**A PROJECT REPORT**

**ON**

**PERSONALITY PREDICTION USING MACHINE LEARNING**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE IN THE

PARTIAL FULFILMENT FOR THE AWARD OF THE DEGREE

**OF**

**BACHELOR OF ENGINEERING IN**

**INFORMATION TECHNOLOGY**

BY

HRITIK KOUL B150078522

RAVIRAJ KOKARE B150078537

ISHWARI PAWAR B150078507

NIKHIL CHAUDHARI B150078558

**UNDER THE GUIDANCE OF**

**PROF.M.R.MAHAJAN**

****

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**PUNE VIDYARTHI GRIHA’S COLLEGE OF ENGINEERING AND TECHNOLOGY & G. K. PATE (WANI) INSTITUTE OF MANAGEMENT**  **44, VIDYA NAGARI, SHIVDARSHAN, PARVATI, PUNE – 411009 MAHARASHTRA – INDIA 2021-2022**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**PUNE VIDYARTHI GRIHA’S**

**COLLEGE OF ENGINEERING AND TECHNOLOGY, PUNE-09.**

**CERTIFICATE**

This is to certify that the project report entitled

**PERSONALITY PREDICTION USING MACHINE LEARNING**

Submitted by

HRITIK KOUL B150078522

RAVIRAJ KOKARE B150078537

ISHWARI PAWAR B150078507

NIKHIL CHAUDHARI B150078588

is a bonafide work carried out by them under the supervision of Prof. M. R. Mahajan and it is approved for the partial fulfilment of the requirement of Savitribai Phule Pune University for the award of the Degree of Bachelor of Engineering (Information Technology) in the academic year 2021-22

Date:

Place: Pune

**Prof. M. R. MAHAJAN Dr. S. A. MAHAJAN**

Seminar Guide HOD Information Technology

This project-based report has been examined by us as per the Savitribai Phule Pune University, Pune, requirements at PVGs COET, Pune

(Name & Signature) (Name & Signature)

**Internal Examiner External Examiner**

**ACKNOWLEDGEMENT**

With immense pleasure, we are presenting this project report as a part of the curriculum of Information Technology. We wish to thank all the people who gave us unending support right from the stage the idea was conceived. We express our profound thanks to our respected head of Department Dr. S. A. Mahajan, whose advice and valuable guidance helped us in making this project successful. We are also grateful to our project guide Prof. M. R. Mahajan, whose valuable guidance helped us in making this project interesting and successful by providing us with immense support and guidance throughout the project.

Hritik Koul Raviraj Kokare Ishwari Pawar Nikhil Chaudhari

**LIST OF FIGURES**

**Figure 1. System Architecture 20**

**Figure 2. DFD Diagrams 21**

**Figure 3. Use Case Diagram 22**

**Figure 4. Class Diagram 23**

**Figure 5. Activity Diagram 24**

**Figure 6. Sequence Diagram 25**

**Figure 7. Sample Code of ANN Model 31-37**

**Figure 8. Confusion Matrix of ANN Model 38**

**Figure 9. Final User Interface 38**

**CONTENTS**

CERTIFICATE II

ACKNOWLEDGEMENT III

LIST OF FIGURES IV

|  |  |  |
| --- | --- | --- |
| CHAPTER | TITLE | PAGE NO |
| 1 | **ABSTRACT** | 6 |
| 2 | **INTRODUCTION** | 7 |
| 2.1 | AIM | 9 |
| 2.2 | MOTIVATION | 9 |
| 2.3 | OBJECTIVES | 10 |
| 3 | **LITERATURE SURVEY** | 11 |
| 3.1 | EXISTING SYSTEM | 11 |
| 3.2 | PROPOSED SYSTEM | 14 |
| 4 | **REQUIREMENT ANALYSIS** | 15 |
| 4.1 | REQUIREMENT SPECIFICATION | 15 |
| 4.1.1 | INTRODUCTION | 16 |
| 4.1.2 | PROBLEM STATEMENT | 16 |
| 4.2 | REQUIREMENTS | 16 |
| 4.2.1 | FUNCTIONAL REQUIREMENTS | 16 |
| 4.2.2 | CONSTRAINTS | 17 |
| 4.2.3 | SOFTWARE AND HARDWARE REQUIREMENTS | 19 |
| 5 | **DESIGN** | 20 |
| 6 | **IMPLEMENTATION** | 26 |
| 7 | **RESULT AND EVALUATION** | 38 |
| 8 | **CONCLUSION AND FUTURE SCOPE** | 39 |
| 9 | **REFERENCES** | 40 |

**1. ABSTRACT**

Personality is an important part of human life. Personality, more specifically, is a subfield of psychology. Personality is made up of factors such as a person's thoughts, feelings, and conduct, all of which change throughout time. It has an impact on how we live, speak, respond, and express ourselves, as well as on our mental health. Personality analysis is a natural human ability that is utilized every day with a wide range of people and for a wide range of objectives.

The modern workplace is made up of a variety of personality types, each with its own approach to communication, engagement, and productivity. When developing plans and solutions, HR must be acutely aware of employee personality. However, this can be difficult for large workforces and newly onboarded teams. As a result, communication becomes ineffective, falling short of anticipated goals and negatively damaging professional interpersonal connections. This is where personality AI can make a significant difference. A personality AI platform generates insights about each employee and recommends the best way to communicate with them. The proposed technique employs Machine Learning and Artificial Intelligence to estimate a person's personality based on CV analysis. For greater accuracy, it employs machine learning algorithms such as Deep Neural Networks, Artificial Neural Networks, and Recurrent Neural Networks. It can be used by recruiters during the job-search process. Some jobs necessitate a certain personality type. This will be beneficial to HR in terms of job selection. Using the proposed method, we may anticipate an individual's personality and categorize it into one of five categories: extroversion, neuroticism, agreeableness, conscientiousness, and openness.

