FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Parsa Sharmila

PREDICTION OF BIG FIVE PERSONALITY TRAITS FROM MOBILE APPLICATION USAGE

Master's Thesis
Degree Programme in Computer Science and Engineering
June 2020

Sharmila P. (2020) Prediction of Big Five Personality Traits from Mobile Application Usage. University of Oulu, Degree Programme in Computer Science and Engineering, 72 p.

ABSTRACT

Smartphones evolved being an integral part of our daily lives and in recent days. Studies show that smartphone usage is correlated to user personality traits. This critical ecosystem is dependent on several variables such as geographic location, demographic traits, ethnic impact or cultural influence and so on. While significant number of demographic, environmental and medical analysis is done based on smartphone usage, there are inadequate amount of study carried out to analyse human personality. All of these information provide pivotal insights for improving user experience, creating recommendations, identifying marketing strategies and for a general overall usage improvement. This study is done with application usage data collected over 6 months from 739 Android smartphone users along with a 50-item Big Five Personality Trait questionnaire. The analysis focuses on the fact that, category-level aggregated application usage is enough for predicting Big Five personality traits achieving 9-14% error which is 86-91% accuracy on average. This study concludes that user personality generates a fundamental impact on people's smartphone application and application category usage. This work reflects the possible personality-driven research in future and depicts the significance and involvement of application categories in achieving proper accuracy in general traits, while pursuing for personality study.

Keywords: Prediction, Personality, Smartphone usage.

TABLE OF CONTENTS

1. INTRODUCTION

Smartphones have become an essential part of people's day today life [1]. In this era of Ubiquitous computing, people are adapting new smart technologies which mostly depend on smartphones. Ubiquitous computing refers to the availability of the computers in everyday physical life [2]. According to the study of Weiser *et al.* [2], ubiquitous computing has been creating significant impact on areas of computer hardware, network protocols, security and of course applications and software. The impact of ubiquitous computing is conspicuous in the smartphone usage. Now smartphones are not just phones. The daily usage of smartphone for everyday facilities like communication, transportation, entertainment, education and so on, have been increasing worldwide. Smartphones are handy, small, mobile and they are capable of providing many functionalities of a personal computer in daily life. Mobile internet technologies have made it easy to communicate with people, locate places, access and share information from anywhere at anytime. With variety of mobile applications, smartphone users are carrying a full sized computer in their pockets. Mobility and simplicity of some popular mobile operating systems like Android, iOS, Windows etc. have made smartphones popular among large age group of users, even children and elderly people.

This thesis focuses on how personality of the smartphone users is impacted by their smartphone application usage. Personality itself is a vast topic and has wide range of versatility depending on person to person as behavioral and emotional pattern evolves based on environmental and biological factors [3]. Personality is something that comes out as a result of someone's psychological interaction with environment [4]. Individual personality study can contribute to the studies of predicting human outcomes in broader range. Such study can capture sociocultural components which can be used to understand in long-term biopsychosocial procedures [5]. The word "Biopsychosocial" is basically the interrelation between biological, psychological and sociological aspects of individuals. According to the study of Martin *et al.* [5] different traits of personality like conscientiousness, extraversion, agreeableness can make relevance to vital human outcomes like life-span mortality risk, physical health related concerns, mental health state etc. For an example, Martin's [5] study states that, one of personality traits - conscientiousness, can play a key role in biopsychosocial aspect because this trait may link to health conscious behavior of individual.

This initial finding can lead to bigger researches like human health behaviors. Personality concepts are yet to be explored more for such biopsychosocial studies. Because of the larger context of human personality, it has always been a challenge to predict the relation between smartphone usage and user personality. There are different methods of measuring personality and in the field of analytical research, self-reported measurements are mostly famous. A well-known example of self reported measurement is various personality measurement questionnaires. This thesis have used Big Five personality trait model which is also called five-factor model (FFM) [6]. This model is defined by multiple researchers who have used various statistical methods of verbal descriptors based on human behavior [7]. Verbal descriptor is simply a series of illustrative phrases. Big Five model has gone through improvements by researchers which finally reached out to five factors of commonly known personality traits which is considered to be representing the base of all personality traits [8]. So each of the

Big Five trait is not just representing one single personality trait but is representing a collective number of personality traits [9].

Big-Five personality traits consists five factors or traits commonly named as Agreeableness, Conscientiousness, Ex-traversion, Openness, and Neuroticism [10]. But the naming of the traits have varied time to time [11]. This thesis have adopted International Personality Item Pool (IPIP), which is a scientific collaboration for advanced measure of personality development and individual differences [12]. According to IPIP the five traits are named as (1) Extraversion, (2) Agreeableness, (3) Conscientiousness, (4) Emotional Stability (instead of Neuroticism), and (5) Intellect/Imagination (instead of Openness).

Studies have suggested that virtual personality measurement can be well categorized under Big-Five personality traits [13]. Judge *et al.* [13] have studied that, in various languages through analysis, the 5-factor have been recaptured and decisions have been made to measure the dimensionality of the 5-factors by expert judges. Extraversion indicates how much outgoing and interacting a person is. The opposite of Extraversion is introverst that is those who keeps their personality to themselves. If a person has the quality Extraversion, it indicates that the person is possibly talkative, energectic, social and expressive. Agreeableness indicates how empathetic and friendly a person is towards others. This trait indicates a person to be cooperative, understanding, trustworthy and optimistic [14]. Conscientiousness is the trait indicating reliability and discipline. A person with a high score for this trait can be said to be methodical, organized and focused on achieving success. Such people have strong goals and are determined fulfilling them.

Emotional Stability, which is also known as Neuroticism relates to the state of emotions of a person and the measure of negative emotions. This trait can indicate emotional instability, negativity, anxiety, mood swings or tension. High score for this trait indicates that a person gets less impacted emotionally that is the person is more calm and is not affected by negative feelings. This does not indicate that the person is only positive thinker but more indicates to the stability to negative feelings. Intellect/Imagination also named as Openness indicates to the enjoyment of experiencing new things and appreciations for unusual ideas. The person with high score in this trait can be said to be imaginative and they have artistic interest. They are more adventurous and liberal to new ideas. People with low score in this trait may have the tendency to be resistant to new ideas and changes and are more closed-off. Overall, Big-Five personality traits individually can give an insight of a person's intellectuality, how a person can react to a situation, how open a person is to new changes or if a person is empathetic to a situation or not. These traits can indicate so many facts of a human personality that research have been done to analyse worker's involvement in an organization based on Big-Five personality traits [15].

Personality traits are one of the suitable ways to understand the smartphone users in order to improve the smartphone user experience in the field of technology. Stachl *et al.* [16] have studied that, in categories like communication, entertainment or gaming, personality traits like extraversion, conscientiousness or agreeableness can predict users' mobile usage better than other basic analytical variables like age, gender, religion, income or health status [16]. Individual behavior differing person to person can sum up to an attested behavior which can lead to justify ideas of user behavior and personality. The availability of cheap technologies like sensors have made it possible

to gather information about people's social and digital life [17]. Personality study has become easy because of the development of such crowdsourced technologies.

Smartphone data collection involves Crowdsourcing which means, gathering data from large number of people where the user's can contribute information about their everyday activity and pattern of mobility [18]. The regular interaction routine of smartphone users have created rising interest among researchers on how smartphone usage influence people's daily routine, personality and behavior. App usage pattern, usage time, types of app usage are variables that can be correlated with other data such as- self reported user behavior [19]. Data gathered by a smartphone users' everyday mobile activity can immensely contribute in research fields like stress, depression, personality and mental health analysis. Such analysis on mental health state can be applied in working policies of companies in order to improve productivity of the workers or can be brought under consideration by educational institutions to understand and improve students' stress level. Reality Mining is something that has been sensed by data analysts a long time ago [20]. Mestry *et al.* [20] described Reality Mining as - machine-sensed environmental data which can be used for user behavior analysis. Analysing people's personality using smartphone data also opens scope for making app recommendation systems based of user's preferences like for example mobile battery consumption, so the user can be recommended the apps which are less battery consuming.

Mobile sensing benefits the research in the field of personality and psychology as mobile usage can collect wide range of information regarding communication, social media interaction, battery usage, time and duration of app usage etc. Millions of apps for different categories like health, communication, entertainment, gaming, education, tools and so are available in the leading app stores [16] which, can reflect variety of user behavior, apparently collecting samples of people's daily activity. The datasets collected by mobile sensing are efficient source of ecological validity [21]. Ecological validity means to popularise a research or study outcome to real-life environment. Schmid *et al.* [21] studied that, the scope of collective ecologically validated data have resulted to personality researches by different approaches like personality and smartphone usage, various aspects of personality facets, intelligence and some rationale demographic factors.

Human-Computer Interaction (HCI) has been gradually formulating reliable research on characterizing users depending on their tasks and experience [22]. The study of Dillon *et al.* [22] concluded that, notable prediction can be done if variation among users regarding one or more human characteristics is related to user analysis in modern system design. Though there has been good number of research on psychology, extensive research on personalization is still detached from psychological research on personality [23]. Arazy *et al.* [23] studied that automatic system models can be constructed with recent advancement of HCI technologies by using psychometric survey tools which collects self described personality of users. Also, because of the popularity of smartphone usage, large amount of information of users is available in larger context. Crowdsourcing has opened a greater opportunity to collect valuable user data and various companies and researchers are now using these data for their analysis and also even for making recommendation systems. A user's personality traits can be analysed using the user's profile [23] and these personality traits can

create scope for improved user experience which eventually can improve the ultimate smartphone user experience more innovative, exciting and personalized.

To get deeper into user behavior related studies like personality, collecting smartphone usage of users is one of the best options because now-a-days smartphones have become very personal. Due to the sensors like GPS - locates users, accelerometer - measures motion/movement speed, gyroscope - measures orientation, ambient temperature sensors - measures temperature around, NFC sensors - can create communication between two electronic devices and a lot more different kind of physical and software sensors have made it possible to track a user's corporeal and online activities. These sensors along with mobile application usage of users can be logged and stored for analysis in real time. Different software are being developed to perform such logging [24, 25].

In this study, Carat platform is used, which has a soup of mobile usage data for a large number of Android application users. Carat is a free app that mainly focuses on explaining users about their mobile battery usage. Carat data have great potentiality for the personality studies like this thesis, because Carat has mobile application usage information like process name (application name), app translated name, priority (i.e. foreground app, background app, visible app etc.) about the users. In this thesis, all the data is fetched first and filtered the needed information for the analysis, mostly data related to application process name, user id, app priority.

For the extensive scope of research in the field of personality study and due to the availability of huge amount of smartphone usage data of individuals, this thesis presents idea of analyzing and predicting Big Five personality traits of 739 Android users based on their mobile application and application category usage. The Big Five personality traits were determined by IPIP 50-item Big Five personality trait questionnaire [12]. The participants who took the questionnaire are the batch of users from Carat data set. Carat has a huge collection of mobile usage data of users and also provides Big Five questionnaires answered by subgroup of these users. A prediction model is implemented to predict the Big Five personality traits of these users and analyse the accuracy of the model. In this thesis, a light-weight method is introduced with meticulous accuracy for predicting Big-Five personality traits of group of users based on their mobile application and application category usage. 96% accuracy is achieved in the best case and 86% accuracy in the worst case for the prediction. Also this thesis has described which application category best outline the Big-Five personality traits and also that the category level accumulation is sufficiently precise for this analysis.

The pipeline of this study follows five steps of DMAIC method for analysing the data analysis problem [26]. Though being defined in different terms in different litigate, the fundamental aspects of data analysis remain in unison. The first step is to define or to interpret the target of the analysis that gives a top view of what it is to be achieved. The following step is 'Measure' – collecting valid data that improves the data quality, meaning getting rid of unnecessary data if possible and if applicable. The third step is 'Analyze', which is the core of the Data Science framework that facilitates process and solution development for the intended outcome. Next comes the step – 'Improve', that is to implement the solution and optimization for better efficiency in terms of accuracy and time. Finally, the step is 'Control' that allows assessment of the solution and to

create proper framework conditions for the long-term use or MVP (Minimum Viable Product) of the targeted solution.

Figure 1. Steps of Data Analysis

Below steps have been followed in this thesis to come up with a prediction model to predict Big-Five personality traits of sample users based on mobile application usage as portrayed in Figure 1:

- Collecting and understanding the Carat mobile application usage data
- Fetching the Big-Five Personality Trait data from Carat AWS platform.
- Scoring the Big-Five personality traits for the sample users
- Understanding the usage of mobile application categories
- Mobile usage data preprocessing
- Dimensionality reduction of the data by applying PCA (Principal Component Analysis)
- Applying multiple machine learning algorithms and comparing the results by accuracy of the model

This thesis answers following research questions:

(**RQ1a**) What are the effects of personality traits on users' application usage?

(**RQ1b**) What are the effects of personality traits on users' application usage based on application category?

(**RQ2**) Application or application category, which one describes Big Five personality more?

This thesis is structured as the following: Section 2 discusses the background and related studies of application usage, Big-Five personality studies, predictive analysis, Carat dataset and so on. Section 3 presents the implementation steps of this thesis briefly along with the analysis. Section 4 presents the outcomes and results from the analysis. Section 5 discusses about the overall study, challenges, limitations, future work. Finally, Section 6 concludes the thesis. All the references are listed in Section 7.

2. BACKGROUND

Smartphones now a days are so personal that they can be used as a key to study personality of users [27]. Due to the rapid growth of wireless technologies all over the world, smartphones are now reachable to mass population [28]. With the availability of Internet access, users are now empowered with smartphone usages for communication, entertainment, health, transportation and many more. Smartphone application usage is making it flexible to collect sensor based usage data which has great potentiality for analysis like user personality.

2.1. Application Usage and Sensor Based Data Analysis

Smartphones are equipped with sensors collecting huge amount of data from users on daily basis. With the growing number of applications from different categories, smartphones are now personal more than ever. Sensors like GPS, accelerometer, camera, microphone etc. are capturing the user interactions and application usage of users (Figure 2). These data are being logged and can be used for extensive level of analysis on human personality, mental health and psychology. Nowadays, various smartphone data logging software are utilized to capture the usage of application and user interactions. Real-time data collection has become easy for such logging software [24, 25, 29]. With the use of these software, now the participants of any analysis related to smartphone usage are not required to be present at the laboratory which has reduced the time and energy of the participants as well as the researchers. Different measurements of mobile application usage such as battery consumption, number of notifications, usage time, timezone, application category names etc. of participants can be collected and analyzed in large extend. Application usage time can contribute to numerous research fields like human psychology, behavior and mental health. Descriptive and contextual statistics can be done to find out which applications are mostly used at which time of the day [30].

These measurements of mobile usage collected by smartphones can be used in expanded study fields of human behaviour, psychology, mental health, environmental science and diseases. There have been significant number of researches which analyze that users behaviour can be understood by studying their application usage. Generalize and reproducible researches have been done to understand users for enhancing smartphone experiences. Such mobile application usage data motivates to drive research related to user motivation and personality. By analyzing a number of user data of application usage, various types of user group can be identified using feature selection methods [31]. According to Zhao *et al.* [31], some studies state that, smartphone users are similar in terms of their behavior. Most smartphone based researches are done by only considering usage data but the variety of user group is ignored. Zhao *et al.* [31] challenges the primary characterization of users and shows how diverse smartphone user groups are by analyzing one month of application usage data from 106,762 Android users. 382 distinct types of users are being discovered from the study of Zhao *et al.* [31]. A study was done on how compulsive YouTube usage can effect personality and motivate users [14]. This study finds out that, the compulsive use of YouTube differs by the personality traits of the users and also plays a role on the

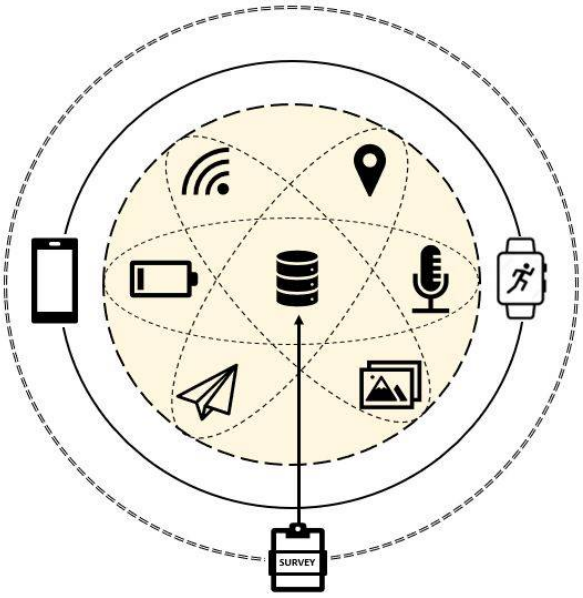

Figure 2. Data Collection Work Flow

motivation of the users. As discussed in the paper, such studies can help implementing possible risk factors when educators suggest students to watch YouTube as one of the means of learning.

With the same platform data set used for this paper, the Carat data set, geographic/demographic/cultural factors have been analyzed significantly [32]. According to this study, among all the demographics available in the data, country dominates to play a vital role in providing more information about application usage. Also, cultural values are influenced by the app usage, which is also demonstrated in this study. Data gathered by Carat platform holds adequate data to contribute in large-scale analysis for cultural and demographic factors of mobile usage. This inspires to study more on user behavior using such rich collection of data set. Mobile battery awareness is also some kind of useful research that has been carried out by using Carat platform [33]. As Carat is a mobile battery awareness application, Athukorala *et al.* [33] studies that, using Carat can bring a significant change in user behavior with long-term use. When the users are concerned about their mobile battery usage with the long-term use of Carat application, they become more conscious on saving battery life of their smartphones without the help of the Carat application itself.

Another study is done to analyse how users' application usage can be improved by battery charging habit and battery usage of the smartphones [28]. The measurements collected by mobile sensors can be sufficiently used in the field of health conditions such as mental health. Though, it should always be kept in mind that, these research outcomes are not claimed as medical solutions/predictions/results but are outcome of research questions to be asked on set of sample data. Some informative studies have been done on depression by analysing mobile sensing data. Depression is a large context now-a-days in terms of studying mental health. Predicting individual depression has been studied based on predictive measurements. But the complexity of such study is so major that, some studies have limited early onset prediction of depression measurement on larger perspective. There is a study that has worked on methods to early detect the major depression [34]. In this study, to implement

an early detection of depressive disorder, entropy analysis method can be used to detect the progression of depressive disorder. A smartphone sensing application was developed which collects meta data like which application is launched at what time etc. Like the Carat platform used in this thesis, the app developed for Asare et.al also collects PHQ9 [35], BDI [36] and Big-5 questionnaires [37]. Entropy analysis [38] and anomaly detection method [39] were developed to have an insight of depression at an early stage. These methods differ this study from other previous studies of analysing depression.

In another study of finding out the correlation between users' behavioral activities collected by mobile and their depressive symptoms, quantitative systematic way for finding such correlation has been done [40]. This study investigates the statistical significance of the correlation between mobile usage and depressive symptoms to find out which features show most favourable result during studies. Such findings can be applied to run mental health studies like depression based on mobile usage data in a systematic way. In this thesis, statistically significant features have been found which directly correlates to mood assessments which concluded to, monitoring a person's mobile usage is a good aspect for behavioral studies. Wearable devices and smartphones can predict mental health situations like depression with a significant accuracy. There have been studies done on higher education students to predict depression using their passive sensing data [41].

In the study of Wang *et al.* [41], the depression dynamics of students are captured by using PHQ-8 and PHQ-4. Such studies can lead to good outcomes in understanding higher education students' mental aspects in order to work on the rising rates of depression among young people. It is a fact that, just analysing different parameters of smartphone usage, how such behavioral aspects can be analyzed and worked on. Mobile phone addiction is a huge concern among young adults. Mobile phone addiction as in irregular usage of mobile applications can drive the emotional facet of individual. Irregular usage can be easily tracked by sensors and usage time which again leads to the same - smartphone usage. A study has been done to find out if there is any relation between mobile phone addiction and negative emotions like social problems among young adults of China [42]. This study has not only find out the interrelation between two aspects but also explored the mechanisms related to such relation. As per this study, mobile phone addiction is positively related to negative emotions and thus social problems are driven by the impact of smartphone addiction level.

Some studies have also been done on complex diseases on the basis of smartphone usage and sensor data. In a study on investigating the driving of smartphone in Parkinson's disease (PD), it was tried to investigate that if the dexterity of PD can be characterized and quantified using smartphones [43]. In this research, the participants were asked to do interactive activities with smartphones, for example, tapping and the participants were mixed of healthy and diseased. In this kind of research, we can see the use of smartphone sensor data. Machine learning models are implemented in this study for predicting movement disorder specialists scores. The study comes up with a result that, tests taken by smartphone sensors for dexterity is feasible and that the data can be used for patients' treatment routine. In another study, Parkinson's diseased patients' motor functions are measured by smartphone drawings [44]. Manual spiral drawing tests are old famous for assessing the severity of PD patients. In this study, the spiral drawing technique is digitized using Android devices. Such implementations

have been contributing to the manual clinical tests with digitized automated process in order to improve the quality of the clinical assessments for crucial diseases.

Another important aspect of mobile sensing is in environmental science and research using smartphone sensors. Health and life quality parameters are monitored and analysed for better living by smartphone sensors. Study has been done to understand the capabilities of smartphone sensors for greater use in the environmental analysis [45]. Nowadays, environmental condition for living, work places and public places have been monitored and controlled by using and analysing smartphone based data collected from the environment. Bluetooth technologies and distributed smart devices are collaborated using smartphone programmable tools by Aram *et al.* [45]. Data can be analyzed and verified using such distribution of smartphone driven tools in order to improve the quality of environment. Building environmental settings by installing innovative sensing system, is the latest target for those institutions who are interested to build a smart environment [46]. But for building such environment, few key factors like quality of the sensor, mobile network coverage, etc. play significant role to the degree of bringing out smart solutions. In the era of crowdsourcing, such sensor based smartphone systems can build environmental fingerprinting [47].

