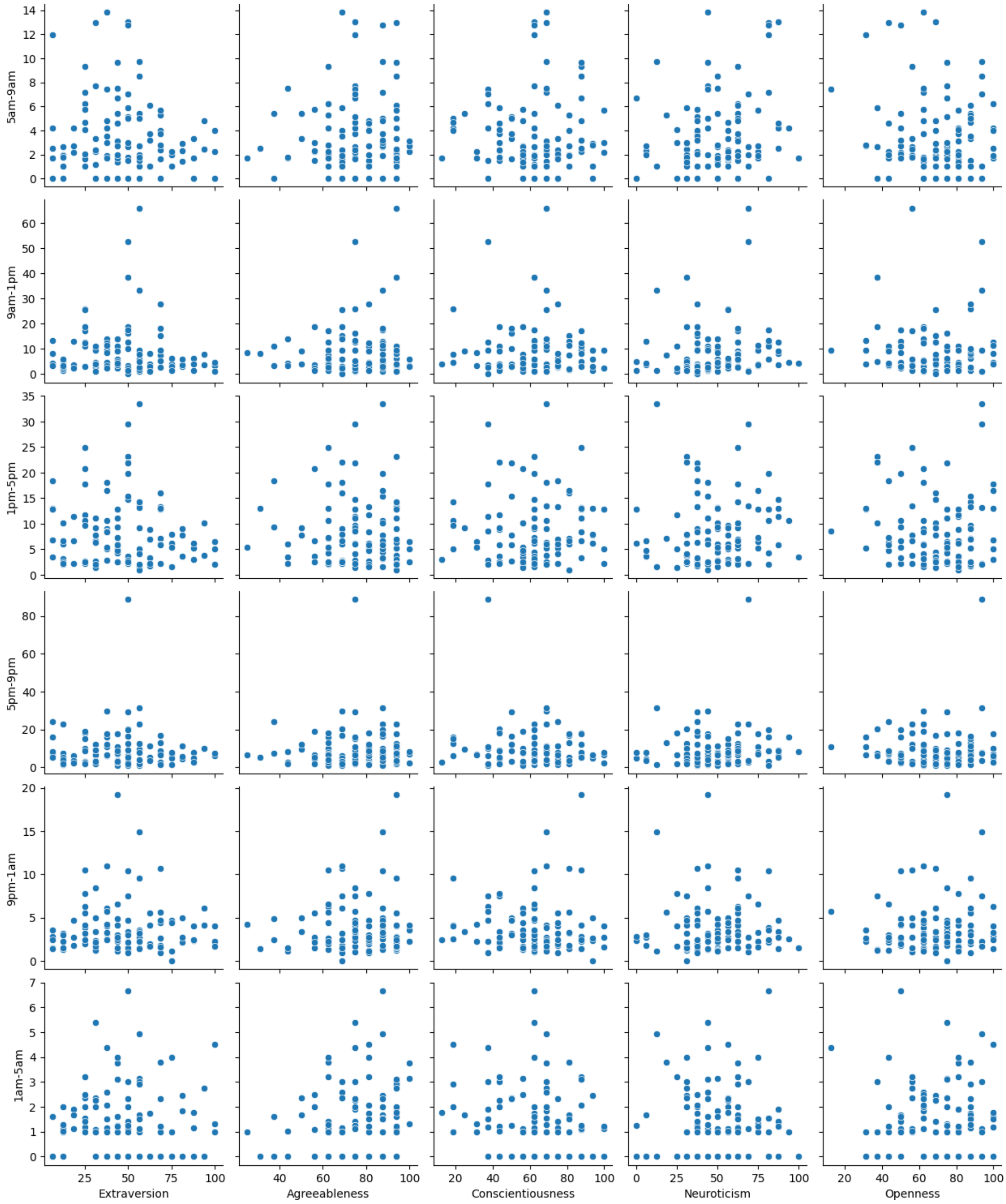


Figure 4.15: Scatter Plot between Average Number of Non-Smartphone Devices Nearby in each time interval and Big-Five traits.

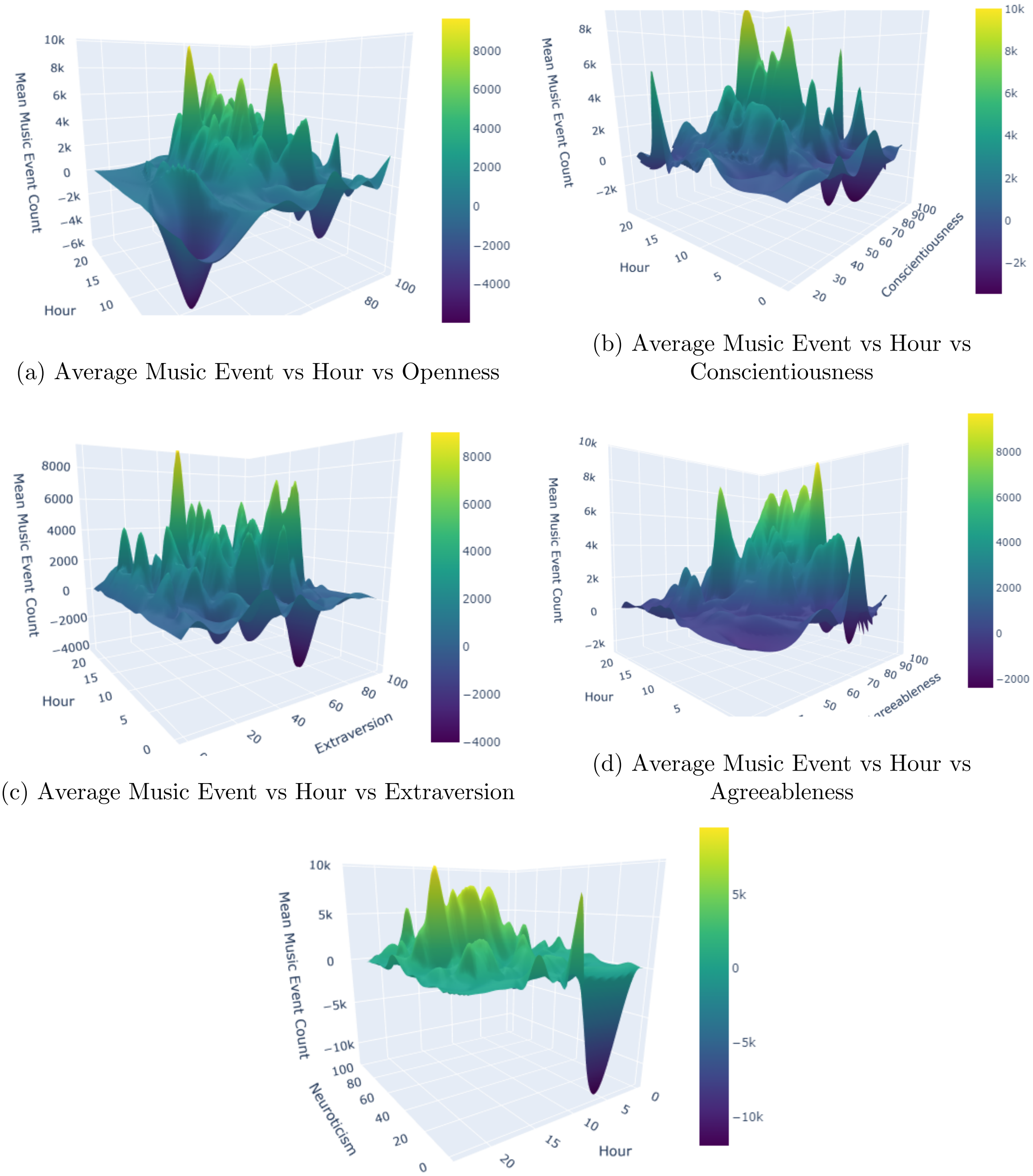
Traits	**App Category**	**r**	**p-value**
Extraversion	Communication	0.16	0.03
Extraversion	Finance	0.14	0.06
Extraversion	Game	-0.13	0.09
Extraversion	Maps and Navigation	0.14	0.05
Extraversion	News and Magazines	-0.22	0.00
Extraversion	Personalization	0.16	0.03
Agreeableness	News and Magazines	-0.13	0.09
Agreeableness	Travel and Local	0.13	0.07
Agreeableness	Video Players and Editors	-0.17	0.02
Conscientiousness	Events	0.16	0.03
Conscientiousness	Health and Fitness	-0.14	0.06
Conscientiousness	Maps and Navigation	-0.18	0.01
Conscientiousness	Sports	0.14	0.06
Neuroticism	Beauty	0.15	0.04
Neuroticism	Finance	-0.25	0.00
Neuroticism	Video Players and Editors	-0.15	0.04
Openness	Art and Design	-0.16	0.03
Openness	Books and Reference	0.14	0.06
Openness	Personalization	0.13	0.07
Openness	Social	-0.13	0.08

Table 4.41: Pearson's Correlation of App Categories with Personality Traits

Trait	**O**	**C**	**E**	**A**	**N**
Openness	-	-0.01	0.14	0.20	-0.01
Conscientiousness	-	-	-0.05	0.06	-0.24
Extraversion	-	-	-	0.36	-0.18
Agreeableness	-	-	-	-	-0.02
Neuroticism	-	-	-	-	-

Table 4.42: Pearson(r) correlations among the Big Five traits. O = Openness, C = Conscientiousness, E = Extraversion, A = Agreeableness, N = Neuroticism

connection between Average Screen Event Per Day, Hour of the Day, and Personality Score. In both instances, it is evident that the time of day plays a pivotal role in shaping how individuals interact with their smartphones. For example, Conscientious individuals exhibit increasing smartphone activity as the day progresses, a trend clearly depicted in the 3D plot. This underscores the significance of time diaries in comprehending behavior and, by extension, people's personalities. Consequently, we will continue to consider the time of day element in subsequent parts of our analysis, especially when employing sensor data in machine learning models for predictive purposes.



(a) Average Music Event vs Hour vs Openness

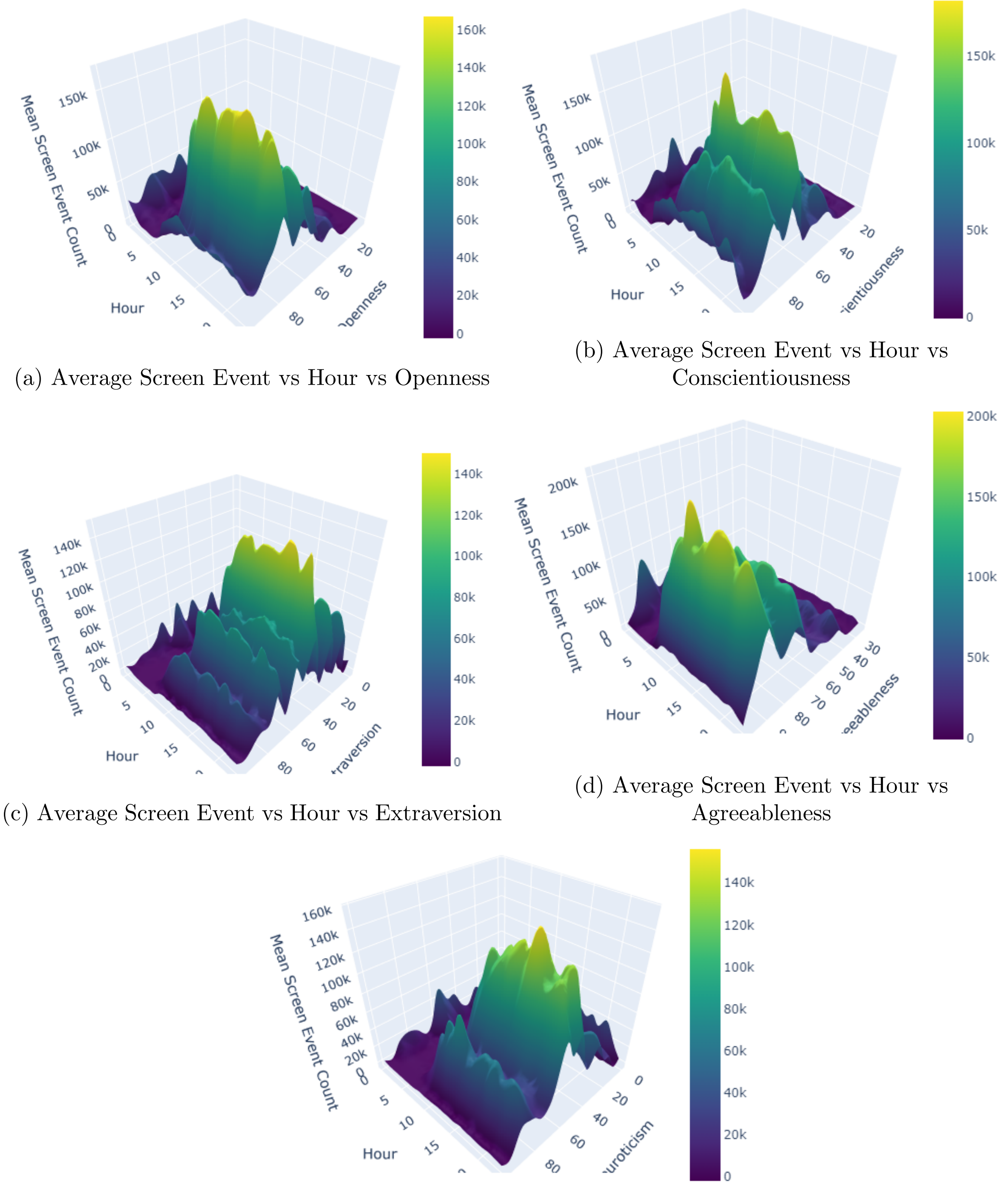
(b) Average Music Event vs Hour vs Conscientiousness

(c) Average Music Event vs Hour vs Extraversion

(d) Average Music Event vs Hour vs Agreeableness

(e) Average Music Event vs Hour vs Neuroticism

Figure 4.16: 3D Plot between Average Music Event Count, Hour of the Day and Personality Score



(a) Average Screen Event vs Hour vs Openness

(b) Average Screen Event vs Hour vs Conscientiousness

(c) Average Screen Event vs Hour vs Extraversion

(d) Average Screen Event vs Hour vs Agreeableness

(e) Average Screen Event vs Hour vs Neuroticism

Figure 4.17: 3D Plot between Average Screen Event Count, Hour of the Day and Personality Score

4.6 Prediction Analysis

In order to predict the Big Five personality traits, supervised machine learning models were built, utilizing the features outlined in Section 4.4. These personality trait values, falling within the continuous range of 0 to 100, framed the problem as a regression modeling task. To facilitate comparisons and showcase the results, five distinct regression algorithms were employed namely Ordinary Least Squares Regression, LASSO Regression, Ridge Regression, KNN and Random Forest Regression. Concurrently, classification models were also constructed, leveraging Support Vector Machine (SVM), Random Forest, KNN and XGBoost algorithms. Model accuracies were meticulously assessed and compared to identify the most effective performing model.

