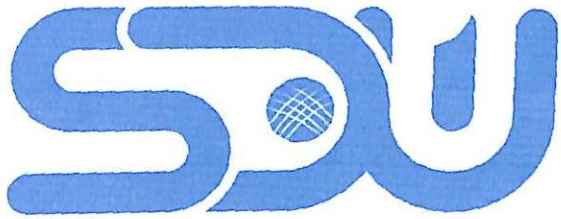
Ministry of Education and Science of the Republic of Kazakhstan
Suleyman Demirel University

UDC 50.10



Aibek Atanbekov

Analysis and development of system for predict psychological portrait of a person

THESIS

Presented in Partial Fulfillment for the
Degree of Master of Science in Computing Systems and Software
(degree code: 6M070400)
Department of Computer Sciences
Faculty of Engineering and Natural Sciences

Supervisor: **Bekmurza Aitchanov**

Kaskelen, 2020

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Aibek Atanbekov

Abstract

Text content is one of the widespread media types. A question which we are trying to find out in this study is the following: can we identify psychological portrait or psychological type of a person by given a short text document written by that person? There are different techniques to identify personality types by given text. In this study, the model has been developed to predict the psychological type of the person based on XGBoost. This study is going to explore machine learning algorithms to predict the psychological portrait of a person based on a given text. The psychological portrait will be described by Myers-Briggs Type Indicators (MBTI) which is the most reliable and popular method. This question is motivated by essays of students at the beginning of their courses to understand their psychological portrait and how it can be met with the profession they choose. The results of this study will help HR specialists and teachers to better understand their students.

Аңдатпа

Мәтін мазмұны - бұқаралық ақпарат құралдарының кең таралған түрлерінің бірі. Осы зерттеуде білгіміз келетін сұрақ: адамның жазған қысқа мәтіндік құжатынан адамның психологиялық портретін немесе психологиялық түрін анықтай аламыз ба? Берілген мәтін бойынша жеке тұлғалық түрлерін анықтаудың әртүрлі әдістері бар. Бұл зерттеуде модель XGBoost негізінде адамның психологиялық түрін болжау үшін жасалды. Бұл зерттеу берілген мәтін негізінде адамның психологиялық портретін болжау үшін машиналық оқыту алгоритмдерін зерттейді. Психологиялық портретті ең сенімді және танымал әдіс болып табылатын Myers-Briggs Type Indicators (MBTI) сипаттайды. Бұл сұрақ студенттердің психологиялық портретін және оны өздері таңдаған мамандықпен қалай сәйкестендіруге болатындығын түсінуге арналған курстардың басында жазған эсселерімен негізделген. Осы зерттеудің нәтижелері HR мамандары мен мұғалімдерге өз студенттерін жақсы түсінуге көмектеседі.

Аннотация

Текстовый контент является одним из распространенных типов медиа. Вопрос, который мы пытаемся выяснить в этом исследовании, заключается в следующем: можем ли мы определить психологический портрет или психологический тип человека по короткому текстовому документу, написанному этим человеком? Существуют разные методы определения типов личности по заданному тексту. В этом исследовании модель была разработана для прогнозирования психологического типа человека на основе XGBoost. Это исследование собирается изучить алгоритмы машинного обучения, чтобы предсказать психологический портрет человека на основе данного текста. Психологический портрет будет описан с помощью индикаторов типа Майерса-Бриггса (MBTI), который является наиболее надежным и популярным методом. Этот вопрос мотивируется эссе студентов в начале их курсов, чтобы понять их психологический портрет и то, как его можно встретить с профессией, которую они выбирают. Результаты этого исследования помогут специалистам по кадрам и преподавателям лучше понять своих учеников.

Acknowledgements

This thesis is going to explore machine learning algorithms to predict the psychological portrait of a person based on a given text. The psychological portrait will be described by Myers-Briggs Type Indicators (MBTI). Text content is one of the widespread media types. A question which we are trying to find out in this paper is the following: can we identify psychological portrait or psychological type of a person by given a short text document written by that person? This question is motivated by essays of students at the beginning of their courses to understand their psychological portrait and how it can be met with the profession they choose. XGBoost will be used for classification model.

Contents

1 Introduction **9**

1.1 Motivation 9

1.2 Personality Types 10

1.3 Thesis Outline 16

2 Literature review **17**

3 Methodology for Prediction **21**

3.1 Tools for Development 21

3.2 Extreme Gradient Boosting 21

3.3 Model Training and Quality Controlling in Machine Learning . . 23

3.3.1 Cross-Validation 23

3.3.2 Regularization 25

3.3.3 Hyperparameter tuning 28

3.3.4 Learning Curves 30

3.3.5 Accuracy-completeness curve, ROC curve 31

3.4 Dataset 33

3.5 Data Analysis 33

3.6 Pre-processing the Dataset 36

3.6.1 Text Vectorization 36

3.7 Prediction 37

4 Results and Discussions **38**

5 Conclusion **40**

References **41**

A Appendix A 43

Nomenclature

MBTI Myers-Briggs Type Indicator

SDU Suleyman Demirel University

XGBoost Extreme Gradient Boosting

1. Introduction

1.1 Motivation

HR, especially those not burdened with psychological education, love the word "psychotype". "This candidate does not fit us in a psychotype!" When you ask what it is and what psychotypes are, you usually hear something about introverts and extroverts in response. They also recall choleric and sanguine people. When asked what psychotype this candidate belongs to and which psychotypes in the company are preferred, there is usually no intelligible answer.

At the same time, with a real selection of personnel, the majority of applicants are no longer on such obvious grounds as specialty, length of service and work experience, place of residence, age, expected income level. When 4 out of dozens of responses to your vacancy remain, and two do not come for an interview, there is no time for a psychotype to find at least someone more or less suitable. We will understand what psychotypes are and how they can be used in personnel management, and we will predict psychotype by only given text.

Traditionally, to identify psychological portrait, the person must take several tests. These tests can be as reports, interviews, or observations conducted by psychologists. This kind of method is costly and less practical. Today, the classification of individuals by the given different criteria is a fascinating study in the psychology field. According to this approach, there is a common view that there are a certain number of types of people in humans. From the perspective of computer science, the modeling of data collected from people computationally or with various machine learning algorithms can make this classification more efficient. Written text can provide some information about the author's personality traits as shown in the previous studies [1]. It is possible to identify the personality of a person by selecting the most commonly used words. The analysis of

personalities makes it possible to make realistic character modeling in the future with the help of engineering disciplines of smarter systems. This, in turn, leads to the emergence of high-level artificial intelligence forms in the future. For example, for a robotic assistant, in social media to analyze and make interpersonal relationships, e-commerce environments, according to customers' personal status, portfolio analysis, or appropriate advertising direction. In order to provide more individual guidance services by using the personal models of students in education, in the health sector, with the help of the identification of the patients in the health, more effective psychological and physiological rehabilitation services, games, etc.

1.2 Personality Types

The first problems in determining a person's psychotype begin with the lack of unified generally accepted approaches to the definition of the term itself — psychotype. The author does not claim to be the only correct one for recognizing his approach to this topic, however, he urges modern specialists working at the intersection of psychology to adhere to international experience and not try to invent the Russian bicycle where it has been invented and defined for a long time. In modern analytical psychology and biometrics, there is still enough space for conducting innovative research and making real scientific discoveries. The unconditional founder of analytical psychology as a modern science is recognized by Carl Jung [2] (Carl Gustav Jung 1875-1961). Among the many achievements and studies, it is sufficient to note the personality characteristics proposed by Jung: introversion and extroversion. Hans Jürgen Eysenck (Hans Jürgen Eysenck 1916-1997) is a British psychologist, one of the leaders in the biological field in analytical psychology, the creator of personality factor theory, and has developed a number of questionnaires and tests to quantify the psycho-physiological parameters of a personality psychotype. Karl Jung and Hans Eisenck are classics of modern psychology, primarily because in their theories they relied on practical and scientific justification of the result, therefore their theories, practical conclusions and tests are widely used in modern psychological research. Our contemporary American psychologist, specialist in the field of clinical psychology and neuropsychology. Howard Gardner [3] (Howard Gardner, 1943) developed the theory of multiple

intelligence. The primary approach of classical psychology was designed to determine a person's psychotype using questionnaires or visually presented tests, for example, the Luscher test [4]. However, any questionnaire is a subjective function, including the real state of a person and his desire to present information about himself in one form or another. Biometric research methods are more objective by definition, as they reflect the real biological characteristics of the object of study.