Now when it comes to personality based studies, understanding how people utilizes the smartphone usage is the key. Previously, there have been studies based on basic factors like demographics on the effect of smartphone usage. Demographics such as age, gender, educational background have been the key for many previous studies based on smartphone usage [48]. The usage time of application among young and old people have been compared and that, what type of application is used by which age range is also studied by Andone *et al.*. There is also another study using age, on the fact of ubiquitous computing being really ubiquitous or not, specially when it comes to the usage of technology by seniors [49]. The study finds out if the trend of adopting new technologies by elderly people is common or not. Hiniker *et al.* [50] studies that, instrumental and ritualistic purpose of using technology differs in terms of users' motivation. People purposefully use instrumental technology which is mostly goal-related use whereas realistic uses are based on users' habit. This study presents an empirical research to compare the instrumental and ritualistic nature of smartphone usage. Such studies help to build recommendation system models for recommending applications to the users based on their preferences.

While there is an abundant of study on wide variety of aspect to understand the smartphone usage, personality studies have not been seen in such extent and somehow are understudied. When it comes to psychological researches, most of the studies are focused on the dangerous aspects of using technology, in general. Studies like understanding the effect of the social media usage on public health has been done. In this era of social networking, the danger of being using social media can lead to addiction [51]. Salehan *et al.* has studied on social network services' network size and intensity of use to understand the effect of mobile usage on individuals' health. Counter studies on the perspective of social media and mobile phone addiction has been done to show how productivity and quality of life has improved with the growing HCI technology [52]. Bødker *et al.* [52] discussed about a 10 years of HCI with three waves, whether the usage of HCI will lead to a beneficial aspect in the long run. In another study on smartphone usage and user compulsive behavior [53], Lee *et al.* conducted an empirical study on participants to find out the impact of smartphone

usage on psychological traits like introvert behavior, materialism and anxiety. The study resulted to find out a positive link between the two aspects, smartphone usage and psychological traits.

Problematic mobile phone usage links to addictive personality and this study was done by Takao *et al.* [54]. In this study, a correlation between problematic mobile phone usage and addictive personality trait has been correlated. Such studies lead to find out the screening of problematic mobile phone usage. Mental health related studies based on mobile phone usage, especially depression has been done by several authors. Well being of a mobile phone user has been studied by the usage time of the day, i.e. night or day time usage rather than the intensity of the usage [55]. Mental health disorder is a complex topic and many studies has been done to link this vast topic using mobile sensing. A study has been done to find out the best suited methods to develop new metrics for an early detection of major depression [34]. As per this study, anomaly detection and entropy analysis has been resulted as best suited for early detection of mental health disorder. However, user behaviour based on mobile phone usage can contribute to develop recommendation systems by bringing change in design and user experience. Such recommendation can contribute to support the personality traits of individuals.

Study has been done to understand individuals' willingness of texting or calling based on their personality [56]. In this study, the personality and self esteem of individuals have been linked to their willingness of calling or texting which resulted to neurotic people being more addictive to social media while lower self-esteemed people being more instant messaging addictive. Similar study has been done by Montag *et al.* [57]. The study linked the personality of users' with WhatsApp usage. WhatsApp is one of the most used apps for communicating with people and usage data collected by WhatsApp can significantly add values to personality and addictive behavioral studies. The study has also done demographic analysis based on WhatsApp usage. The result obtained by this study has stated that, females use WhatsApp more than males. Personality trait like Extraversion is positively linked to daily usage of WhatsApp and Conscientiousness is inversely correlated to the daily usage length of the app. This kind of study brings clear indication on how much social media has been dominating our day-to-day life. Some popular apps like Facebook, WhatsApp are dominating the social interaction of people and thus their personalities.

For such mentioned studies which has brought effective result on people's behavioral aspects with different types of application, it is seen that the usage of communication applications are the dominant among other app categories. This brings the idea of app category based analysis for the personality studies. An user's application preference is an important factor which can define the person's personality. Also talking about recommendation systems or development of new applications, understanding people's app preference study plays a key role for understanding people's preferences based different demographics like age group or gender. Talking about the commercialization of application development, Lane *et al.* [58] studied personality traits based on smartphone usage in order to understand the commercial success of various smartphone applications. Various regressions were applied to conclude the link of personality traits with application of smartphone technologies.

Technostress is a relevant word now a days into which many researchers are interested in. Information technology plays vital role in everyday life but the impact

of such diverse and regular use of technology leads to technostress that is the psychological impact on someone for the usage of technology in daily life. Personality and technostress is linked in the study by Hsiao *et al.* [59]. This study tries to expand the focus of application usage outside the boundary of interface and user experience. The study focuses on how compulsive usage of technology effects the technostress. The survey based study was done by a software and result shows link of personality with compulsive usage of mobile applications. Along with compulsive usage, materialism and external locus also have impact on technostress as per the result of this study.

In this thesis, a prediction model is implemented for predicting Big Five Personality Traits based on peoples' smartphone usage. The work emphasised on the error calculation of the model in order to understand the impact of mobile usage on the five factors of personality named as Big Five. The next section is going to discuss about the background study on understanding the Big Five Personality Traits in depth.

2.2. In Depth of Big Five Personality Traits

In the study of human personality, whether in sociology or technology, Big Five personality traits is one of the most commonly used factors. Even if Big Five personality traits are famous but are also criticized by some researchers who suggest there are fundamental problems with the concept of the five traits [60]. Behavioral studies are always complicated due to the complex human behavior pattern for various variables of life. This study focuses on the mobile usage as previously discussed. The five factors of this model - Agreeableness, Conscientiousness, Extraversion, Openness and Neuroticism (Figure 3) are self explainable terms [61]. John *et al.* [61] described the factors with label as below -

1. Extraversion - being talkative, energetic
2. Agreeableness - being cooperative, confiding
3. Conscientiousness - being dependable, responsible
4. Neuroticism or oppositely Emotional Stability - high values tend to be more calm and not neurotic
5. Openness - being intellectual or being able to have independent mind or imagination

The traits mentioned above by John *et al.* [61] is known as "Big Five" (Goldberg 1981) [62] and that the traits are not just individual words but has broader sense. The format of Big Five traits does not implicit to reduce the human personality to only five traits. These are the five dimensions of personality that summarize distinct and wider personality traits which are specific and large. The names of the traits have differed time to time [11]. This thesis have adopted International Personality Item Pool (IPIP). IPIP is a scientific collaboratory which does advanced measurement of personality [1]. Goldberg *et al.* [37] plead for a change in the way of constructing measures for

[1] https://ipip.ori.org/

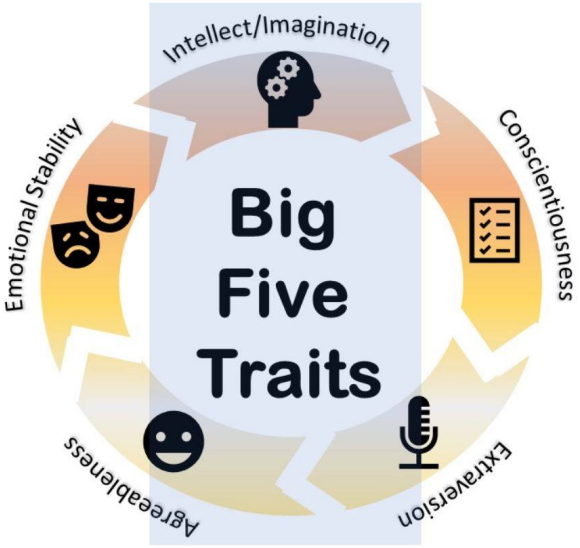

Figure 3. Big Five Personality Traits

personality traits in his study [37] to establish IPIP scale as the more predictive one compared to original inventory scales by regression analysis. Goldberg *et al.* [37] stated three problems with the original scale:

1. Need for *taxonomic framework* to organize the wide range of individual differences.
2. Common *item format* which is manageable to amend in diverse languages.
3. *Mode of communication* - a logistic approach to easily find out the previous studies.

This study by Goldberg *et al.* [37] have approached to solve these problem statements with the IPIP measurement scale. Though Big Five is famous and well accepted, also it is not without condemnation. Because of the vast concept of personality, it is hard to define the concept into five traits. A handbook on personality traits has discussed about the broadness of the trait [61]. According to this handbook, Big Five is obtained by researching the natural language terms which is used by people for describing themselves. Big Five taxonomy is a diversified system and it brings all the personality study under one framework. The word "Big" represents how vigorous the trait is rather than it's perfection. The dimensions of Big Five personality does not come from a single theory rather it is more than that.

John *et al.* [61] described how natural language has been used in various taxonomies before Big Five and how it has inspired researchers to consider for describing study like personality. Natural language provides a set of feature which is derived from the dictionaries of natural language that has been found useful by the users who speak that language in their daily interactions. The concept of personality comes from various theoretical perspectives and breadth. In the very beginning of defining personality traits by the researchers, the aim was to design a descriptive model. Taxonomy allows to work on specific territory of research rather than separate attributes of human individuality. General taxonomy makes it easy for the researchers to narrow down the

diversity of personality to a single domain. This is what Big Five personality traits serve, it brings the broadness of the personality traits into one taxonomy which offers a nomenclature.

Human personality may have link to real life behavior of individuals and as Big Five is one of the famous traits to understand human personality traits, it can also be thought if Big Five has anything to do with real life performance of individuals, for example social activities, stress or anger management, group work - in total organizational behavior and psychology [63]. Judge *et al.* [63] studied on the contribution of personality from this perspective. The study considered below summarised points for finding out the outcome of the effect of Big Five personality traits:

1. Job: performance, attitude, motivation, leadership
2. Influence: power and politics
3. Conflict: negotiation, anger management
4. Coping up- stress management, adaptability

Judge *et al.* [63] have found relationship between Big Five personality traits and organizational behaviors like job performance, work motivation, job attitudes, leadership, power and politics, adaptability in various situations, team effort and effectiveness, productivity, conflicts and negotiation etc. Though Big Five has such broad correlation with organizational behavior and performance, Big Five is not above criticism. Judge *et al.* [63] have also discussed the criticisms of Big Five in two points:

1. Lack of quality validity evidence, that is, the statistical significance of the validity of prediction variables while predicting the relation of Big Five for various organizational behavior (job performance etc.).
2. Faking the personality test, that is, people can fake the answers.

Based on such criticism points, it can be said that, personality traits may not be above imperfection and thus if any organization considers personality trait scores as one of the criteria of job recruitment, the decision making may be haphazard. However, considering the broadness of Big Five, it has been considered as one of the most accepted personality trait system by a wide range of researchers globally.

Individual's personal values can be analyzed by Big Five personality factors [64]. Roccas *et al.* [64] studied on the correlation between agreeableness and kindness, openness correlates to accessibility, extroversion with energizing and prompt, conscientiousness with success and authorized entity. As per this study, the positive effect of the correlation between values and personality traits is significant. Big Five personality traits have such comprehensive aspect of understanding human behavior that the model itself can play an influential factor to understand human values.

Internet addiction is another aspect where Big Five has been used to find out a correlation [65]. Kayics *et al.* [65] summarize the five factor relationship with internet addiction - positively or negatively. The summary looks like:

1. Openness, conscientiousness, extraversion, agreeableness negatively associated with internet addiction

2. Neuroticism positively associated with internet addiction

Here, the naming of the Big Five is different from what has been followed in this thesis but as mentioned before, the naming has gone through variety. Understanding internet addiction is an important aspect of research in this growing age of technology. The addiction and usage time may impact the personal behavior as well as social values of individual users. Internet addiction, the name itself says about the unhealthy usage of internet. The study of Kayics *et al.* [65] enlightens on how to analyse internet addiction based on Big Five personality traits using meta-analysis method.

In the study of Kayics *et al.* [65], the main concern is to understand and find out if personality traits have any link to individual's social media and application usage. Some previous studies have researched on finding out the effect of personality in individual's usage behavior (smartphone and social media usage).

Klobas *et al.* [14] studied on how YouTube usage effects on motivation and personality [14]. Compulsive usage of any social media platform can negatively effect on human personality in terms of motivation, agreeableness, emotional stability. Klobas *et al.* [14] explored the compulsive usage of YouTube among university students and compared the motivation of usage based on information, entertainment which then is examined how the motivational factor of using YouTube effects human personality. Usage for study and information is associated with lower usage and usage for entertainment is associated to higher compulsive usage of YouTube. Though the effect of motivation is independent of the effect of personality in this study. But personality has independent effect on compulsive YouTube usage, according to this study. The higher compulsive YouTube usage is related to disagreeableness and lower compulsive usage is linked to emotional stability. Measure of agreeableness and emotional stability are also Big Five traits. The study has not found any effect of extraversion or intellect/imagination with the compulsive usage of YouTube.

A similar study has been done by Quercia *et al.* [66] with Twitter usage. In this study, a background of linking the personality with digital world like music genre has been explained which drives this study of linking personality with different types of twitter users - popular or influential. The research was done on 335 users' personality data resulting popular and influential users being extroverts and emotionally stable. Popular users came out to be high in openness and influential being high in conscientiousness. This study also predicts user personality using three types of profiles - following, followers and listed counts. These three types of profile can predict Big Five personality with RMSE 0.88. The result looks promising and shows the impact of personality in regards to Twitter usage is high.

It is not only the mobile application usage that drives the user behavior, but also the willingness of checking mobile phone is dependent on occurrences like notifications. A study on understanding people's willingness to use phone when notification comes, has been done by Mehrotra *et al.* [67]. Mobile notifications are very important but they can hamper user concentration and interrupt personal and physical factors. In this study, the disruption caused by notifications has been done in order to find out the effect on mental and physical action by the user. The study has used automated logging system to collect smartphone data samplings and questionnaire related to notification perception has been conducted by the sample users. The study has found that disruption and response time from notifications can be influenced by the type of

notification, in case of SMS - relationship between the sender and receiver, complexity of the task the user has to deal with, etc. Useful contents in a notification may also cause disruption in cases. In addition to these studies, Mehrotra *et al.* [67] also find out that, psychological traits have significant role on the response time and disruption caused by notifications.

Chittaranjan *et al.* [68] analyzed personality traits with smartphone usage. In the study of Chittaranjan *et al.* [68], for 83 users, they collected smartphone data for continuous 8 months. The study shows that, smartphone derived aggregated features can indicate Big Five personality traits. The study finds out office apps are related to conscientious and low openness along with no emotional stability. Introverts are the prime users of internet, communicative media is less used by conscientious, conscientious and neurotic behavior is observed among mail users. The study used supervised learning to train the model. Significance of the result is 75.9% accurate. This study was done in 2011 with 83 Nokia N95 users. Nokia N95 is not full smartphones in the current understanding are is limited in functionalities. This thesis adds value in terms of utilizing smartphone usage data from more advanced technology smartphones with variety of application usage.

In 2013, Chittaranjan *et al.* [69] continue the study with 117 users of the same phone model users. This study finds out that the lesser use of internet, games and camera applications is observed among extroverts. Any applications are seldom used by agreeable users, Music apps are inversely correlated to conscientiousness, office application usage is inversely correlated to emotional stability. The summary of this study is, various collected features obtained from the logging of smartphone usage can indicate the Big Five traits. Machine learning methods can be applied to predict the personality traits by training the models. The study also focuses on gender-specific models and it also facilitates research on personalized services based on personality study.

Another study has been done to build a psycho graphic model from smartphone usage [70]. In order to design individualized user interface, personality has been used as an advantage by de Oliveira *et al.* [70]. In this study, mobile phone call behavior has been used as a basis of predicting personality traits. Call detail records and social network analysis have deduced the factors of Big Five personality traits, in this study. Another study similar to this has been done by de Montjoye *et al.* [71]. This study has also used mobile phone call logs to predict personality traits. The mobile phone call data carriers have vast data that can be used for personality studies. The accuracy rate of predicting personality based on phone calls is 42% in this study.

Such studies show that, smartphone driven data have huge potentials to carry researches on personality traits. The number of mobile phone usage is increasing as well as the logging system of data collected from the smartphones are now more advanced and accessible by the researchers. Various cloud platforms like AWS, Google Cloud, Azure etc. are developing advanced techniques to store smartphone driven data in more convenient shape for researchers. On the other hand, personality questionnaires are one of the most cost-effective ways of studying human personality in a huge scale. The broadness of Big Five personality traits along with available smartphone data samples have brought many research outcomes to understand human computer interaction.

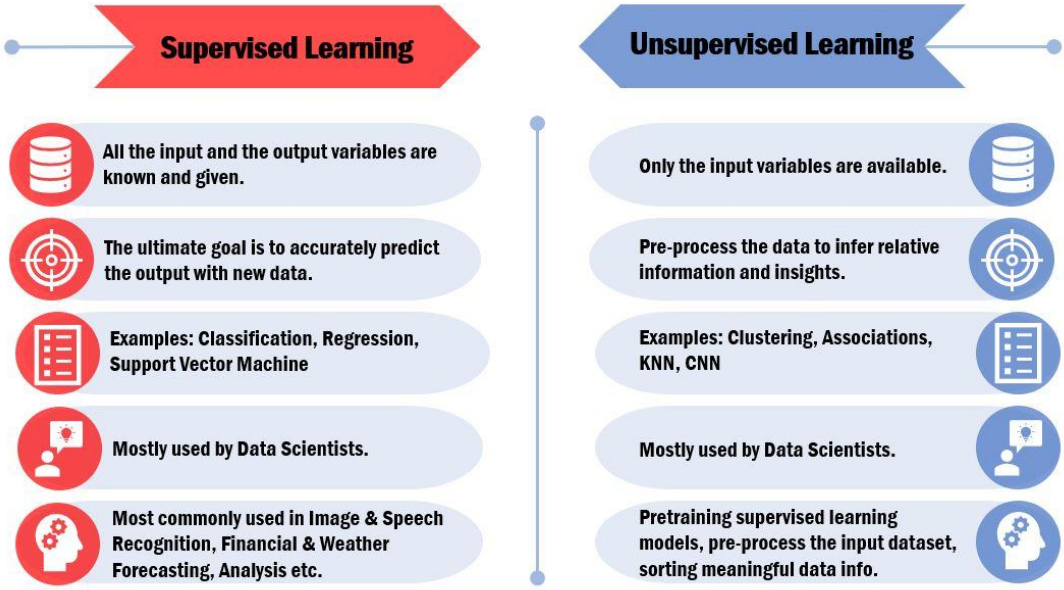

Figure 4. Supervised and Unsupervised Learning

Recent studies have placed a roadmap of possible potentiality to research on personality prediction using mobile application usage. Stachl *et al.* [16] have focused on relationship between personality and factors like demography of smartphone users, app usage etc. The study was done on 137 users (both male and female) with average age of 24 and the study was done in lab sessions. The data used in the study suggested, categories like communication, hobbies, entertainment, transportation etc. can be used to predict personality traits. In comparison to demographics and some Big Five traits, personality traits like extraversion, conscientiousness and agreeableness are more predictable than the demographic factors. Additionally demographics are useful for predicting categorical app usage. This study concludes that, individual differences of usage behavior can play significant role in understanding behavioral aspects of personality. This thesis does not consist any lab sessions but by using crowdsourcing paradigm, this thesis can capture effects of smartphone usage on personality study in the wild.

2.3. Supervised Learning in Predictive Analyses

Machine learning algorithms and statistical models are the technologies that have been used to develop concise models in the field of data science and technology. The new era of technology is overwhelmed with data. Among the vast implementation of Machine Learning (ML), data analysis and data mining is one of the most significant ones. Machine learning algorithms build models using training data to make predictive decisions. The categorization of machine learning is vast. The main two approaches of machine learning algorithms are: supervised and unsupervised learning. It is necessary to understand the differences between supervised and unsupervised learning in order to train the model based on expected outcome. In Figure 4, the basic differences between supervised and unsupervised learning is summarized.

The main difference between supervised and unsupervised learning is the presence of input and output variables. In supervised learning, the instances have known outputs, that is the instances are labeled, whereas in unsupervised learning the instances are unlabeled [72]. In a nutshell, supervised learning has a guide to train the data by using the link between input and output by the algorithms. The model developed by supervised learning can be applied to new data for similar expected outcome like the model data set. Both of them are used in machine learning problems, data mining, big data processing, analysis and are used by data scientists. Mostly supervised learning is the one which are used in forecasting, neural networks, image and speech recognition, decision trees and so on.

To understand the concept of supervised learning and methods using supervised learning, Caruana *et al.* [73] have compared ten supervised learning methods. In order to achieve excellent performance, the study of Caruana *et al.* [73] has adopted *calibration* with either Platt's method or Isotonic Regression. Calibration is done to improve the accuracy of the model so that the behavior of the prediction model is similar to the behavior of the training model [74]. The study has compared: logistic regression, neural networks, SVMs (Supprot Vector Macine), random forests, memory based learning, naive bayes, boosted stumps, boosted trees, bagged trees and decision trees. A summary of the study is given based on the performance analysis in Table 1.

Table 1. Summary of the study of Caruana *et al.* [73]

Boosting, Random Forests, Bagging, and SVMs	The improvement in the performance is outstanding compared to 15 years ago
Neural nets	Compared to earlier learning methods, neural nets is proven to be the best performed and is competitive with some new methods
Boosted trees, SVMs, Boosted stumps, Naive bayes	Calibration with relevant methods improve the performance dramatically
Neural nets, Bagged trees, Logistic regression, Memory based methods	Calibration with relevant methods does not improve the performance

Caruana *et al.* [73] state that, with appropriate scaling, boosted trees are the best learning algorithm, followed by random forest. After that, uncalibrated bagged trees, calibrated SVMs and then uncalibrated neural nets. As per this study, the algorithms that performed least good are naive bayes, logistic regression, decision trees and boosted stumps.