In all of the experiments, 5-fold cross-validation was employed to evaluate the quality of the

Datasets	Sensor Combinations	Userid Count
Dataset 1	App Usage, Touch Event, Step Counter, Gender, Department	131
Dataset 2	Screen Event, Charge Event, App Usage, Ring Event, Gender, Department	115
Dataset 3	Screen Event, App Usage, Ring Mode, Gender, Department	120
Dataset 4	Screen Event, Charge Event, Doze Event, Ring Mode, Gender, Department	101
Dataset 5	Screen Event, App Usage, Doze Event, Gender, Department	130
Dataset 6	Screen Event, Charge Event, App Usage, Step Counter, Gender, Department	130

Table 4.43: List of Combinations of Sensors used in Machine Learning Models

machine learning models. The entire dataset was divided into 5 folds, with one set reserved as a test set, while the model was trained using the remaining 4 sets of data. This method allowed for testing against unseen data without potentially biasing the results by randomly choosing a favorable or unfavorable test set. Because this is a regression modeling problem, Root Mean Squared Error (RMSE) was used as the metric of success. For classification models, accuracy was used as the success metric. RMSE values and accuracy values were converted into percentages to aid interpretation. Python programming language was used for creating the machine learning models. The Scikit-learn 1.3 module was used in the creation of Lasso regression, Ridge regression, Decision Trees, Random Forests, K-Nearest Neighbour and SVM models, while statsmodels library was used for Ordinary Least Square Regression (OLS).

Multiple sensor combinations were attempted, specifically six distinct sensor combinations and the results were evaluated using predictive models. This approach was taken due to the presence of uncommon user IDs among the datasets. Combining data from all the sensors we have, resulted in a significantly reduced dataset, comprising approximately 40 users. However, when combining data from three to five sensors, a more substantial dataset with over 100 user IDs was consistently available for analysis. Table 4.43 refers to all the sensor combinations used to build our models. When these set of sensors were combined to be used in machine learning models, the features extracted from those sensors were used which has been reported under the Table 4.36 in Section 4.4.

Detailed explanations about training and tuning the machine learning models are provided in the following sections.

4.6.1 Regression Models

As previously mentioned, five regression models—OLS, KNN, Lasso, Ridge Regression and Random Forest Regression—were applied in our analysis, utilizing the Statsmodels and Scikit-learn APIs. The performance of these regression models was assessed using several metrics, including RMSE, MSE, R-squared, and adjusted R-squared. Fit Line Plot Analyses and Residual Plot Analyses were performed as well. It's noteworthy that lower RMSE and MSE values are indicative of superior model performance, while R-squared measures the proportion of variance in the dependent variable explained by the model.

Initially, regression was performed on individual sensor data, but the results indicated poor performance. This was characterized by substantial errors, and in some instances, negative R-squared values, implying that the models performed less effectively than a mere horizontal line. Later on multiple sensors were combined and along with gender and department information added as input features, regression analysis was performed which resulted in much improved performance as compared

	OLS									
	O		**C**		**E**		**A**		**N**	
	RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
Dataset 1	17.08	0.42	18.18	0.41	23.21	0.44	14.73	0.42	19.60	0.40
Dataset 2	17.34	0.51	18.43	0.54	23.91	0.45	15.35	0.42	21.75	0.40
Dataset 3	17.05	0.46	18.30	0.48	24.07	0.36	15.33	0.36	21.01	0.38
Dataset 4	18.75	0.29	21.11	0.22	23.05	0.26	14.89	0.32	20.25	0.31
Dataset 5	19.03	0.32	18.66	0.47	23.67	0.39	15.78	0.36	19.49	0.40
Dataset 6	18.12	0.42	17.95	0.47	23.43	0.47	13.99	0.52	19.22	0.46

Table 4.44: Results from OLS run on all Sensor Combinations

	LASSO									
	O		**C**		**E**		**A**		**N**	
	RMSE	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**
Dataset 1	13.88	0.41	14.77	0.40	18.83	0.43	12.27	0.38	16.16	0.38
Dataset 2	13.16	0.49	14.01	0.52	17.89	0.44	11.84	0.38	16.53	0.38
Dataset 3	13.73	0.44	14.76	0.46	19.04	0.36	12.60	0.31	16.93	0.35
Dataset 4	16.55	0.27	18.59	0.21	20.31	0.24	13.27	0.29	17.84	0.30
Dataset 5	15.29	0.32	15.14	0.45	18.93	0.39	12.96	0.32	16.12	0.36
Dataset 6	14.20	0.41	14.05	0.46	18.56	0.45	11.30	0.48	15.10	0.45

Table 4.45: Results from LASSO run on all Sensor Combinations

to baseline models.

OLS Regression

OLS regression models were applied to all six datasets listed in Table 4.43. Various metrics, including RMSE, MSE, R-Squared, and Adjusted R-Square values, were utilized to evaluate the results. The outcomes of the OLS model, when applied to the sensor combinations, are presented in Table 4.44. Notably, the lowest error rates were observed with the outcome variable Agreeableness, with values ranging from 13.99% to 15.78%. The R-Square values for all outcome variables consistently hovered around 0.40.

LASSO Regression

LASSO regression models were applied to all six datasets listed in Table 4.43. Various metrics, including RMSE, MSE, R-Squared, and Adjusted R-Square values, were used to evaluate the results. The outcomes of the LASSO model, when applied to the sensor combinations, are presented in Table 4.45. Notably, the lowest error rates were observed with the outcome variable Agreeableness, with values ranging from 11.30% to 13.27%. The R-Square values for all outcome variables consistently hovered around 0.40.

Ridge Regression

Ridge regression models were applied to all six datasets. Various metrics such as RMSE and R-Squared values, were used to assess the results. The outcomes of the Ridge regression model, when applied to different sensor combinations, are presented in Table 4.46. The lowest RMSE value, at 10.84%, was reported for Dataset 6 with the outcome variable Agreeableness. Conversely, the highest RMSE value was observed for Dataset 4, with the outcome variable Extraversion, at 20.13%.

	Ridge									
	O		C		E		A		N	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Dataset 1	13.76	0.42	14.46	0.41	18.70	0.44	11.87	0.42	15.92	0.40
Dataset 2	12.93	0.51	13.75	0.54	17.84	0.45	11.45	0.42	16.23	0.40
Dataset 3	13.48	0.46	14.46	0.48	19.03	0.36	12.12	0.36	16.61	0.38
Dataset 4	16.37	0.29	18.44	0.22	20.13	0.26	13.00	0.32	17.68	0.31
Dataset 5	15.21	0.32	14.91	0.47	18.91	0.39	12.61	0.36	15.58	0.40
Dataset 6	14.03	0.42	13.91	0.47	18.15	0.47	10.84	0.52	14.89	0.46

Table 4.46: Results from Ridge Regression run on all Sensor Combinations

	KNN									
	O		C		E		A		N	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Dataset 1	16.32	0.19	17.50	0.16	22.00	0.22	14.12	0.18	18.02	0.23
Dataset 2	17.44	0.10	17.57	0.25	21.08	0.23	13.62	0.19	18.76	0.20
Dataset 3	16.94	0.16	17.71	0.23	20.92	0.22	13.84	0.16	19.03	0.18
Dataset 4	16.91	0.24	19.65	0.12	20.69	0.21	14.41	0.17	18.64	0.24
Dataset 5	16.98	0.16	17.04	0.31	21.70	0.19	14.14	0.19	17.96	0.21
Dataset 6	16.65	0.19	19.03	0.02	21.30	0.27	14.20	0.18	18.27	0.19

Table 4.47: Results from KNN Regression run on all Sensor Combinations

K-Nearest Neighbour Regression

KNN was selected due to its capability to model non-linear relationships. In case of KNN models we got the highest RMSE score for the Extraversion outcome variable. The outcomes of applying the KNN regressor to all six datasets are detailed in Table 4.47.

Random Forest Regression

The Random Forest model can be effectively applied for both classification and regression tasks. In our analysis, we utilized this tree-based model to fit our sensor combinations. The results obtained from the Random Forest are summarized in Table 4.48. It is evident that the Random Forest model consistently demonstrated significantly lower error percentages across all sensor combinations, signifying this as the most favorable outcome achieved thus far.

	Random Forest									
	O		C		E		A		N	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Dataset 1	7.08	0.84	7.52	0.84	9.30	0.86	5.55	0.87	7.94	0.84
Dataset 2	6.27	0.88	7.54	0.86	9.35	0.84	5.88	0.84	8.53	0.83
Dataset 3	7.12	0.85	7.81	0.85	9.53	0.83	6.08	0.83	8.54	0.83
Dataset 4	6.70	0.88	7.67	0.86	9.35	0.84	6.20	0.84	8.69	0.83
Dataset 5	7.58	0.83	8.06	0.84	9.12	0.85	5.98	0.85	7.46	0.86
Dataset 6	7.01	0.85	7.55	0.84	9.76	0.84	5.74	0.86	8.03	0.84

Table 4.48: Results from Random Forest run on all Sensor Combinations

	Baseline Mean Model									
	O		C		E		A		N	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Dataset 1	18.19	0	19.13	0	25.01	0	15.62	0	20.63	0
Dataset 2	18.49	0	20.33	0	24.09	0	15.14	0	21.04	0
Dataset 3	18.48	0	20.21	0	23.81	0	15.18	0	21.14	0
Dataset 4	19.49	0	21.01	0	23.41	0	15.86	0	21.42	0
Dataset 5	18.56	0	20.58	0	24.25	0	15.77	0	20.25	0
Dataset 6	18.53	0	19.26	0	25.05	0	15.77	0	20.39	0

Table 4.49: Results from Baseline run on all Sensor Combinations

Trait	High (%)	Low (%)
Extraversion	42.57	57.43
Agreeableness	49.00	51.00
Conscientiousness	49.00	51.00
Neuroticism	43.37	56.63
Openness	40.96	59.04

Table 4.50: Class Distribution Statistics after binning

Mean Model

The baseline model for the regression problem is the Mean Model, as discussed in the Section 3.2.1. The results obtained when the Mean Model was applied to all sensor combinations have been documented in Table 4.49. These results will function as the baseline, enabling us to make comparisons with the performance of our other models.

4.6.2 Classification Models

In the course of our analysis, an array of classification models, namely Decision Tree, SVM, Random Forest, XGBoost and KNN were employed. The Scikit-learn API served as the foundation for our model implementation. A host of performance metrics, encompassing accuracy, precision, recall, and f-1 score, was employed to assess the efficacy of these models. For model comparison, however, we resorted solely on accuracy metric. Also we had to bin the continuous outcome variables into categorical types. For that I used the mean score for every personality trait as splitting criteria and segregated the users as high or low in each personality trait. After converting all the data, as can be seen in Table 4.50, class distributions were almost equally distributed across the class labels for all five traits. This ensured that there was no skew or class bias.

Decision Tree

Dataset 1 through Dataset 6 of Table 4.43 were subjected to decision tree using the scikit-learn API. An 80-20 split was employed for the training and testing data partitions, and the reporting of test accuracies was carried out. Initially, all the decision tree models were run with default parameters. Subsequently, an attempt was made to adjust the parameters through the application of the grid search cross-validation method, which permitted the evaluation of the model across a range of parameter combinations, ultimately selecting the one with the highest accuracy metric.

Hyper-parameters for the decision tree model used were:

- criterion: ['gini', 'entropy']
- max_depth: [None, 10, 20, 30, 40, 50]
- min_samples_split: [2, 5, 10]

	E	**A**	**C**	**N**	**O**
Dataset 1	0.56	0.48	0.56	0.52	0.59
Dataset 2	0.70	0.61	0.74	0.57	0.57
Dataset 3	0.63	0.46	0.54	0.67	0.58
Dataset 4	0.67	0.62	0.43	0.67	0.48
Dataset 5	0.50	0.62	0.42	0.46	0.69
Dataset 6	0.38	0.50	0.58	0.62	0.69

Table 4.51: Accuracy from Decision Tree models run on all the sensor combinations

	E	**A**	**C**	**N**	**O**
Dataset 1	0.41	0.52	0.52	0.44	0.67
Dataset 2	0.57	0.66	0.57	0.57	0.57
Dataset 3	0.33	0.50	0.54	0.46	0.50
Dataset 4	0.66	0.62	0.67	0.57	0.57
Dataset 5	0.35	0.69	0.58	0.42	0.58
Dataset 6	0.46	0.42	0.62	0.42	0.77

Table 4.52: Accuracy from SVM models run on all the sensor combinations

- min_samples_leaf: [1, 2, 4]

Several experiments were carried out with different ranges of values set for these parameters as mentioned above. Table 4.51 are the accuracy scores for all the datasets.

Support Vector Machine

Support vector classifiers were applied to Dataset 1 through Dataset 6 of Table 4.43 using the scikit-learn API. A 5 fold cross validation was utilized for the training and testing data partitions, and the reporting of test accuracies was conducted. Initially, all the support vector classifier models were run with default parameters. Subsequently, an attempt was made to adjust the parameters through the application of the grid search cross-validation method with 5-folds, enabling the model's evaluation across a variety of parameter combinations, and the selection of the one with the highest accuracy metric.

Hyper-parameters for the SVM model used were:

- C: [0.1, 1, 10]
- kernel: ['linear', 'rbf', 'poly']

Several experiments were carried out with different ranges of values set for these parameters as mentioned above. Table 4.52 are the accuracy scores for all the datasets.

Random Forest Classifier

The application of the Random Forest Classifier to Dataset 1 through Dataset 6 was carried out. A 5 fold cross validation was employed for the training and testing data partitions, and the reporting of test accuracies was conducted. Initially, all the models were executed with default parameters. Subsequently, an effort was made to modify the parameters through the utilization of the grid search cross-validation method with 5-folds, allowing the model to be evaluated across various parameter combinations.