Typology of Myers-Briggs is a personality typology based on Jung typology [2] and became widespread in the USA and Europe. This is an introspective self-report questionnaire indicating preferences in how people perceive the world and make decisions [5]. Using the Myers-Briggs typology, we can determine the propensity for the type of specialist activity, the nature of problem-solving, and other behavioral features. MBTI contains 4 scales for personality research (see Figure 1.1).

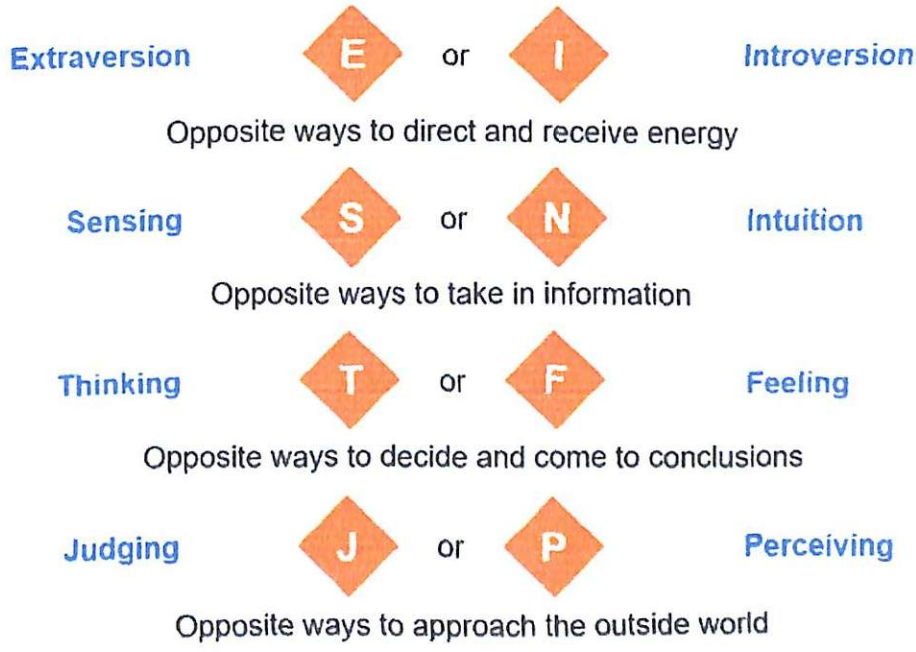


Figure 1.1: The MBTI Preference Pairs

EI scale: type of energy Extroverts prefer to work within a team and get energy from others. They are very talkative and oriented to the outside world. Unlike introverts work independently or in a small team. Charged with energy alone. They prefer thoughts rather than words and oriented on the inner world.

SN scale: type of thinking Sensorics love specifics in everything, including work. They are interested in real experience and what is happening here and now. Orienting to facts and experience gained. Very accurate and information

no matters to them if it is not confirmed. Intuitions prefer theories and ideas. Intuition seeks opportunities and looks at things holistically. Often they have a tremendous imagination and see a lot of opportunities.

TF Scale: Style of behavior Thinkers strive for the work where you can use your intellect. Make decisions guided by logic and analysis. Able to rationally weigh the pros and cons, analyze information well and fairly. Feelers prefer to work where they share their values. They love and know how to work with people, always put themselves in the place of another person. Make decisions emotionally, human feelings are most important than anything else.

JP scale: lifestyle Judger loves organization and structure, is result-oriented, and strives to control what is happening around. Feel uncomfortable when something unplanned happened. Like planning and ordering. Perceiver strives for freedom and flexibility. Prone to spontaneity, it is difficult for them to make decisions. In other words, the perceiver desire to navigate to circumstances, flexible, and able to adapt.

Those four scales form a combination of 16 personality types based on the Myers-Briggs Type Indicator. This type of uncertainty is the type of uncertainty expressed to determine which class of input data (survey responses, etc.) is more suitable for the degree of membership. I will use NLP methodologies to detect uncertain expressions within natural language. The Myers-Briggs Type Indicator (MBTI) is a generally utilized measure of personality identity. It uses constrained decision inquiries to distinguish respondents' inclinations on four frames of mind and capacity sets dependent on Jung's brain science [6] (i.e., Extraversion-Introversion, Sensing-Intuition, Thinking-Feeling, and Judging-Perceiving). These four singular inclinations are consolidated to shape respondents' four-letter identity types. The MBTI identity types are called psycho-types, which demonstrate that people with similar types have a similar approach toward interacting with others and the world [7]. In the MBTI show, all individuals have mental sorts, what's more, there are no sorts other than the 16 distinguished mixes showed in Figure 1.2.

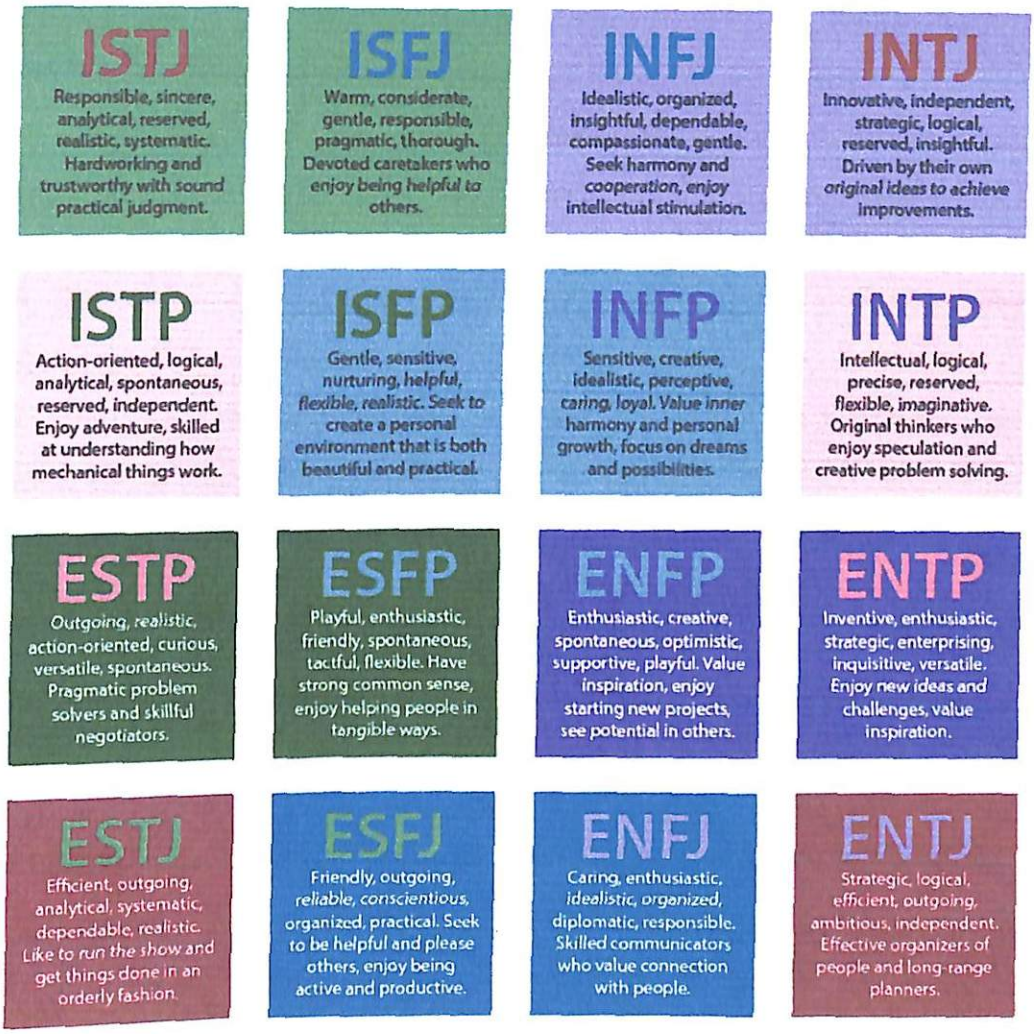


Figure 1.2: MBTI Types

ESTJ: extrovert, sensoric, logic, rational A capable employee who perceives the world "as it is." He is inclined to plan and complete the work begun. It takes care of the close environment, good-natured, but at the same time, it can be quick-tempered, sharp, and sometimes stubborn. Professions: managerial positions, general director, chef.