This thesis is associated with supervised learning as the goal is to do predictive analysis of Big Five personality traits. Kotsiantis *et al.* [72] have briefly reviewed supervised machine learning technique. Every machine learning algorithm constitute of same set of features which can be continuous, categorical or binary. As mentioned previously as per the study of Kotsiantis *et al.* [72], supervised learning has given instances with known labels. For predictive analysis to determine Big Five personality traits, supervised learning has brought out some success in previous studies.

In order to study personality based on Twitter, supervised learning has been used by Carducci *et al.* [75]. The basis of such study is what people do or say is what they are. This one idea is behind all the researches done to understand human personality based on people's smartphone and application usage. Data gathered by social media platforms have made it possible to achieve such research with the help of supervised learning.

Information such as favorite music and movies, time spent on applications, responsive behavior towards particular category of application and so on. are inputs of supervised learning which can result effective and accurate prediction of personality traits. This study notes that, the algorithms may be accurate for the expected outcome but due to the lack of enough data which is needed for such accuracy can make it unfeasible to complete the research with a good success. Also, some data may not be accessible for use due to access restrictions and privacy.

Carducci *et al.* [75] propose a supervised approach to analyze the personality of a Twitter user based on what the user posts publicly. This study is a combination of natural language processing and machine learning, as word tokens and word vector representations have been used to feed in the supervised learning classifier. The accuracy is calculated by the classic mean squared error.

Human behavior analysis is not limited to the samples of the youth or elderly but also to analyze children behavior. Lizzeri *et al.* [76] have done a study on parental guidance using supervised machine learning approach. There is an old debate on nature vs nurture theory, that is if a human behavior growth from childhood to rest of the life is connected to environment or genes or parental guidance. Lizzeri *et al.* [76] propose a theoretical model using supervised learning to explain empirical relevance to this nature nurture debate. The study captures the idea of trade-off between a child learning from mistakes and allowing the child to learn from experience. The study implements a model to accomplish this idea. The optimal parenting policy is being characterized in this study and the features of these policies are interpreted for finding out the behavioral genetics. Supervised learning, even in theoretical research have been proven to be effective on such studies.

Another study based on supervised learning is done by Gunes *et al.* [77] to assess facial beauty by image processing using supervised learning approach. Universal beauty standard proportions has been is debated among researchers for ages. Psychologists and anthropologists have been studying to find out what is the standard of beauty based on regions and areas. This study experiments to evaluate the universal beauty standard by a survey done on diverse referees who graded a collection of female facial images. The summary of the result of this study states that, human grades on beauty has a strong central tendency which proves agreement on beauty assessment. Average human grades are trained using classifier and then the model is used to classify an independent set of test facial images. This is a good example of supervised learning. The classifier can be used to automatically classify the female facial beauty. Such classifiers can be used in virtual media and plastic surgery industry, Gunes *et al.* [77] emphasize.

Adeyemi *et al.* [78] studied on understanding online behavior by exploring online personality trait probability. This study has also used supervised machine learning approach. Online anonymity is referred to the identification of network and system identifiers, not the real user's identity. Personality behind the anonymity can play

both negative and positive role in online media platforms. A recent exploration of personality traits in a form of human identity has been done in various studies. Internet usage and human personality traits have correlation and regression and human personality traits can be analyzed based on media platform usage, for example social media. According to Adeyemi *et al.* [78], some previous studies have been done on the relationship between human personality and Internet based interest. A person's interest is a key indicator of his/her personality.

However, in the study of Adeyemi *et al.* [78], the main focus is to explore the personality traits in perspective of platform independent digital fingerprint. Platform independent technologies are basically implementing a technology in one machine which can be used in various machines, that is the platform is independent of the implementation machine. Digital fingerprint leaves mark of the user to monitor by the owner of the system to track or identify the user of the system. Anonymity in one hand is important to maintain privacy but for some cases, it can create problems like bullying, spamming, phishing. Understanding the user identity in the perspective of user personality and behavior is important. Adeyemi *et al.* [78] consider the Internet usage as an extension of daily communication by users which leaves digital behavior signature. So the main focus of this study is to explore the client-server interaction as the basis of online communication to predict the digital personality traits of the users.

Big Five model is used for the personality measurement in the study of Adeyemi *et al.* [78]. The data samples are server-side network traffic which has been collected from 43 respondents for over 8 months. The data is analyzed using supervised machine learning techniques. Among the five personality traits, this study suggests that the conscientiousness behavior can be observed in online communication among the users. The analysis of this study has brought out that online platforms are novel for exploring the online identity of the users.

Another study which is nearly similar to the study of Adeyemi *et al.* [78] is done by Christopher *et al.* [79]. The study is on reviewing authenticity verification and learning reviewer personality traits using supervised learning [79]. Online product review is very common in online shopping and the reviews are used as marketing policy as well as defines the quality of shopping experience. Reviews boost up the urge in betterment of shopping experience. But reviews can be also unnecessarily negative, considering various types of Internet users. Accurately verified review analysis is important to meet the goal of improving the product quality based on customer reviews. Supervised machine learning can extract the useful reviews by identifying the fake and unnecessary reviews and filtering them out. That is, supervised machine learning approach can classify reviews by authentic and fake. For performing this task, personality prediction plays the role of identifying the key personality of the fake reviewers. By understanding the reviewer traits, Christopher *et al.* [79] used the Big Five personality model to track people who fakes the review through their associated social media accounts.

Web is overwhelmed with data and information. Among such vast information, finding out the trusted ones is a challenge. Trust plays a very important role for helping users collect reliable information from the Social Web [80]. Zolfaghar *et al.* [80] studied on the evolution of trust networks in social web. The study has used supervised learning. Trust evolution is a side by side research topic with trust related related online applications. In order to evolve the online trust networks, Zolfaghar *et al.* [80] studied

on how to move time-aware trust prediction. In order to achieve this, the impact of trust network evolution is investigated to predict tasks by using supervised learning. The information used in such study is the history information available on the trust links. The study concludes that, the accuracy of the predicted future trust relations significantly improves based on past trust relations used as training data.

Gilpin *et al.* [81] used supervised learning to predict Big Five personality traits by speech signals. This study focuses on how human look or sound plays an important role in unconscious communication behavior of an individual. This is a complex process and Gilpin *et al.* [81] tried to (1) develop a model by building SVM and HMM classifiers to predict Big Five personality using speech signals, (2) correlation between feature and speaker subgroups and (3) assessment of SVR classifier on new set of speech signals. This study is an informative example of how supervised learning can be used to train a model with one set of input and output and can be used on new set of data. This study was applied on 640 speech corpus based on 11 Big Five assessments to train the model. Then the prediction models are tested with 15 new speech records, labeled with same Big Five inventory. The accuracy of the predictive analysis is impressive. It resulted the accuracy of 70.15% for Extraversion, 66.72% for Agreeableness, 90.78% of Conscientiousness, 77.66% of Emotional Stability and 78.98% of Intellect/Imagination.

Staiano *et al.* [82] studied on how personality traits can be inferred from social network structure using supervised learning. In this study Staiano *et al.* [82] analysed the relationship between personality and social network structure. The performances of various structural network feature subsets has been assessed in predicting Big Five personality traits. Such studies have been done in extensive manner by the researchers to predict the driving force of social networks on the diffusion of behavior and effect. Supervised learning makes a way of understanding the definition of the classes specifically and is able to train the model or the classifier with perfect boundary of decision to differentiate the classes with accuracy.

Another similar study on social media using semi-supervised learning is done by Nie *et al.*[83] where unlabeled samples are used for the prediction accuracy improvement. Social media and personality is a burning topic among the researchers because the relationship between human behavior, personality and social media is salient to be analyzed due to the rapid growth of social media usage. But because of the lack of labeled samples, predictive analysis gets tougher to implement. Social media data is mostly huge, messy and unorganized at the raw level. Doing predictive analysis using social media is challenging due to the unlabeled samples.

Nie *et al.* [83] aimed to research on unlabeled samples. The study used 1792 users' data and semi-supervised regression is used to predict the personality traits of the users based on Micro blog usage. In semi-supervised learning, it combines labeled and unlabeled samples. So in this study also, with a few labeled users alongside a set of Micro blog users, the prediction model is built to predict the personality of unlabeled users. This study is a good example of how the usage of unlabeled data can be used to improve the prediction accuracy.

Supervised learning can be used to predict targets which are labeled, that is there are sample target values which are used by the algorithms as examples to train the prediction model. This thesis predicts each five traits of Big Five based on mobile usage and the Big Five personality score for each trait are corwdsourced for this study.

So there are sample scores for each trait which can be used by the algorithm for prediction. This is why supervised learning algorithms like Random Forest Regression and Support Vector Regression are considered for this study.

2.4. Data Storage and AWS

Designing a data collection infrastructure is the prerequisite of data science and analysis. Data engineers are up for designing architectures to fetch and store data automatically without any interruption and delay. ETL (Extract, Transform, Load) processes need to be smooth for storing data for further analysis by data scientists. Smartphone data are large in scale and sometimes messy. Storing data in a conventional format is essential for further analysis. Gray *et al.* [84] analyzed some rules of thumb for data engineers. According to this study, the main focus points of data storage are: (1) storage medium, (2) processing, (3) cost of network, (4) cost of platform and (5) performance. Disk bandwidth and network bandwidth need to be saved by caching the storage, ensuring smooth flow of data.

There are various data storage cloud based platforms available in the technology market which are providing smooth ETL services for data scientists to further use the data in convenient format. AWS (Amazon Web Services) is one of the most favorite cloud based data storage platform among the data scientists. Baron *et al.* [85] described the services of AWS which is offered to store data in a conventional way for further analysis. According to this paper, traditional data storage provide following potentiality:

1. Memory caches, RAM disks
2. Message queues, asynchronous transmission of data between various application components
3. Option of backup, data retain and archiving
4. Traditional databases like SQL relational database, NoSQL non-relational database

These traditional storage resources differ from the perspective of performance, cost and some lack interfaces. Data engineers' and architects' main concern now a days is these performance factors. Architects mostly want multiple storage technologies under one platform. AWS offers multiple cloud storage options. The options have their own performance factors, durability and most importantly, interface. Baron *et al.* [85] described some AWS cloud storage options in the journal: Amazon S3, Amazon EC2 Instance Storage, Amazon DynamoDB, Amazon Redshift etc. are some of the popular ones.

Such large extend of cloud storage systems have made it possible to store huge data and information from devices like smartphones for further analysis by data scientists. AWS is famous among other cloud storage systems for below benefits:

1. AWS services are low cost specially from the prespective of small organizations which are just start ups and are looking for a cheap yet fine data storage platforms.

2. AWS instances and services can be kept in halt when unused and that can save cost.

3. Setting up a new server or environment is a matter of few minutes in AWS

4. AWS needs to be payed for as much the user wants to use it

5. AWS is flexible to use because of it's easy interfaces and simple infrastructure

2.5. Carat Dataset

The data set used for this study is collected by Carat [29]. Carat is an application that tells the application user about the battery information of the smartphone. The application recommends the users on how to improve the battery life of the smartphone [2]. Carat dataset contains factors like system settings, variables and energy rates. The key information contained in a Carat dataset are: Wi-Fi information, mobile network usage, mobile network quality, CPU usage, battery temperature, screen brightness, voltage measurements etc. The best part about Carat dataset is, some data can be used for research purposes freely by maintaining the privacy policy.

The system variables collected by Carat can be used in building prediction models from various perspectives. When Carat holds some useful variables to work with for research purposes, Carat uses AWS for the storage of such huge extent of data along with Spark. The functionality of Carat application is automatic. From the beginning of the installation to one week, it starts showing the results. The user has to just install the app, open the app every now and then and keep using the device in usual manner. It collects application usage and battery level information with every 1% battery drainage. Opening the app will let the app communicate with the servers. Carat gives the user a personalized report of battery usage, for example, which application instances are using more energy. Carat tends to improve the result over the time of usage.

Carat app has some preferred Action List like killing app or upgrading app which suggests the user which app should be dealt in which way. This option also states if the user takes the suggested action, how much battery improvement will take place. User can compare his/her device's battery life with three out of four other devices running Carat. Bug Report reports the apps which are using excessive energy.

Carat maintains the security policy of not collecting user's personal information such as names, email or any private data from any app. As Carat is a research project so the data is used for academic research purposes and some data are open and free. But the data is fully anonymous. Carat itself uses very few resources of the device and constantly transmits data to the servers without making the user bother about the app. The commercial usage of open Carat data is against the Carat privacy policy. A summary of what Carat data contains is given below as per the official Carat website [3]:

1. uuid (anonymous and unique user id)

[2] http://carat.cs.helsinki.fi/

[3] https://www.cs.helsinki.fi/group/carat/data-sharing/

2. timestamp (UNIX timestamp of sample collection time)
3. batteryLevel (battery percentage)
4. timeZone (user's timezone in text format)
5. mobileCountryCode (mobile network MCC with which the user was connected)
6. apps (app usage)

All these data are in JSON format. The app usage has nested JSON elements of below variables:

1. processName (Android package name of the app)
2. priority (background, foreground, system etc.)

The data consists some more information like:

1. model (Android device model)
2. osVersion (current OS version number)
3. categoryName
4. numberOfSamples

Some significant work have been done with the Carat dataset. Sigg *et al.* [86] studied on the fact that, only number of downloads, installations or user ratings are not enough to define the sovereignty of the apps. The number of downloads or installations does not directly indicate the exact usage of the app. The apps can be installed but then left unused and unopened by the user. This study works on the retention rate and user behavior trends, that is the actual user interaction with an app installed. The study has used large-scale app usage data from 339,842 users.

According to this study, 70% of the users are lost within first week of application usage whereas for popular applications, the user loss is 45%. Complex usage behavior is observed for some applications due to the factors like marketing. The fame of an app is measured by understanding novel-app-usage behavior. The trends of app popularity and app classification is being identified in this study. The summary of this study states that almost 40% of the apps do not gain enough users, less than 0.4% of the remaining 60% remain popular, 1% is flops after a sheer rise in usage and only 7% continues to rise popularity wise.

Such studies of usage behavior trend can contribute in developing mobile app recommendation systems. App installation and download count do not necessarily show the actual sovereignty of an application. It is the trend of usage behavior that counts. Usage based behavior analysis of retention and trend are able to shift the supremacy of an application. The actual statistics is what matters in order to understand the real popularity of an app.

Carat data have contributed in understanding the challenges of sharing large data over common mediums like email [87]. Peltonen *et al.* [87] noted that, Carat servers

store massive data of Carat users which contain large-scale information about device details. As Carat is an open source platform for researchers to use the data for educational research purposes, it is a matter of challenge to share such large-scale data for analysis. To understand how big Carat data is, in 2012, Carat has collected over 1.5 TB of data from all over the world with 850,000 count of smart device users. Peltonen *et al.* [87] have presented some way forward to solve the challenges of sharing such huge data for further research.

Another study on recommendation system is done by Peltonen *et al.* [88] using Carat data set. System setting recommendation is important to smartly use the regular devices in order to save battery consumption. Applications used in smartphones have made life easy but the background resources applications use is not as easy as it seems, some have really complex system settings which is running in the background of the application. Because of such complex settings, battery optimization becomes a tough task to manage. Peltonen *et al.* [88] develop an approach of making energy models using crowdsourced measurements collected by Carat.

This study focuses on simultaneous capture of relationships between various application system factors and shows an united perspective of mobile device energy state. Battery models can be developed by using such huge-scale corwdsourced data. In this study, an extensive analysis has been done with the data set containing system state measurements and battery charge/discharge information. The study result is compared to previous battery consumption studies, and the outcome resulted a cost efficient way of developing energy models when the application user is doing normal operations with the device.

The study also analyzed other insights of battery consumption. It states that, high CPU activity and automatic screen brightness combined have higher effect than the effect of medium CPU load and manual screen brightness. Wi-Fi signal strength can also impact on battery life. This study shows if the Wi-Fi strength is dropped of one bar, it can shorten the battery life of the device by 13%. Presence of sunlight also plays a role on shortening battery life. Outdoor usage of devices shorten battery life by 50% than indoor usage.

A system recommender is also developed based on the crowdsourced models and the name of the system is Constella. Recommendation systems are one of the best real life life implementation of such analyses. Constella offers recommendations on how to adjust system settings in order to reduce battery drainage, in human readable and actionable form. The effectiveness of this system has also been analyzed in this study through hardware power measurement experiment. The recommender can save up to 61% battery life, according to the study. Battery life consumption is a matter of headache for the users, specially those who use their devices very regularly for day today activities. There are some apps which needs to be activated for the purpose it should solve, i.e. health apps. Understanding and optimizing is a big challenge for such regular users of smart device applications. A recommender of battery usage can bring an ease for such users.

Another study done with Carat data set is about malware infection and risk factors [89]. Mobile malware is a topic which is not much talked about like computer malware. This study note that in public domains, a very little information is present about mobile malware. Truong *et al.* [89] studied on malware infection rates and possible risk factors by crowdsourced data of over 55,000 Android users. Compared to previous

independent estimates, this study finds out the infection rate is significantly higher as per the outcome of the results.

There is a hypothesis which states that, some application stores consist high volume of malicious applications, based on this hypothesis, Truong *et al.* [89] investigate on whether it is possible to indicate the infection of a device by the applications used in that device. The analysis infers that while the root cause or the point zero can not be detected with the highest accuracy, the number of applications that are required to be monitored and maintained properly with the most sophisticated tools can be found out. The analysis consists of two separate malware datasets and makes a promising effort identifying applications prone to be affected by malware than other random checking methods. This analysis also sets a base for detecting previously unknown or ignored infections and hence, it can be taken as a complement to the scanning solutions that is available today. In another note, the study also brings out a relative battery usage measurement comparison between infected and clean device.

In the study of user behavior, Carat data has contributed to find out how it affects user behavior [33]. In the analysis done by Athukorala*et al.* [33], examines the effect of Carat in long-term usage from the perspective of device users. The study was is done by a surveying users with Carat application in their Android devices. The analysis resulted in an hypothesis, that demonstrates that, users with Carat application in their devices for a long time, eventually subjects themselves to have a better view on battery usage, which results in less charging, more battery saving, even without or a minimal assistance from Carat. Considering the findings stated, the study proposes a set of recommendations for mobile battery awareness applications that suggests that the application needs to be comprehensible to the application users, structured to be sued for a long time by the users, consider the user point of view and perspective while generating feedback and be able to recognize the in-built system applications and the third-party applications.

Carat collects battery usage data, battery status information, application usage, timezone of user and also timestamp of data sampling. Using Carat data, researches like mobile battery energy modeling and long term smartphone usage behavior of users are done. There are some noteworthy research achieved by using battery information of Carat data. Though study regarding usage behavior of users have been completed with Carat but user personality related analysis is limited. This thesis aims to fill the gap of inadequate user personality related study with Carat and utilize the application usage data for predictive analysis of human personality which adds a new value to the previous analyses done with Carat data.

3. IMPLEMENTATION

This section describes the step by step implementation process of the prediction model. The model predicts Big-Five personality traits of 739 users based on their application or application category usage. The implementation steps are summarized in Figure 5. The roadmap to implementation follows below steps:

1. Carat data pre-processing from JSON format to dataframes (data in columns) in order to make the data suitable for analysis using Python libraries.
2. Create binary matrices from the data set, one for application another for application category.
3. Split the data set into Train, Test and Validate sets.
4. Make base models using default parameters of the chosen algorithms for predicting the five traits of personality.
5. Determine the best parameters of the chosen algorithms by hyperparameter tuning in order to improve the performance and accuracy of the model by eventually comparing the performance of the model with the base model developed using default parameters. The accuracy of the prediction is calculated using Root Mean Square Error (RMSE).
6. Validate the model performance with Validation dataset.
7. Make predictions with Training data set which is tested with Test data set.

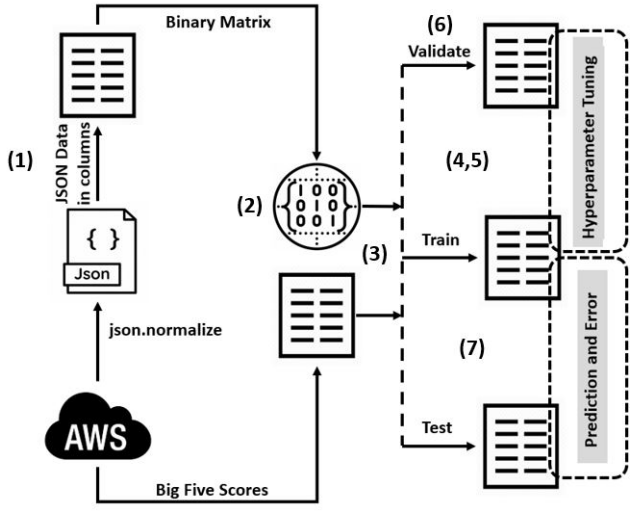

Figure 5. Flow of the implementation steps

Each of the roadmap steps are described in details in the following sections. Section 3.1 describes about the Carat data, how the data is stored and fetched for analysis,

what is the format of the mobile data, some demographic analysis of the users, Big-Five personality data and scoring of five traits. Section 3.2 describes how the mobile usage data and Big-Five scoring data is pre-processed. Section 3.3 describes how the dimensionality is reduced for the mobile application data. Section 3.4 describes the analysis process for predicting the Big-Five traits.

3.1. Carat Data Description

The data used for this thesis is from Carat [29], as mentioned previously. Carat [4] data set is being made public for research purposes. The identity of the participants are kept anonymous. The data is collected from voluntary participants' smartphones where Carat application was installed. Carat is an energy awareness application which is free and tells the user what is consuming the battery life of the mobile device. Carat not only finds out the reason of battery consumption but also suggest if it is normal or what can be done about excessive battery consumption. Carat learns the behavior of the battery usage of a user over the time.