The hyper-parameters for the Random Forest model used were:

- n_estimators: [100, 200, 300]
- max_depth: [None, 10, 20, 30]

	E	**A**	**C**	**N**	**O**
Dataset 1	0.74	0.66	0.44	0.48	0.59
Dataset 2	0.65	0.61	0.65	0.61	0.70
Dataset 3	0.76	0.50	0.46	0.54	0.58
Dataset 4	0.67	0.67	0.62	0.57	0.52
Dataset 5	0.73	0.73	0.46	0.58	0.42
Dataset 6	0.62	0.62	0.54	0.58	0.50

Table 4.53: Accuracy from Random Forest models run on all the sensor combinations

	E	**A**	**C**	**N**	**O**
Dataset 1	0.63	0.56	0.44	0.48	0.59
Dataset 2	0.52	0.57	0.52	0.61	0.43
Dataset 3	0.46	0.63	0.50	0.46	0.58
Dataset 4	0.62	0.57	0.67	0.52	0.43
Dataset 5	0.58	0.58	0.69	0.46	0.62
Dataset 6	0.65	0.62	0.46	0.69	0.58

Table 4.54: Accuracy from K-NN models run on all the sensor combinations

- min_samples_split: [2, 5, 10]
- min_samples_leaf: [1, 2, 4]
- max_features: ['sqrt', 'log2']

Table 4.53 denotes the accuracy results obtained from the random forest models.

KNN Classifier

Similarly the same set of data were run on the K-Nearest Neighbour algorithm too. KNN identifies the K training data points with the closest feature resemblance and assigns the class label of the majority of these nearest neighbors to the new data point; we needed to test with varying degrees of k values. With a 5 fold cross validation and below set of hyper-parameters tuned we obtained the accuracies for all our datasets.

- n_neighbors: [3, 5, 7, 9]
- weights: ['uniform', 'distance']
- p: [1, 2]

Here p is a distance metric for which the values 1 signify Manhattan distance whereas 2 signifies Euclidean distance. The results after the cross validation are in Table 4.54.

XGBoost Classifier

XGBoost is an ensemble method which basically aggregates multiple decision trees in order to get an improved result. We implemented this algorithm with hyper-parameter tuning and 5 fold cross validation. The results obtained are denoted in Table 4.55.

Hyper-parameters used were:

- n_estimators: [100, 200, 300]
- learning_rate: [0.01, 0.1, 0.2]
- max_depth: [3, 4, 5]

The accuracy results are denoted in Table 4.55.

	E	**A**	**C**	**N**	**O**
Dataset 1	0.67	0.74	0.63	0.56	0.56
Dataset 2	0.61	0.74	0.61	0.56	0.57
Dataset 3	0.63	0.54	0.67	0.58	0.71
Dataset 4	0.62	0.81	0.52	0.57	0.52
Dataset 5	0.54	0.62	0.65	0.63	0.65
Dataset 6	0.54	0.58	0.69	0.65	0.65

Table 4.55: Accuracy from XGBoost models run on all the sensor combinations

5 Results

5.1 Regression Results

To analyze, compare, and illustrate the predictive power and quality of regression models, a statistical analysis was conducted. It was revealed by the statistical analysis that the best results were achieved through the application of Random Forest Regression methods. Comprehensive details pertaining to this are elaborated here.

5.1.1 Statistical Analysis (RMSE)

The comparison of regression models was based on RMSE values, and Figure 5.1 illustrates this comparison. On the x-axis, you can observe the root mean squared error, while the y-axis represents the Big Five personality traits. To facilitate interpretation, the error values are presented as percentages.

In the figures from Figure 5.1a to Figure 5.1f, we can observe the outcomes of the regression models. Notably, the random forest models consistently exhibited outstanding performance across all personality traits. The lowest RMSE error recorded was 6%, with the highest RMSE of 10% observed in datasets 3 and 6. These results are notably impressive when compared to the baseline model. Ridge Regression models emerged as the second-best performers after random forest. It's important to note that all datasets yielded relatively consistent results. In every scenario, Random Forest models outperformed other approaches by a margin of about 50%. Ridge Regression appeared as the second-best model in terms of RMSE percentage, albeit with nearly double the error percentage compared to the best model.

5.1.2 Fit Line Plots Analysis

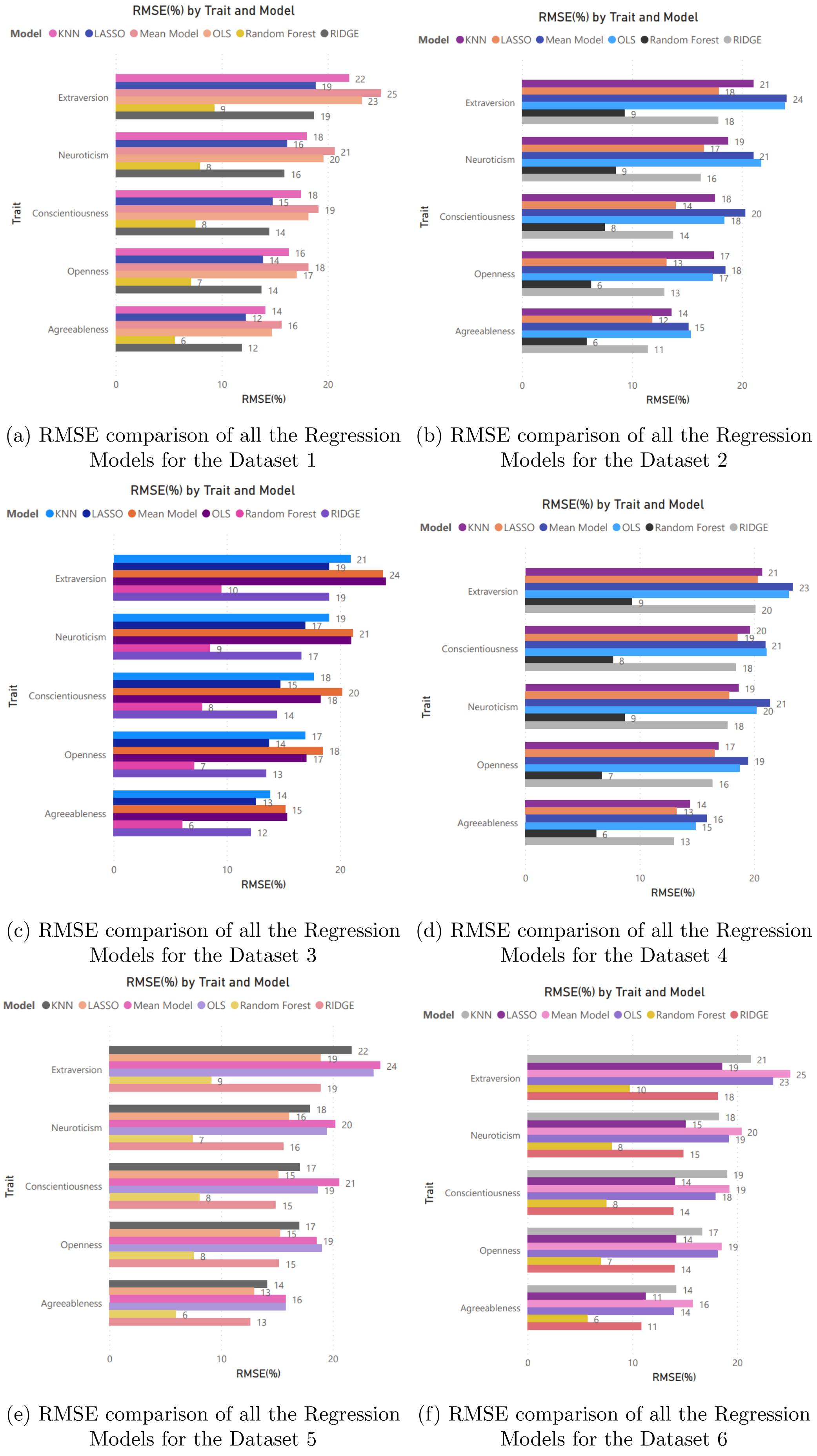
For the fit line plots, the x-axis displays the actual personality trait values (ground truth), while the y-axis shows the values predicted by the machine learning models. These fit line plots are presented from Figure 5.2 to Figure 5.8. In each of these fit line plots, the x-axis represents the actual personality trait, the y-axis represents the predicted values generated by the machine learning models, and the dotted diagonal line illustrates the regression line. A machine learning model that aligns closely with the regression line, featuring many data points near it, is considered the most effective.

Figure 5.2 present the fit line plots for Mean Models corresponding to all Big Five Personality traits for all 6 datasets. These plots display all five traits collectively within a single image. Given that these are mean models, the points are uniformly dispersed as flat lines along the fit line. It is evident that there are only a few points in close proximity to the fit line, signifying that baseline models exhibited poor generalization.

Figure 5.3 to Figure 5.8 illustrate the fit line plots for the Random Forest models, with each Big Five personality trait having its dedicated plot. In comparison to the baseline models, there is a noticeable increase in the number of points situated near the fit line for all five traits. In summary, the Random Forest models exhibited superior performance when contrasted with the baseline models.

5.1.3 Residual Plots Analysis

Residual plots depict the relationship between actual values and residual values. Residual values represent the disparity between predicted values and actual values. These plots include a horizontal line, representing the performance line where the error or residual is zero. The distance of each point from this line indicates the error or deviation from the actual value for that specific point. Figure 5.9 to Figure 5.14 display the residual plots for the Random Forest models applied on all 6 sensor combinaton. In all residual plots, the x-axis signifies the actual values, while the y-axis represents the residual values. Typically, baseline models exhibit residuals that increase consistently with the independent



(a) RMSE comparison of all the Regression Models for the Dataset 1

(b) RMSE comparison of all the Regression Models for the Dataset 2

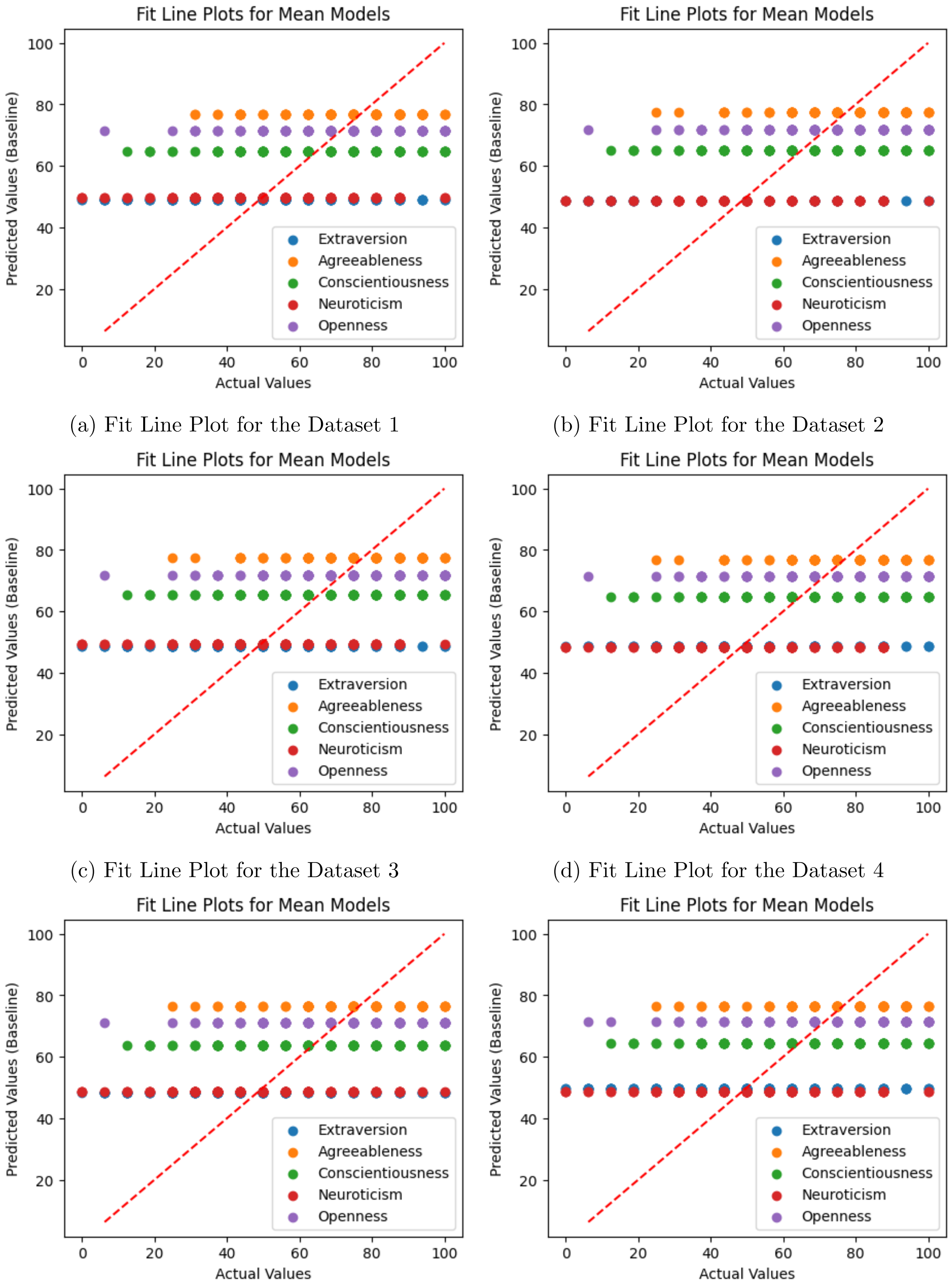
(c) RMSE comparison of all the Regression Models for the Dataset 3