ENTJ: extrovert, intuition, logic, rational The employee is easily carried away, takes risks, relies on intuition. He safely introduces new technologies, is able to deeply analyze himself and the world. For such a person, life is a struggle in which he feels at ease. It is open to new possibilities but has a need for control. Professions: doctor, engineer, lawyer, prosecutor.

ISTJ: introvert, sensoric, logician, rational Responsible, strict, pedantic - such qualities are possessed by a person with this combination. He focuses on objective reality and is inclined to analyze information. He takes up business only when he is

confident in his abilities and a positive result. Professions: system administrator, accountant, office manager.

ENFJ: extrovert, intuition, ethics, rational Such an employee is emotional and empathetic. His facial expressions are pronounced, and he himself is eloquent. Inclined to self-organization, which is why his fantasies and ideas are realized. He intuitively feels what decision needs to be made in a given situation. Professions: teacher, mentor, PR-specialist.

ESFJ: extrovert, sensory, ethics, rational An employee with such a combination knows how to influence people, takes care, and is inclined to sacrifice himself for the sake of another. He knows how to easily establish a connection with anyone and turn the situation into the direction necessary for himself. Professions: educator, cosmetologist.

INTJ: introvert, intuition, logician, rational He knows how to highlight the main thing, he speaks clearly and on the case, the practitioner. This person likes to improve everything that he does, seeks to realize the task even better. He does not like empty talk, therefore avoids big noisy companies and it is difficult to make contact with people. Professions: programmer, technical copywriter, surgeon.

INFJ: introvert, intuition, ethics, rational About such people, they say "you can trust him." He is very sensitive, attaches great importance to relations between people, knows how to give valuable advice, and "discover others." Developed intuition provides not only an endless stream of ideas but also helps to organize yourself. Professions: counselor, psychologist, pediatrician.

ISFJ: introvert, sensory, ethics, rational This employee is inclined to analyze himself and others, recognizes falsehood, and keeps a psychological distance. He is executive, caring, serves others. He draws strength and energy from internal resources and always relies solely on his own experience. Professions: social worker, educator.

ENTP: extrovert, intuition, logician, irrational Inventive, proactive, and adaptable - this is how you can characterize this specialist. He is the very generator of ideas, a pioneer who hates routine. Constant movement and intuitive decision making are his constant companions. Professions: entrepreneur, real estate agent, producer.

ESTP: extrovert, sensoric, logic, irrational Such a person achieves victory at all costs. Obstacles only strengthen his desire to achieve a goal. He strives for

leadership positions and does not know how to obey. It usually develops a specific action plan and clearly follows it. Professions: officers, construction contractors.

ENFP: extrovert, intuition, ethics, irrational Creative person and dreamer. The combination of qualities allows him to cooperate effectively with other people, to be open and sociable. And also participate in various events, resolve issues, and be flexible. Professions: department head, tour guide.

ESFP: extrovert, sensory, ethics, irrational Identifies the weaknesses of people, which helps him to manipulate and guide. He is guided by his own interests in communicating with people and "lives in the present." Such a person often does not finish what he has begun, seeks immediate results. But at the same time focuses on harmonious relationships with others. Professions: holiday organizer, administrator, politician.

ISTP: introvert, sensoric, logician, irrational He cognizes the world through sensations. By nature, he is empathic, but he is mostly focused on himself. His ability to objectively make decisions and analyze speaks of a technical mindset. It always fits into the deadlines, but sometimes it can act unpredictably. Professions: operational analyst, mechanic, policeman.

INTP: introvert, intuition, logician, irrational This employee is a real scholar who has a philosophical mindset. He analyzes his decisions, strives for objectivity and impartiality. The violent manifestations of emotions are not about him. On the other hand, it feels stress due to a large amount of data being analyzed and its variability. Professions: software engineer, mathematician, analyst.

INFP: introvert, intuition, ethics, irrational This person is a sensitive lyricist who is well versed in people and disposes them to himself. He has a good sense of humor and attaches great importance to appearance. Strives for self-knowledge, to live in harmony with himself, and to benefit others. Professions: artist, creator, psychologist.

ISFP: Introvert, Sensory, Ethics, Irrational Able to enjoy life in the conditions of uniformity and routine. It interacts well with people, avoids conflicts. He loves to feel his worth and help. Such a person does not seek to lead others or remake them, he respects their space and demands the same for himself. By nature, he is a mundane practitioner who you can rely on. Professions: artisan, consultant, jeweler.

1.3 Thesis Outline

To summarize, the contribution of this study is: Machine Learning technique to predict MBTI type of a person; better accuracy than previous studies; prediction performing in real-time. Scientific novelty of current study is the best tool for human resource managers to filter out candidates before the interview only by analyzing the given text. Also, it is very effective tool to gather the productive team. There is a lot of use-cases where it can be helpful. Using this model make it possible to predict psycho-type of real persons, i.e. to gather effective team to achieve better communication and interaction between members. One of the contributions to mention here is prediction in real-time. The rest of the paper is structured as follows. The next chapter briefly reviews related works and accomplishments. The chapter 3 describes tools for development, dataset, data analysis for dataset, pre-processing the dataset and prediction. Chapter 4 concludes the paper and outlines. hole process except inputting data from Speech Recognition System implemented in this study. This step will be developed in future works.

SDU will stand for Suleyman Demirel University, Kaskelen.

2. Literature review

Since ancient times, attempts have been made to divide people into various psychological types. Knowledge of the psychological type of a person was used to determine his character and tendency to various diseases. The first attempt to classify psychophysiological types goes back thousands of years and is reflected in the treatises of the ancient Indian medical system of Ayurveda. According to Ayurveda, in nature there are three gunas - light, passion and darkness, various combinations of which provide mental heterogeneity of people.

In ancient China, 2200 years before the birth of Christ, a system for selecting officials was created that encompassed a very wide range of "manifestations of personality" - from writing skills to behavior in everyday life [8].

The idea of dividing people into 4 psychological types goes back to the 5th century BC, to the teachings of Hippocrates that the human body consists of four elements: air, water, fire and earth. It was assumed that the combination of these elements forms four substances in the human body: blood - a combination of fire and air, mucus - a combination of water and air, black bile - a combination of water and earth, yellow bile - a combination of fire and earth. The Greek philosopher Claudius Galen (130-200 BC), based on the teachings of Hippocrates, proposed to distinguish four main psychological types: sanguine, phlegmatic, choleric and melancholic. It was believed that sanguine predominates in blood (lat. "Sanguis"), which is associated with joy; phlegmatic mucus predominates (lat. "phlegm"), which is associated with calm; in a melancholic black bile predominates (lat. "melancholy"), which is associated with longing; in choleric yellow bile predominates (lat. "chole"), which is associated with anger.

This division of people into psychological types has stood the test of two thousand years. Many centuries passed before the psycho-types found physiological justification. I.P. Pavlov identified four types of higher nervous activity in terms of

strength, balance and mobility of excitatory and inhibitory processes and showed the dependence of four psychotypes on the type of central nervous system. By I.P. Pavlov, sanguine - this is a strong balanced mobile type; choleric - a strong unbalanced (excitable) mobile type; phlegmatic - a strong balanced inert type; melancholic is a weak unbalanced type, nervous processes are inactive.

It is known that each psychological type is more prone to certain diseases. For example, a melancholic prone to cardiovascular disease, vegetative-vascular dystonia, depression. Choleric prone to dysfunction of the gastrointestinal tract, genitourinary system, spasms, cramps.

In the XX century, the study of psycho-types received a new development on the basis of psychoanalysis. American psychologist James McKeen Cattell (1860 - 1944) [8] became the initiator of the testological movement in psychology. For the first time, he used the complex method of modern statistics in a psychological study of personality - factor analysis, which minimizes a lot of different indicators and personality assessments and allows you to identify 16 basic personality traits (Cattel's 16-factor personality questionnaire). The Cattel questionnaire identifies such basic personality traits as rationality, secrecy, emotional stability, dominance, seriousness (frivolity), conscientiousness, caution, sensitivity, credulity (suspicion), conservatism, conformity, controllability, tension. Cattel's questionnaire contains more than 100 questions, the answers to which (affirmative or negative) are grouped according to the "key" - a certain way of processing the results, as a result of which the severity of one or another factor is determined.