After a week of the app installation the user starts receiving recommendations for battery life improvement. Carat app can be used in mobile phones, tablets or even smart music players and it supports both iOS and Android. The data collection by this app is not continuous but intermittent, the measurements collected by Carat is sent to Carat servers fro statistical analysis and report making. Data storage and processing is managed by AWS and Spark does the statistical analysis and report making. The report is sent back to the user as a report in Carat app and recommends actions for increasing battery life and betterment of user experience. The work flow of Carat is simply summarized below and also in Figure 6:

1. Records variable measurements about the device and the usage data
2. Measurements are transferred to the servers (AWS and Spark)
3. Raw data is statistically analyzed
4. Results of the statistical analyses is sent back to the servers (AWS and Spark)
5. The analysis report is sent back to the device application for the visualization of the users

Carat follows strict privacy policy. Table 2 lists what and what not Carat collects from user device. Information like Name, Email address or private data from any other app is not collected by Carat. There are unique user ids which represents each user and is enough to map a certain user for analysis without knowing the identity of the user.

According to the privacy policy of Carat mentioned in the official website [5], Carat is a research project so the data can be published in academic publications and public domains without releasing the user identity. As Carat itself is also an app, it is built in such a manner that it consumes as few resource as possible. It periodically transfers

[4] http://carat.cs.helsinki.fi/
[5] http://carat.cs.helsinki.fi/

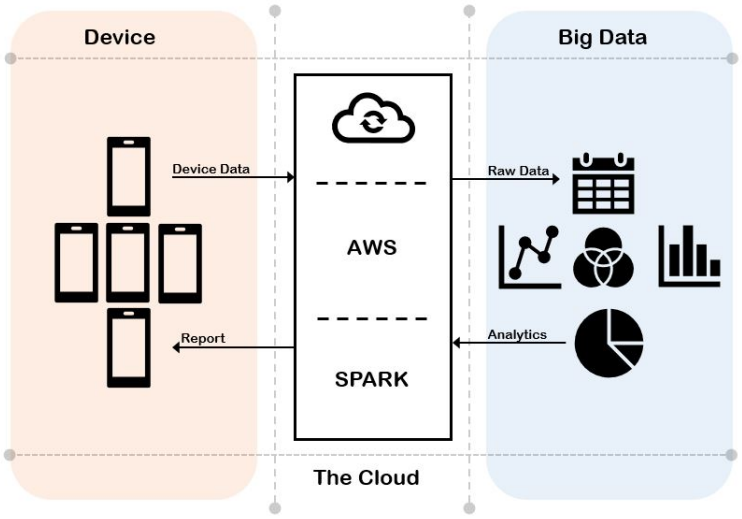

Figure 6. Carat data flow

Table 2. Carat Data Collection and Privacy

Collects	Does not collect
Running apps	
Remaining battery percentage	
CPU utilization	Name of the user
Memory utilization	Email address of the user
Device ID which is unique	Any private data from various apps
OS version of the device	
Model of the device	
Battery state - like charging, discharging etc.	

data to the server for mathematical analysis. Carat is suggested to be run every now and then so that it can keep collecting battery information from the device.

The data used in this thesis was collected by Carat in between 14 March 2018 and 25 August 2018. In this particular timestamp, the volunteering Carat users also answered the Big Five personality trait questionnaires which was given to them via user questionnaire tool. The tool was a part of the Android version of the platform. In the following two sections, the description of mobile usage data and Big Five questionnaire score data is discussed elaborately.

3.1.1. Mobile Usage Data

Mobile usage data of an initial count of 843 participants were collected directly from the AWS server instance of Carat. For the purpose of this thesis, before accessing the data, a privacy agreement is signed as per the Carat policy. The raw data is in nested

JSON format. Nested JSON means JSON object inside another JSON object. For example, a nested JSON looks like below:

{ "uuid":"xxxx", "time":1479772764, "batteryLevel":88, "batteryState":"charging", "apps":[{ "priority":"Service" }, "networkStatus": "wifi", "androidFields": { "screenBrightness":-1, "battery":{ "charger":"ac", "health":"good" }, "networkInfo":{ "networkType":"wifi" }, "timeZone":"America/Toronto"] } }

In Carat raw data, each user has a time series of following noteworthy data attributes in JSON format mentioned in Table 3. uuid is the unique identity of each user which is anonymous. To map the mobile usage of an user with Big-Five scoring data, uuid has been used to understand which user's data it is. timestamp is the UNIX time at when the sample was collected, for example 1479772764 is an UNIX timestamp which is Monday, November 21, 2016 11:59:24 PM as per GMT. batteryLevel is the battery percentage at the time when the sample was collected, for example 80 (%), 90 (%). The plugging status of the battery is batteryStatus which states if the mobile was charging or discharging at the time of sample collection. apps contains information about which application is being used by the user and if the app is running in foreground or background.

With the initial count of 843 users, the sample is cut to 739 users which have at least 10 days of application usage data. The participants who have less than 10 days of data are insignificant because the data is too small to train a model. Samples with less than 10 days of data holds very little information about the users. So data of these 104 users are not considered in this thesis work. The users have self reported demographics from which an idea can be built about the participants' age, educational background, gender and country (Figure 7, 8 9 and 10).

From the demographic graphs of Figure 7 and 8, it is visible that, around 31.1% participants are from age group 25-34 which is the major percentage among the total percentage. The difference between Male and Female participants is huge - 87.8% Male and 10.3% Female. Others did not disclose their gender or belongs to other gender than Male/Female. So the data set is dominant by Male participants. From the country perspective, 55.3% people did not disclose their country name whereas the second highest number of participants are from US (16.5%) followed by Finland (6.1%) and UK (3.5%). The insignificant data (countries with sample less than 0.4%) is not shown in the graph for Country in Figure 9. Figure 10 shows the educational background of the participants. Majority (37.3%) of the participant is undergraduate degree students, that is Bachelor's or equivalent. Second highest number of participants (30.9%) are Professional graduate/Master's or equivalent degree students. 13.9% are from high school, 9.7% are from vocational school, 5.8% are PhD students.

In this thesis, the Big Five personality prediction is done based on application and application category. The categorization of the applications has been done by considering Android's Google Play categories mapped with the application processName from the Carat mobile data. The total number of applications in the data set is 7852 which is categorized in 41 app categories. There are categories like Tools, Communication, Education, Transportation and so. Figure 11 and 12 show which application categories are most and least used by the participants.

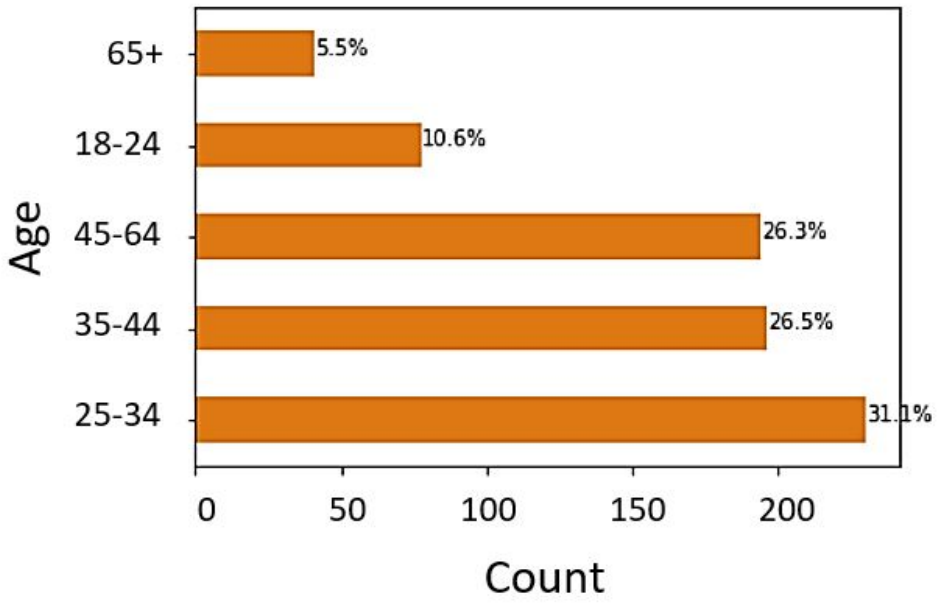

Figure 7. Demography of participants (age)

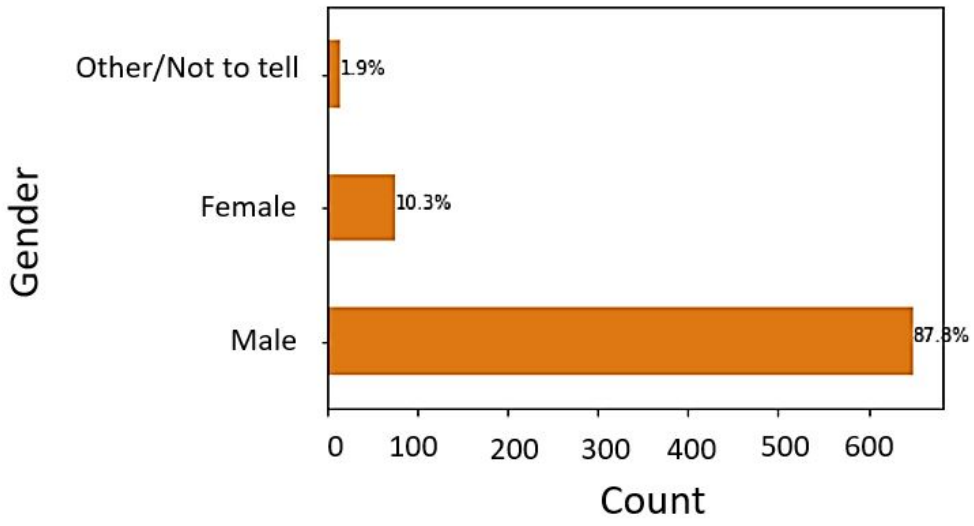

Figure 8. Demography of participants (gender)

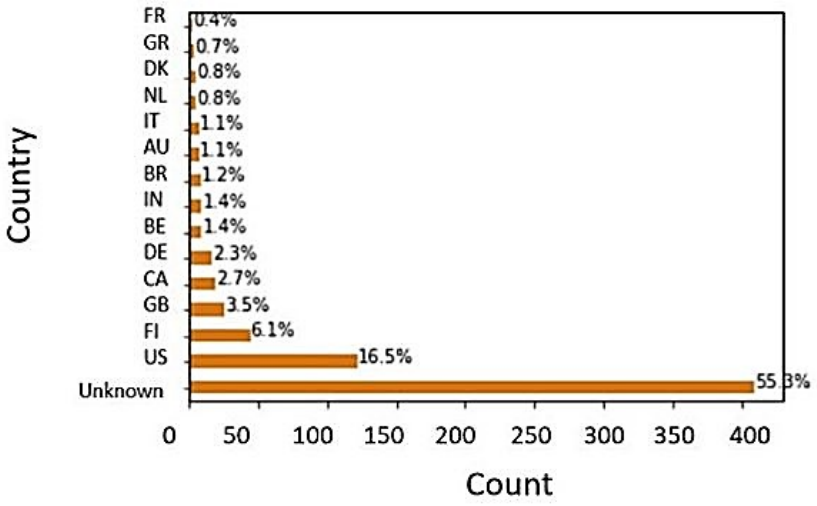

Figure 9. Demography of participants (country)

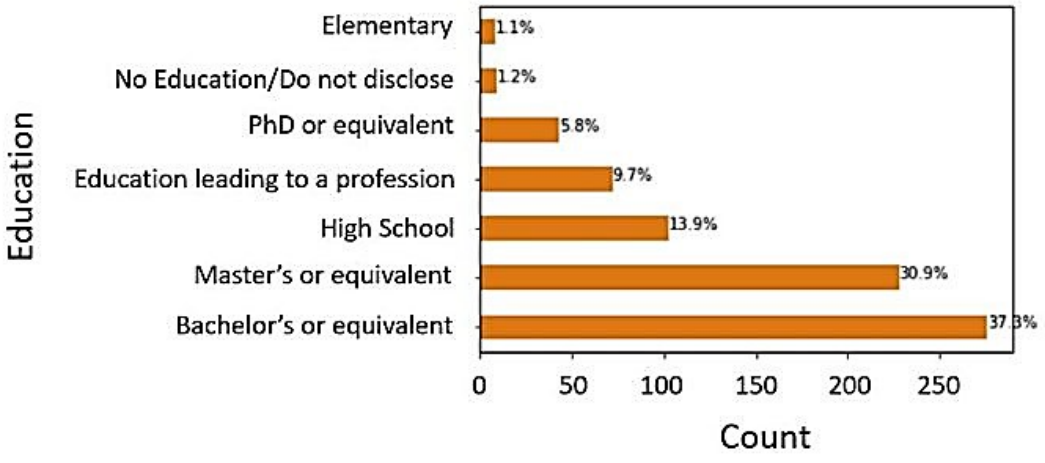

Figure 10. Demography of participants (education)

Table 3. Description and example of noteworthy Carat attributes

Carat Mobile Usage Data Attributes		
Attribute Name	**Description**	**Example**
uuid	Anonymous unique user ID	00fe639859d5c222
timestamp	The UNIX time at when the sample was taken	1479772764
batteryLevel	Battery percentage at the time of sample collection	88, 90 etc.
batteryStatus	The plugging status of the battery during the time of sample collection	charging, discharging etc.
timeZone	User's time zone	America/Toronto
apps	App usage data. It contains attributes like processName and priority which is described in the next two rows.	-
processName	Android package name of the apps	com.google.android.music:main, com.instagram.android:mqtt etc.
priority	App priority - whether the app is actively being used or is running in the background	foreground, background, service etc.

The top five most used application categories are Tools (100% users), Communication (99.5% users), Productivity (92% users), Social (80.5% users) and Travel_local (66.3% users). Tools category includes Android tool applications like Google Assistant, Time and Stopwatch, Anti-virus, File Manager and so on. Also, Carat itself is a Tool so 100% people were using Tools anyway. Followed by the second most used application category, Communication, includes Instant Messaging, Email, Contacts, Call Recorder etc. Among the five least used app category, all of them are under Gaming. In Android app categorization, there are several categories under Gaming. So overall the Game category has been sub categorized into several groups like Game_Sports, Game_Trivia, Game_Educational and so on.

3.1.2. Big Five Personality Trait Data

The volunteering Carat platform users have answered 50-item Big Five personality questionnaires. The questionnaire was a part of the Android version of the platform. The Big Five personality questionnaire answers for 739 users of this study has been collected from the server and the score is calculated as per the official website of IPIP

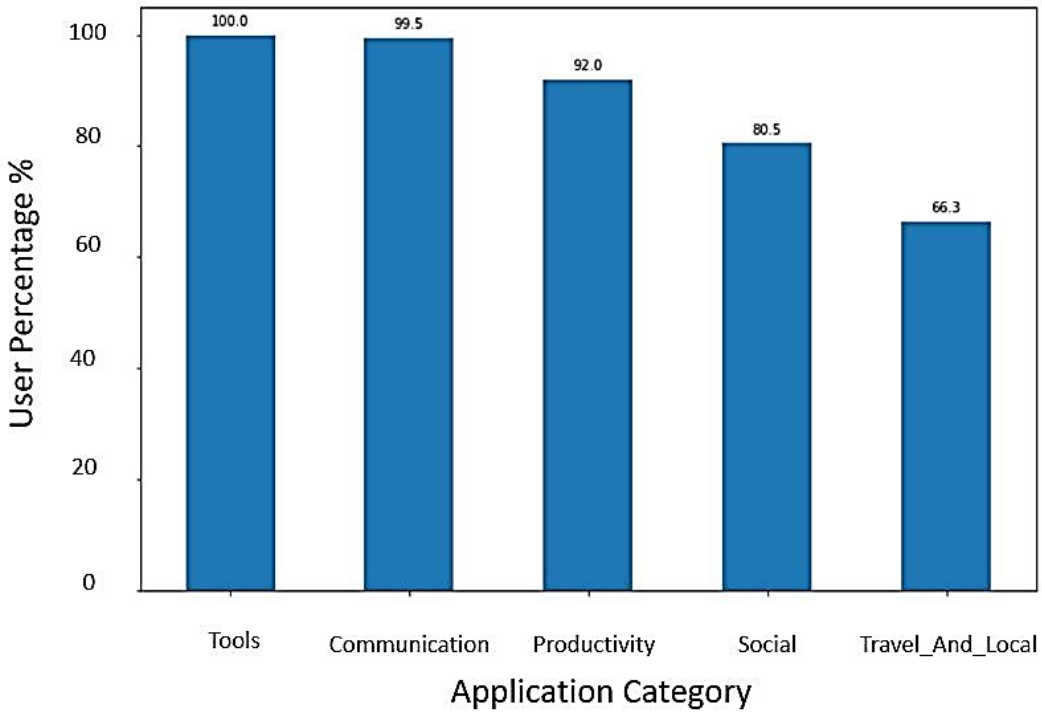

Figure 11. Percentage of users with 5 most used application category (739 users)

50-item scale [90]. The user id (uuid) is same for both mobile usage data and Big Five questionnaire which is how the questionnaire is matched with the mobile usage data for 739 subject users. The IPIP standard scoring of Big Five personality traits provide an uniform psychological assessment measure for each time Big Five is used.

The Big Five questions are grouped into "+keyed" and "-keyed" items with five factors as - Factor 1 is Extraversion, Factor 2 is Agreeableness, Factor 3 is Conscientiousness, Factor 4 is Emotional Stability and Factor 5 is Intellect_Imagination. For the responses, "Very Inaccurate" is assigned a value of 1, "Moderately Inaccurate" a value of 2, "Neither Inaccurate nor Accurate" a 3, "Moderately Accurate" a 4, and "Very Accurate" a value of 5. So for example, if a question is grouped as (+1) and a participant has responded with "Moderately Accurate", a 4 will be added to the "Extraversion" score of the participant. The scores for each traits are summed individually to get the total scores for the five personality traits.

From the descriptive statistics of the overall Big Five personality scores of 739 users (Figure 13), it is visible that the mean value for Extraversion is the lowest (27.6) and for Intellect/Imagination it is highest (39). If the coefficient of the variation (CV = standard deviation/mean) is calculated for each traits, for Extraversion it is 0.291, Agreeableness it is 0.178, Conscientiousness it is 0.176, Emotional Stability it is 0.244 and Intellect Imagination it is 0.151. All the coefficient of variance values are less than 1 which indicates the values are small and thus the distribution of data is centered around mean and not spread out over a wide range [91].

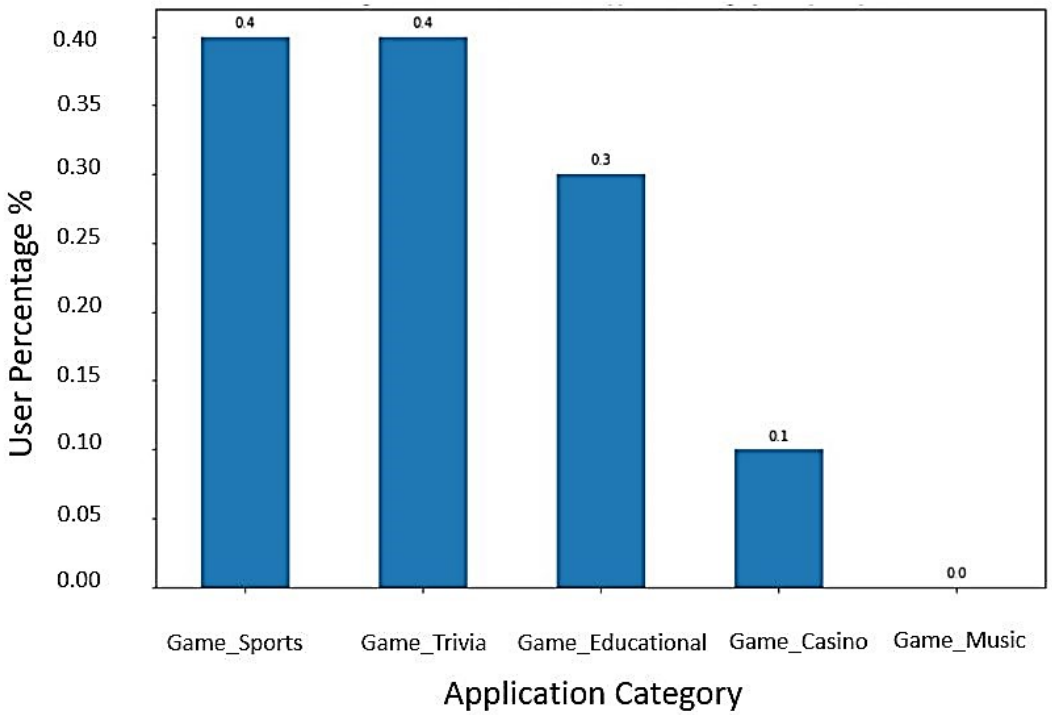

Figure 12. Percentage of users with 5 least used application category (739 users)

3.2. Data Pre-Processing

3.2.1. Data Collection from Carat AWS Server

The mobile usage data are stored in Carat AWS server in gunzip files. For this thesis, only data of people who have answered the Big Five questionnaire was needed to fetch from the large amount of data stored in Carat AWS server. For this, a bash script is used to fetch the necessary data from the server and is saved in JSON files for further processing.

3.2.2. Create Binary Matrix from Raw Mobile Usage Data

The raw mobile usage data is pre-processed into dataframes in order to run the analysis. In this study, Python Pandas is used to create the dataframe from JSON format using JSON normalization. Initially there are three dataframes for application categories, applications and Big Five personality traits. Then a huge binary matrix is created from app category and app dataframes in a way that if a user has used an app or app category there will be 1, if not then 0. So there are two binary matrices, one for app and another for app category having column names as app and app category respectively with row values as 0 (did not use the app or app category) or 1 (used the app or app category). Figure 14 and 15 shows a snap of how the binary matrices look like.