**Chapter 2**

**INTRODUCTION**

Human beings tend to learn from previous experiences. The same thing is done nowadays digitally, and the technology is known as Machine Learning. Machine Learning is a branch of Artificial Intelligence or in short AI, the hot cake of today’s technology. Using Machine Learning we can do a lot of stuff, for instance, predicting Bitcoin prices, movie ticket price prediction, or people you may know the section on Facebook. Even ML Is used in medical science also. Nowadays, researchers are also able to predict personality using Machine Learning.

Personality is defined as the distinctive sets of cognitions, behaviors, and passionate examples that emerge from biological and ecological components. It reflects the people's differences in thoughts, conduct, and emotions. Personality traits are constant in the universe because they provide a sense of high and low explicit features in an individual on a consistent quality rather than exhibiting a specific personality. The difference in personality should be expected to prevent the cause of deterrents or friction in the workplace. Personality identification of a human being by their nature is an old technique. Earlier these were done manually by spending a lot of time predicting the nature of the person. Data mining is primarily used today by companies with a strong consumer focus- retail, financial, communication, and marketing organizations.

The exponential development of new information and communication technologies in recent years, as well as their incorporation into people's daily lives, has dramatically increased the amount of information available at all levels of their social world. Job seekers are continually using networks like LinkedIn as well as job search portals like Glassdoor, Naukri, and others. More company organizations, on the other hand, employ their online management system. Personality is the most important component that reflects the personality, which continues to change. Applicants typically submit their CVs to online recruitment platforms in the form of a loosely structured paper that must be reviewed by a professional recruiter. The importance of manual interviews and resumes in human resources has grown substantially in recent years. It is vital to devise a strategy for reducing or expediting the HR department's workload.

The Big Five Model was used in the study to forecast important personality features, i.e. characteristics or qualities that are unique to an entity. The OCEAN model (also known as the Big Five Personality Model) has been used to predict the candidate's personality, which includes Openness, Conscientiousness, Extraversion, Acceptability, and Neuroticism.

Personality can also affect his/her interaction with the outside world and his/her environment. Personality can also be used as an additional feature during the recruitment process, career counseling, health counseling, etc. Predicting personality by analyzing the behavior of the person is an old technique. This manual method of personality prediction required a lot of time and resources. Analyzing personality based on one’s nature was a tedious task and a lot of human effort would be required to do such analysis. Also, this manual analysis did not give accurate results while analyzing the personality of a user from their nature and behavior. Since the analysis was done manually, it affects the accuracy of the results as humans are prone to be prejudiced and generally see things accordingly.

Methods used to analyze the data include surveys, interviews, questionnaires, classroom activities, shopping website data, and social network data about the user experiences and problems they are facing. But these traditional methods are time-consuming and very limited in scale. The proposed system will automate the process and reduce dependency on humans thereby reducing human-induced errors. Nowadays while conducting interviews, the interviewer does not have enough time to pen down all the details for each candidate that he interviews. In such scenarios, we can provide the interviewer with a platform where he just has to enter the text from the dropdown menu pertaining to the few behavior traits that are listed on UI. The input from UI goes to the AI model running in the backend and it makes the necessary predictions. The final personality label is displayed on UI again and the interviewer can decide whether to select the candidate or reject them on the basis of his behavioral traits.

**1.1 Aim**

The project is based on identifying the personality of an individual using machine learning algorithms and big 5 models. The personality of a human plays a major role in his personal and professional life. Nowadays, many organizations have also started shortlisting candidates based on their personality as this increases the efficiency of the work because the person is working on what he is good at than what the candidate is forced to do.

The 5 types of personality for classifying the person are:

1. **Open to Experience:** It involves various dimensions, like imagination, sensitivity, attentiveness, preference for variety, and curiosity.
2. **Conscientiousness**: This trait is used to describe the carefulness and diligence of the person. It is the quality that describes how organized and efficient a person is.
3. **Extraversion:** It is the trait that describes how the best candidates can interact with people that is how good are his/her social skills.
4. **Agreeableness:** It is a quality that analyses the individual behavior based on generosity, sympathy, cooperativeness, and the ability to adjust to people.
5. **Neuroticism:** This trait usually describes a person having mood swings and extreme expressive power.

**1.2 Motivation**

It’s extremely important for recruiters to assemble all the relevant information regarding candidates during the [recruitment process](https://www.skeeled.com/blog/turn-your-digital-recruitment-into-a-digital-experience/?hsLang=en) so that the best hiring decisions can be made. But the traditional methods for gathering information are very subjective and unreliable, which is why pre-selection assessments play an important part in the recruitment process. A [personality assessment](https://www.skeeled.com/blog/the-beginners-guide-to-personality-assessment-software/?hsLang=en) can provide recruiters with deep insights that combined with other evaluation elements (profile matching, interviews, etc.) will significantly improve the candidate selection.

Although cognitive ability is commonly linked to the likelihood to succeed professionally, studies have proven that personality traits like achievement striving, cooperation, or leadership orientation, among others, play an important part in one’s job performance. The personality label allows to quickly assess if a candidate is suitable for a job position or not, given that it measures workplace personality traits and occupational scores that are predictive of job performance.

**1.3 Objectives of the project**

* The major objective is to predict an individual's personality, which will aid in evaluating the person in numerous aspects.
* The goal is to automate the system so that it can identify the personality traits displayed by the candidate and gather information about them without meeting them.
* The company will have a better understanding of the applicant and will be able to make a more informed judgment about hiring the best candidate for the job.