Deeper studies of psychotypes based on psychoanalysis were made by Jung [2], [8]. According to the typology of C.G. Jung, people are divided into 8 psychotypes. Four of them are basic: mental, sensitive, sensory and intuitive. Each of them can manifest itself through extraversion and introversion. According to C.G. Jung, an extroverted type of personality is distinguished by a certain alienation of the subject from himself, a reduced significance for him of the subjective world. This type of personality is characterized by increased social adaptation, an active impact on the social environment, an adequate reflection of their social connections, and an optimal level of claims. The introverted personality type is characterized by the predominant orientation of the personality consciousness to the inner world, giving it the highest value, increased introspection, and insufficient adaptation to the social environment.

The Myers-Briggs Type Indicator (MBTI) is one of the most generally utilized measures of personality identity. It uses constrained decision inquiries to distinguish respondents' inclinations on four frame of mind and capacity sets dependent on Jungian brain science (i.e., Extraversion-Introversion, Sensing-Intuition, Thinking-Feeling, and Judging-Perceiving). These four singular inclinations are consolidated to shape respondents' four-letter identity types. The MBTI identity types are called psycho-types, which demonstrate that people with similar types have similar approach toward interacting with others and the world [6]. In the MBTI show, all individuals have mental sorts, what's more, there are no sorts other than the 16 distinguished mixes. Dr. Carol Ritberger classifies human personality according to different colors in her work [7]. Written text can provide some information about author's personality traits as shown as in previous studies [1]. There is growing interest in prediction of personality types using new technologies. From another hand traditional psycho-tests is limited. In study of Komisin and Guinn [9] used Naïve Bayes and SVM technique to predict MBTI type based on their word choice. They used database of samples taken from students along with their MBTI type. They compared two techniques and discovered that Naïve Bayes showed better accuracy on their small dataset than SVM. Classification performed by division to emotional, social, cognitive, and psychological dimensions by analysis software, Linguistic Inquiry and Word Count (LIWC). Golbeck [10] in his study, tried to predict personality type by presented information in Twitter. They monitored user's posts and interactions to get more information about user preferences. They showed that is will be useful to predict job satisfaction, professional and private live success. They tried to predict personality type by publicly available information only.

Two years later, Wan [11] used a machine learning method to predict the massive Five personality sort of users through their texts in Weibo, a Chinese social network, and they were able to successfully predict the personality sort of the users. Li, Wan and Wang [12] used the grey prediction model, the multivariate analysis model and also the multi-tasking model to predict the user personality type supported the massive Five model and their text samples. They compared the performance of those three models and located that the grey prediction model performs better than the two other models. In another study, Tandera [13] used the massive Five personality model and a few deep learning architecture to predict

a person's personality supported the user's information on their Facebook. They compared the performance of their method with other previous studies that used classical machine learning methods and also the results showed that their model successfully outperformed the accuracy of previous similar studies. Furthermore, in another study, Hernandez and Knight [14] used various varieties of recurrent neural networks (RNNs) like simple RNN, gated recurrent unit (GRU) which is gating mechanism in recurrent neural networks, long memory (LSTM) which is a man-made recurrent neural specification utilized in the sector of deep learning, and Bidirectional LSTM to make a classifier capable of predicting people's MBTI personality type supported text samples from their social media posts. Recent research by Cui and Qi [15] used Baseline, Logistic Regression, Naïve Bayes, and SVM to predict an individual's MBTI personality type from one amongst their social media posts. They compared the results of of these methods and located that SVM performed better.

3. Methodology for Prediction

3.1 Tools for Development

In this research as a main machine learning algorithm used Extreme Gradient boosting which was explained chapter above. For the development process used XGBoost library in Python. Also utilized other libraries like Pandas, numpy, seaborn and etc. Model was trained on laptop with 8GB RAM and Intel Core i5-6300 CPU.

3.2 Extreme Gradient Boosting

XGBoost[16] is one of the most popular and effective implementations of the gradient boosting algorithm on trees for 2019.

XGBoost is based on the gradient boosting algorithm of decision trees. Gradient boosting is a machine learning technique for classification and regression that builds a prediction model in the form of an ensemble of weak predictive models, usually decision trees. The ensemble is trained sequentially in contrast to, for example, bagging. At each iteration, deviations of the predictions of the already trained ensemble in the training set are calculated. The next model to be added to the ensemble will predict these deviations. Thus, by adding the predictions of the new tree to the predictions of the trained ensemble, we can reduce the average deviation of the model, which is the target of the optimization problem. New trees are added to the ensemble until the error decreases, or until one of the "early stop" rules is fulfilled [17].

XGBoost supports all the features of libraries like scikit-learn with the ability to add regularization. Three main forms of gradient boosting are supported:

- Standard gradient boosting with the ability to change the learning rate.

- Stochastic gradient boosting [18] with the ability to sample on rows and columns of the dataset.

- Regularized gradient boosting [19] with L1 and L2 regularization.

The standard boosting method, like the random forest method, is to combine many classifiers, which do not have sufficient quality separately, into a common classifier showing a good classification result. In contrast to the random forest method, individual classifiers are constructed not independently from each other, but sequentially, iteratively. The idea of the gradient boosting method is to iteratively improve the current classifier F_k. At the first step, you can take any classifier for F_1, for example, constant. The constant classifier has the form $\arg\min_{c \in C} L(y_i, c)$, where y is the desired class vector, L is the loss function, C is the set of classes in the classification problem.

$$\arg\min_{h \in \mathcal{H}} \sum_{i=1}^{n} L\left(y_i, F_k\left(x_i\right) + h\left(x_i\right)\right) \tag{3.1}$$

Then the base classifier is trained on the original sample with new r_{im} marks. After that, the coefficient γ_m is calculated as a solution to the following optimization problem:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i\right) + \gamma h_m\left(x_i\right)\right) \tag{3.2}$$

The model is updated using the newly built base classifier:
$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$

Note that at the moment there are no generally accepted guidelines regarding the choice of the loss function. In the literature, there are a large number of loss functions, the use of which generates an algorithm, slightly different from the rest. Often, each new feature is accompanied by a publication. Examples of categories of loss functions are scalar, vector, structural loss functions. As a loss function of the gradient boosting method, any differentiable loss function. Considering decision trees as basic classifiers, $h_m(x)$, that is, the classifier trained in step m, takes the form $h_m(x) = \sum_{j=1}^{J_m} b_{jm} I\left(x \in R_{jm}\right)$, where b_{jm} is the value predicted by the tree in the R_{jm} region.

3.3 Model Training and Quality Controlling in Machine Learning

An important component of the correct application of machine learning methods is the quality control of models and the observance of some fundamental principles in their training. It is necessary to competently control the breakdown of the available data into the training and test subsets, select quality assessment metrics relevant to the task, and make sure that the trained models are well generalized to the new data. Also, any problem to be solved brings its own specifics to the training of models; therefore, it is necessary to carefully evaluate the full flow of research stages from training models to quality assessment and commissioning. Let us dwell on some concepts and principles important for research. Note that all subsections of this chapter are not independent techniques for improving the quality of trained models, but mechanisms that are used together and must work seamlessly within the same pipeline.

3.3.1 Cross-Validation

In any task where machine learning models are used, the task of evaluating this model arises at almost every stage. The naive approach is to calculate the overall assessment relative to the initial data used in the training, using some metric of the quality of the assessment. The disadvantage of this approach is that it does not give any idea of how well a trained model will give predictions for data not used in model training, that is, how well the model generalizes training data. The main method of overcoming this drawback is a better approach to model assessment - the cross-validation method. With this approach, for training the model, not all the available data sample is used, but only a part of it. Thus, before starting the model training, a subset of the data is removed from the source data. Then, after the model is trained, the data that was originally deleted can be used as data for testing the model. This data is for the model. are "new" because they were not used in training, which means that through such testing one can assess the ability of the model to predict the result on arbitrary data. Cross-validation methods are divided into several categories:

- The data separation method is the simplest method, according to which all

the initial data is divided once into a training and test sets, after which the model is trained on a training set. To evaluate the model, its predictions on the sample elements are considered, fell into the test set, and therefore not present in the training set, that is, previously not encountered by the classifier. The predictions of the classifier are compared with the reference ones and some metric is calculated relative to the predicted and reference predictions characterizing the accuracy of the classifier. The advantage of this method is that it requires no more time to execute than the method without cross-validation, and at the same time gives a more accurate estimate of the effectiveness of model predictions. Nevertheless, the result of the work of this method depends on the partitioning of the sample into the training and test ones and may have a large dispersion, that is, it may produce different results for different partitionings of the sample.