(d) RMSE comparison of all the Regression Models for the Dataset 4

(e) RMSE comparison of all the Regression Models for the Dataset 5

(f) RMSE comparison of all the Regression Models for the Dataset 6

Figure 5.1: RMSE Comparison



(a) Fit Line Plot for the Dataset 1

(b) Fit Line Plot for the Dataset 2

(c) Fit Line Plot for the Dataset 3

(d) Fit Line Plot for the Dataset 4

(e) Fit Line Plot for the Dataset 5

(f) Fit Line Plot for the Dataset 6

Figure 5.2: Fit Line Plots for Mean Models

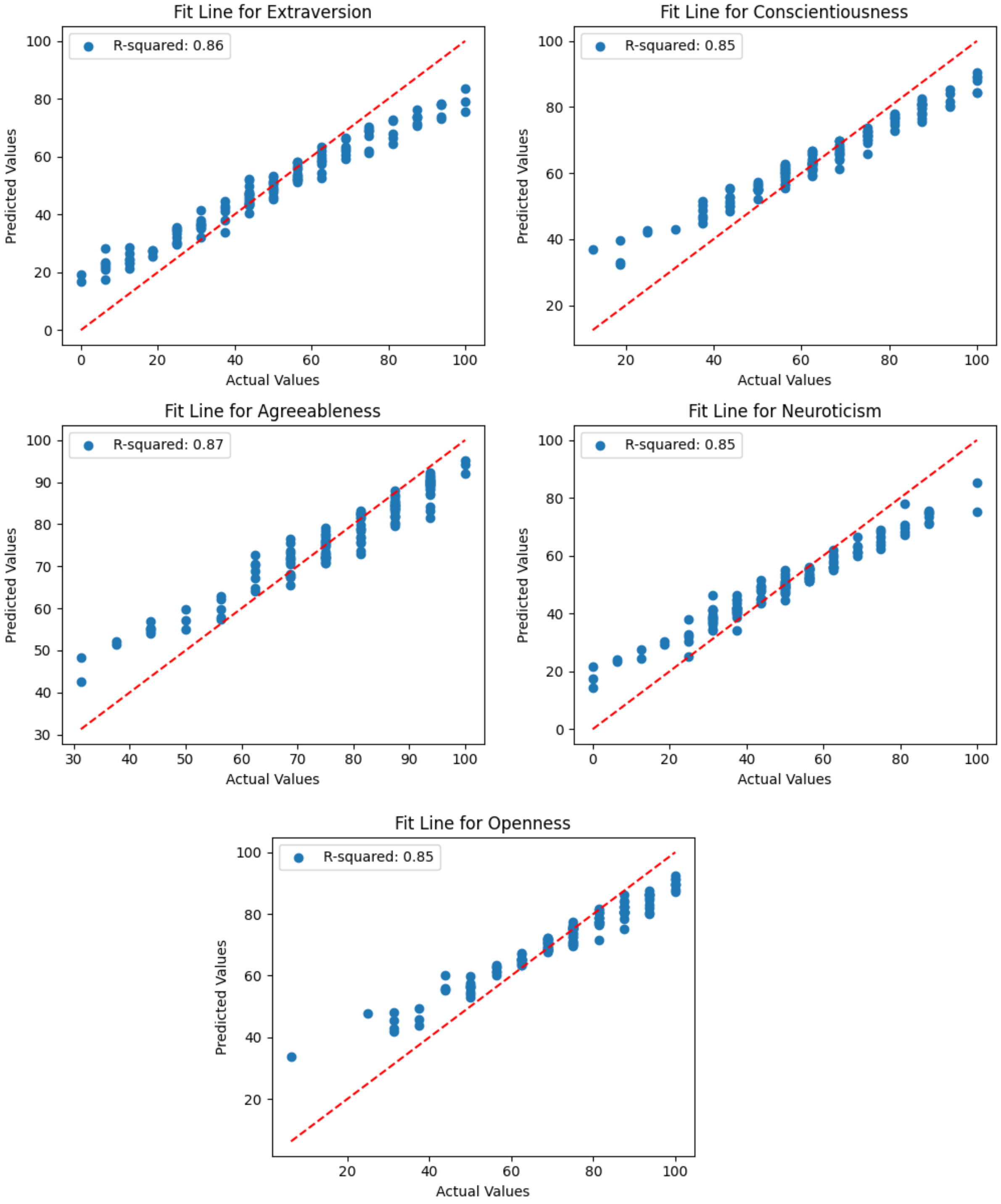


Figure 5.3: Fit Line Plots for Dataset 1

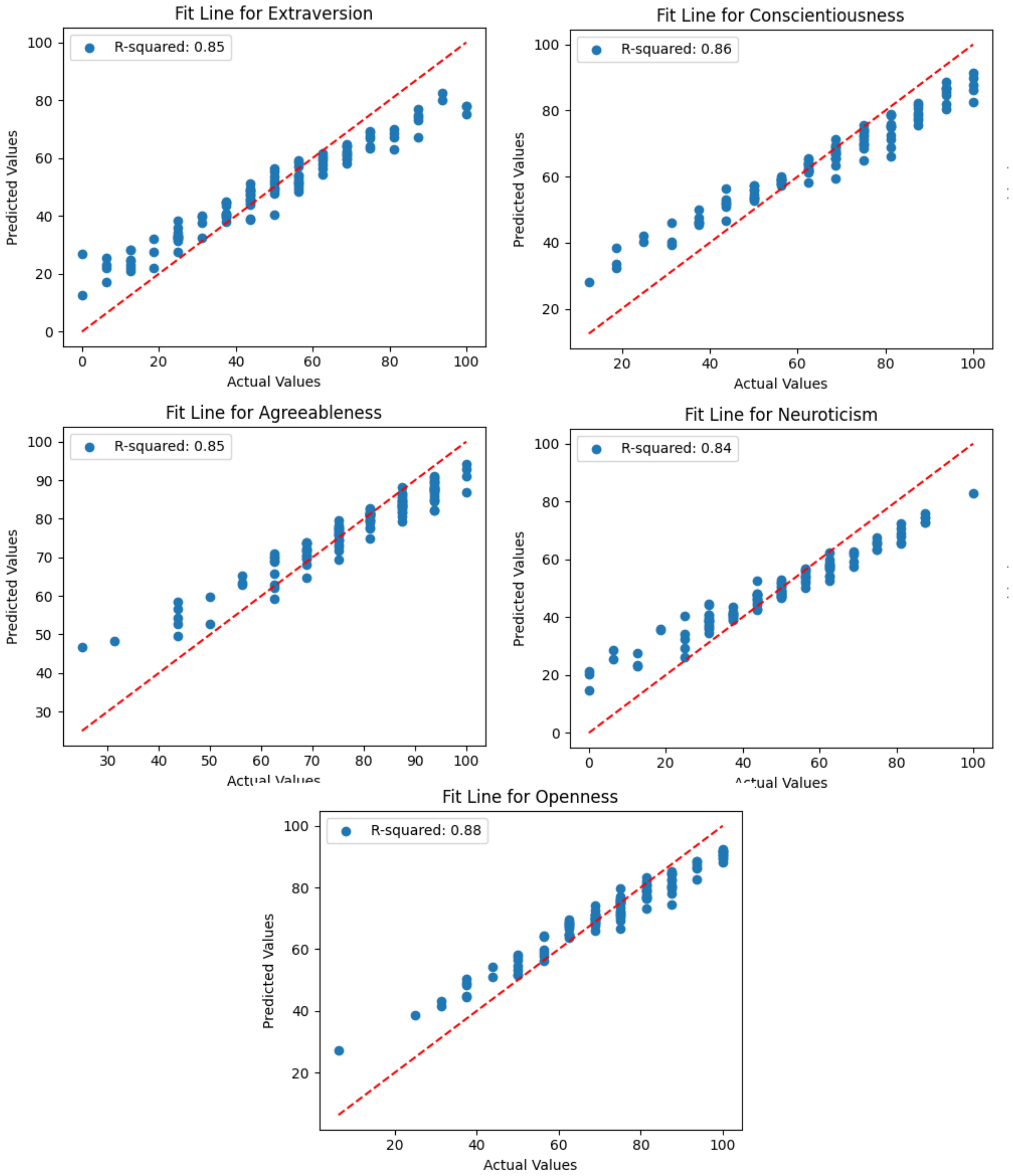


Figure 5.4: Fit Line Plots for Dataset 2

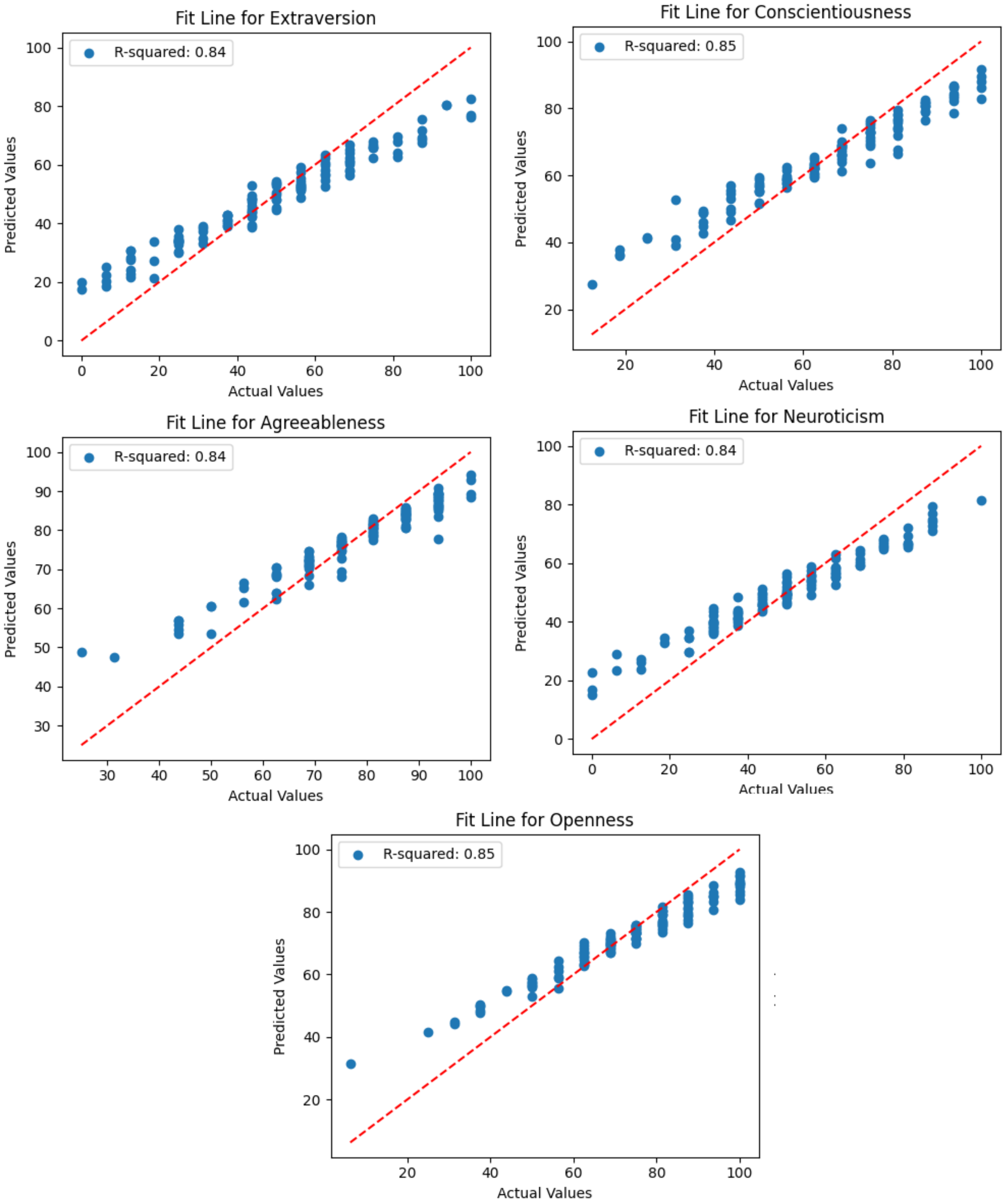


Figure 5.5: Fit Line Plots for Dataset 3

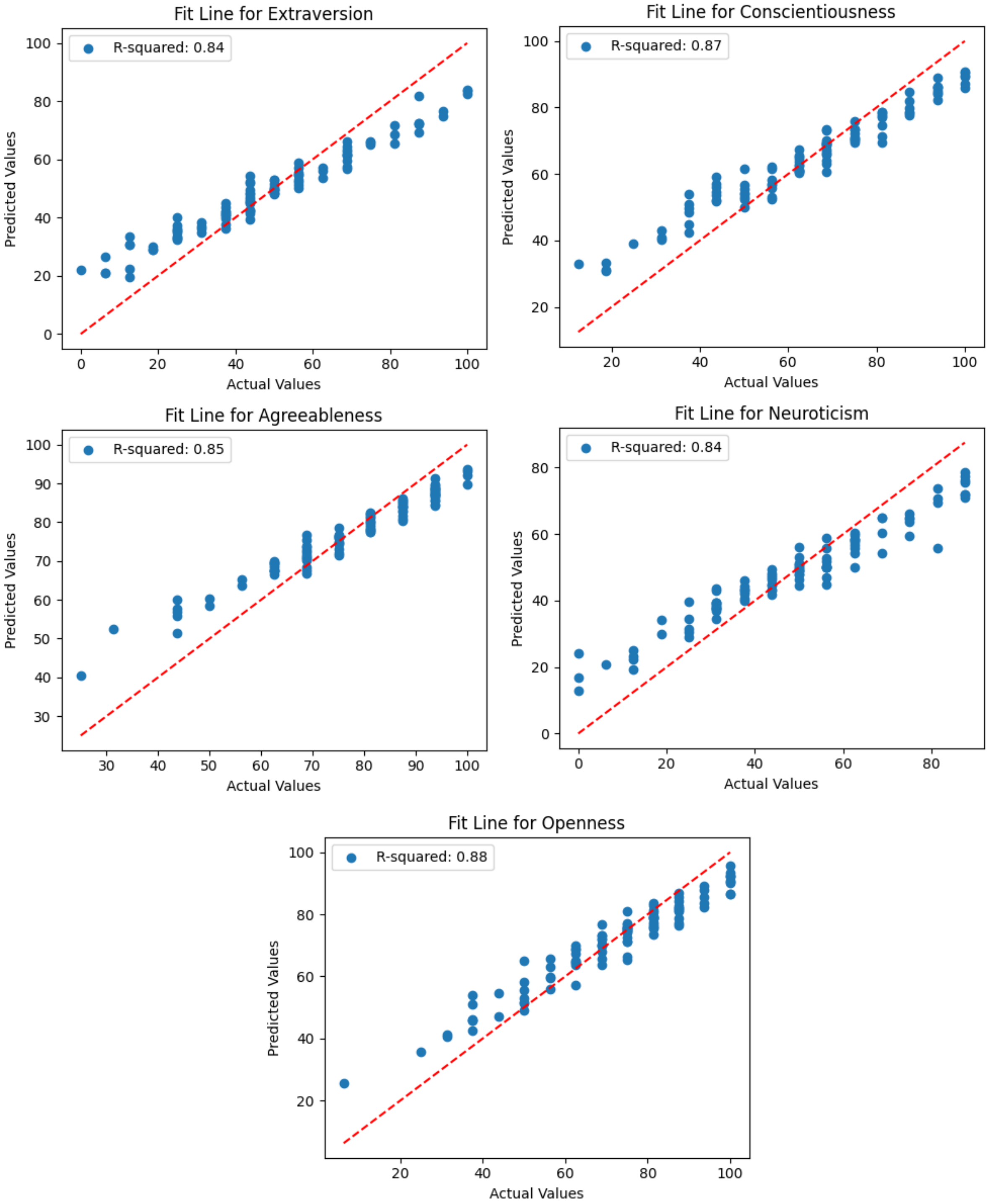


Figure 5.6: Fit Line Plots for Dataset 4

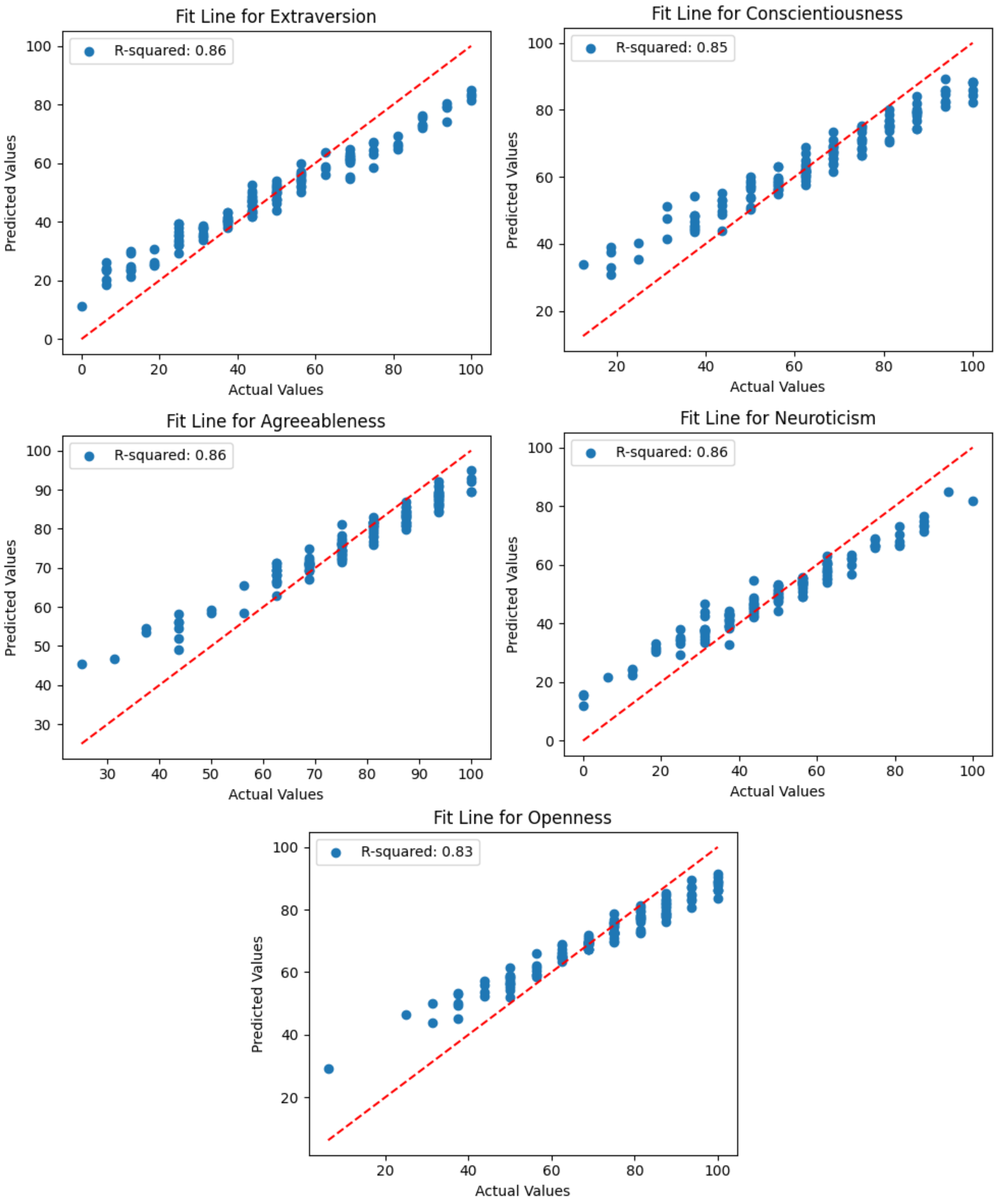


Figure 5.7: Fit Line Plots for Dataset 5

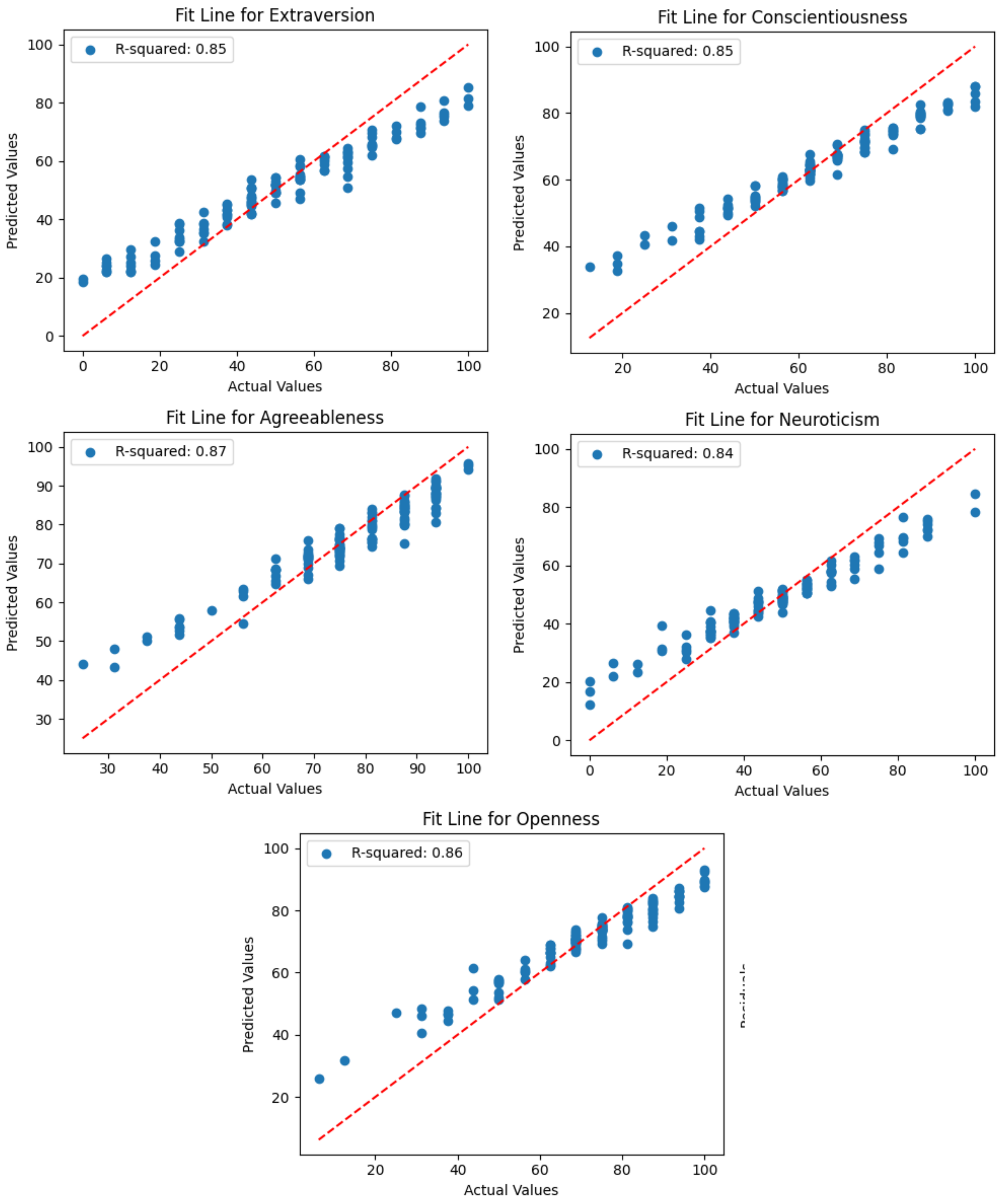


Figure 5.8: Fit Line Plots for Dataset 6

parameter, revealing a model with substantial systematic error. In contrast, Random Forest models display random residuals, suggesting that the model is more effective at capturing the underlying patterns based on independent variables but remains sensitive to noise.

5.2 Classification Results

In order to compare and illustrate the prediction of the classification models, statistical analysis was performed. Full details are mentioned in the following sections.

5.2.1 Statistical Analysis (Accuracy)

Figure 5.15 present a comparison of classification model accuracies for the datasets. In Figure 5.15a, SVM exhibits the highest accuracy, achieving 67% in predicting openness. Random Forest classifier outperforms other models in predicting extraversion, with an accuracy of 74%. For agreeableness and conscientiousness, the XGBoost model demonstrates the best accuracy at 74% and 63%, respectively. Unfortunately, none of the models achieve higher accuracy than the baseline Zero Rule model for predicting neuroticism.