**Chapter 2**

**Literature Review**

**2.1 Existing System**

A number of technological advancements during the last decade have enabled academics to establish a variety of novel approaches for gathering data in personality psychology. Consumer electronics advancements (e.g., smartphones and wearables) and the development of mobile sensing systems have permitted the longitudinal in vivo collection of highly detailed multidimensional data on behaviors and situations (Harari et al., 2016). Furthermore, behavioral residue collected from websites and online social media platforms has shown to be a rich source of data on behavior associated with personality traits (Gosling and Mason, 2015). Along with advancements in the collecting and availability of such data, advancements have been made in the analytic approaches that may be used to model these complex data. A plethora of new algorithms, in particular, are available that use existing data to make predictions about previously unknown data, detect patterns, or find groupings of similar cases.

Personality prediction is an extremely far-reaching and fluctuated field of study. In the previous ten years, there has been a lot of work done on computerized personality discovery and evaluation of online media clients. Prior chips away at personality prediction from web-based media information generally utilized AI procedures. (Golbeck et al., 2011) proposed a technique where the personality of clients could be anticipated through the information gathered from their Facebook profiles. For this review, two AI calculations, M5'Rules, and Gaussian Processes were utilized. The outcomes showed that the exhibition and connections created by M5'Rules were more grounded when contrasted with those in Gaussian Process where it showed no relationships. However, their review was powerful, they didn't zero in on the organization thickness between the clients, which would have contributed tremendously to the exploration in the space of personality prediction.

Around the same time, (Golbeck et al., 2011) again proposed a model to foresee the personality utilizing the information clients share on the most famous microblogging webpage, Twitter. To gather information for this examination, a personality test for 50 Twitter clients was directed and information was gathered from the Twitter profiles of similar clients utilizing the Twitter API. LIWC (semantic Inquiry and word count) apparatus was utilized to include extraction and insights on 81 highlights of text were created. In this review, a feeling examination of tweets was additionally performed and two relapse calculations, Xero and Gaussian Processes were utilized. The outcomes showed that the methods utilized in this model can be utilized to foresee the personality of Twitter clients from their tweets yet the model can show better outcomes on a bigger dataset.

A lot of researchers have contributed to the task of personality prediction. Few people have used psychological tests to identify personality labels, while others have used machine learning algorithms like Naive Bayes to predict outcomes It employed linear regression and support vector regression to ascertain Facebook user personality. The My Personality dataset is used in this investigation. The study's findings show that linear regression is a superior prediction method (Kunte and Panicker, 2019). Others have utilized K nearest neighbor, simple logistics, and SVM to classify personality based on Twitter text. The Twitter streaming API was utilized to collect the data, and the Martingale framework was used to forecast the results. This is a one-of-a-kind technique (Pratama and Sarno, 2015).

(Quercia et al., 2011) introduced a review to break down the personality of various kinds of Twitter clients (well-known clients, forces to be reckoned with, and so forth) and the connection between these clients in light of how they cooperate on web-based stages. Relapse investigation with a 10-overlay cross approval procedure was performed and the root means square mistake was determined which ends up being 0.88 which was significantly low. This study saw that there are various essential likenesses notwithstanding contrasts among the various Twitter clients.

The soundness of personality attributes is the most contended point among therapists and scientists. A study conducted by (van der Linden, te Nijenhuis and Bakker, 2010) titled "The General Factor of Personality": A criterion-related validity review and a meta-analysis of Big Five intercorrelations has been published. This study examined the interrelationships between the Big Five personality traits to determine the presence of a GFP (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). The meta-analysis provides support for GFP at the most operational level, and this article indicated that the GFP is relevant since it is associated to supervisor-rated job performance.

(Najib and Nawab, 2015) proposed a framework to recognize age, orientation, and personality characteristics through the tweets of a client. For the examination, the PAN-2015 dataset was utilized which is assembled corpus in which the tweets are characterized by creator and time of creator, language, orientation, and relating personality attributes. This study underlined the significance of content-based elements which can be utilized to separate texts composed by individuals having a place with unmistakable profiles. Four AI classifiers: J48, Random Forest, SVM, and Naïve Bayes were utilized for preparing the model. The outcomes showed that the framework didn't play out as well on the testing information as it did on the preparation information. (Kalghatgi, Ramannavar and Sidnal, 2015) suggested A Neural Network Approach to Personality Prediction Based on the Big-Five Model. The parallelism between the personality qualities of an individual's linguistic information is investigated for analytics. The Big Five paradigm enables the identification of personality traits through the use of language information. This demonstrates personality attributes that can be used in a variety of industries, including business intelligence, marketing, and psychology.

(Faliagka, Tsakalidis and Tzimas, 2012) proposed an Integrated recruitment System for Automated Personality Mining and Applicant Ranking. This approach now includes automatic candidate rating. It was decided based on objective principles that the candidate's information would be collected from the candidate's LinkedIn profile. The applicant's rank was calculated using the Analytical Hierarchy Process (AHP) from individual selection criteria, with the recruiter controlling their weight.