- The K-fold method is one way to improve the data separation method. The data in this approach is divided into k parts, after which the data separation method is applied k times. At each iteration, the next of k subsets of data is used as a test set, and the remaining k - 1 subsets are merged together into one set and used as a training set. Thus, k quality results of the model are obtained, which are then averaged to obtain a single quality assessment. The advantage of this method is that it is no longer so important how the source data will be divided into subsets. Each object from the data set appears exactly once in the test set, and k - 1 time in the calibration set. The variance of the resulting estimate decreases with increasing k. The disadvantage of this method is that the model must be trained from the very beginning, that is, the whole process takes k times more time. One of the options for this method is to randomly divide the data into training and test sets k times. The advantage of the latter is that you can independently choose how large the current test suite is and how many rounds of training to complete. When training models in training samples with unbalanced classes, they often use the approach of maintaining the relative distribution of class shares in each of the k groups into which the training set is divided.

- The testing method at one object is an extreme case of the K-fold method with the number k equal to the number of elements of the available training

set N. This means that the model is trained N times except for one object, and a prediction is made for this object. As before, the average result is calculated taking into account the selected metric. The disadvantages of this method include a long time learning models. An important advantage is that due to the large number of experiments, the variance of the averaged result of the quality assessment of the selected model decreases and allows you to have a more accurate quality assessment. Also, when training by this method, the model is trained on the N - 1 sample element, which is the best approximation to the full sample, on which the final model will be trained before it is put into operation. When using techniques for choosing hyperparameters of machine learning algorithms, this gives an improvement in the quality of the choice of hyperparameters, since many of them can depend on the number of elements in the training set.

An important point in training models using the cross-validation method is the selection of features and other stages that require analysis of the entire sample, only on the training data set of iteration of cross-validation, since if these steps are performed before the cross-validation begins, there is a risk of biasing the cross -validations, since even before evaluating the model on the test sample, the model was trained with some implicit knowledge about the test sample.

3.3.2 Regularization

Regularization in machine learning in the broad sense refers to the technique Regularization in machine learning in the broad sense refers to the technique of avoiding retraining models, based on the observation that when retraining in particular linear models, the weights of attributes become unreasonably large. The idea of the technique is to include information on the absolute value of the weighting coefficients in the loss function of the model. For simplicity, we consider the loss function equal to:

$$L(X, y) = \sum_i |y_i - f(X_i)|^2 \tag{3.3}$$

where X is a matrix of attributes, y is a vector of true values, f is a matched function representing a linear model. When applying regularization to the loss function, an additional term is added equal to the norm of the function f with a

certain coefficient, thus, the loss function turns into:

$$L(X, y) = \sum_i |y_i - f(X_i)|^2 + \lambda \|f\| \tag{3.4}$$

Moreover, depending on the type of norm used in regularization, the result of training the model depends. The coefficient λ is called the regularization coefficient. The larger this constant, the more large modulo coefficients will be fined, which means that the model will have lower variability, but higher bias. The choice of this coefficient depends on the task and therefore there are no generally accepted formulas or practices for choosing this coefficient, except for enumerating several values of different order. We also note that often the sum of several norms with different coefficients is used as the regularization term. Consider the most commonly used regularization methods:

- The *Ridge* method is to use the $\mathbf{L}_2$ norm. If there are signs that have a high level of correlation among themselves, this method seeks to align the weights of these two signs. This method penalizes large modulo weights strictly, since in the corresponding norm the weight is taken to the second degree. On the contrary, small modulo coefficients are penalized loosely.

- The *Lasso* method is to use the $\mathbf{L}_1$ norm. It is believed that with this method, features are selected in a sparse manner, that is, the number of nonzero weights will be less than that of the Ridge method. In the presence of correlated features, in the general case, only one of them will have nonzero weight in the final model. This method penalizes large modulo and small modulo weights more uniformly than Ridge.

- The *Elastic net* method is a compromise between the previous two and represents the use of a linear combination of $\mathbf{L}_1$ and $\mathbf{L}_2$ norms with some independent coefficients. This method simultaneously seeks to equalize the weights and select the correlated features in a sparse manner at the same time.

To mathematically substantiate the difference between trained weights without and with regularization, we use a simple case of the loss function and the formula

of the most widely used method of training a linear model - gradient descent:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x_j^{(i)} \tag{3.5}$$

where θ is the approximation of the calculated weight vector, α is the parameter responsible for the rate of convergence of the gradient descent, h_θ is the model with weights θ, m is the number of objects in the training set, j is the number of the calculated weight corresponding to the attribute with the same number. When regularization is added to the loss function, for example, regularization by the Ridge method, the gradient descent formula will take the form:

$$\theta_j := \theta_j \left(1 - \alpha \frac{\lambda}{m} \right) - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x_j^{(i)} \tag{3.6}$$

The difference from the fact that at each iteration, the initial value of θ is taken slightly less in norm than the value calculated at the previous iteration, so the method tries to converge to the smallest possible value of θ. The above calculations are valid for linear teaching methods. Regularization of nonlinear methods is carried out using other techniques. Nevertheless, it is worth noting right away that, with correctly selected parameters, many non-linear methods do not need to be regularized as much as linear methods, since often they have the idea of averaging a large number of predictions with a large bias, as happens in a random forest. As a regularization of the random forest model, the depth of the decision trees under construction is usually limited; ceteris paribus, the preference is for a partition according to the features by which the partition was already performed in other trees of the forest; pruning trees. Also, to increase the independence of individual classifiers, and therefore, for a smaller bias on average, for training individual trees, a subsample of the training sample is often taken (typical volumes are 70-90%), and when deciding to split a set in a tree node, a subset of features is considered (typical sizes of a subset of features are again 70-90%, however, others may also occur).

When learning the gradient boosting model, the regularization parameter is the learning coefficient, which usually takes a value from 0 to 1 and indicates a decrease in the contribution of each new base predictive model added to the general gradient boosting model at the next iteration. When learning the gradi-

ent boosting model to avoid overfitting, they usually either reduce the number of iterations of training or decrease the learning coefficient. It makes sense to adjust these parameters at the same time, while the larger the value of one parameter, the smaller the value of the other parameter to obtain good results of model predictions. As in the case of a random forest, use the technique of selecting subsets of objects for learning the next tree and subsets of features to split the set into two in the tree node, if the decision tree is used as base predictor. In models of the support vector method, the regularization parameter is the parameter C, which is responsible for the compromise between the width of the indent (corridor), which must be maximized and the number of objects in the original sample that did not fall in the right direction from the corresponding corridor and their degree remoteness from the corridor.

3.3.3 Hyperparameter tuning

For many machine learning algorithms, it is typical to have 5-10 or more hyperparameters, that is, parameters that are set at the model architecture level and not optimized during training. Such hyperparameters are the regularization coefficients in the loss function, the number of basic classifiers in complex models, the depth of decision trees, the size of the training subsample for constructing basic classifiers, and many others. These hyperparameters can significantly affect the characteristics of the constructed model and its ability to generalize. Despite the fact that many implementations of machine learning algorithms have reasonable default hyperparameter settings, the selection of hyperparameters is an important component of the study of the use of machine learning models, since for different types and distributions of the data under study, different hyperparameters can give models significantly different in their characteristics. The optimization of hyperparameters aims to find such parameters of the algorithms for which the model trained in the training set gives the best result in an independent test set. Cross-validation techniques are often used to optimize hyperparameters. At the same time, it is important that the initial data set is divided into three sets each time: first, before training the model and selection of hyperparameters, into training and test ones, so that the test set does not explicitly or implicitly participate in the construction of the model. Further, the training set, usually during cross-validation, is divided into the actual training and validation sets, where the

training set serves directly for optimizing the parameters by the machine learning algorithm, and the validation set corresponds to one set of hyperparameters and serves to assess the quality of the algorithm with these hyperparameters. Hyperparameter optimization approaches:

- Grid search is a traditional approach consisting in enumerating the hyperparameters of interest from some given set by means of exhaustive search. Thus, the hyperparameter vectors from the Cartesian product are sorted over the sets of values of interest. Since many parameters come from continuous space, they consider the discretization of space with some step, hence the name of the method. However, there are no restrictions on the regularity of the step. Since the set of interesting values of hyperparameters is often large, and the Cartesian product of individual sets as a result gives a sufficiently large space, this method is quite laborious. Nevertheless, the search can be performed using the parallel method, since the results of each set of hyperparameter values are independent.