In Figure 5.15b, the decision tree provides the highest accuracy for predicting extraversion and conscientiousness. XGBoost yields the best result for agreeableness at 74%. Neuroticism is predicted with similar accuracies across all models, with KNN classifier and random forest both achieving 61%. Random Forest leads the way in predicting openness with an accuracy rate of 70%.

Figure 5.15c highlights that XGBoost is the most accurate predictor for openness (71%) and conscientiousness (67%). Decision trees perform well in predicting extraversion and neuroticism, both achieving an accuracy score of 67%. In the case of extraversion, other models, such as Random Forest or XGBoost, are not far behind.

In Figure 5.15d, XGBoost delivers an exceptional classification accuracy of 81% for agreeableness. For conscientiousness, both KNN and SVM models achieve equal accuracy. However, all models underperform compared to the baseline in predicting openness. Decision trees offer the best accuracy of 67% for neuroticism and extraversion.

Figure 5.15e shows that random forest is the most effective classifier for predicting openness and conscientiousness, both with an accuracy of 73%. SVM provides 69% accuracy for agreeableness. XGBoost outperforms other models in predicting neuroticism with an accuracy rate of 63%. Unfortunately, none of the models can predict extraversion better than the baseline model.

In Figure 5.15f, a comparison of model accuracy is provided for Dataset 6 from Table 33. The KNN classifier offers the highest accuracy for predicting neuroticism, agreeableness, and extraversion. It is also the second-best model for conscientiousness. XGBoost achieves a classification accuracy of 69% for agreeableness. Only SVM attains an accuracy as high as 77% for classifying openness traits.