For predicting MBTI personality types, classic machine learning approaches and neural networks have been employed successfully. (Golbeck et al., 2011) conducted one of the first research on personality prediction using machine learning techniques. Based on the MBTI personality type indication and the information offered on Twitter, they were able to reliably determine a user's personality type. (Komisin and Guinn, 2012) employed the Nave Bayes and Support Vector Machine (SVM) methodologies to identify an individual's personality type based on their word choice in another investigation. Their database was created using in-class writing examples from 40 graduate students, as well as their MBTI personality type. On their limited dataset, they examined the performance of these two algorithms and discovered that the Naive Bayes technique outperforms SVM. Two years later, (Wan et al., 2014) employed a machine learning method to predict the Big Five personality types of users using their words on Weibo, a Chinese social network, and they were effective in predicting the users' personality types. Based on the Big Five model and their text samples, (Li, Wan, and Wang, 2017) employed the grey prediction model, multiple regression model, and multi-tasking model to predict the user personality type. They examined the performance of these three models and discovered that the grey prediction model outperforms the other two. (Tandera et al., 2017) used the Big Five personality model and some deep learning architecture to estimate a person's personality based on information from their Facebook profile. They compared their method's performance to that of other earlier studies that used classical machine learning approaches, and the results showed that their model beat the accuracy of past similar studies.

**2.2 Proposed System**

To overcome the problems of the existing system an Automated personality classification system is proposed which uses some data mining techniques and machine learning algorithms are used to classify the personalities of different users. By identifying the past data and their patterns it is easy to identify the personality by applying new techniques, so it overcomes the existing system. The dataset consists of the 7 features namely Gender, Age, Openness, Neuroticism, Conscientiousness, Agreeableness, and Extraversion. These will be used to predict the personality label of a person. The model will be trained with an Artificial Neural Network algorithm and it will be supervised learning. The ANN will be tuned to obtain optimum results.

Steps involved in building the system:

1. The user will select the text from the dropdown on UI. The score associated with each text will be given as input from User Interface.
2. Input will be passed to the AI model instance and in turn AI model will predict the output label for the candidate
3. On the model building front, Dataset acquired from Kaggle is used to train the AI model.
4. Data visualization and Data pre-processing will be performed to make data ready and in the acceptable form
5. AI model will be trained by using a deep learning algorithm named ANN
6. Tuning for the AI model will be done using advanced Keras tuning techniques aiming to achieve the highest possible accuracy.
7. Once the model attains the best accuracy, it is pickled and stored for future reference.
8. AI models will be integrated with UI to make future predictions.

**Chapter 3**

**Requirement and Analysis**

**3.1 Requirement specification**

**3.1.1 Introduction**

The relevant dataset is required to perform analysis and train the ANN model on it. Once the dataset is finalized then there come the feature selections, pre-processing the data, and visualizing the data.

**Data collection:** The first step for the prediction system is data collection and deciding about the training and testing dataset. In this project, we have imported a dataset from the Kaggle website. Data collection is defined as the procedure of collecting, measuring, and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach to data collection is different for different fields of study, depending on the required information.

**Attribute selection:** Attributes of datasets are properties of datasets that are used for systems and for personality many attributes are like heart gender of the person, age of the person, Big five traits like Openness, Neuroticism, Extraversion, Agreeableness, Consciousness (value 1 -10). The importance of feature selection can best be recognized when you are dealing with a dataset that contains a vast number of features. This type of dataset is often referred to as a high-dimensional dataset. Now, with this high dimensionality, comes a lot of problems such as - this high dimensionality will significantly increase the training time of your machine learning model, it can make your model very complicated which in turn may lead to Overfitting.

**Data Preprocessing:** Pre-processing is needed for achieving the best result from the machine learning algorithms. In this, we gathered the dataset and it was pre-processed before it is sent to the training stage. Sampling is a very common method for selecting a subset of the dataset that we are analyzing. In most cases, working with the complete dataset can turn out to be too expensive considering the memory. Using a sampling algorithm can help us reduce the size of the dataset to a point where we can use a better, but more expensive, machine learning algorithm. When we talk about data, we usually think of large datasets with a huge number of rows and columns. While that is a likely scenario, it is not always the case — data could be in so many different forms: Structured Tables, Images, Audio files, Videos, etc. Machines don’t understand free text, image, or video data as it is, they understand 1s and 0s. So we pre-process the data.

**Predictions:** In this, system we used deep learning algorithms that gives the best accuracy for personality prediction. By applying all these modules finally, the personality is predicted and the final result is the personality of the user.

**3.1.2 Problem Statement**

Personality assessment has become the most used test to hire many employees. Classifying the personality of the user based on the big five personality traits using data mining is a convenient way to judge the candidate. We need a strong model that predicts the personality of people based on their imagination and thoughts. The goal of this project is to build a model that predicts the personality of the people. The main aim of the proposed system is to predict the personality of the user on the feedback given by the interviewer. The project is aimed to develop software that will be helpful in identifying the personality of the person. It uses the concept of machine learning algorithms. An advanced artificial neural network algorithm is used to build the system. After classification is done the user type personality will be displayed.

**3.2 Requirements**

**3.2.1 Functional requirements**

**Interoperability / Open Architecture:**  There is no standard or uniform IIoT infrastructure platform. The key consideration is whether the analytics solution works with multiple platforms or is a closed add-on to one platform.

**Asset and Sensor Neutrality:** The key consideration is whether the solution functions in heterogeneous plant environments with data from all production assets.

**Asset Visualization:** At a facility level, technicians accessing the user-interface will not be trained in Artificial Intelligence and Big Data. The key considerations when defining this requirement are the visualization of machine behavior and the ability to depict the health of machinery or the entire facility, and take specific action as a result.

**Scalability:** Analytics platforms must be applicable to a machine or facility of any size. The solution must be able to add assets without a need for any incremental investment in hardware, software, or dedicated labor hours

**Performance:** The objective of an industrial analytics platform is to provide the production facility with accurate and timely data.

**3.2.2 Constraints**

There are a lot of challenges that machine learning professionals face to inculcate ML skills and create an application from scratch.