	Extraversion	Agreebleness	Conscientiousness	Emotional_Stability	Intellect_Imagination
count	739.000000	739.000000	739.000000	739.000000	739.000000
mean	27.596752	36.744249	34.975643	32.456022	39.096076
std	8.038174	6.545925	6.158428	7.936684	5.937134
min	11.000000	10.000000	17.000000	10.000000	17.000000
25%	22.000000	33.000000	31.000000	27.500000	35.000000
50%	28.000000	37.000000	35.000000	33.000000	39.000000
75%	33.000000	42.000000	39.000000	38.000000	43.000000
max	49.000000	50.000000	50.000000	50.000000	50.000000

Figure 13. Snap of Descriptive statistics of Big Five questionnaire for 739 users

	uuid	.NetradarService	.NettitutkaService	.bbmonserver	.com.czjk.ibraceletplus.fitristpulzz.BluetoothLeService	.dataservices	.esfm
0	00fe639859d5c222	0	0	0	0	0	0
1	10788bb43960a383	0	0	0	0	0	0
2	10aa1f562c6fd571	0	0	0	0	0	1
3	10f2ec9d9f80f1e3	0	0	0	0	0	1
4	11aa13e2b647773c	0	0	0	0	0	0

Figure 14. Snap of App Binary Matrix

3.2.3. Convert Big Five Data File into Dataframes

Big Five personality score data was saved in a csv file in the Carat server, which is fetched and converted to dataframe using Python Pandas. The personality traits are scored as per the IPIP instruction described in section *3.1.2*. Figure 16 shows a snippet of the Big Five personality traits score dataframe.

3.3. Principal Component Analysis (PCA) for App Matrix

The binary matrix of the application has 7852 number of applications which makes it a large-scale matrix. It is important to reduce insignificant data from the large-scale matrix to reduce the complexity of the task [92]. When a large number of feature is input to a prediction algorithm, there can be irrelevant data which has no significance to train the model. Dimensionality reduction is done in such a way that the irrelevant data are reduced but the overall structure of the data points are intact [92]. For this purpose, Principal Component Analysis (PCA) is used. PCA is a an explanatory data analysis process which helps to reduce dimensionality of a matrix by finding out the values which are closest to the best fit [93]. PCA shows the large variance among the significant data by choosing the basis vectors of the transformed space [92]. In layman's term, PCA reduces the dimensions of the matrix while retaining the properties of the data.

In order to find out the number of components (n_components parameter of PCA) which preserve most of the total variance data, the cumulative sum (cumsum) of explained variance ratio for each attribute is calculated. Explained variance tells how much information (variance) can be assigned to each principal components. When

	uuid	BOOKS_AND_REFERENCE	BUSINESS	COMICS	COMMUNICATION	EDUCATION	ENTERTAINMENT	FI
0	00fe639859d5c222	0	0	0	1	0	0	
1	10788bb43960a383	1	1	0	1	1	1	
2	10aa1f562c6fd571	0	0	0	1	0	0	
3	10f2ec9d9f80f1e3	0	1	0	1	0	0	
4	11aa13e2b647773c	1	0	0	1	0	1	

Figure 15. Snap of App Category Binary Matrix

	uuid	Extraversion	Agreebleness	Conscientiousness	Emotional_Stability	Intellect_Imagination
0	00fe639859d5c222	30	34	33	18	38
1	10788bb43960a383	23	39	34	28	38
2	10aa1f562c6fd571	48	41	37	33	40
3	10f2ec9d9f80f1e3	18	35	43	43	42
4	11aa13e2b647773c	22	32	42	31	41

Figure 16. Snap of Big Five personality traits scores

PCA is applied for dimensionality reduction, a number of information can be lost. From the explained variance ratio, it can be visualized what percentage of data is preserved by the principal components.

Figure 17 shows that 600 components preserve 99% of the total variance. This reduces the dimensionality of our data to under 10% and enables rapid iteration for automatically selecting the best parameters for the model.

3.4. Analysis

The aim is to predict the Big Five personality traits from the mobile application and application category binary matrix (if used 1, if not used 0). The nature of the data requires an algorithm that takes multidimensional input (the binary matrix) and one target output (each trait). Like a conventional machine learning approach, this study has input dataset X (binary matrix of app and app categories) which are the features and target data y (each Big Five trait) which is to be predicted. As per the formal statistical learning framework [94], if training data is considered as a set, then it is a set of pairs of X X y. This training set is an input to the learner which provides a prediction output h: $X \rightarrow y$, a function of predictor or hypothesis. This predictor can be used to predict the targets of new feature sets. The success of a prediction model is assessed by error calculation. An error states the probability of not being able to predict a target data on random new feature. If a random instance is a, then error of h is the probability to predict a such that h(a) is not equal to f(a).

This thesis has split the data set into train, test and validate sets as X_train, X_test, X_validate, y_train, y_test and y_validate, splitting with the ratio of 60% train data, 20% test data and 20% validate data of the whole data set. The splitting is done randomly and there is no repetitive entry from one data set to another. The predictions are made with Training data set which is tested with Test data set. The Validation data set is for checking and validating the model performance. The success measurement or

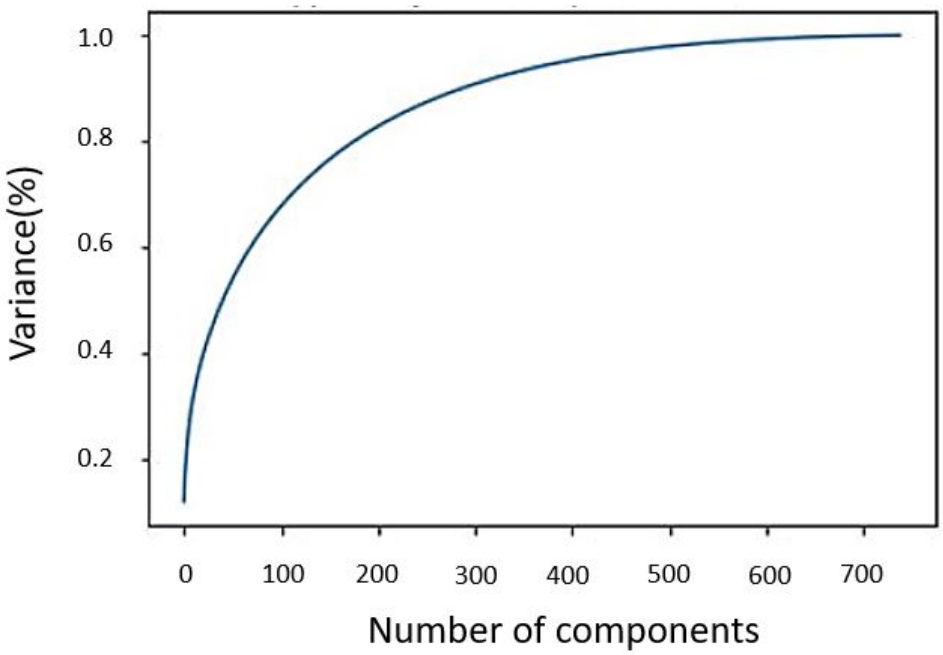

Figure 17. App Binary Matrix Explained Variance Ration

accuracy of the model is measured by Root Mean Square Error (RMSE) [95]. RMSE is the square root of the difference between the predicted values of a model and the actual values observed for the data set. RMSE describes how data is condensed around the best fit. The mathematical notation of RMSE is -

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - d_i)^2} \tag{1}$$

where n is the total number of samples, p is the predicted value of the data, and d is the actual value of the model. The equation sums up the square of the difference between predicted and actual value for all the samples, the result is divided by the total sample and then the square rooted.

Initially, four different algorithms were considered for this thesis - Logistic Regression (LR), Multinominal LR, Random Forest Regression (RF) and Support Vector Regression (SVR). But considering the nature of the data and expected prediction model, RF and SVR are finally considered for this study. To specify the nature of the data and expected prediction model, this study requires algorithm which takes multi column input/feature/independent variable (individual binary matrix of the apps and app categories) and predict a single column output/target/dependent variable (scores of each Big-Five trait). LR are appropriate for the multidimensional target prediction [96]. The model performances for RF and SVR are compared to find the best algorithm for this model based on their performance accuracy.

3.4.1. Random Forest Regression

Random forest regression is a troupe technique which can perform classification and regression both. This thesis has target data (each Big-Five prediction scores) which is continuous not discrete, this is the reason why regression is used. Also the target data (each Big-Five prediction scores) are labeled that is, the prediction model has training examples to use for evaluating its performance accuracy on training the data. RF is a supervised learning algorithm. RF uses Bootstrap Aggregation or bagging with multiple decision trees [97]. Decision trees are a type of model which sequentially routes down to an ultimate result by fulfilling if else conditions. All the possible solutions to a particular decision is explored in a decision tree based on conditions to yield a specific result. In Random Forest regression the target variable of a decision tree is continuous. Figure 21 shows how a decision tree flows down to a result by fulfilling True or False conditions. Bagging is a method for training each decision tree with different sample data. In combination of multiple decision trees, the final output is determined. Figure 18 shows a simple picture of RF algorithm flow.

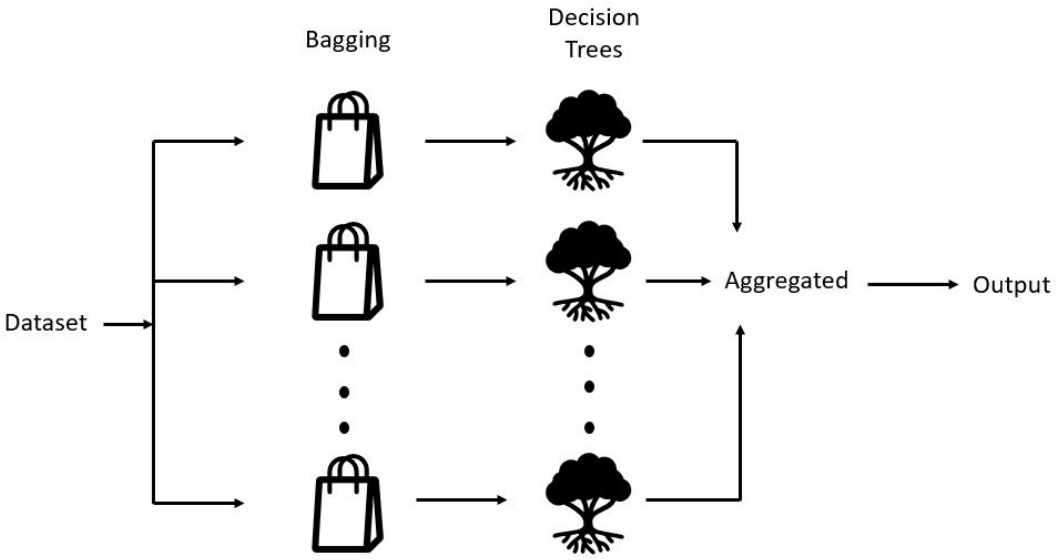

Figure 18. Simple representation of Random Forest Algorithm Flow

As per Fawagreh *et al.* [98], the tree of random forest grows following below rules:

- If the number of cases in training data set is N, then N is randomly sampled which is used as training data for growing the tree.
- When the number of variables are M, the variable m is selected as if $m \ll M$ at each node.
- m is randomly selected from M.
- Best split is applied to m.
- The tree is grown as much as possible.

A pseudo-code is given in Algorithm 1 for more detailed understanding on how RF works. The pseudo-code of RF is described as: if S is a bootstrap sample collected from each tree in the forest, then S(n) is the nth bootstrap sample. N is the number of trees in the forest, F is the features. The RandomForest function takes training set S and features F as input. H represents the decision tree which is initially null or empty. RandomForest function runs a loop till the number of trees to get the decision trees by calling DecisionTreeLearning function for each bootstrap sample S(n). All the learned decision trees returned by DecisionTreeLearning function are aggregated to achieve the final output H. The DecisionTreeLearning grows as: at each node of the tree, all possible best features F is examined till the decision has been made and is returned to RandomForest function.

Algorithm 1. Random Forest

Data: Number of trees in the forest N, training set S, features F
Function `RandomForest`(S, F)**:**
    H $\leftarrow \phi$
    for $n \in 1$ **to** *N* **do**
        S(n) $\leftarrow$ Bootstrap sample from S
        h(n) $\leftarrow$ DecisionTreeLearning(S(n), F)
        H $\leftarrow$ H $\cup$ h(n)
    end
    return H
Function `DecisionTreeLearning`(S, F)**:**
    for *each node* **do**
        split on best feature in F
    end
    return The learned decision tree

Hyperparameter Tuning of Random Forest Algorithm Parameters

Before implementing the RF model, finding out the best parameters of RF is important so that the accuracy of the model performance can be improved. There are default parameters for each machine learning algorithm which can be adjusted for better model performance. Hyperparamters need to be tuned before training the data. Hyperparameter tuning means to find the best values for the parameters used in the algorithm in order to achieve the best performance than the default parameters. In RF, number of decision trees can be tuned as well as number of features for consideration during node splitting. In this thesis, Scikit-learn (a Python library) is used for machine learning. Scikit-learn provides "rational" default parameters already [99], though a tuning has to be done to check if there can be any more improvement done to the model performance.

As mentioned earlier, the total data set is divided into Train, Test and Validate with 60%, 20% and 20% ration respectively. Hyperparameter tuning is trained and validated by Train and Validate data. The parameters which are considered to be tuned are below (the definitions are taken from official website of Sckit-learn [6]) -

1. `n_estimators` = number of trees in the forest

[6] `http://bit.ly/2klLtTw`

2. `max_features` = max number of features considered for splitting a node
3. `max_depth` = max number of levels in each decision tree
4. `min_samples_split` = min number of data points placed in a node before the node is split
5. `min_samples_leaf` = min number of data points allowed in a leaf node
6. `bootstrap` = method for sampling data points (with or without replacement)

The flow of the hyperparameter tuning is given described below:

- Create parameter grid, that is making arrays of possible values for the parameters considered to be tuned.
- Create a base model to tune using RF.
- Randomly search for best parameters using RandomizedSearchCV. Here 3 fold cross validation has been used. This type of cross validation is used over Grid search to make the searching random instead of exhausted search. With the n_iter parameter along with the parameter "cv" (cross validation) of Randomized Search Cross Validation, the number of combinations can be specified for the search. Here n_iter=100 and cv=3, fits 3 folds for each of 100 candidates, total 300 fits
- Fit the random search model.
- Find the best parameters using best_params_, a parameter setting which gives the best results on the data that is hold out.
- Check if the best parameters give the better performance than default parameters by validating with Validate data. RMSE is used to measure the performance.

The comparison of RMSE values for default and tuned parameter is shown in Figure 19 and 20 for respectively - application and application category on each Big Five trait. The orange bars are for best parameters and blue bars are for default parameters. Lower RMSE score is an indication of good fit. For further prediction, best parameters are used for RF.

Prediction model with Random Forest Regression

After the best parameters are found by training with Train dataset and validating with Validate dataset, prediction model for predicting each Big Five trait using app and app category matrices is implemented. The accuracy of each prediction is calculated by RMSE. For the implementation of the RF algorithm, Scikit-learn library is used. RF implementation by Scikit-learn refers to the study of Random Forest by Breiman *et al.* [100]. Breiman *et al.* defines RF as a combination of prediction trees which is dependent on independent random vector samples.

In Figure 21 a reduced sized tree with depth of 3 is shown from the model for predicting 'Extraversion' based on app category for showing how Scikit-learn works

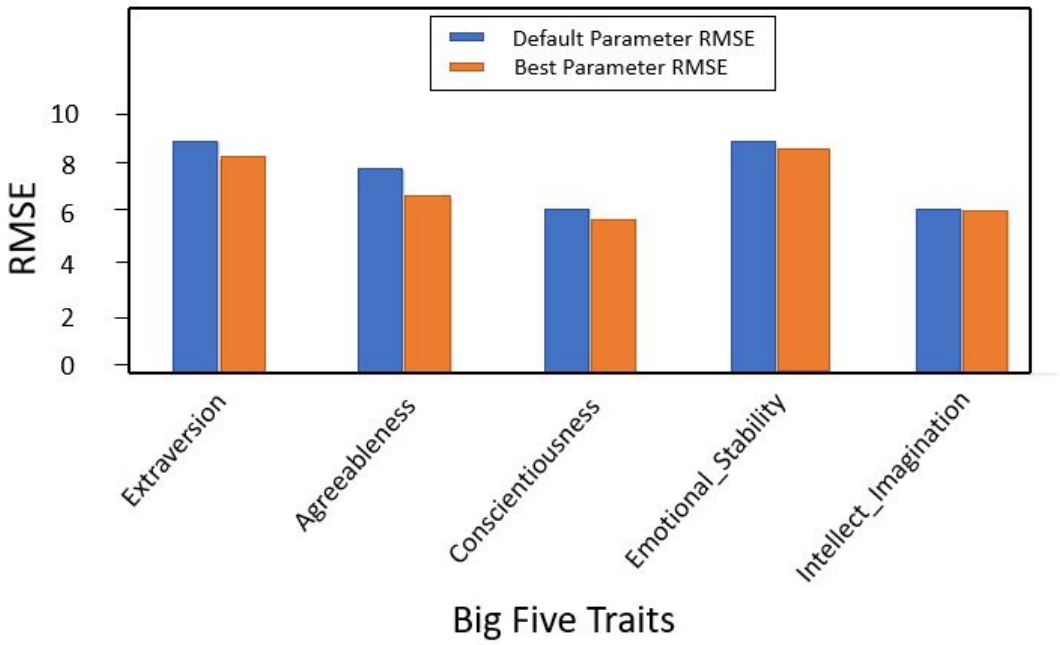

Figure 19. RMSE comparison between default parameter and best parameter for application

on RF. As the app categories are in binary format, the condition value for splitting a node is 0.5. For an example, if a sample has 0 in MEDIA_AND_VIDEO category, then the tree will flow to the left because it fulfills MEDIA_AND_VIDEO <= 0.5 condition. And thus the tree flows similar way by fulfilling the conditions to the leaf node and the 'value' will be the prediction value for the sample. Random sampling of data together with random sampling of features (input data) at each node is thus called Random Forest.

3.4.2. Support Vector Regression

In this study, the Big Five traits are also predicted using Support Vector Regression (SVR). SVR is a supervised learning algorithm that can take multiple inputs as features to predict target data. SVR trains the data in symmetrical manner which penalizes high and low misestimates [101]. The computation of SVR can provide high prediction accuracy, independent of the dimensionality of the input.

Awad *et al.* [101] has described the mathematical explanation of SVR. SVR introduces Vapnik's ε-insensitive region around the function which is called ε-tube. In layman's term ε represents violations which SVR aims to limit to fit as many instances as possible. SVR optimizes the problem by first minimizing ε-insensitive loss function and then finding the tube that contains most of the training instances.

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + b \quad x, w \in \mathbb{R}^{M+1} \tag{2}$$

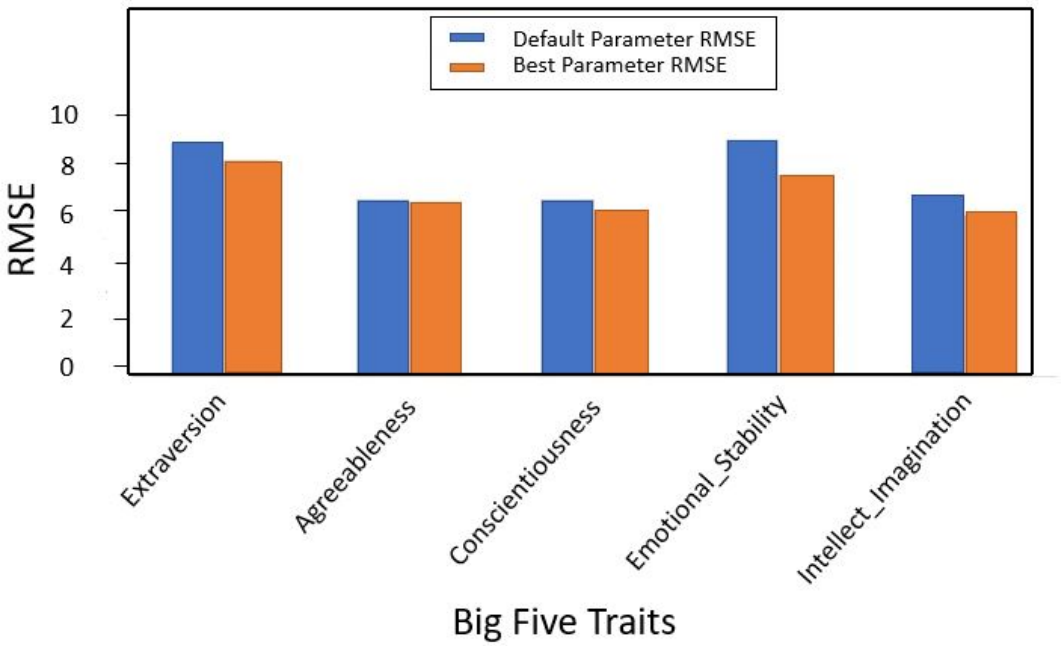

Figure 20. RMSE comparison between default parameter and best parameter for application category

Equation (2) defines the SVR formulation for multidimensional data. As per Awad *et al.* [101], the equation (2) defines as, x is augmented by one and include b in the ω vector to simplify the mathematical notation to finally obtain multivariate SVR. SVR is a complex algorithm to be applied but provides high accuracy for prediction models.

In this study, the steps that is followed to implement SVR are:

- Split the dataset by Train, Test and Validate
- Find the best kernel parameter for SVR and validate with Validate data
- Fit the model
- Predict the model by Test data

Finding the best kernel parameter for SVR

The parameter that needs to be tuned to improve the performance of the model is 'kernel' for SVR. SVR is characterized by Kernel space so it is important to find out the appropriate value for kernel parameter. Scikit-learn provides 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' or a callable kernel option [7]. The best parameter for kernel is determined by following the same steps used for RF in this study (mentioned in section 3.4.1).