- Random search is similar to the previous approach, however, in this case, the parameters are selected randomly from a given space, and a special distribution can be specified according to which hyperparameters will be selected. The method can show itself better than a grid search in cases where a small number of hyperparameters affects the result for the data being studied, or when for some reason the hyperparameters have a hidden relationship between themselves. This method can also be run in parallel. However, it lacks some systematic character.

- Gradient optimization first calculates the quality value of the trained model at some points corresponding to certain values of hyperparameters, and then determines the gradient of the quality function from the values of hyperparameters and calculates the next potentially interesting value of the hyperparameters vector for which a quality assessment is performed. This method works best for continuous spaces of individual hyperparameters, but it also holds for discrete sets of values.

All the described approaches have their advantages and disadvantages and all of them are used in practice. There are other, more complex approaches that require even more significant processing power.

3.3.4 Learning Curves

The learning curve is a graph whose abscissa axis corresponds to the number of copies in the sample under consideration, and the ordinate axis to the quality of training in the corresponding test sample. For the first time, learning curves were presented in works on educational and cognitive psychology [20]. Back in 1936, the effect was described and a mathematical model of the learning curve associated with labor productivity in the aviation industry was proposed. Over time, this term began to be used in many areas, including the field of mathematical statistics and machine learning. In terms of general machine learning, the learning curves show the model's ability to generalize relative to the size of the training data. Often, different training models are compared on a sample of raw data of a fixed size. Such a comparative analysis has an important drawback, which follows, in particular, from the fact that for different models on the same data set, the learning curves of these models can intersect, including repeatedly. This drawback is that studies on a fixed sample size do not provide a sufficient idea of the comparative characteristics of methods in larger samples or smaller [21]. Thus, a more preferable method of comparing learning models is generally the method of comparing learning curves. Considering the learning curve of a separately selected model, one can make an assumption about whether the quality of the prediction of the model will improve when adding new data to the training set, as well as whether the current model suffers from large variability or from a large bias. Typically, as part of the study of one model, two curves are considered - the result of the model on the training sample and on the test sample. If both curves with an increase in the size of the training set converge to the same value with not very high quality, this indicates that adding new data to the training set will almost certainly not improve the quality of the model classification. To improve the quality of model predictions in this case, a different classifier, or another parameterization of the classifier in question, is usually required, which could learn more complex concepts of knowledge about the source data, i.e. would have a lower bias. If the result on the training set is much higher than on the test set for the maximum number of objects in the sample, then this indicates that adding new objects to the training set is highly likely to improve the model's ability to summarize data, that is, improve the quality of classification.

3.3.5 Accuracy-completeness curve, ROC curve

For any result of data classification, in the general case, four important characteristics are considered, presented in absolute terms:

- the number of true-positive predictions - the number of test sample objects from the positive class for which the classifier signaled that the object belongs to the positive class
- number of false positive predictions - the number of test sample objects from the negative class for which the classifier signaled that the object belongs to the positive class
- the number of true negative predictions - the number of test sample objects from the negative class for which the classifier signaled that the object belongs to the negative class
- number of false-negative predictions - the number of test sample objects from the positive class for which the classifier signaled that the object belongs to the negative class

In areas related to information processing, accuracy is a measure of the relevance of a result, while completeness is a measure of how many really relevant instances are obtained as a result. The accuracy-completeness curve is used to evaluate the quality of the classifier. The high area under the curve simultaneously corresponds to a high value of accuracy and completeness, where high accuracy is associated with a low number of false-positive predictions, and high completeness is associated with a low number of false-negative predictions. Large values of both quantities indicate that the classifier gives accurate predictions and returns most of the positive results. A system with high completeness but low accuracy classifies many instances as positive, but a large number of predictions are incorrect compared to current set labels. A system with high accuracy but low completeness, on the contrary, classifies a small number of results as positive, but most predictions are correct. In an ideal system, all returned positive labels are valid and cover many objects that are truly positive labels. More formally, accuracy (P) is defined as a fraction, where the number of true positive predic-

tions (T_p) is in the numerator and the sum of true positive and false positive (F_p) predictions is in the denominator:

$$P = \frac{T_p}{T_p + F_p} \tag{3.7}$$

Completeness (R) is defined as the number of true positive predictions divided by the sum of true positive and false negative predictions (F_n):

$$R = \frac{T_p}{T_p + F_n} \tag{3.8}$$

Formulas 3.7 and 3.8 are associated with the F1-measure, calculated as follows:

$$F1 = 2\frac{PxR}{P + R} \tag{3.9}$$

The accuracy-completeness curve is a set of accuracy and completeness values for different set thresholds of the classifier, that is, confidence values at which the classifier signals that the object belongs to a positive class. It is important to note that accuracy cannot be reduced along with completeness. It follows from formula 3.7 that a decrease in the classifier threshold can increase the denominator due to an increase in the number of positive predictions.

If the previous threshold value was too high, all new results after lowering the threshold may turn out to be truly positive, which will increase accuracy. If the previous threshold value was close to correct, or too low, then a further decrease in the threshold value will give false-positive predictions, reducing accuracy.

Completeness, in accordance with formula 3.8, is a fraction whose denominator does not depend on the threshold value of the classifier, which means that lowering the threshold value of the classifier can increase the completeness by due to the increase in the number of true positive predictions, but may leave the completeness unchanged, while the accuracy will change.

Accuracy-completeness curves are widely used in the diagnosis of predicted values of binary classification models, but it is worth noting that these methods can also be used in the diagnosis of multiclass classification methods.

The ROC curve is another method for studying binary classification models regarding the variation of the decision threshold by the classifier. In the ROC curve, the concept of the share of true positive predictions regarding the total number of instances belonging to the positive class is used, this concept is denoted

by TPR, as well as the concept of the share of false positive predictions regarding the total number of instances belonging to the negative class, this concept is designated as FPR.

The value of TPR is also called the sensitivity of the model. Note that this value coincides with the previously defined concept of completeness. A value of 1 - FPR is called specificity, or the fraction of true negative predictions.

A mane graph is a set of values (FPR; TPR) for various classifier threshold values. It shows the relationship between an increase in sensitivity and a decrease in specificity with a change in the classification threshold.

The area under the ROC curve, in contrast to the accuracy-completeness curve, has a clear interpretation. It can be interpreted as the value of classification consistency, that is, as the probability that for a random instance of a positive class the value of the classifier's confidence in the positivity of the class is greater than the corresponding value for a random instance of a negative class.

3.4 Dataset

The main problem of personality prediction is the lack of labeled dataset. The train dataset was collected through the PersonalityCare forum and uploaded from Kaggle[22], as it provides a large selection of people and their MBTI personality type, as well as what they have written on their posts or comments under the posts. What makes Kaggle particularly suitable because of content is available for public that many users can use it.

The dataset contains over 8600 rows of data on each row is person's:

- type (these persons 4 letter MBTI type)
- section of each of the last 50 things they have posted (each entry separated by 3 pipe characters)

3.5 Data Analysis

Overall process flow diagram showed in Figure 3.1. MBTI labeled dataset requires data analysis before starting to data pre-processing stage. However data analysis is hidden in this diagram.