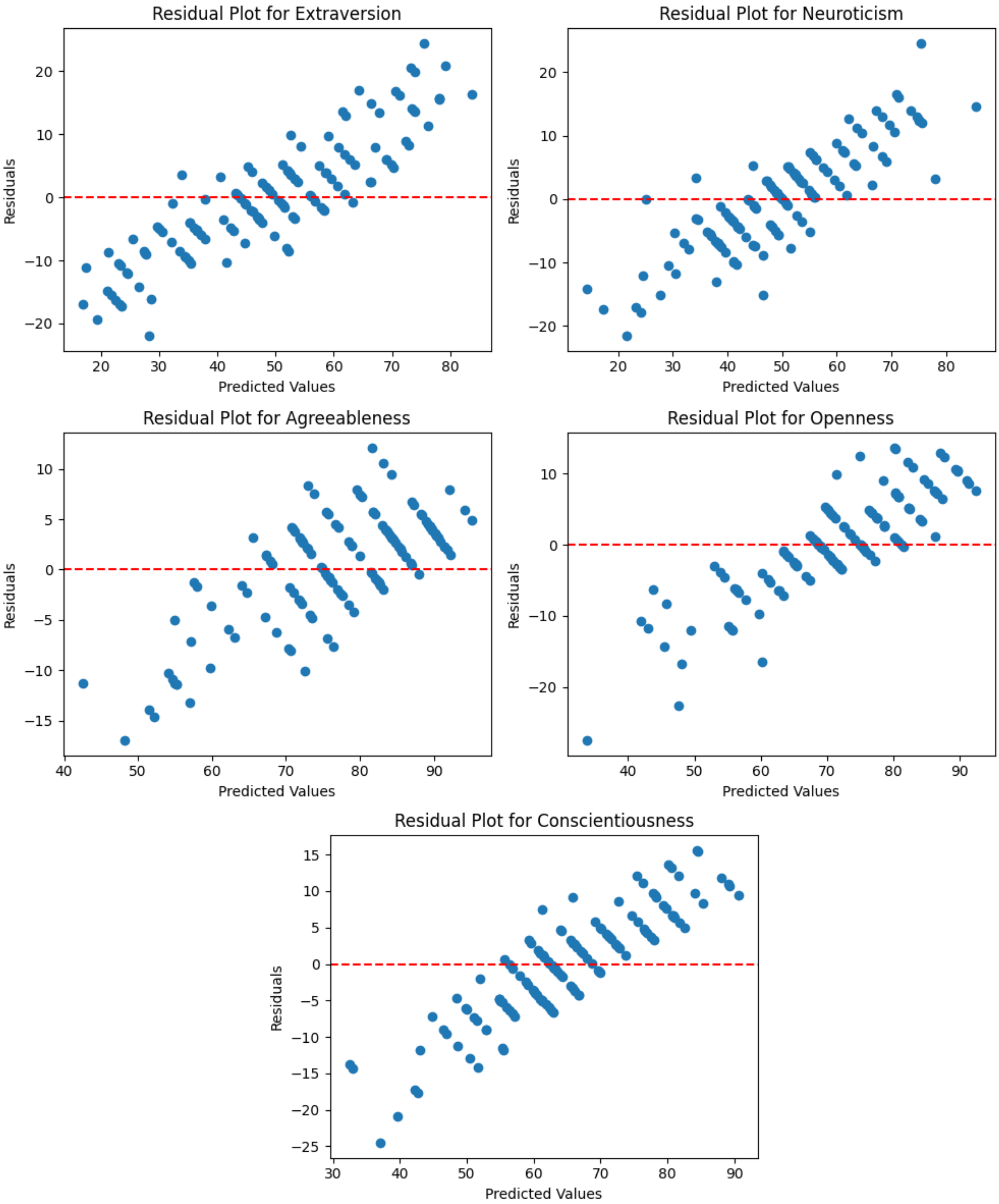


Figure 5.9: Residual Plots for Random Forest Models on Dataset 1

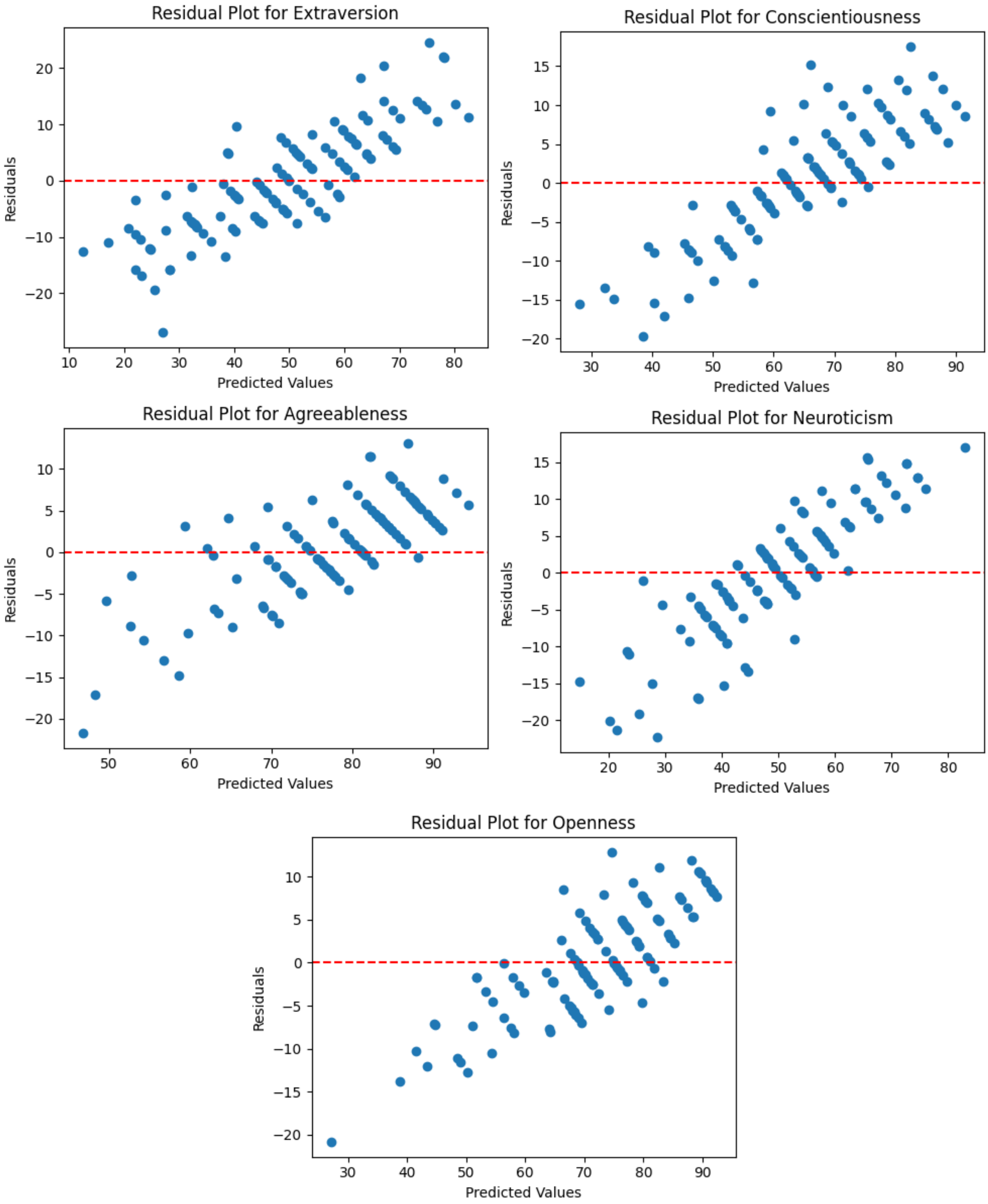


Figure 5.10: Residual Plots for Random Forest Models on Dataset 2

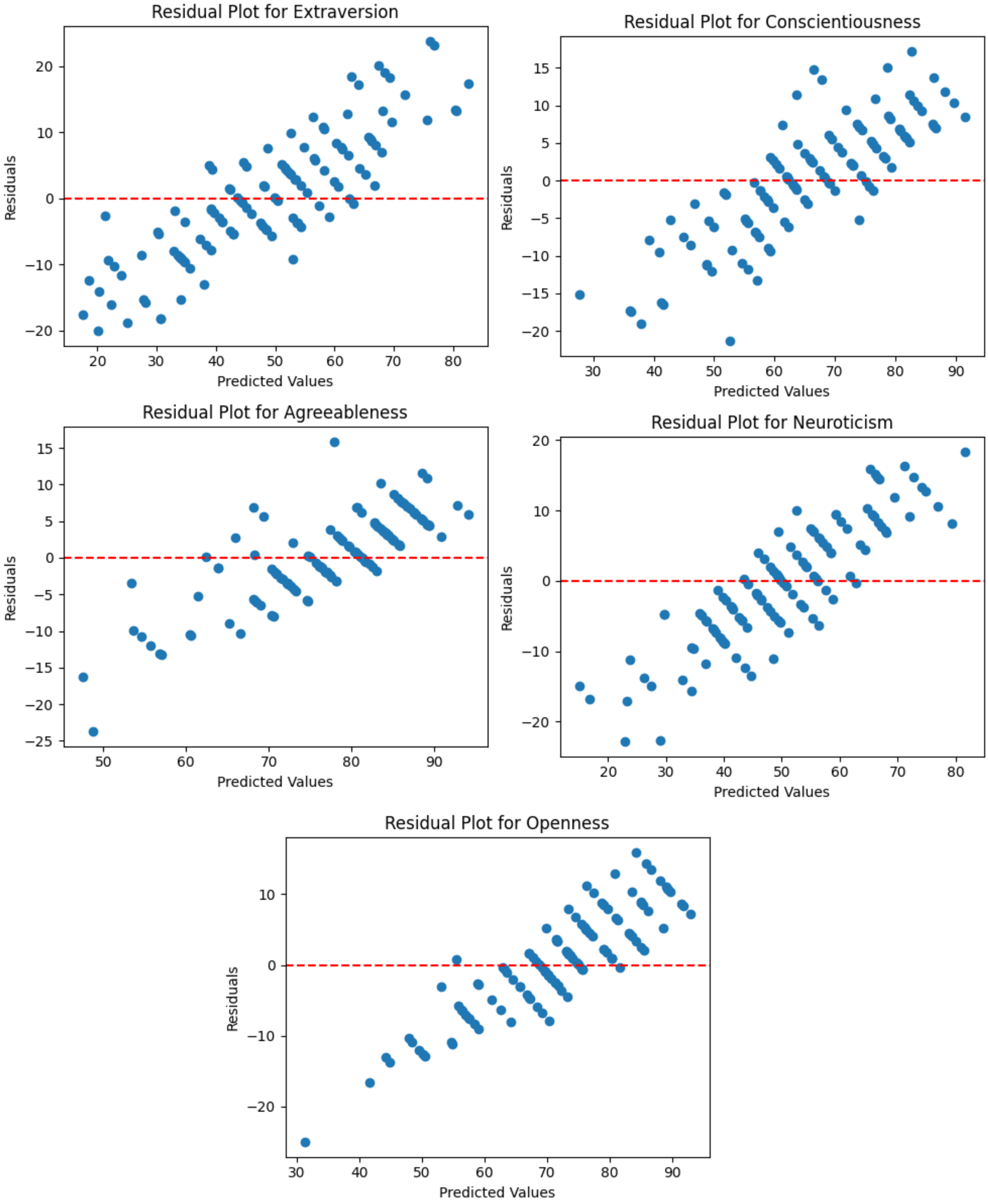


Figure 5.11: Residual Plots for Random Forest Models on Dataset 3

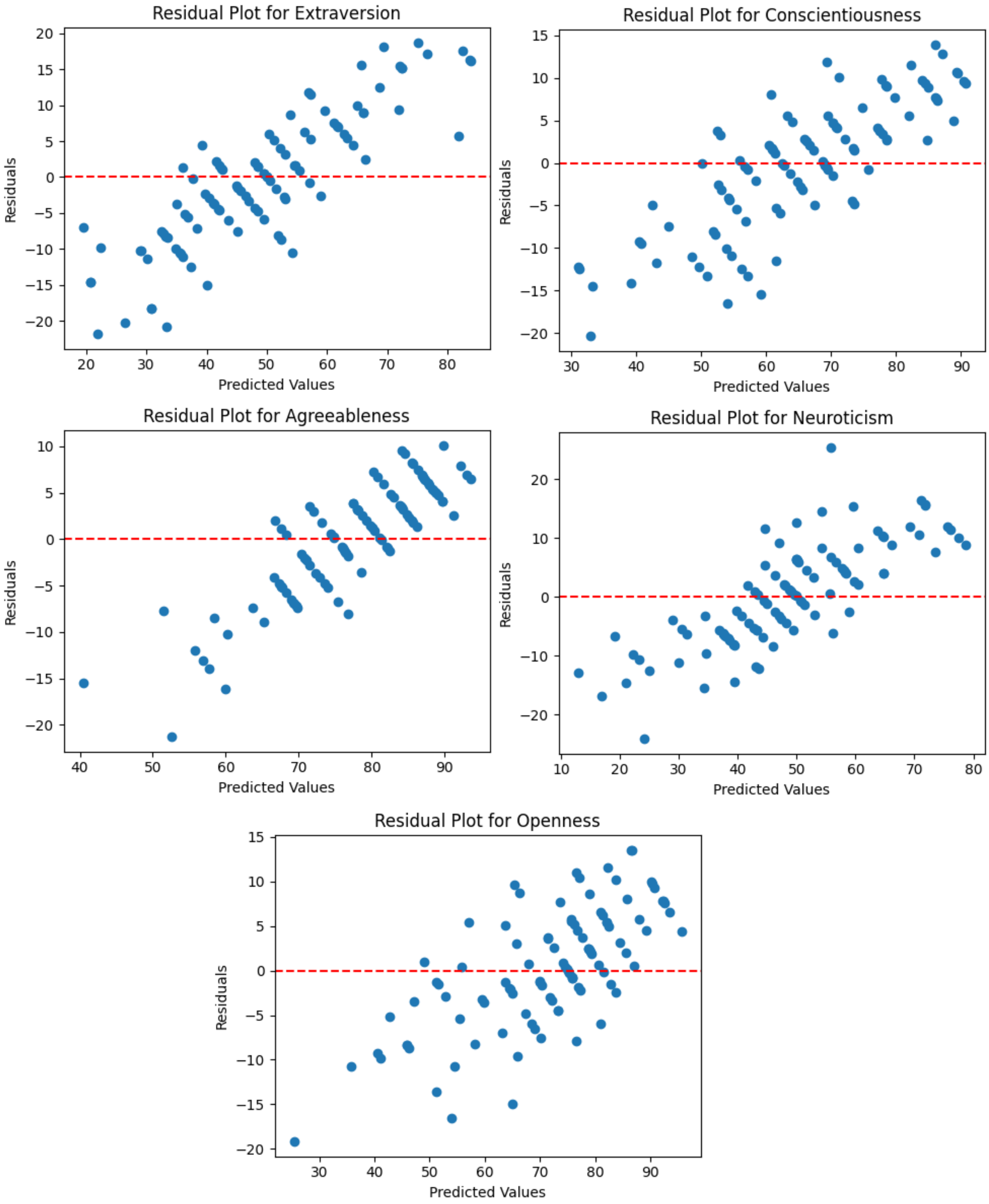


Figure 5.12: Residual Plots for Random Forest Models on Dataset 4

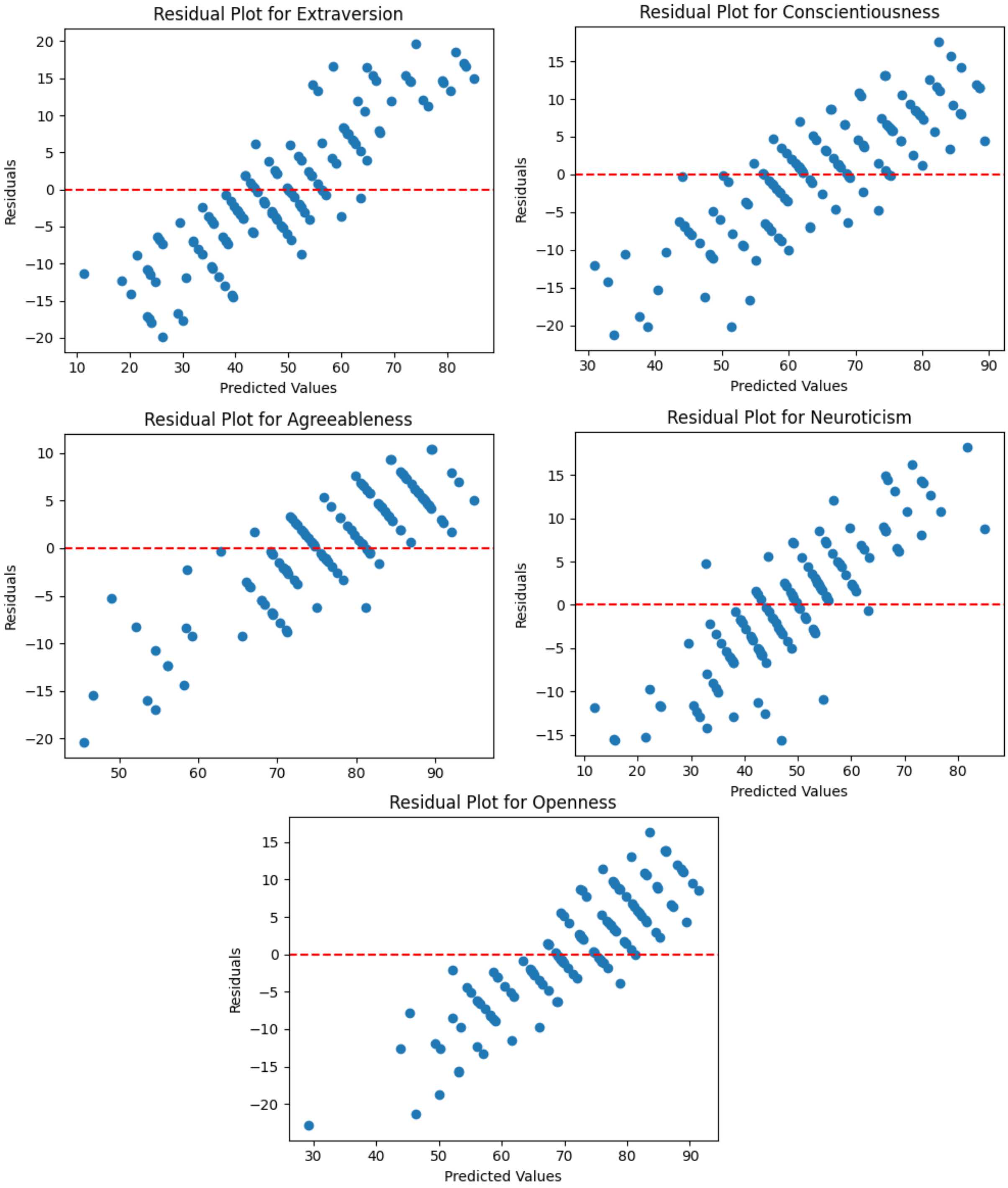


Figure 5.13: Residual Plots for Random Forest Models on Dataset 5

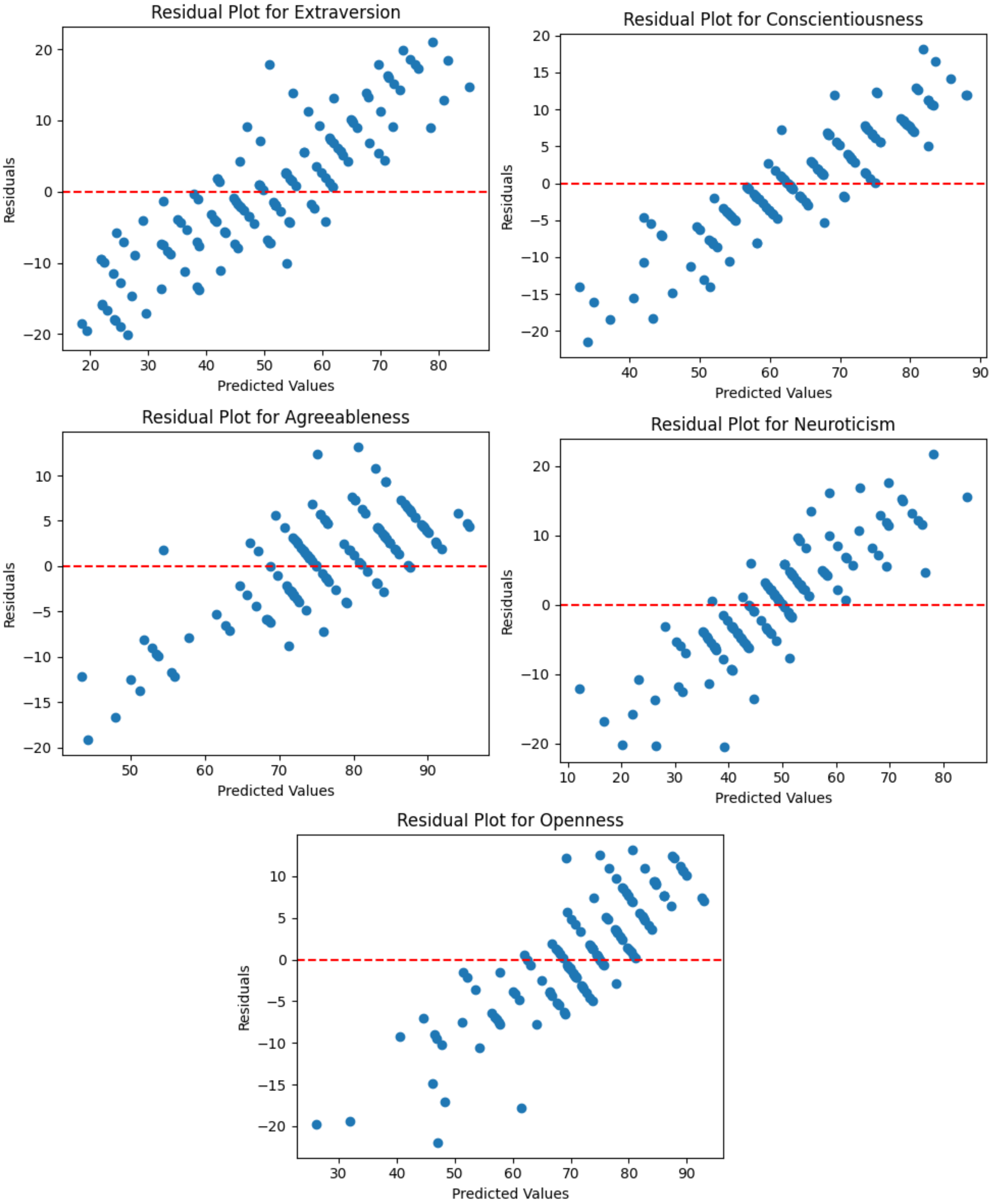
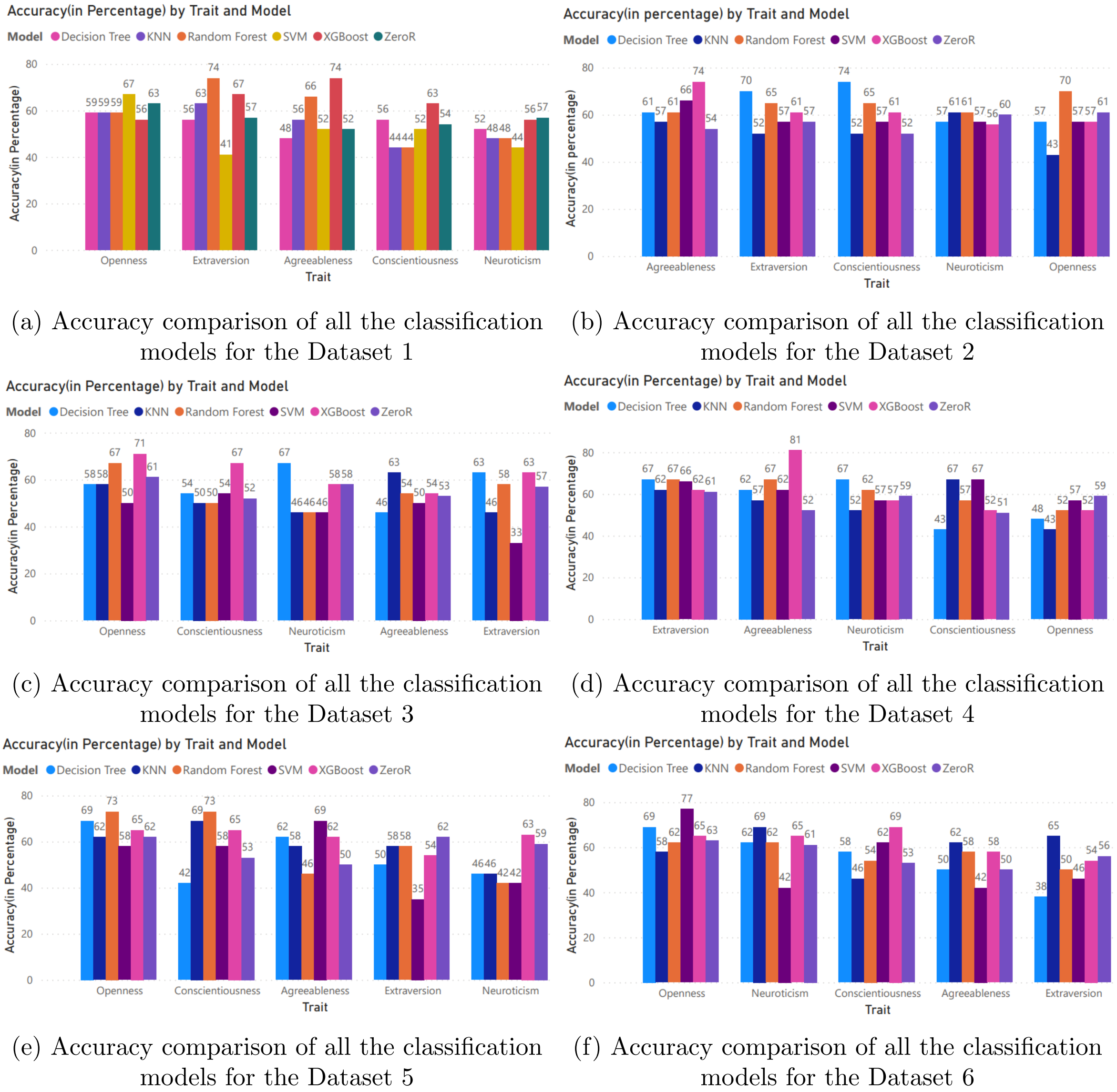


Figure 5.14: Residual Plots for Random Forest Models on Dataset 6



(a) Accuracy comparison of all the classification models for the Dataset 1

(b) Accuracy comparison of all the classification models for the Dataset 2

(c) Accuracy comparison of all the classification models for the Dataset 3

(d) Accuracy comparison of all the classification models for the Dataset 4

(e) Accuracy comparison of all the classification models for the Dataset 5

(f) Accuracy comparison of all the classification models for the Dataset 6

Figure 5.15: Accuracy Comparison

6 Discussion and Future Work

This work aimed to create machine learning models capable of predicting personality traits based on real-world activities, habits, and interaction behavior. Ensemble models such as XGBoost and random forest were utilized, and they yielded promising results. In regression problems, our models achieved errors ranging from 6% to 25%, which is comparable to previous research [24, 44] in the field and in some cases better than previous research [2]. Also all our results in classification models were comparable to previous research [6, 53].

Importantly, our findings did not exclusively rely on app usage data, a common focus in prior studies. Instead, we incorporated diverse behavioral parameters, including charging time, doze events, and ring mode, in conjunction with demographic characteristics such as gender and academic department enrollment. This investigation into the relationship between personality and behavior, along with the influence of variables like gender and academic discipline on model performance, constitutes a noteworthy secondary contribution of this study. It also serves as a potential point of reference for fellow researchers in this field.

Our approach to this problem encompassed both classification and regression perspectives. The error metrics we attained were significantly better compared to the baseline. Although the accuracy metrics we attained were akin to, and sometimes slightly superior to, the baseline, it is essential to acknowledge that this work represents a significant stride in reporting the reliability of estimates. In the existing body of literature, the predominant portion of research studies that address personality as a classification problem have disclosed accuracies spanning from 50% to 83% when personality traits are grouped into two categories [1, 7]. However, such classification studies come with certain limitations.

Firstly, model performance is influenced by the number of personality bins used, with studies employing three-class categorizations reporting accuracies between 50% and 70% [31, 50, 53, 54], whereas those adopting a two-bin approach achieved higher accuracy rates of up to 83%. Furthermore, some studies have dichotomized personality traits into merely "low" and "high" categories, which may not effectively distinguish between individuals. In contrast, regression modeling provides a more precise means of addressing this challenge. Among the limited number of studies that have tackled personality prediction as a regression problem, the best-performing study reported root mean squared errors of approximately 28% [2]. Our regression models exhibited 70% better results compared to the best performing previous research. While a couple of regression studies have reported mean absolute error(MAE) values around 11%, it's important to note that these errors cannot be directly compared to root mean squared errors due to their inherent difference in properties [15]. The limitation of the mean absolute error is that it balances negative errors with positive errors, resulting in lower error values. In contrast, the root mean squared error, squares the errors, thereby preserving the magnitude of the error.

In the case of the Extraversion trait, a previous study reported a higher accuracy of 89.1%, though with a smaller population size of 4 [39]. The next best study for the Extraversion trait achieved an accuracy of 76.8% which is slightly better than ours (70%) [50]. For the Neuroticism trait, prior research outperformed this study with an accuracy of 80.2%, whereas this study achieved an accuracy of 69% [50]. From a pragmatic perspective, tree based models used in this work such as Decision Tree and XGBoost as well as non-linear models like KNN are more useful and practical and hence should be used as a starting point for future research. The baseline model, operating under the assumption of population homogeneity, lacks utility for distinguishing individuals based on personality traits. Nevertheless, it is important to emphasize that this form of analysis, where performance is assessed in comparison to a sensible baseline, holds significance when dealing with non-linear models of intricate behavior. This is crucial because linear statistical metrics can yield deceptive outcomes.