Comparison between RF and SVR

The RMSE comarison in terms of accuracy is described in table 4 and 5. The difference between the RMSE accuracy is very minimum in between RF and SVR.

[7]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

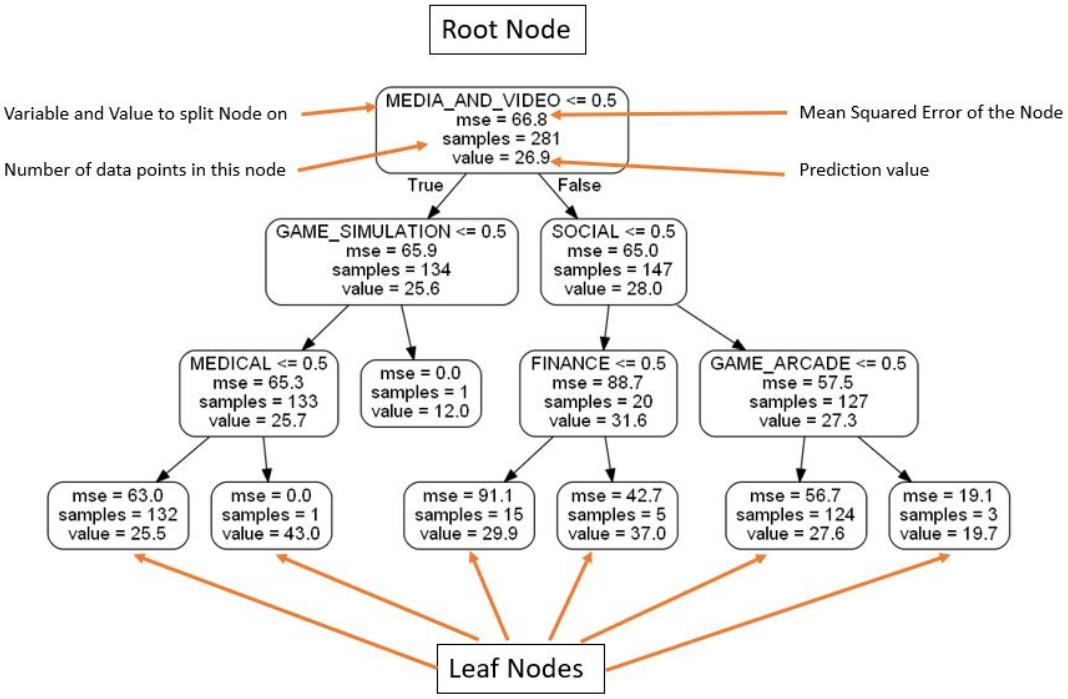

Figure 21. A sample of Random Forest nodes for predicting Big Five trait for app category

For application usage, for the trait Extraversion the accuracy is 90% for SVR and 91% for RF, for the trait Agreeableness the accuracy is 93% for both SVR and RF, for the trait Conscientiousness the accuracy is 93% for SVR and 94% for RF, for the trait Emotional Stability the accuracy is 91% for SVR and RF both, for the trait Intellect/Imagination the accuracy is 93% for both SVR and RF. For application category usage, for the trait Extraversion, Agreeableness and Emotional Stability the model performance is better for RF than SVR. For Conscientiousness and Intellect/Imagination, the performance accuracy is same for SVR and RF. Random Forest has better model performance than SVR overall for training the data set of this thesis. For further analysis, RF is used over SVR to achieve better accuracy.

Table 4. Comparison of the RMSE accuracy between RF and SVR in percentage: Application

Big Five Trait	**Accuracy based on RMSE for RF (%)**	**Accuracy based on RMSE for SVR (%)**
Extraversion	91%	90%
Agreeableness	93%	93%
Conscientiousness	94%	93%
Emotional Stability	91%	91%
Intellectual/ Imagination	93%	93%

Table 5. Comparison of the RMSE accuracy between RF and SVR in percentage: Application Category

Big Five Trait	**Accuracy based on RMSE for RF (%)**	**Accuracy based on RMSE for SVR (%)**
Extraversion	91%	90%
Agreeableness	93%	92%
Conscientiousness	93%	93%
Emotional Stability	92%	91%
Intellectual/ Imagination	93%	93%

4. RESULTS

As per the roadmap, the prediction model is tested using test data set by applying Random Forest Regression. The test data consists 148 number of samples. To understand the error distribution of the model, Figure 22 shows the RMSE distribution of each Big Five trait for both app and app category. The graph shows that, the distribution of RMSE for app and app category are very similar for each trait. A big outlier can be observed for Agreeableness whereas other traits are showing similar range of outliers. But the population of RMSE is dense under 10 for all the traits which is an indication of at least 90% accuracy for each trait, both for application and application category.

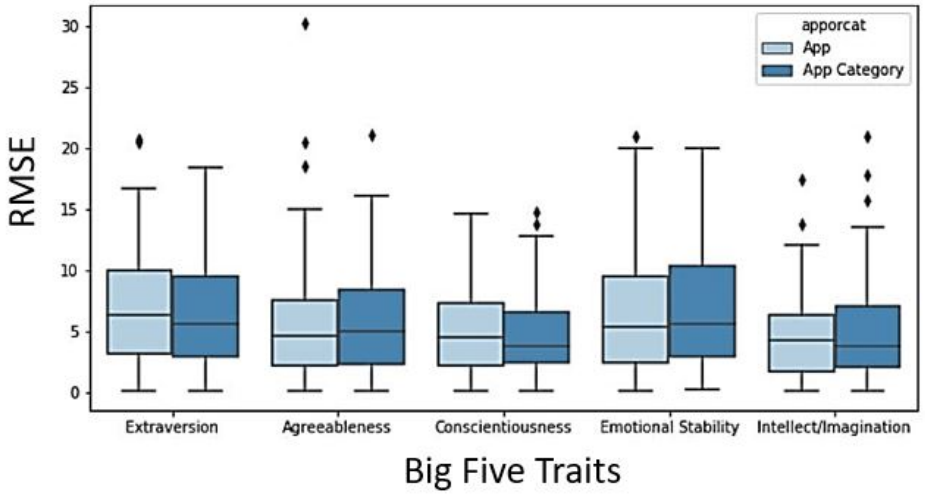

Figure 22. RMSE value distribution of Big Five traits for App and App Category

Describing Figure 22 based on percentiles, the boxplot shows the distribution of the RMSE for each trait on first quartile (25th percentile), median (50th percentile) and third quartile (75th percentile). Table 6 and 7 shows the three percentile values of RMSE for app and app category respectively. 25th percentile is the middle value between smallest value and median of the RMSE score, 50th percentile represents the middle value of RMSE distribution and 75th percentile represents the middle value between the median and highest value of RMSE distribution. For app, the 50th percentile value is between 4-7 which is 96-93% accuracy. For app category, it is 3-6 which is 97-94% accuracy. For the first quartile (25th percentile) of the data the accuracy is almost 98-97% for both app and app category. For the third quartile (75th percentile) of the data the accuracy is 91-94% for app and 89-94% for app category.

Figure 23 shows the distribution of Big Five prediction scores for app. Extraversion has the lowest predicted score distribution among all Big Five traits which means that the Extraversion personality is low among the Test data. Emotional Stability score distribution is approximately between 30-34, Conscientiousness prediction score distribution is between 34-36, Agreeableness prediction score distribution is between 36-38 and Intellect/Imagination prediction score distribution is the highest, between

Table 6. RMSE percentiles for each Big Five trait: App

Big Five Trait	**25th percentile**	**50th percentile**	**75th percentile**
Extraversion	3.12	6.32	9.95
Agreeableness	2.15	4.60	7.47
Conscientiousness	2.10	4.44	7.28
Emotional Stability	2.43	5.28	9.48
Intellect/ Imagination	1.63	4.15	6.27

Table 7. RMSE percentiles for each Big Five trait: App Category

Big Five Trait	**25th percentile**	**50th percentile**	**75th percentile**
Extraversion	2.85	5.52	9.44
Agreeableness	2.20	4.98	8.37
Conscientiousness	2.36	3.78	6.55
Emotional Stability	2.81	5.59	10.32
Intellect/ Imagination	2.01	3.75	6.99

38-40. The higher the score is, the higher that personality exists among the Test data sample.

A similar graph for app category is shown in Figure 24. The score distribution for each trait for app category is similar to app. Extraversion has the lowest predicted score distribution among all Big Five traits. Emotional Stability score distribution is approximately between 32-34, Conscientiousness prediction score distribution is between 34-36, Agreeableness prediction score distribution is between 36-38 and Intellect/Imagination prediction score distribution is the highest, between 38-40.

The similarity between the prediction score distribution of app and app category is an indication that, the study for user personality based on app category is enough. Study with app category is more optimized than studying individual application. Because the number of application used by a user is huge, which results to huge matrices as feature data for any algorithm. Large-scale feature data as an input to any algorithm needs dimesionality reduction in order to find out the insignificant data. If this study is compared to a study by Ortigosa *et al.* [102] which is done on predicting personality by social media (Facebook) interaction (how many post is posted by the user, how many friends the user has etc.), the study has achieved accuracy of 60-80% in predicting personality traits which is lower than the study result of this thesis.

Figure 25, 26, 27, 28, 29 shows the actual vs predicted values for the prediction of the test data for App. Figure 30, 31, 32, 33, 34 shows the same for app category. All of these figures have a common representation of how the model has predicted the lower and higher Big-Five scores. For both low and high values of Big-Five scores, the gap between actual (dark blue line) and predicted (red crosses) values are higher than the scores for mid range (mid range for Big-Five scores are different for each traits (Figure 23 and 24)). For the mid range, the error is very minimum. The reason behind this is the distribution of the Big-Five scores for 739 users. The distribution of the score is not uniform enough to train the model for the lower and higher scores. Figure 35 shows

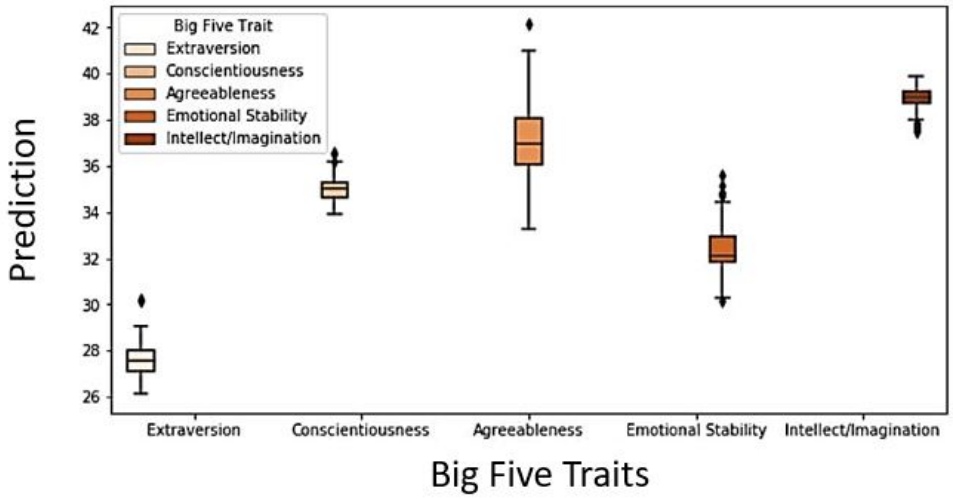

Figure 23. Big Five personality prediction score distribution: App

the distribution of the Big-Five scores for 739 users where it is visible that the score distribution is high in the mid ranged scores. This is an important finding for future analysis where the distribution of the low and high scores or the less distributed data can be manually manipulated in train, test and validate data uniformly, in order to train the model with equal range of data. For now, this thesis shows that the model provides satisfactory training accuracy for the larger population or sample numbers, which is the nature of Random Forest - the more sample for bagging, the more accurate the outcome is.

4.1. Answering the Research Questions

In this section, the research questions defined in the beginning of the thesis is discussed briefly.

RQ1a: Is there any effect of personality traits on users' application usage?

From the results of the RMSE score of the prediction model for application, upto 94% accuracy is achieved for predicting Big Five personality traits for third quartile (75th percentile) of the users. Among the five traits, Intellect/Imagination has the lowest RMSE error (6.27, accuracy: around 94%) followed by Conscientiousness with RMSE error 7.28 (accuracy: around 93%), Agreeableness with RMSE error 7.47 (accuracy: around 93%), Emotional Stability with RMSE error 9.48 (accuracy: around 91%) and Extraversion with RMSE error 9.95 (accuracy: around 90%) is achieved in the third quartile of the RMSE distribution. With such high accuracy values, it can be stated that application usage has effect on Big Five personality traits with low RMSE error. Also, from the predicted scores for each trait, application users tend to score lowest for Extraversion, followed by lowest to highest for Emotional Stability, Conscientiousness, Agreeableness and Intellect/Imagination.

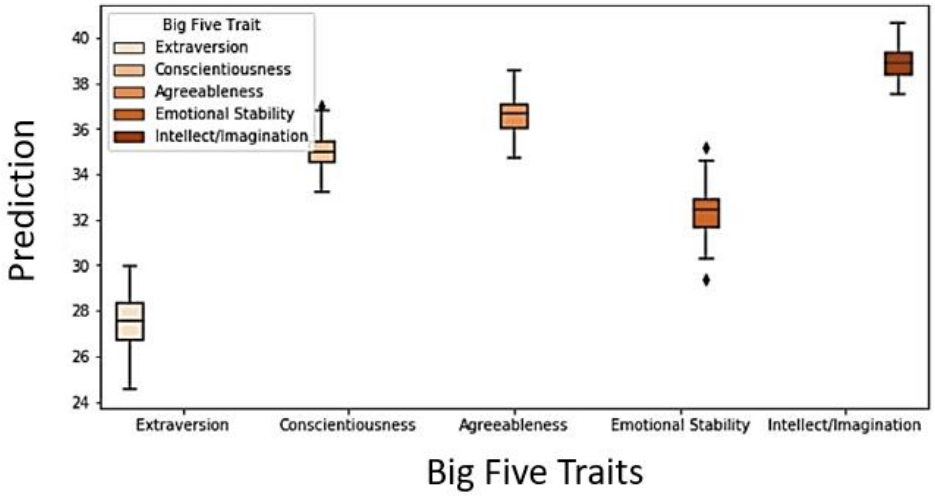

Figure 24. Big Five personality prediction score distribution: App Category

RQ1b: Is there any effect of personality traits on users' application usage based on application category?

Similar to application usage, the results of RMSE score of the prediction model for application, upto 94% accuracy is achieved for predicting Big Five personality traits for third quartile (75th percentile) of the users. Among the five traits, Conscientiousness has the lowest RMSE error (6.55, accuracy: around 94%) followed by Intellect/Imagination with RMSE error 6.99 (accuracy: around 93%), Agreeableness with RMSE error 8.37 (accuracy: around 92%), Extraversion with RMSE error 9.44 (accuracy: around 91%) and Emotional/Stability with RMSE error 10.32 (accuracy: around 90%) is achieved in the third quartile of the RMSE distribution. With such high accuracy values, it can be stated that application category usage has effect on Big Five personality traits with low RMSE error. Also, from the predicted scores for each trait, application category users tend to score lowest for Extraversion, followed by lowest to highest for Emotional Stability, Conscientiousness, Agreeableness and Intellect/Imagination.

RQ2: Application or application category, which one describes Big Five personality more?

The RMSE results from Figure 22 shows that the results are largely similar for application and application category usage for predicting Big Five traits. However, there are 7852 applications which needed to be pre-processed and reduced with dimensionality reduction for finding out the significant data only. Whereas there are 41 Google app category which did not need any reduction. For further studies to avoid reducing the dimensionality of feature matrix, it can be stated that, only application category level analysis is enough for predicting Big Five personality traits as there is no significant difference in the results between application or application category.