1. Poor Quality of Data

Data plays a significant role in the machine learning process. One of the significant issues that machine learning professionals face is the absence of good quality data. Unclean and noisy data can make the whole process extremely exhausting. We don’t want our algorithm to make inaccurate or faulty predictions. Hence the quality of data is essential to enhance the output. Therefore, we need to ensure that the process of data preprocessing which includes removing outliers, filtering missing values, and removing unwanted features, is done with the utmost level of perfection.

2. Under-fitting of Training Data

This process occurs when data is unable to establish an accurate relationship between input and output variables. It simply means trying to fit in undersized jeans. It signifies the data is too simple to establish a precise relationship. To overcome this issue:

* Maximize the training time
* Enhance the complexity of the model
* Add more features to the data
* Reduce regular parameters
* Increasing the training time of model

3. Overfitting of Training Data

Overfitting refers to a machine learning model trained with a massive amount of data that negatively affect its performance. It is like trying to fit in Oversized jeans. Unfortunately, this is one of the significant issues faced by machine learning professionals. This means that the algorithm is trained with noisy and biased data, which will affect its overall performance.

We can tackle this issue by:

* Analyzing the data with the utmost level of perfection
* Use data augmentation technique
* Remove outliers in the training set
* Select a model with lesser features

4. Machine Learning is a Complex Process

The machine learning industry is young and is continuously changing. Rapid hit and trial experiments are being carried on. The process is transforming, and hence there are high chances of error which makes the learning complex. It includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, and a lot more. Hence it is a really complicated process which is another big challenge for Machine learning professionals.

5. Lack of Training Data

The most important task you need to do in the machine learning process is to train the data to achieve an accurate output. Less amount of training data will produce inaccurate or too biased predictions. But on the other hand, a machine-learning algorithm needs a lot of data to distinguish. For complex problems, it may even require millions of data to be trained. Therefore, we need to ensure that Machine learning algorithms are trained with sufficient amounts of data.

6. Slow Implementation

This is one of the common issues faced by machine learning professionals. The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time. Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results. Further, it requires constant monitoring and maintenance to deliver the best output.

7. Imperfections in the Algorithm When Data Grows

If you have found quality data, trained it amazingly, and the predictions are really concise and accurate. But the model may become useless in the future as data grows. The best model of the present may become inaccurate in the coming Future and require further rearrangement. So you need regular monitoring and maintenance to keep the algorithm working. This is one of the most exhausting issues faced by machine learning professionals.

**3.2.3 Software and Hardware requirements**

**Software requirements**

* Python: Most suitable programming language for carrying out all Machine Learning Tasks.
* Google Colab: Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis, and education.
* TensorFlow and Keras: To train our Machine Learning model on Personality Prediction.
* Streamlit: open-source app framework in Python language. It helps us create web apps for data science and machine learning. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib, etc.

**Hardware requirements**

A computer or a laptop with Linux operating system and Ubuntu 20.04.