Statistical metrics such as error and accuracy cannot be used alone to assess or compare the quality of machine learning models. In some instances the baseline model was outperforming our other models. This behavior can be attributed to less hyperparameter tuning and use of a very limited number of simple models. The results highlighted the need for additional analyses, including confusion matrices and ROC curves, to validate the quality of classification models. These analytical methods were lacking in the existing literature, where the quality of models was primarily reported using simple statistical metrics. Overall, when examining the classification model results, it became evident that non-linear models exhibited a relatively higher capacity to generalize the data. The primary factors contributing to the suboptimal generalizability of our regression models can be attributed to two main issues. Firstly, the level of effort dedicated to fine-tuning the models was notably low. Secondly, suboptimal hyper-parameter selection, such as choosing kernels for SVM models, which inherently dictate the structure of the input data, played a crucial role. A more rigorous and comprehensive approach to hyper-parameter selection, coupled with extensive training, could significantly enhance the accuracy of the classification models.

This work makes three important contributions to the literature and smartphone sensing research:

Greater Sensor Diversity: Our utilization of a total of nine sensors, coupled with various demographic and additional variables like gender and department, introduces a significant departure from previous endeavors in terms of the factors taken into consideration, which may have relevance to personality traits. The incorporation of a broader range of sensor data ensures that our predictive factors closely mimic real-life scenarios.

Continuous Personality Assessment: Previous models have predominantly focused on discretizing each personality trait into two or three distinct categories and subsequently treated the problem as a classification task. This study stands out as one of the very few research initiatives that offer a continuous representation of Big Five personality traits by establishing a regression model utilizing easily accessible smartphone sensor data with reliable results.

Enhanced Methodological Approach: A substantial proportion of previous works have not furnished a comprehensive performance analysis or conducted comparisons with baseline algorithms. In contrast, our study not only employed a baseline algorithm known as 'assume population mean' or 'naive baseline' but also engaged in a performance comparison employing both aggregate metrics and residuals. This approach sets a new benchmark for the reporting of personality estimation from mobile devices.

The following observations are made in this study:

Throughout our analysis, a pattern emerged, indicating that the average step count tends to be higher for introverted individuals consistently throughout the day. In contrast, disagreeable individuals exhibit an inclination to walk more during the latter part of the day. Furthermore, conscientious individuals display a preference for listening to music in the morning hours, while neurotic individuals tend to do so in the evening, especially between 5 pm and 9 pm. These findings align with prior research, such as [6], providing additional support to their conclusions. Another significant observation pertains to individuals scoring high in conscientiousness, who demonstrate increased smartphone activity throughout the day, as evidenced by the strong positive correlation between conscientiousness and average screen event data. This finding corroborates the results of earlier research conducted by [24]. Our analysis also unveiled the correlation between the normal ring mode and conscientiousness, suggesting that more organized individuals tend to prefer the normal ring mode. This observation adds a valuable dimension to our understanding of how personality traits influence smartphone usage patterns. Furthermore, our research extended to the analysis of app usage and its correlation with personality traits. It was evident that extroverted individuals tend to utilize communication, finance, maps and navigation, and personalization apps more frequently. In contrast, introverts display a greater inclination towards using game apps, reflecting their preference for solitary activities, in line with the findings of [38]. Highly agreeable individuals show a propensity for engaging with travel and local apps while exhibiting a reduced preference for news and video player applications. Conscientious individuals, on the other hand, demonstrate a proclivity for events and sports apps, possibly influenced by their orderliness dimension, and a decreased preference for maps, navigation, and health and fitness apps. Emotionally stable individuals exhibit reduced usage of beauty apps but show a greater affinity for finance and video apps. Open-minded individuals tend to employ art and design apps more frequently, aligning with their openness to new experiences and learning. Additionally, open individuals are more likely to engage with personalization and social apps, suggesting their inclination for creatively customizing smartphone interfaces, indicative of their creative dimension.