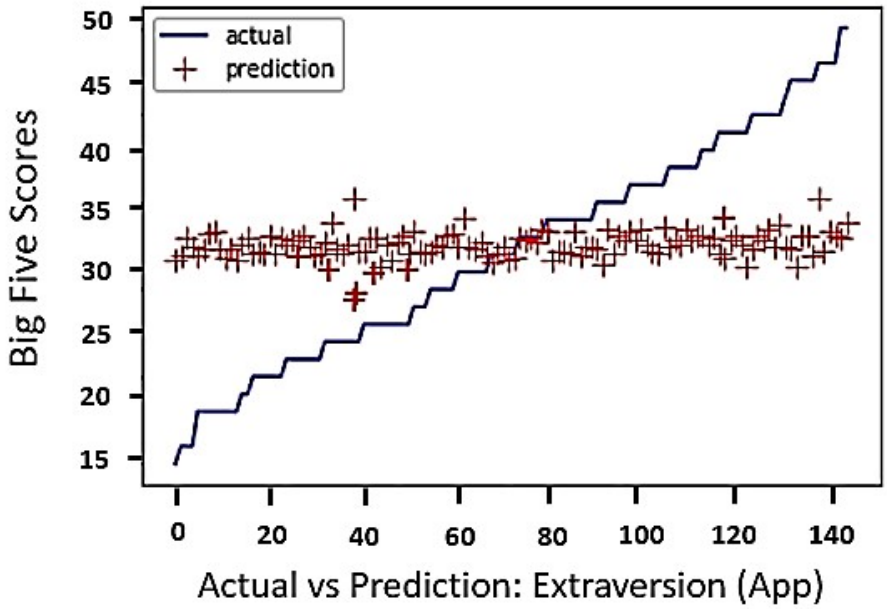

Figure 25. Actual vs Prediction values for Extraversion: App

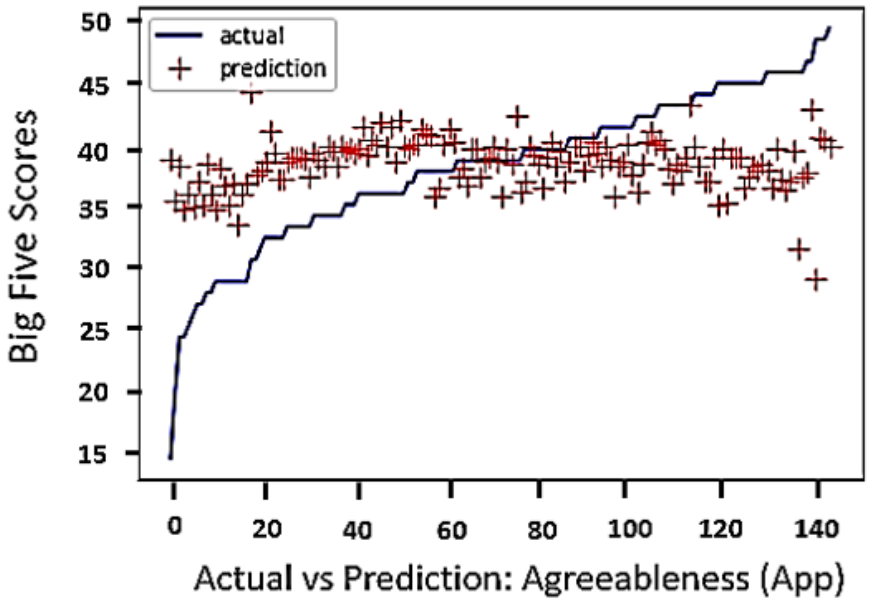

Figure 26. Actual vs Prediction values for Agreeableness: App

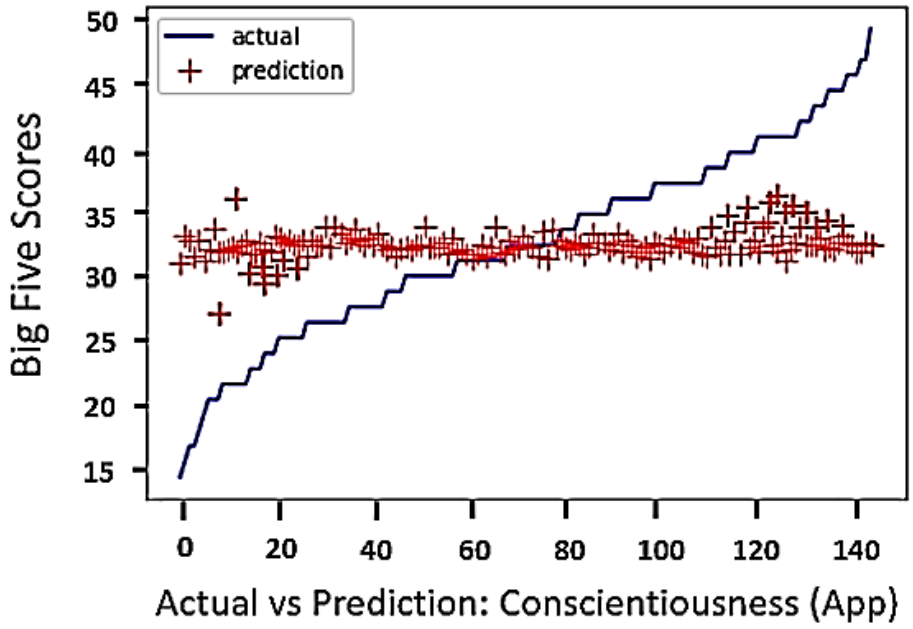

Figure 27. Actual vs Prediction values for Conscientiousness: App

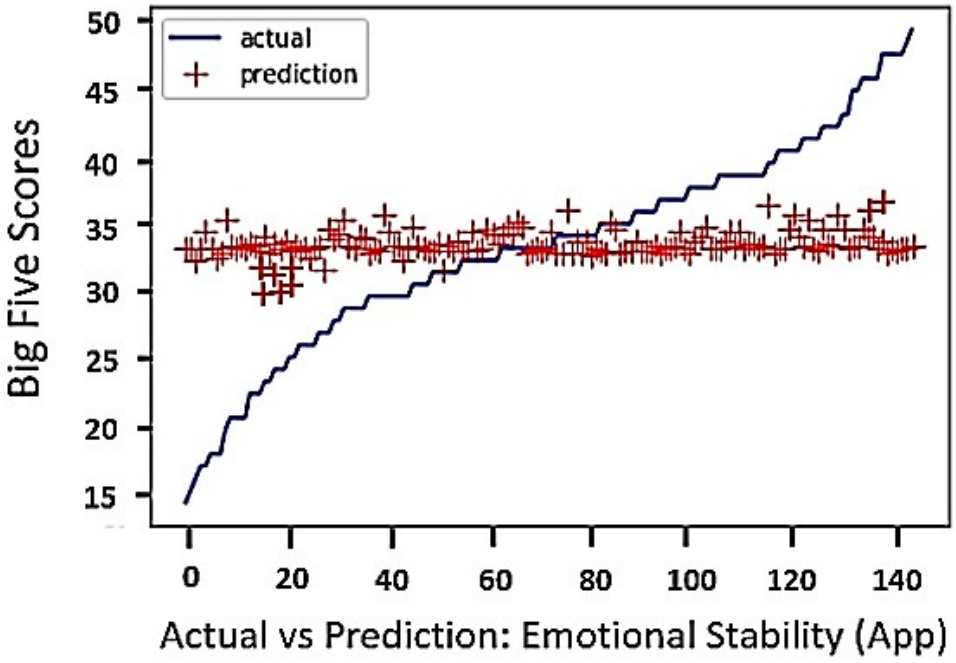

Figure 28. Actual vs Prediction values for Emotional Stability: App

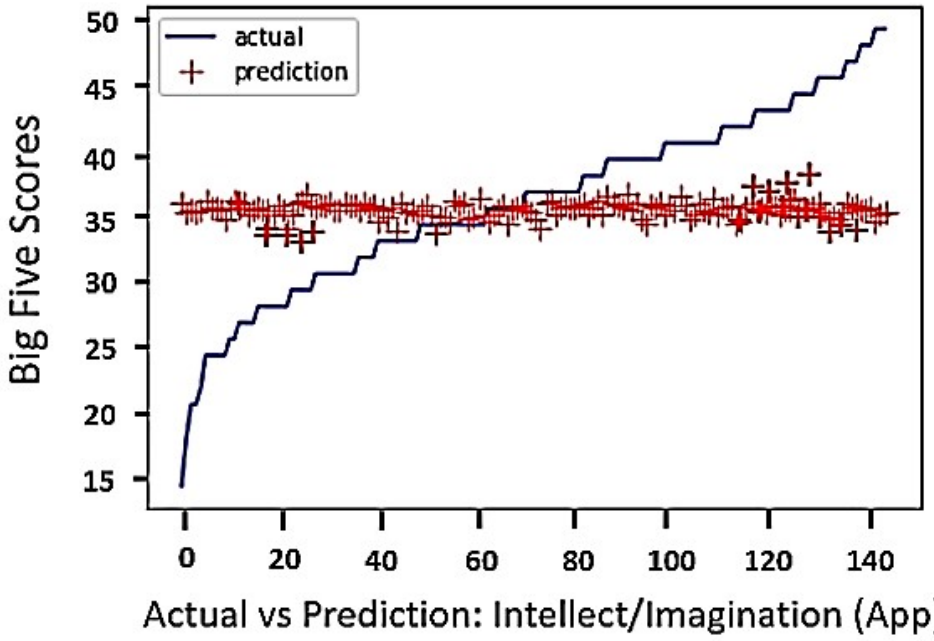

Figure 29. Actual vs Prediction values for Intellect/Imagination: App

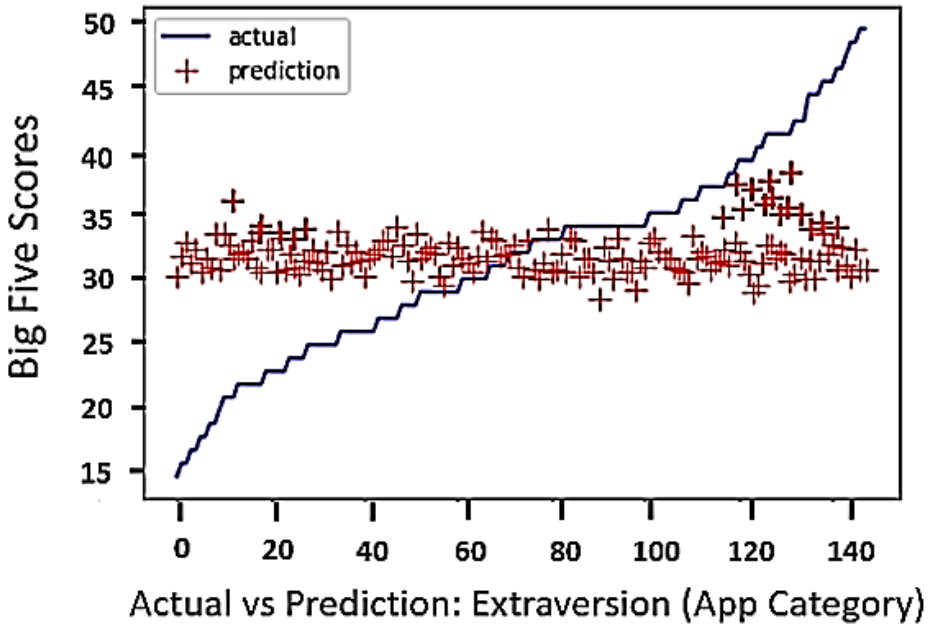

Figure 30. Actual vs Prediction values for Extraversion: App Category

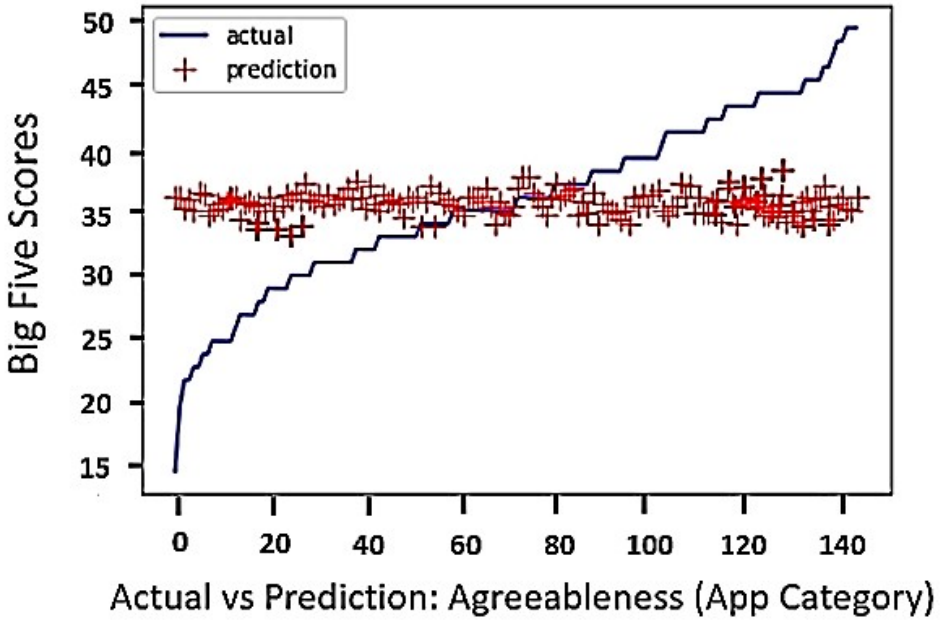

Figure 31. Actual vs Prediction values for Agreeableness: App Category

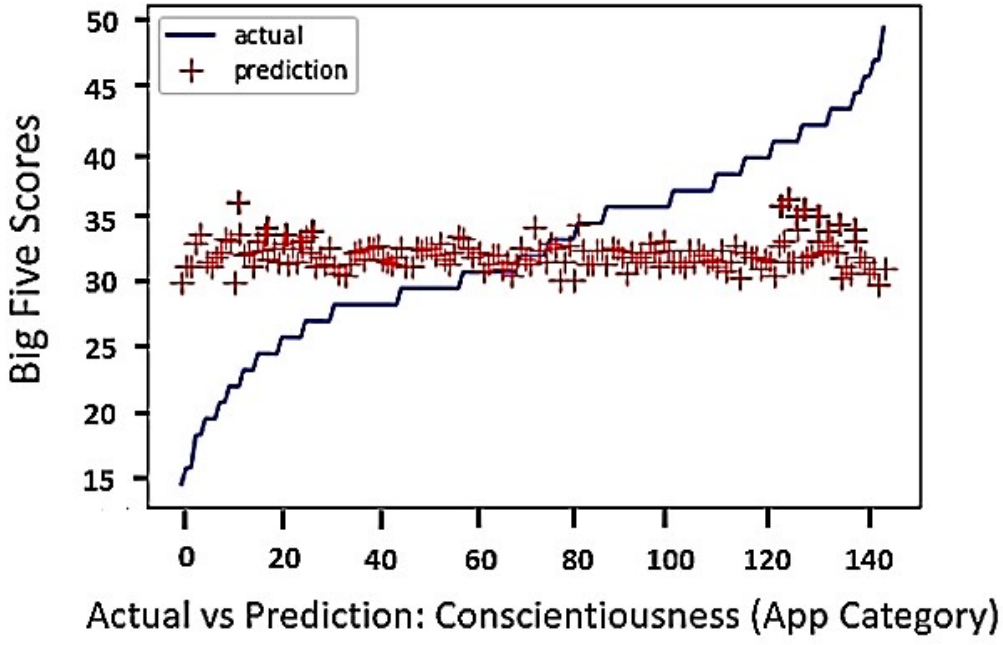

Figure 32. Actual vs Prediction values for Conscientiousness: App Category

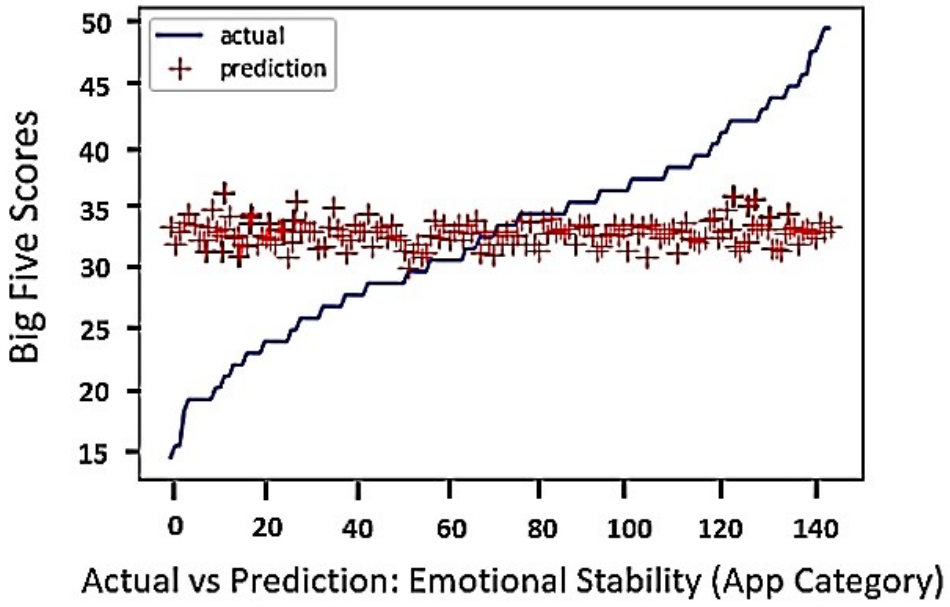

Figure 33. Actual vs Prediction values for Emotional Stability: App Category

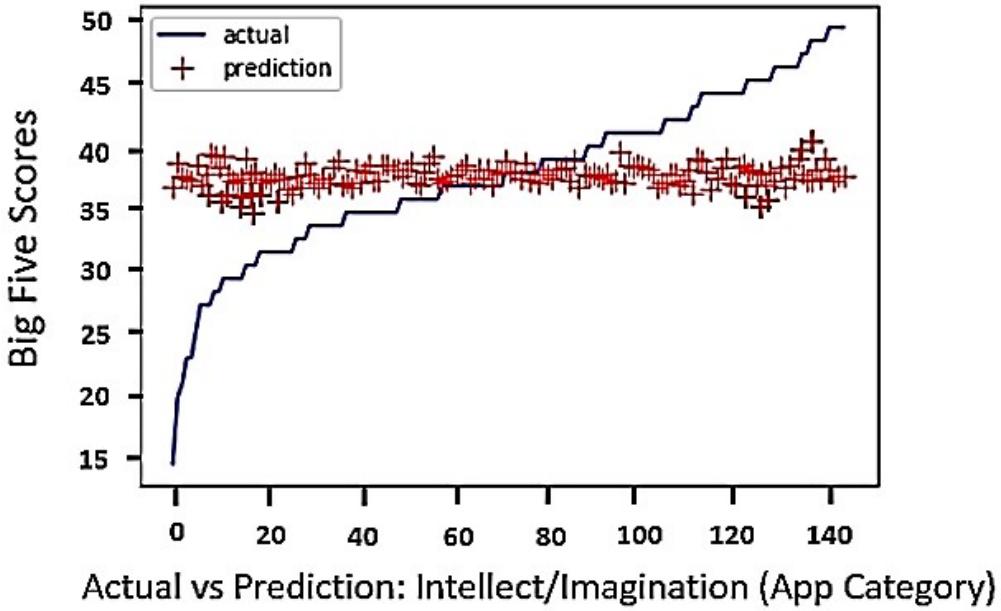

Figure 34. Actual vs Prediction values for Intellect/Imagination: App Category

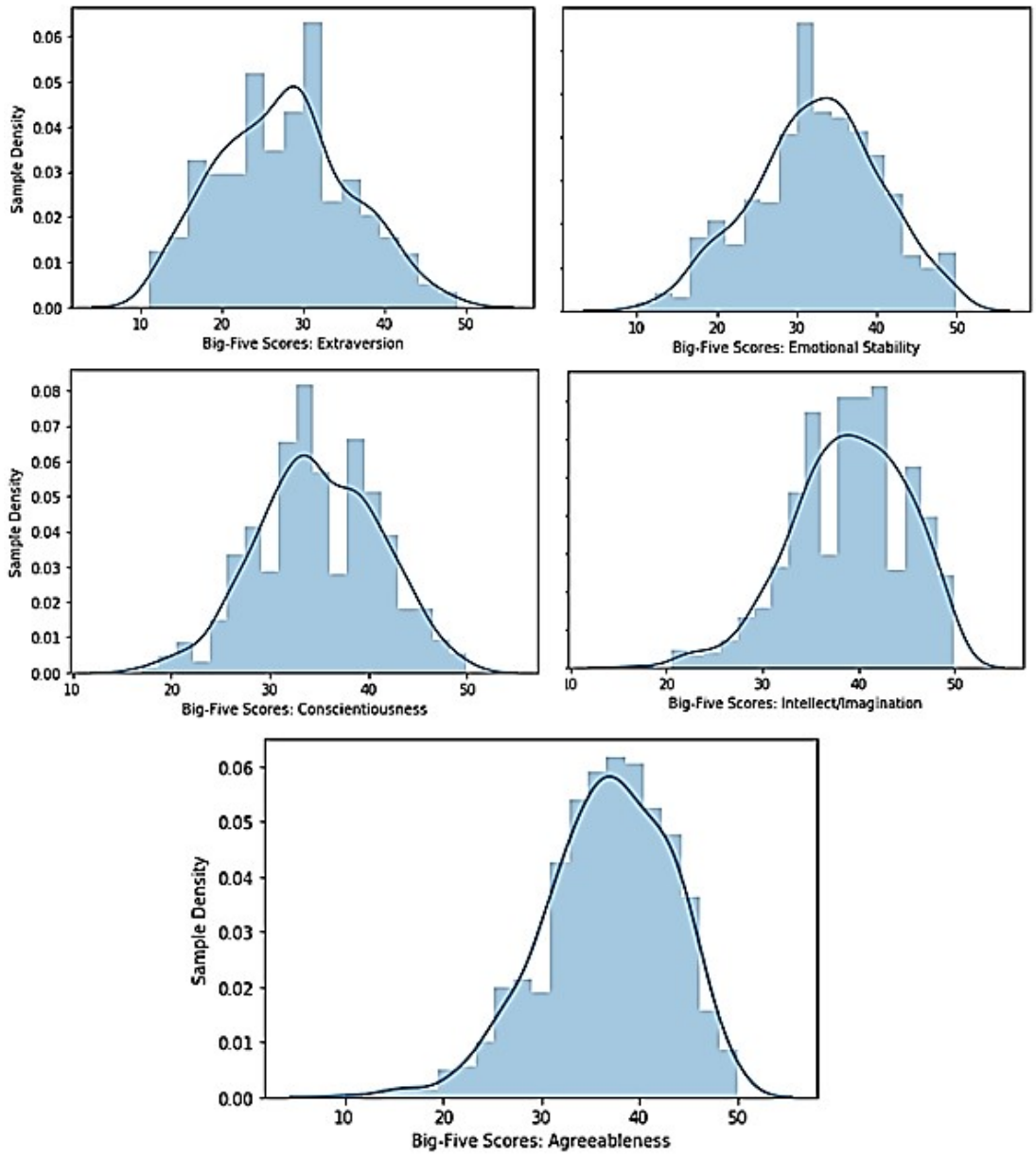

Figure 35. Distribution of the Big-Five scores for 739 users

5. DISCUSSION

Carat data collection challenges

Carat data is very large and is divided into gunzip files in AWS server. It was challenging to collect data of those users who have answered 50-item personality survey from this soup of files. A bash script was used to fetch the data but it took very long time to run the script for initial 843 users because of the type of the file - gunzip. The shell command used for fetching the data from gunzip files is an expensive command and takes long time to execute. Also, the connection between the local machine and server needed to be uninterrupted throughout the time of the execution of the bash script to collect the whole set of required data. To ensure uninterrupted data collection and to reduce the hassle of running script from local machine, Linux 'screen' command is used in this case. Screen multiplexes physical terminal into interactive virtual terminal where user can write commands which will virtually run in the remote machine. Screen makes it possible to run a shell script and then there is no need to think about internet connection timeout. Screen command has made it possible to collect the data without thinking about internet connection interruption for this thesis.

Challenges of normalizing the data into binary matrices

The raw data collected from the server is very huge in size and is in JSON format. In order to normalize the data in dataframe, a high capacity machine was needed. Otherwise, a normal capacity computer takes very long time to process such huge data into dataframes for further analysis. Working with big data is always challenging at the data processing phase. Once the data is in good shape, applying any algorithm becomes very handy. Normalizing the data is a part of data pre-processing and is an important step of data analysis which is called data transformation [103]. Carat data consists enormous amount of information about the smartphone usage of an user. The data is in nested JSON format which is tricky to normalize in dataframes. The JSON objects create column values of a dataframe. For nested JSON, when the JSON is normalized, the columns may consist objects with embedded arrays which need to be flattened also because each component of the array can be an entry of row. Dealing with such big amount of data in nested JSON format was a tricky task to do and needed much attention as well as machine capacity to run.

The demography of the participants

The educational background of the participants shows that, majority of the participants are undergraduate and professional graduate students. Rest are from high school, vocational school, elementary school, researchers etc. This shows that the result of this study is slightly biased for undergraduate and professional graduate students because of the higher number of participant from these two groups. Also the ratio of Male and Female participants is significantly high for Male. Which indicates that this study does not describe a population which has equal distribution of Male and Female. This is an important consideration for an analysis which is reflecting human personality because personality and gender may have impact on each other.

However, the aim of this study is not to correlate the personality with gender rather to discuss the impact of smartphone usage on personality overall. The majority of the age group for the data set is between 25-64 years, which is a big age range for research. The educational background of the participants shows that there are very small amount of users who can be categorized under no education (also do not disclose option). So the data for this study has participants who are students or regular professionals. However, personality is a broad study and is not limited to educational or professional background overall. For now on, this study has participants who use smartphones and have some kind of educational or professional background. But it is always interesting to study the personality of smartphone users who have little educational background. This can be a great future work to be considered.

Various aspects of the data collection and user population

The mobile usage data for this study is all about Android users. It does not include the mobile usage of iOS, Windows and some other mobile OS users. So this study does not cover all human population for predicting personality study. This study is narrowed down to the population who are Android users, Carat users and those who answered the Big Five questionnaire. Though smartphone penetration is high globally, still there are a good number of population who do not use smartphone and are out of scope of this study. Smartphone based studies are always limited to the people who use smartphones or some particular application which is used to collect the data. When an application like Carat is used to collect mobile usage data, it's important to make sure the application itself does not consume much capacity and is not bothersome for the user. Carat is an application which the user just needs to run and that's it. It runs as a background app and keeps collecting data without consuming much energy.

Ample amount of application

There are 7852 number of applications as feature for this study which is huge for any algorithm to predict a target for. Such huge number of feature data may contain insignificant information which may mislead the prediction model outcome. That is why dimensionality has been reduced for application dataset. But dimensionality reduction needs to be done carefully so that any significant information do not get lost with the tuning. Choosing the right number of component is important and cumulative sum of the explained variance helps finding the best number of components for dimensionality reduction. However, this study results that application category is enough to predict personality based on smartphone usage. So the application can be kept aside in future for further analysis.

Smartphone data and Big Five questionnaire for personality studies

As discussed in the background studies, there are studies that has been done for predicting personality based on social media usage [102, 14]. Also various demographic factors are linked with personality in some studies [31]. But personality is a very broad topic to cover. Human personality study can not be covered fully by only self filled questionnaires like 50-item Big Five questionnaire. However, it is a well established and accepted process used by researchers to study human personality

and good amount of study has been done using these traits. This study has found up to 96% of accuracy to predict human personality keeping in consideration that the data has some influence by demographics or OS type. Self filled questionnaires have some chance of being fake or random, so running the survey multiple times is a good way forward to achieve the nearly real and true data about users.