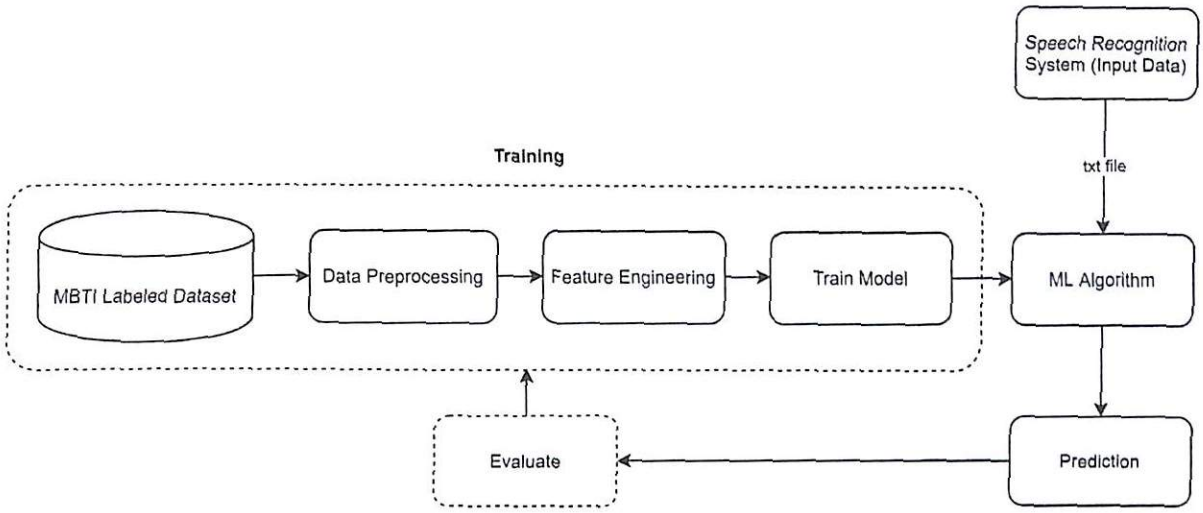


Figure 3.1: Process flow diagram

Count of MBTI types withing the dataset showed in Figure 3.2. In this dataset types with Introvert scale are more prevalent. My assumption it is because of introverts tend to have much time online than extroverts. Also we can say that types in this dataset distributed abnormally and our prediction accuracy will depend on that.

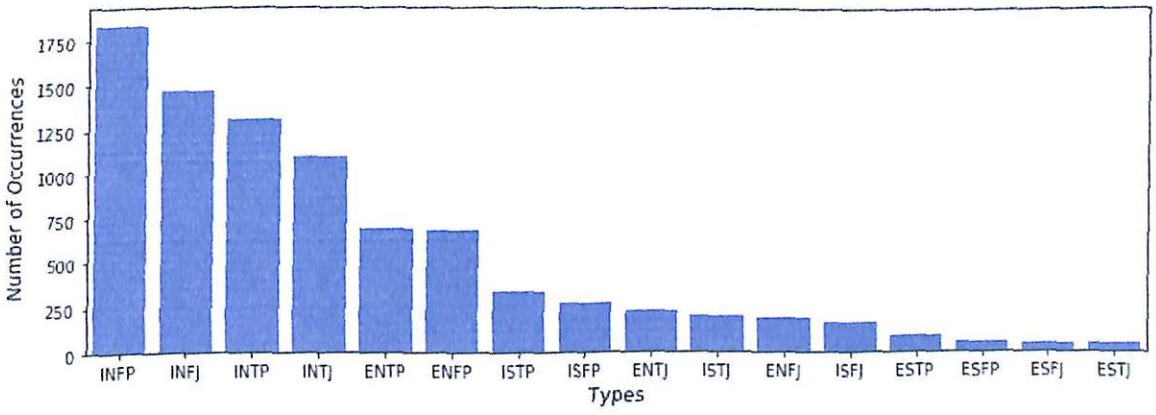


Figure 3.2: MBTI count in dataset

Figure 3.3 show percentage of MBTI scales which was categorized for four dimensions. According to this figure, distribution of Introverts is much greater than Extroverts. The same distribution for the Intuition/Sensing scale where Intuition distribution is much greater than Sensing. Feeling/Thinking scale distributed near to 50/50 but Feeling distribution is slightly greater than Thinking. This is one of the reasons to predict on that scale in previous work. And last scale, Perceiving/Judging, the distribution of Judging is greater than Perceiving.

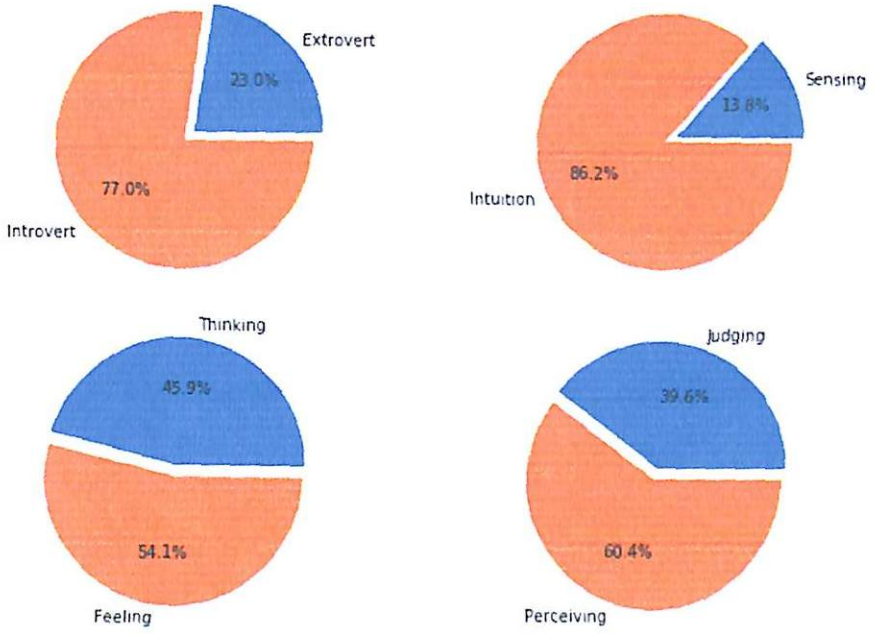


Figure 3.3: MBTI types scatter in dataset

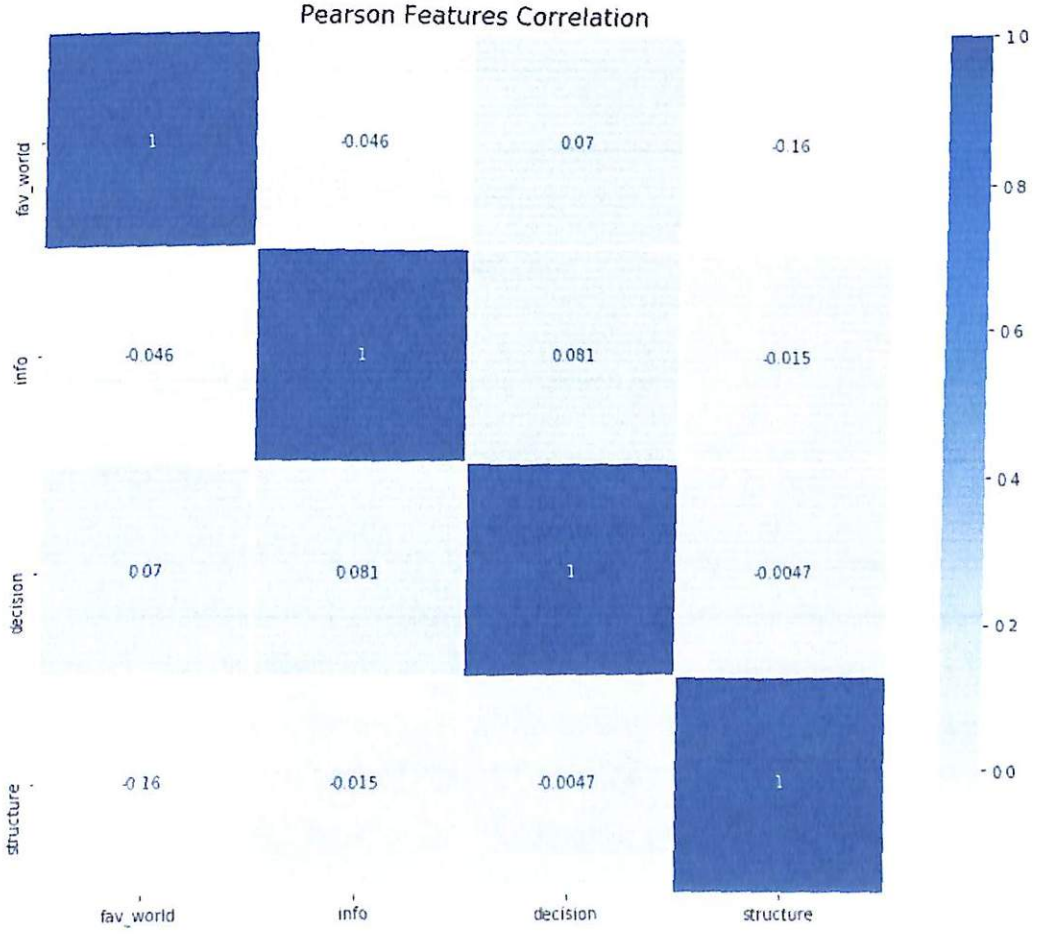


Figure 3.4: Pearson Features Corellation

3.6 Pre-processing the Dataset

As mentioned before dataset was collected from Personality Care forum, so data require some cleaning steps before applying machine learning technique. Because of this forum was about personality types dataset contain a lot of discussions about personality types. It will affect to accuracy of prediction. Thus decided to clean dataset from MBTI types.