This work leveraged an important dataset of sensor information, socio-demographic data and BFI Scores collected from 149 participants over a 2 months period to establish its findings. This suffices for establishing the proof of concept and demonstrating feasibility, which serves as the primary contribution of this thesis. However, it is worth noting that the size and composition of the participants' personality profiles may not entirely mirror the characteristics of the general population. Specifically, the participant group exhibited a skew towards higher conscientiousness and agreeableness, a phenomenon commonly expected in a university population. Moreover, there was a noticeable class skew towards the mean of the sample, with limited representation at the extremes of any of the personality dimensions. This skew had the effect of inflating the effectiveness of the baseline algorithm since the higher density of average participants reduced the error contributions from the outliers. Additionally, it's important to acknowledge that the population under study consisted of individuals from the university community, which imposed similar constraints on their daily schedules and lifestyles. These factors collectively impede the generalizability of the models generated within this study. However, the approach itself, designed and tested here, should remain applicable and transferable to larger and more diverse populations, where the limitations observed in this specific sample may not be as pronounced.

This work is subject to several limitations, which, in turn, offer valuable directions for future research. Firstly, the generalizability of our findings is constrained by the homogeneity of the demographic factors, encompassing both the personalities represented and the behaviors exhibited. To extend the validity of this approach, further studies should encompass a broader cross-section of demographics.

Secondly, there is room for improvement in the consistency of data collection across all sensors to ensure a more comprehensive dataset. The merging of multiple datasets resulted in the loss of nearly one-third of the participants' data due to data unavailability for certain sensors. Enhancing data consistency could alleviate this issue.

Thirdly, the machine learning algorithms employed in this study, especially classification models, yielded limited success. While ensemble regression models demonstrated low error rates compared to previous research, the predictability of the classification models remained suboptimal. Consequently, there is potential for the implementation of more advanced statistical models that may offer similar precision and instill greater confidence in delineating causal pathways. In the absence of such models, sensitivity analysis could be conducted on the existing model to ascertain the impact of including various features in the estimation of specific personality traits

6.1 Summary

This study has showcased the remarkable potential to discern an individual's personality traits over the course of several days, all through the subtle analysis of behavioral data gathered from everyday smartphones. This innovative approach introduces a fresh paradigm, utilizing smartphone usage data as the key to unlocking one's personality, all without the need to delve into app-specific and social media content.

Furthermore, the inclusion of a broader array of sensors and the consideration of specific factors like gender and educational background have significantly enhanced the performance compared to previous findings in the literature. This research holds promise for a wide spectrum of applications, spanning the realms of social sciences and public health, where it can serve as a means to automatically detect potential confounding variables. Additionally, in the field of marketing, it offers a streamlined method for extracting more comprehensive consumer behaviors, while in the service sectors, it can be harnessed to discern preferences for personalized experiences.

Although smartphones currently stand as the most prevalent devices boasting considerable sensing capabilities, it's worth noting that they are not the sole technology at our disposal for such investigations. A myriad of other devices are gaining increased acceptance, some of which are purposefully engineered for ongoing data collection. For instance, smartwatches and fitness trackers, explicitly designed for the continuous monitoring of sleep patterns and physical activity, are now becoming more commonplace. As technology continues to evolve, this research has the potential to expand its horizons, encompassing a wider array of unobtrusive sensing techniques.

While this work marks a promising initial step and a proof of concept, it underscores the need for extensive further research to expand, validate, and apply these findings across a diverse range of disciplines.

Bibliography

[1] Margit Antal, László Zsolt Szabó, and Gy{\textbackslash}Hoz{\textbackslash}Ho Nemes. "Predicting user identity and personality traits from mobile sensor data". In: *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings 22.* Springer, 2016, pp. 163–173.

[2] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. "Personality and patterns of Facebook usage". In: *Proceedings of the 4th annual ACM web science conference.* 2012, pp. 24–32.

[3] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. "Facebook profiles reflect actual personality, not self-idealization". In: *Psychological science* 21.3 (2010). Publisher: Sage Publications Sage CA: Los Angeles, CA, pp. 372–374.

[4] EG Boring. "A history of experimental psychology. New York: Appleton-Century-Crofts, 1950". In: *Sensation and perception in the history of experimental psychology. New* (1950).

[5] Sarah Butt and James G Phillips. "Personality and self reported mobile phone use". In: *Computers in human behavior* 24.2 (2008). Publisher: Elsevier, pp. 346–360.

[6] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. "Mining large-scale smartphone data for personality studies". In: *Personal and Ubiquitous Computing* 17 (2013). Publisher: Springer, pp. 433–450.

[7] Gokul Chittaranjan, Blom Jan, and Daniel Gatica-Perez. "Who's who with big-five: Analyzing and classifying personality traits with smartphones". In: *Proceedings - International Symposium on Wearable Computers, ISWC.* Type: Conference paper. 2011, pp. 29 –36. DOI: 10.1109/ISWC.2011.29. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-80051958495&doi=10.1109%2fISWC.2011.29&partnerID=40&md5=b1052d3d7fa261664713996b64960e79.

[8] *Choose a category and tags for your app or game - Play Console Help.* Google Play Console Help. URL: https://support.google.com/googleplay/android-developer/answer/9859673?hl=en#zippy=%2Capps%2Cgames (visited on 10/11/2023).

[9] Paul T Costa and Robert R McCrae. "The revised neo personality inventory (neo-pi-r)". In: *The SAGE handbook of personality theory and assessment* 2.2 (2008), pp. 179–198.

[10] M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. "The mini-IPIP scales: tiny-yet-effective measures of the Big Five factors of personality." In: *Psychological assessment* 18.2 (2006). Publisher: American Psychological Association, p. 192.

[11] Nathan Eagle and Alex Pentland. "Reality mining: sensing complex social systems". In: *Personal and ubiquitous computing* 10 (2006). Publisher: Springer, pp. 255–268.

[12] Elisabeth Engelberg and Lennart Sjöberg. "Internet use, social skills, and adjustment". In: *Cyberpsychology & behavior* 7.1 (2004). Publisher: Mary Ann Liebert, Inc., pp. 41–47.

[13] David C Funder. "Accurate personality judgment". In: *Current Directions in Psychological Science* 21.3 (2012). Publisher: Sage Publications Sage CA: Los Angeles, CA, pp. 177–182.

[14] Fausto Giunchiglia et al. "A worldwide diversity pilot on daily routines and social practices (2020)". In: (2021). Publisher: University of Trento.

[15] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. “Predicting personality from twitter”. In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing.* IEEE, 2011, pp. 149–156.

[16] Lewis R Goldberg. “From Ace to Zombie: Some explorations in the language of personality”. In: *Advances in personality assessment* 1 (1982), pp. 203–234.

[17] Lewis R Goldberg. “The development of markers for the Big-Five factor structure.” In: *Psychological assessment* 4.1 (1992). Publisher: American Psychological Association, p. 26.

[18] Samuel D Gosling. “Personality in non-human animals”. In: *Social and Personality Psychology Compass* 2.2 (2008). Publisher: Wiley Online Library, pp. 985–1001.

[19] Ronald K Hambleton and Russell W Jones. “Comparison of classical test theory and item response theory and their applications to test development”. In: *Educational measurement: issues and practice* 12.3 (1993), pp. 38–47.

[20] Yair Amichai Hamburger and Elisheva Ben-Artzi. “The relationship between extraversion and neuroticism and the different uses of the Internet”. In: *Computers in human behavior* 16.4 (2000). Publisher: Elsevier, pp. 441–449.

[21] Gabriella M. Harari, Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. “Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges”. In: *Perspectives on Psychological Science* 11.6 (2016). _eprint: https://doi.org/10.1177/1745691616650285, pp. 838–854. DOI: 10.1177/1745691616650285. URL: https://doi.org/10.1177/1745691616650285.

[22] Peter Hills and Michael Argyle. “Uses of the Internet and their relationships with individual differences in personality”. In: *Computers in Human Behavior* 19.1 (2003). Publisher: Elsevier, pp. 59–70.

[23] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. “Predicting students’ happiness from physiology, phone, mobility, and behavioral data”. In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 2015, pp. 222–228.

[24] Naveen K Kambham, Kevin G Stanley, and Scott Bell. “Predicting personality traits using smartphone sensor data and app usage data”. In: *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON).* IEEE, 2018, pp. 125–132.

[25] Alan E Kazdin. “Unobtrusive measures in behavioral assessment”. In: *Journal of Applied Behavior Analysis* 12.4 (1979). Publisher: Wiley Online Library, pp. 713–724.

[26] Breiman Leo. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–23.

[27] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. “Moodscope: Building a mood sensor from smartphone usage patterns”. In: *Proceeding of the 11th annual international conference on Mobile systems, applications, and services.* 2013, pp. 389–402.

[28] Wim J van der Linden and Ronald K Hambleton. *Handbook of modern item response theory.* Springer Science & Business Media, 2013.

[29] Robert R McCrae and Paul T Costa. “Empirical and theoretical status of the five-factor model of personality traits”. In: *The SAGE handbook of personality theory and assessment* 1 (2008), pp. 273–294.

[30] Peter F Merenda. “Toward a four-factor theory of temperament and/or personality”. In: *Journal of personality assessment* 51.3 (1987). Publisher: Taylor & Francis, pp. 367–374.

[31] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Pentland. “Predicting personality using novel mobile phone-based metrics”. In: *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings 6.* Springer, 2013, pp. 48–55.

[32] Isabel Briggs Myers. *The Myers-Briggs Type Indicator: Manual (1962).* Consulting Psychologists Press, 1962. URL: `https://books.google.it/books?id=EV55tgAACAAJ`.

[33] Bjarke Mønsted, Anders Mollgaard, and Joachim Mathiesen. "Phone-based metric as a predictor for basic personality traits". In: *Journal of Research in Personality* 74 (2018). Publisher: Elsevier, pp. 16–22.

[34] Rodrigo de Oliveira, Alexandros Karatzoglou, Pedro Concejero Cerezo, Ana Armenta Lopez de Vicuña, and Nuria Oliver. "Towards a psychographic user model from mobile phone usage". In: *CHI'11 Extended Abstracts on Human Factors in Computing Systems.* 2011, pp. 2191–2196.

[35] Daniel J Ozer and Veronica Benet-Martinez. "Personality and the prediction of consequential outcomes". In: *Annu. Rev. Psychol.* 57 (2006). Publisher: Annual Reviews, pp. 401–421.

[36] Ella Peltonen, Parsa Sharmila, Kennedy Opoku Asare, Aku Visuri, Eemil Lagerspetz, and Denzil Ferreira. "When phones get personal: Predicting Big Five personality traits from application usage". In: *Pervasive and Mobile Computing* 69 (2020). Publisher: Elsevier, p. 101269.

[37] Leif E Peterson. "K-nearest neighbor". In: *Scholarpedia* 4.2 (2009), p. 1883.

[38] James G Phillips, Sarah Butt, and Alex Blaszczynski. "Personality and self-reported use of mobile phones for games". In: *CyberPsychology & Behavior* 9.6 (2006). Publisher: Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA, pp. 753–758.

[39] Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. "Multimodal recognition of personality traits in social interactions". In: *Proceedings of the 10th international conference on Multimodal interfaces.* 2008, pp. 53–60.

[40] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. "Our twitter profiles, our selves: Predicting personality with twitter". In: *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing.* IEEE, 2011, pp. 180–185.

[41] Georg Rasch. "Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests." In: (1960). Publisher: Nielsen & Lydiche.

[42] Craig Ross, Emily S Orr, Mia Sisic, Jaime M Arseneault, Mary G Simmering, and R Robert Orr. "Personality and motivations associated with Facebook use". In: *Computers in human behavior* 25.2 (2009). Publisher: Elsevier, pp. 578–586.

[43] Tracii Ryan and Sophia Xenos. "Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage". In: *Computers in human behavior* 27.5 (2011). Publisher: Elsevier, pp. 1658–1664.

[44] Dominik Rüegger. "Studying Personality Change with Smartphone Data". PhD thesis. ETH Zurich, 2021.

[45] Gerard Saucier and Lewis R Goldberg. "What is beyond the Big Five?" In: *Journal of personality* 66 (1998). Publisher: DUKE UNIVERSITY PRESS, pp. 495–524.

[46] A Schizzerotto and F Denti. "Perduti e in ritardo. L'esperienza dell'abbandono e dell'irregolarità degli studi in cinque leve di immatricolati all'ateneo di Milano-Bicocca. Rapporto di ricerca". In: *Rapporto di ricerca* (2005).

[47] Carsten Schmitz and others. "LimeSurvey: An open source survey tool". In: *LimeSurvey Project Hamburg, Germany. URL http://www. limesurvey. org* (2012).

[48] Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. "Predicting user traits from a snapshot of apps installed on a smartphone". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 18.2 (2014). Publisher: ACM New York, NY, USA, pp. 1–8.

[49] Sonu Shamdasani. *Jung and the making of modern psychology: The dream of a science.* Cambridge University Press, 2003.

[50] Jianqiang Shen, Oliver Brdiczka, and Juan Liu. “Understanding email writers: Personality prediction from email messages”. In: *User Modeling, Adaptation, and Personalization: 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013 Proceedings 21*. Springer, 2013, pp. 318–330.

[51] *The WeNet Project*. WeNet. Feb. 21, 2019. URL: `https://www.internetofus.eu/project/` (visited on 10/11/2023).

[52] Eugene J Webb, Donald T Campbell, Richard D Schwartz, and Lee Sechrest. *Unobtrusive measures*. Vol. 2. Sage Publications, 1999.

[53] William R Wright and David N Chin. “Personality profiling from text: introducing part-of-speech N-grams”. In: *User Modeling, Adaptation, and Personalization: 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings 22*. Springer, 2014, pp. 243–253.

[54] Runhua Xu, Remo Manuel Frey, Elgar Fleisch, and Alexander Ilic. “Understanding the impact of personality traits on mobile app adoption–Insights from a large-scale field study”. In: *Computers in Human Behavior* 62 (2016). Publisher: Elsevier, pp. 244–256.

[55] Mattia Zeni, Ivano Bison, Britta Gauckler, and Fernando Reis Fausto Giunchiglia. “Improving time use measurement with personal big data collection–the experience of the European Big Data Hackathon 2019”. In: *arXiv preprint arXiv:2004.11940* (2020).