**Chapter 4**

**DESIGN**

**4.1 ARCHITECTURE OF THE SYSTEM**

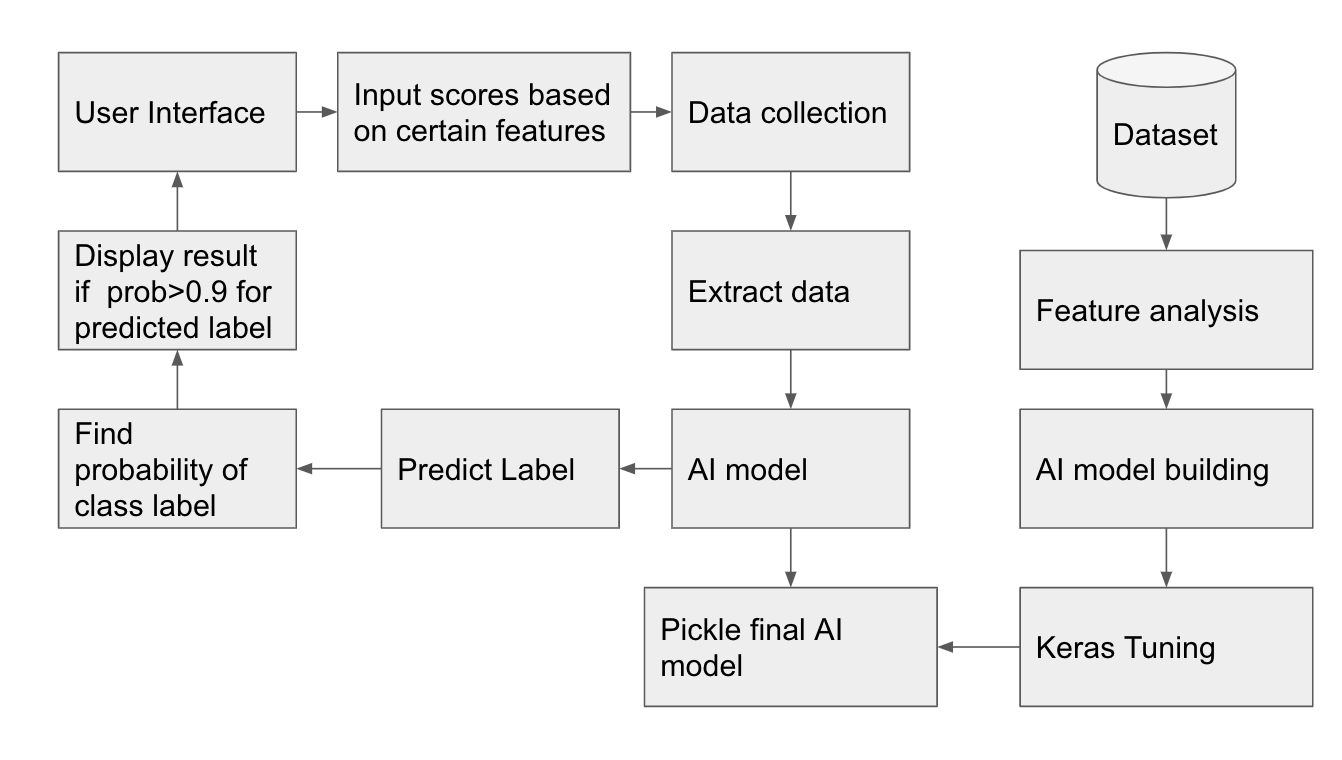
****

Figure 4.1.1 System Architecture

**4.2. DFDS**

**4.2.1 DFD LEVEL 0**

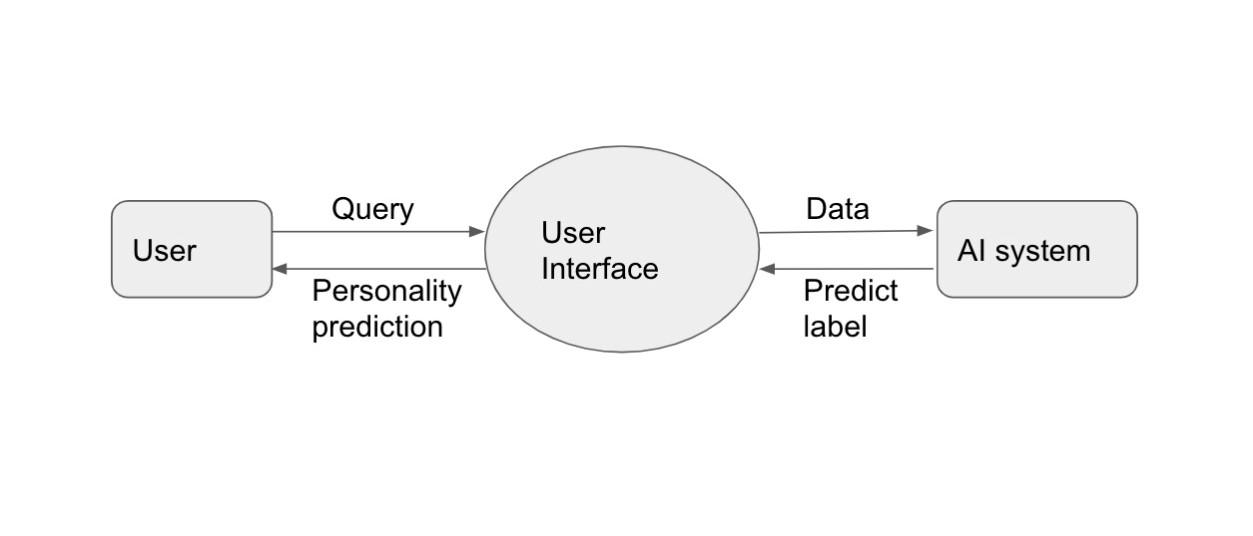