Idea is to define which words are more usable by MBTI personalities. Our distinctive feature of dataset is labelled type of personality. We suppose to predict that they have to use the different words in their vocabulary. To get high accuracy in prediction necessary to achieve high precision in both datasets train and test. The data cleaning in steps:

1. Next used regex-based pattern to:
   (a) remove links;
   (b) strip punctuation's;
   (c) remove non-words;
   (d) remove multiple letter repeating words;
   (e) transform all words to lower case.

3.6.1 Text Vectorization

The Bag of Words

The Bag of Words Representation could be a general process of converting a text document into numerical features. the method of converting documents into vectors is named text vectorization. By representing every word as a integer and counting the frequency of appearence, a document is marked as a vector.For most of the documents, only a subset of the dictionary might be used, thus the matrix of text vectorization would mostly be like sparse matrix.

CountVectorizer

CountVectorizer converts the input text into a matrix whose values are the number of occurrences of this key (word) in the text. Unlike FeatureHasher, it has more

configurable parameters (for example, you can set the tokenizer), but it works more slowly.

Tf-idf

In some documents, there are certain words which are quite frequently appearing thus they rarely have useful information on the particular content of the document. If we put of these words into classifier, these frequently appearing words may mask the words rarely-presented yet more meaningful. So on recalculate the weights of every word, usually, tf-idf transformation would be processed.The measure TF-IDF is a product of two factors:

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \tag{3.10}$$

3.7 Prediction

The classification will be multi-labeled because of persons might have dominant trait or several distributed traits. Multi-labeled problem represents a different classification task, but the tasks are somehow related. That means a person can have features of several personalities in the same time. Multi-labeled classification is trying to predict the label sets through analyzing training instances with known label sets. In our research we have predefined 16 classes and those classes was considered as for binary classification tasks. Each binary classes shows scale of personality and trained individually as per their relationship to relevant scale. Training set and test set was divided by proportion of 80% to 20%. Early stopping was monitored when model was fit by training set and evaluated by test set. Initially deep of trees was configured in range of 2 to 8. After some configuration parameters of XGBoost was setup as follow:

- n_estimators = 100
- max_depth = 6
- learning_rate = 0.3

4. Results and Discussions

After creating XGBoost model MBTI scales were trained separately, data was divided to train and test sets. Then the model was fit by training set and prediction performed on the test set. Predictions were evaluated. Tree_depth was configured and predictions evaluated again.

Binary Class	Accuracy after Configuration	Accuracy before Configuration
IE 0.84	79.01%	78.17%
NS 0.1	85.96%	86.06%
FT 2.41	74.19%	71.78%
JP 0.28	65.42%	65.70%

Other existing strategies which utilized the equivalent dataset were talked about in Literature review. Accordingly, the precision of expectation after arrangement was contrasted with the most recent and best existing strategy. This technique was presented by Hernandez and Knight [27] in 2017. The equivalent dataset was utilized in their examination and the pre-handling step was actually equivalent to this exploration. Henceforth, the correlation between their strategy and the introduced technique in this exploration was on similar information. They utilized different kinds of repetitive neural systems (RNNs, for instance, basic RNN, GRU, LSTM, and Bidirectional LSTM to assemble their classifier. For assessment, they utilized two distinct strategies, which were (1) a post-grouping approach and (2) a client arrangement technique. For post-grouping, they prepared the test set and anticipated the class for each individual post. They at that point delivered a precision score and disarray lattice for each MBTI measurement. Then again, so as to group clients, they expected to discover a method of tuning the class forecasts of individual posts all wrote by a person into an

expectation for the class of the creator. Accordingly, they took the mean of the class likelihood expectations for the entirety of the posts in a client's corpus and adjusted either to 0 or 1. The exactness of the intermittent neural system model utilizing client the grouping strategy was better than the repetitive neural system model utilizing the post-arrangement approach. In this manner, the precision of Extreme Gradient Boosting was contrasted with their repetitive neural system classifier utilizing the client characterization approach. Actually, a similar proce-dure for assessing the outcomes was utilized in this exploration, as we needed to look at the performance of the models similarly.

5. Conclusion

This study has developed prediction model with better accuracy to identify MBTI psychological types in comparison with the previous studies that shown in results section. Especially it is related to Intuition-Sensing and Introversion-Sensing personality scales. This model was implemented using XGBoost library which represents Extreme Gradient Boosting algorithm. Scientific novelty of this study is that it can be useful tool for human recourse managers and teachers to get a better understanding of their prospects. Study requires to fulfill dataset with normal distributed MBTI types.

References

[1] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text", *J. Artif. Int. Res.*, vol. 30, no. 1, pp. 457–500, Nov. 2007, ISSN: 1076-9757.

[2] C. Jung, "Psychological types", 1921.

[3] H. Gardner, "Multiple intelligence", 1993.

[4] M. L. Ian Alastair Scott, "The lüscher colour test", 1971.

[5] (). Mbti wikipedia, [Online]. Available: `https://en.wikipedia.org/wiki/Myers%5C%E2%5C%80%5C%93Briggs_Type_Indicator`. (accessed: 02.04.2020).

[6] I. B. Myers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer, *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press Palo Alto, CA, 1998, vol. 3.

[7] C. Ritberger, *What Color Is Your Personality?: Red, Orange, Yellow, Green...* Hay House, Inc., 2009.

[8] L. Burlachuk, "Dictionary reference for psychodiagnostics", 2001.

[9] M. Komisin and C. Guinn, "Identifying personality types using document classification methods", *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25*, pp. 232–237, Jan. 2012.

[10] J. Golbeck, M. Edmondson, and K. Turner, "Predicting personality from twitter", pp. 149–156, Oct. 2011. DOI: 10.1109/PASSAT/SocialCom.2011.33.

[11] D. Wan, C. Zhang, M. Wu, and Z. An, "Personality prediction based on all characters of user social media information", *Communications in Computer and Information Science*, vol. 489, pp. 220–230, Nov. 2014. DOI: 10.1007/978-3-662-45558-6_20.

[12] C. Li, J. Wan, and B. Wang, "Personality prediction of social network users", pp. 84–87, Oct. 2017, ISSN: 2473-3636. DOI: 10.1109/DCABES.2017.25. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/DCABES.2017.25.

[13] "Personality prediction system from facebook users", *Procedia Computer Science*, vol. 116, pp. 604–611, 2017, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.10.016.

[14] K. I. Hernandez R., "Predicting myers-bridge type indicator with text classification", *In Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017. [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6839354.pdf.

[15] B. Cui and C. Qi, "Survey analysis of machine learning methods for natural language processing for mbti personality type prediction", 2017.

[16] (). Scalable, portable and distributed gradient boosting, [Online]. Available: https://github.com/dmlc/xgboost. (accessed: 02.04.2020).

[17] (). Xgboost wiki, [Online]. Available: https://neerc.ifmo.ru/wiki/index.php?title=XGBoost#cite_note-1. (accessed: 02.04.2020).

[18] (). Stochastic xgboost, [Online]. Available: https://neerc.ifmo.ru/wiki/index.php?title=XGBoost#cite_note-8.

[19] (). Regularized xgboost, [Online]. Available: https://neerc.ifmo.ru/wiki/index.php?title=XGBoost#cite_note-8.

[20] C. Perlich, "Learning curves in machine learning", Jan. 2011. DOI: 10.1007/978-0-387-30164-8_452.

[21] P. L. D. Kibler, *Machine Learning as an Experimental Science*. Proceedings of the Third European Working Session on Learning., 1988, pp. 81–92.

[22] M. J. (). (mbti) myers-briggs personality type dataset, [Online]. Available: https://www.kaggle.com/datasnaek/mbti-type.

A. Appendix A

Figure A.1: ENFJ