Application of personality studies in real world

Smartphone application market is bringing out new smart applications which are becoming part and parcel of tech users' life. Based on the usage behavior, advanced applications can be developed by targeting the need of various user group. Based on the personality of users, recommendation systems can be developed. Recommendation systems are now a days a latest demand in the world of technology because tech world has a lot to offer but not all of what is offered is needed by the users. Users need more filtering while searching for the right technology - in this case mobile applications. Personality studies based on mobile usage can together bring better services and products for the right user.

Limitations

The 50-item Big Five personality questionnaire is a self filled survey which represents participant's own thought about thyself. Such self assessment type of questionnaires remain limited to user's own perception. Though, unlike studies like depression, human personality is considered to be persona which is developed within times. So it can be said that, taking the personality questionnaire once should be enough. However, if the participant fakes or put random answers, then the study may get manipulated by data which is unreal. 50-item personality questionnaire is quite efficient but is also long and takes time to fill up. So there is always a chance of vague data in this kind of self filled surveys and questionnaires. The data sample of this study is quite big but as discussed earlier, the study does not guarantee the result for whole human population. There are constraints like OS and smartphone dependency, which do not cover a wider range of audience. Also the data collection is dependent on Carat itself. Also, in this thesis, only the smartphone usage is considered for predicting the personality of the users. Other factors like demographics, health condition, personal background, culture, mental condition during fulfilling the questionnaire are not considered.

Another important aspect of such personality related studies is the proper distribution of score data. If the scores are biased to a certain range, the prediction model will also be biased towards the same range. For this, the train, test and validate data can be uniformly distributed manually before applying any algorithm. But then again, manual manipulation is always a challenge for a huge sample. This is something which can be studied more to find a better solution of distributing data uniformly.

Future work

As this study narrows down the fact that, only category based analysis is enough for studying personality traits, in future a more concrete prediction model can be developed only considering the application category. Though Big Five is well

established for personality studies, some other personality measurements like Myers-Briggs Type Indicator, Hogan personality inventory and so, can be used to do similar kind of studies in future.

6. CONCLUSION

This thesis concludes the Big Five personality prediction analysis based on users' smartphone application and application category usage. The platform for collecting the data is Carat. Carat is an application that takes measurements of the usage data and reports to the Carat server for analysis and report generation to suggest the user about which application is consuming how much energy and how to reduce the energy consumption. Carat data consists enormous information about smartphone usage. In this thesis, only the application and application category usage is considered for analysing human personality.

The users have taken 50-item Big Five personality questionnaire. Big Five personality traits have five traits - Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Intellect/Imagination. These five traits describe five different aspects of human personality and is famous among researchers for this kind of personality studies. The personality study is done with 739 users who have taken the questionnaire and are Carat users. The data used for this study is six months long.

There are 7852 applications and 41 Google play application categories in this study. The usage of application and application category is arranged in a binary matrix defined as - if used then 1, if not used then 0. The application binary matrix needed dimensionality reduction for finding out the most significant data among huge number of columns. The input or feature of the prediction model is the application or application category binary matrix and the output or target data is the five traits. Supervised regression algorithms are considered for developing the prediction model for this thesis. Based on the nature of the data, Random Forest Regression (RF) and Support Vector Regression (SVR) are applied in this study. Compared to SVR, the model performance accuracy is slightly better for RF.

The result shows that, whether it is application or application category, the accuracy is not significantly different rather largely similar. Which suggests that, application category level analysis is enough for predicting Big Five personality traits. The model achieved 9-14% error which is 86-91% accuracy on average. For the first quartile (25th percentile) of the data the accuracy is upto 98% whereas for the third percentile (75th percentile) of the data the accuracy is upto 89-94%. The results indicate that there are fundamental effect of personality on people's smartphone application and application category usage.

7. REFERENCES

[1] Wagner D.T., Rice A. & Beresford A.R. (2014) Device analyzer: Large-scale mobile data collection. SIGMETRICS Perform. Eval. Rev. 41, pp. 53–56. URL: http://doi.acm.org/10.1145/2627534.2627553.

[2] Weiser M. (1993) Ubiquitous computing. Computer , pp. 71–72.

[3] Corr P.J. & Matthews G. (2009) The Cambridge handbook of personality psychology. Cambridge University Press Cambridge.

[4] Sadock B.J., Sadock V.A. & Ruiz P. (2000) Comprehensive textbook of psychiatry, vol. 1. lippincott Williams & wilkins Philadelphia.

[5] Martin L.R., Friedman H.S. & Schwartz J.E. (2007) Personality and mortality risk across the life span: the importance of conscientiousness as a biopsychosocial attribute. Health Psychology 26, p. 428.

[6] Rothmann S. & Coetzer E.P. (2003) The big five personality dimensions and job performance. SA Journal of Industrial Psychology 29, pp. 68–74.

[7] Digman J.M. (1990) Personality structure: Emergence of the five-factor model. Annual review of psychology 41, pp. 417–440.

[8] O'Connor B.P. (2002) A quantitative review of the comprehensiveness of the five-factor model in relation to popular personality inventories. Assessment 9, pp. 188–203.

[9] Jia Y., Xu B., Karanam Y. & Voida S. (2016) Personality-targeted gamification: a survey study on personality traits and motivational affordances. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, pp. 2001–2013.

[10] Resnik P., Garron A. & Resnik R. (2013) Using topic modeling to improve prediction of neuroticism and depression in college students. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 1348–1353.

[11] Goldberg L.R. (1993) The structure of phenotypic personality traits. American psychologist 48, p. 26.

[12] Goldberg L.R., Johnson J.A., Eber H.W., Hogan R., Ashton M.C., Cloninger C.R. & Gough H.G. (2006) The international personality item pool and the future of public-domain personality measures. Journal of Research in personality 40, pp. 84–96.

[13] Judge T.A., Higgins C.A., Thoresen C.J. & Barrick M.R. (1999) The big five personality traits, general mental ability, and career success across the life span. Personnel psychology 52, pp. 621–652.

[14] Klobas J.E., McGill T.J., Moghavvemi S. & Paramanathan T. (2018) Compulsive youtube usage: A comparison of use motivation and personality effects. Computers in Human Behavior 87, pp. 129–139.

[15] Bozionelos N. (2004) The big five of personality and work involvement. Journal of Managerial Psychology 19, pp. 69–81.

[16] Stachl C., Hilbert S., Au J.Q., Buschek D., De Luca A., Bischl B., Hussmann H. & Bühner M. (2017) Personality traits predict smartphone usage. European Journal of Personality 31, pp. 701–722.

[17] Makri M., Hitt M.A. & Lane P.J. (2010) Complementary technologies, knowledge relatedness, and invention outcomes in high technology mergers and acquisitions. Strategic management journal 31, pp. 602–628.

[18] Stevens M. & D'Hondt E. (2010) Crowdsourcing of pollution data using smartphones. In: Workshop on Ubiquitous Crowdsourcing, pp. 1–4.

[19] Ferdous R., Osmani V. & Mayora O. (2015) Smartphone app usage as a predictor of perceived stress levels at workplace. In: 2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), IEEE, pp. 225–228.

[20] Mestry M., Mehta J., Mishra A. & Gawande K. (2015) Identifying associations between smartphone usage and mental health during depression, anxiety and stress. In: 2015 International Conference on Communication, Information & Computing Technology (ICCICT), IEEE, pp. 1–5.

[21] Schmid Mast M., Gatica-Perez D., Frauendorfer D., Nguyen L. & Choudhury T. (2015) Social sensing for psychology: Automated interpersonal behavior assessment. Current Directions in Psychological Science 24, pp. 154–160.

[22] Dillon A. & Watson C. (1996) User analysis in hci—the historical lessons from individual differences research. International Journal of Human-Computer Studies 45, pp. 619–637.

[23] Arazy O., Nov O. & Kumar N. (2015) Personalityzation: Ui personalization, theoretical grounding in hci and design research. AIS Transactions on Human-Computer Interaction 7, pp. 43–69.

[24] Ferreira D., Kostakos V. & Dey A.K. (2015) Aware: mobile context instrumentation framework. Frontiers in ICT 2, p. 6.

[25] Felix I.R., Castro L.A., Rodriguez L.F. & Banos O. (2019) Mobile sensing for behavioral research: A component-based approach for rapid deployment of sensing campaigns. International Journal of Distributed Sensor Networks 15, p. 1550147719874186.

[26] De Mast J. & Lokkerbol J. (2012) An analysis of the six sigma dmaic method from the perspective of problem solving. International Journal of Production Economics 139, pp. 604–614.

[27] Lane W. & Manner C. (2011) The impact of personality traits on smartphone ownership and use. International Journal of Business and Social Science 2.

[28] Ferreira D., Dey A.K. & Kostakos V. (2011) Understanding human-smartphone concerns: a study of battery life. In: International Conference on Pervasive Computing, Springer, pp. 19–33.

[29] Oliner A.J., Iyer A.P., Stoica I., Lagerspetz E. & Tarkoma S. (2013) Carat: Collaborative energy diagnosis for mobile devices. In: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, pp. 1–14.

[30] Böhmer M., Hecht B., Schöning J., Krüger A. & Bauer G. (2011) Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In: Proceedings of the 13th international conference on Human computer interaction with mobile devices and services, pp. 47–56.

[31] Zhao S., Ramos J., Tao J., Jiang Z., Li S., Wu Z., Pan G. & Dey A.K. (2016) Discovering different kinds of smartphone users through their application usage behaviors. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 498–509.

[32] Peltonen E., Lagerspetz E., Hamberg J., Mehrotra A., Musolesi M., Nurmi P. & Tarkoma S. (2018) The hidden image of mobile apps: Geographic, demographic, and cultural factors in mobile usage. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–12.

[33] Athukorala K., Lagerspetz E., Von Kügelgen M., Jylhä A., Oliner A.J., Tarkoma S. & Jacucci G. (2014) How carat affects user behavior: implications for mobile battery awareness applications. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1029–1038.

[34] Opoku Asare K., Visuri A. & Ferreira D.S. (2019) Towards early detection of depression through smartphone sensing. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, pp. 1158–1161.

[35] Kroenke K., Spitzer R.L. & Williams J.B. (2001) The phq-9: validity of a brief depression severity measure. Journal of general internal medicine 16, pp. 606–613.

[36] Beck A.T., Ward C.H., Mendelson M., Mock J. & Erbaugh J. (1961) An inventory for measuring depression. Archives of general psychiatry 4, pp. 561–571.

[37] Goldberg L.R. et al. (1999) A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. Personality psychology in Europe 7, pp. 7–28.

[38] Bereziński P., Jasiul B. & Szpyrka M. (2015) An entropy-based network anomaly detection method. Entropy 17, pp. 2367–2408.

[39] Xiao X., Wu Z.C. & Chou K.C. (2011) A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PloS one 6.

[40] Rohani D.A., Faurholt-Jepsen M., Kessing L.V. & Bardram J.E. (2018) Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. JMIR mHealth and uHealth 6, p. e165.

[41] Wang R., Wang W., daSilva A., Huckins J.F., Kelley W.M., Heatherton T.F. & Campbell A.T. (2018) Tracking depression dynamics in college students using mobile phone and wearable sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2, pp. 1–26.

[42] Chen L., Yan Z., Tang W., Yang F., Xie X. & He J. (2016) Mobile phone addiction levels and negative emotions among chinese young adults: The mediating role of interpersonal problems. Computers in Human behavior 55, pp. 856–866.

[43] Aghanavesi S., Nyholm D., Senek M., Bergquist F. & Memedi M. (2017) A smartphone-based system to quantify dexterity in parkinson's disease patients. Informatics in Medicine Unlocked 9, pp. 11–17.

[44] Kuosmanen E., Kan V., Visuri A., Boudjelthia A., Krizou L. & Ferreira D. (2019) Measuring parkinson's disease motor symptoms with smartphone-based drawing tasks. In: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, pp. 1182–1185.

[45] Aram S., Troiano A. & Pasero E. (2012) Environment sensing using smartphone. In: 2012 IEEE Sensors Applications Symposium Proceedings, IEEE, pp. 1–4.

[46] Nemati E., Batteate C. & Jerrett M. (2017) Opportunistic environmental sensing with smartphones: a critical review of current literature and applications. Current environmental health reports 4, pp. 306–318.

[47] Opoku Asare K., Leikanger T., Schuss C., Klakegg S., Visuri A. & Ferreira D. (2018) S3: environmental fingerprinting with a credit card-sized nfc powered sensor board. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, pp. 298–305.

[48] Andone I., Błaszkiewicz K., Eibes M., Trendafilov B., Montag C. & Markowetz A. (2016) How age and gender affect smartphone usage. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: adjunct, pp. 9–12.

[49] Berenguer A., Goncalves J., Hosio S., Ferreira D., Anagnostopoulos T. & Kostakos V. (2016) Are smartphones ubiquitous?: An in-depth survey of smartphone adoption by seniors. IEEE Consumer Electronics Magazine 6, pp. 104–110.

[50] Hiniker A., Patel S.N., Kohno T. & Kientz J.A. (2016) Why would you do that? predicting the uses and gratifications behind smartphone-usage behaviors. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 634–645.

[51] Salehan M. & Negahban A. (2013) Social networking on smartphones: When mobile phones become addictive. Computers in human behavior 29, pp. 2632–2639.

[52] Bødker S. (2015) Third-wave hci, 10 years later—participation and sharing. interactions 22, pp. 24–31.

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

[54] Takao M., Takahashi S. & Kitamura M. (2009) Addictive personality and problematic mobile phone use. CyberPsychology & Behavior 12, pp. 501–507.

[55] Katevas K., Arapakis I. & Pielot M. (2018) Typical phone use habits: Intense use does not predict negative well-being. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–13.

[56] Ehrenberg A., Juckes S., White K.M. & Walsh S.P. (2008) Personality and self-esteem as predictors of young people's technology use. Cyberpsychology & behavior 11, pp. 739–741.

[57] Montag C., Błaszkiewicz K., Sariyska R., Lachmann B., Andone I., Trendafilov B., Eibes M. & Markowetz A. (2015) Smartphone usage in the 21st century: who is active on whatsapp? BMC research notes 8, p. 331.

[58] Lane W. (2012) The influence of personality traits on mobile phone application preferences. Journal of Economics and Behavioral Studies 4, pp. 252–260.

[59] Hsiao K.L. (2017) Compulsive mobile application usage and technostress: the role of personality traits. Online Information Review .

[60] Pervin L.A. (1994) A critical analysis of current trait theory. Psychological Inquiry 5, pp. 103–113.

[61] John O.P., Srivastava S. et al. (1999) The big five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research 2, pp. 102–138.

[62] GOLDBERG L., Goldberg L., GOLDBERG L., Goldberg L., Goldberg L. & Goldberg R. (1981) Language and individual differences: The search for universals in personality lexicons .

[63] Judge T.A., Klinger R., Simon L.S. & Yang I.W.F. (2008) The contributions of personality to organizational behavior and psychology: Findings, criticisms, and future research directions. Social and Personality Psychology Compass 2, pp. 1982–2000.

[64] Roccas S., Sagiv L., Schwartz S.H. & Knafo A. (2002) The big five personality factors and personal values. Personality and social psychology bulletin 28, pp. 789–801.

[65] Kayiş A.R., Satici S.A., Yilmaz M.F., Şimşek D., Ceyhan E. & Bakioğlu F. (2016) Big five-personality trait and internet addiction: A meta-analytic review. Computers in Human Behavior 63, pp. 35–40.

[66] Quercia D., Kosinski M., Stillwell D. & Crowcroft J. (2011) 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, pp. 180–185.

[67] Mehrotra A., Pejovic V., Vermeulen J., Hendley R. & Musolesi M. (2016) My phone and me: understanding people's receptivity to mobile notifications. In: Proceedings of the 2016 CHI conference on human factors in computing systems, pp. 1021–1032.

[68] Chittaranjan G., Blom J. & Gatica-Perez D. (2011) Who's who with big-five: Analyzing and classifying personality traits with smartphones. In: 2011 15th Annual international symposium on wearable computers, IEEE, pp. 29–36.

[69] Chittaranjan G., Blom J. & Gatica-Perez D. (2013) Mining large-scale smartphone data for personality studies. Personal and Ubiquitous Computing 17, pp. 433–450.

[70] de Oliveira R., Karatzoglou A., Concejero Cerezo P., Armenta Lopez de Vicuña A. & Oliver N. (2011) Towards a psychographic user model from mobile phone usage. In: CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 2191–2196.

[71] de Montjoye Y.A., Quoidbach J., Robic F. & Pentland A.S. (2013) Predicting personality using novel mobile phone-based metrics. In: International conference on social computing, behavioral-cultural modeling, and prediction, Springer, pp. 48–55.

[72] Kotsiantis S.B., Zaharakis I. & Pintelas P. (2007) Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 160, pp. 3–24.

[73] Caruana R. & Niculescu-Mizil A. (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning, pp. 161–168.

[74] Bella A., Ferri C., Hernández-Orallo J. & Ramírez-Quintana M.J. (2010) Calibration of machine learning models. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, IGI Global, pp. 128–146.

[75] Carducci G., Rizzo G., Monti D., Palumbo E. & Morisio M. (2018) Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning. Information 9, p. 127.

[76] Lizzeri A. & Siniscalchi M. (2008) Parental guidance and supervised learning. The Quarterly Journal of Economics 123, pp. 1161–1195.

[77] Gunes H. & Piccardi M. (2006) Assessing facial beauty through proportion analysis by image processing and supervised learning. International journal of human-computer studies 64, pp. 1184–1199.

[78] Adeyemi I.R., Abd Razak S. & Salleh M. (2016) Understanding online behavior: exploring the probability of online personality trait using supervised machine-learning approach. Frontiers in ICT 3, p. 8.

[79] Christopher S.L. & Rahulnath H. (2016) Review authenticity verification using supervised learning and reviewer personality traits. In: 2016 International Conference on Emerging Technological Trends (ICETT), IEEE, pp. 1–7.

[80] Zolfaghar K. & Aghaie A. (2011) Evolution of trust networks in social web applications using supervised learning. Procedia Computer Science 3, pp. 833–839.

[81] Gilpin L.H., Olson D.M. & Alrashed T. (2018) Perception of speaker personality traits using speech signals. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–6.

[82] Staiano J., Lepri B., Aharony N., Pianesi F., Sebe N. & Pentland A. (2012) Friends don't lie: inferring personality traits from social network structure. In: Proceedings of the 2012 ACM conference on ubiquitous computing, pp. 321–330.

[83] Nie D., Guan Z., Hao B., Bai S. & Zhu T. (2014) Predicting personality on social media with semi-supervised learning. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, IEEE, vol. 2, pp. 158–165.

[84] Gray J. & Shenoy P. (2000) Rules of thumb in data engineering. In: Proceedings of 16th International Conference on Data Engineering (Cat. No. 00CB37073), IEEE, pp. 3–10.

[85] Baron J. & Kotecha S. (2013) Storage options in the aws cloud. Amazon Web Services, Washington DC, Tech. Rep .

[86] Sigg S., Lagerspetz E., Peltonen E., Nurmi P. & Tarkoma S. (2016) Sovereignty of the apps: There's more to relevance than downloads. arXiv preprint arXiv:1611.10161 .

[87] Peltonen E., Lagerspetz E., Nurmi P. & Tarkoma S. (2016) Too big to mail: On the way to publish large-scale mobile analytics data. In: 2016 IEEE International Conference on Big Data (Big Data), IEEE, pp. 2374–2377.

[88] Peltonen E., Lagerspetz E., Nurmi P. & Tarkoma S. (2016) Constella: Crowdsourced system setting recommendations for mobile devices. Pervasive and Mobile Computing 26, pp. 71–90.

[89] Truong H.T.T., Lagerspetz E., Nurmi P., Oliner A.J., Tarkoma S., Asokan N. & Bhattacharya S. (2014) The company you keep: Mobile malware infection rates and inexpensive risk indicators. In: Proceedings of the 23rd international conference on World wide web, pp. 39–50.

[90] Goldberg L.R. (1992) The development of markers for the big-five factor structure. Psychological assessment 4, p. 26.

[91] Brown C.E. (1998) Coefficient of variation. In: Applied multivariate statistics in geohydrology and related sciences, Springer, pp. 155–157.

[92] Hemanth D.J. & Estrela V.V. (2017) Deep learning for image processing applications, vol. 31. IOS Press.

[93] Alpaydin E. (2020) Introduction to machine learning. MIT press.

[94] Shalev-Shwartz S. & Ben-David S. (2014) Understanding machine learning: From theory to algorithms. Cambridge university press.

[95] Willmott C.J. (1982) Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society 63, pp. 1309–1313.

[96] Peduzzi P., Concato J., Kemper E., Holford T.R. & Feinstein A.R. (1996) A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology 49, pp. 1373–1379.

[97] Liaw A., Wiener M. et al. (2002) Classification and regression by randomforest. R news 2, pp. 18–22.

[98] Fawagreh K., Gaber M.M. & Elyan E. (2014) Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal 2, pp. 602–609.

[99] Buitinck L., Louppe G., Blondel M., Pedregosa F., Mueller A., Grisel O., Niculae V., Prettenhofer P., Gramfort A., Grobler J. et al. (2013) Api design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238 .

[100] Breiman L. (2001) Random forests. Machine learning 45, pp. 5–32.

[101] Awad M. & Khanna R. (2015) Support vector regression. In: Efficient Learning Machines, Springer, pp. 67–80.

[102] Ortigosa A., Carro R.M. & Quiroga J.I. (2014) Predicting user personality by mining social interactions in facebook. Journal of computer and System Sciences 80, pp. 57–71.

[103] García S., Luengo J. & Herrera F. (2015) Data preprocessing in data mining. Springer.