****

Figure 4.2.1: Level 0 Data Flow Diagram

**4.2.2 DFD Level 1**

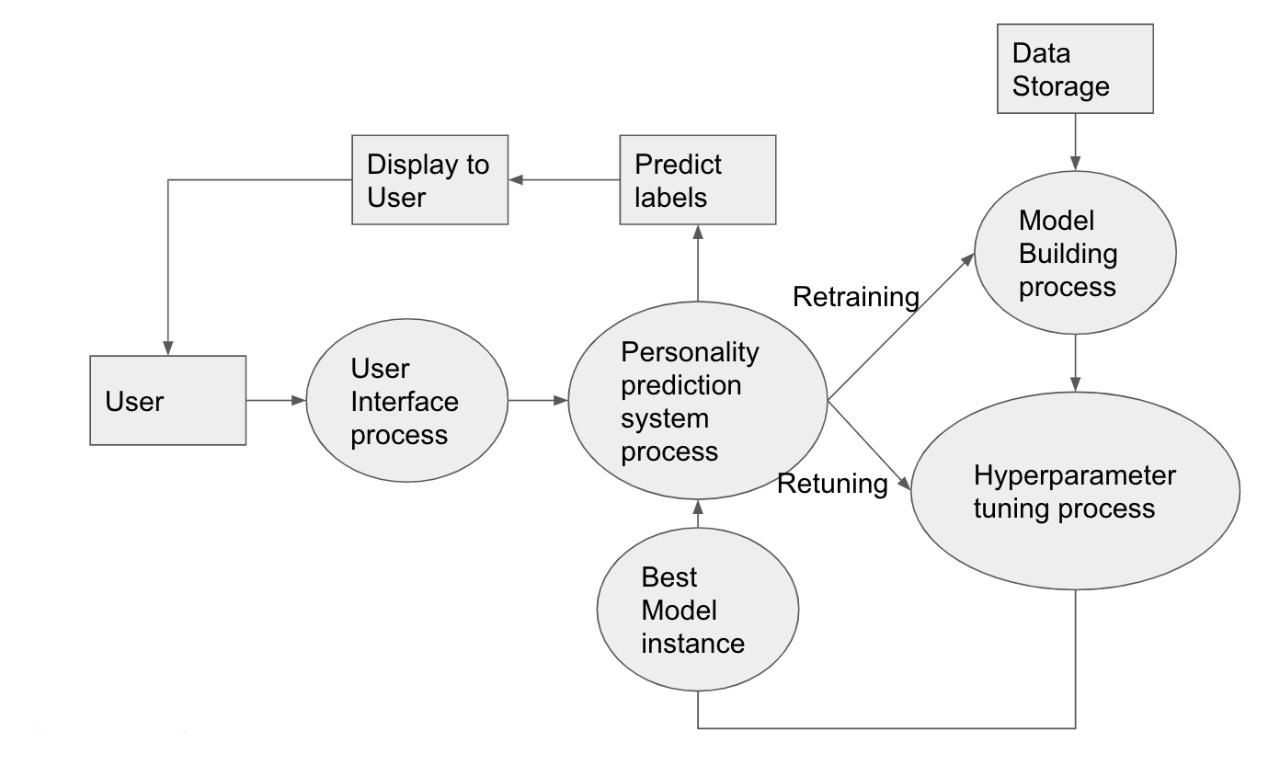
****

Figure 4.2.2: Level 1 Data Flow Diagram

**4.3 USE CASE DIAGRAM**

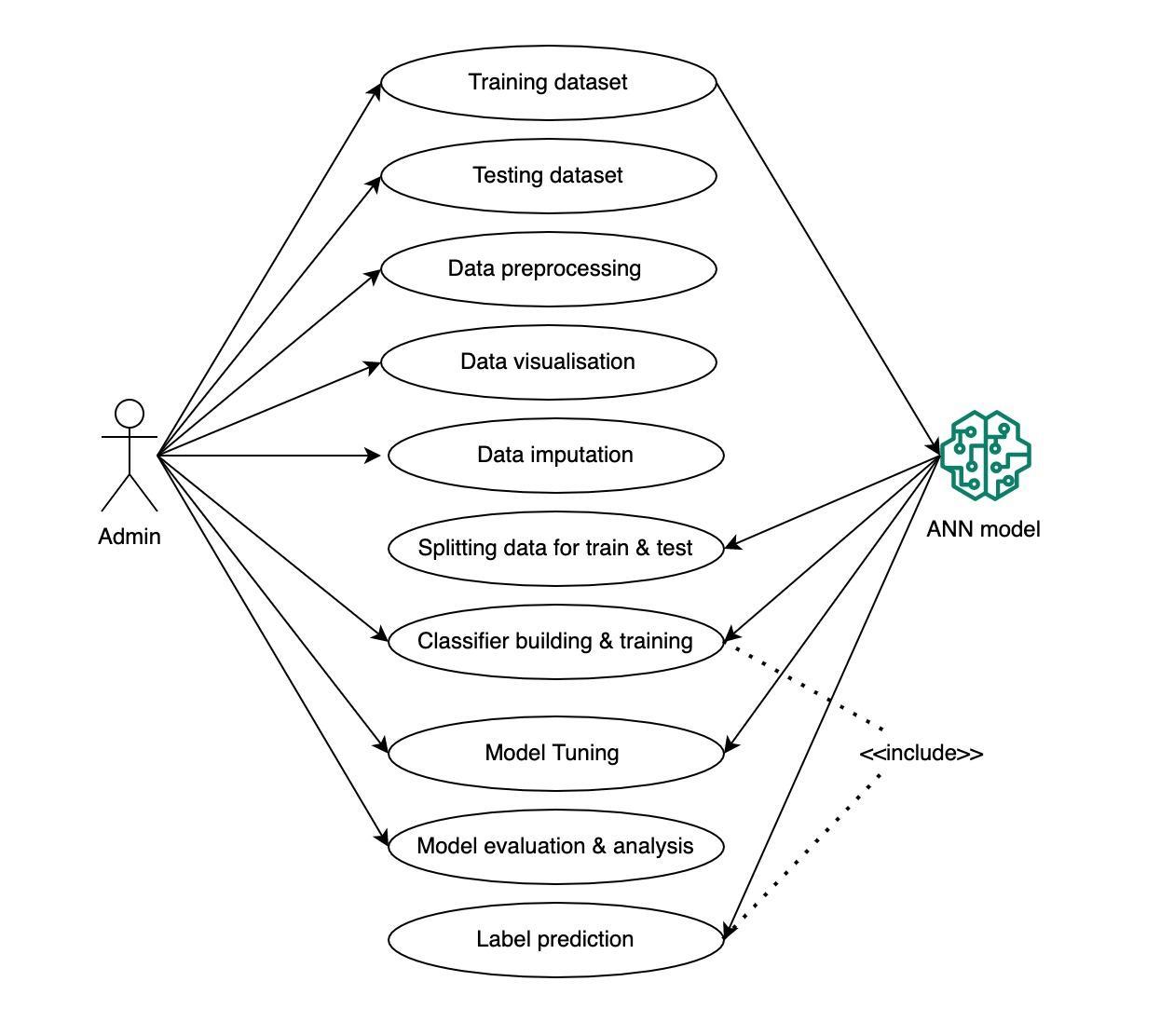
****

Figure 4.3.1: Use case Diagram

**4.4 CLASS DIAGRAM**

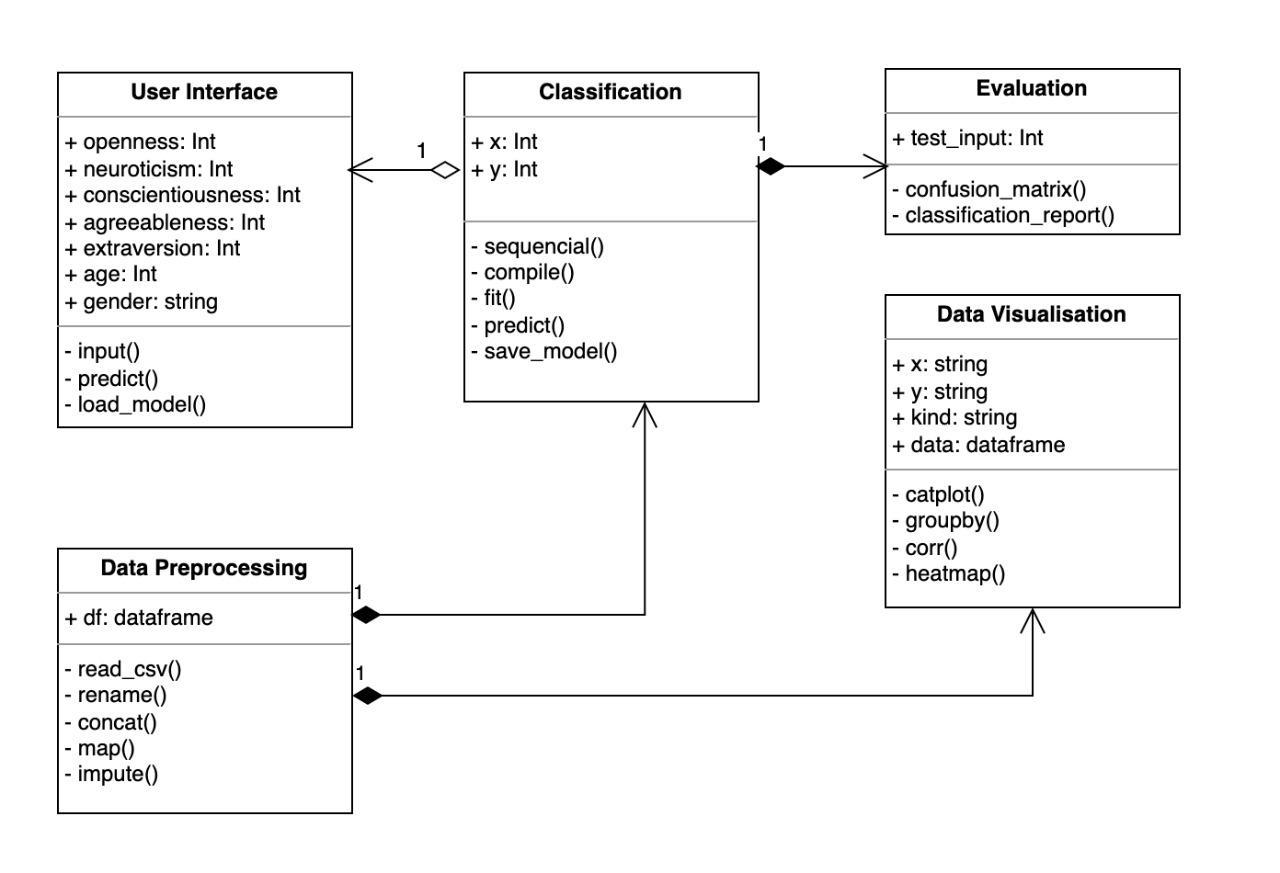
****

Figure 4.4.1: Class Diagram

**4.5. ACTIVITY DIAGRAM**

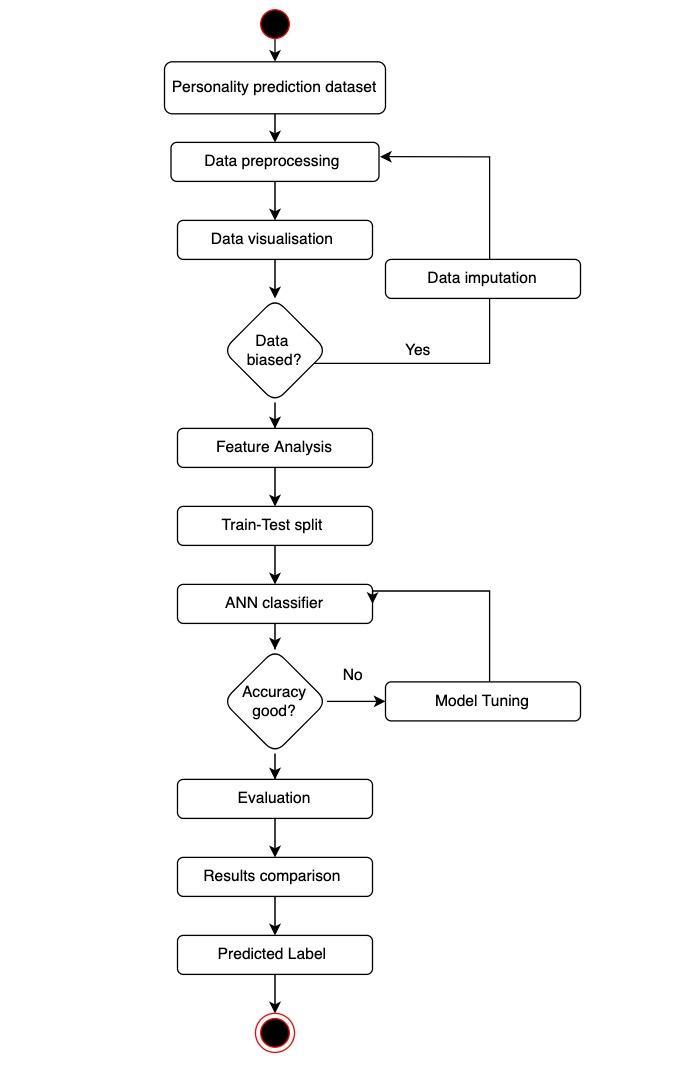
****

Figure 4.5.1: Activity Diagram

**4.6 SEQUENCE DIAGRAM**

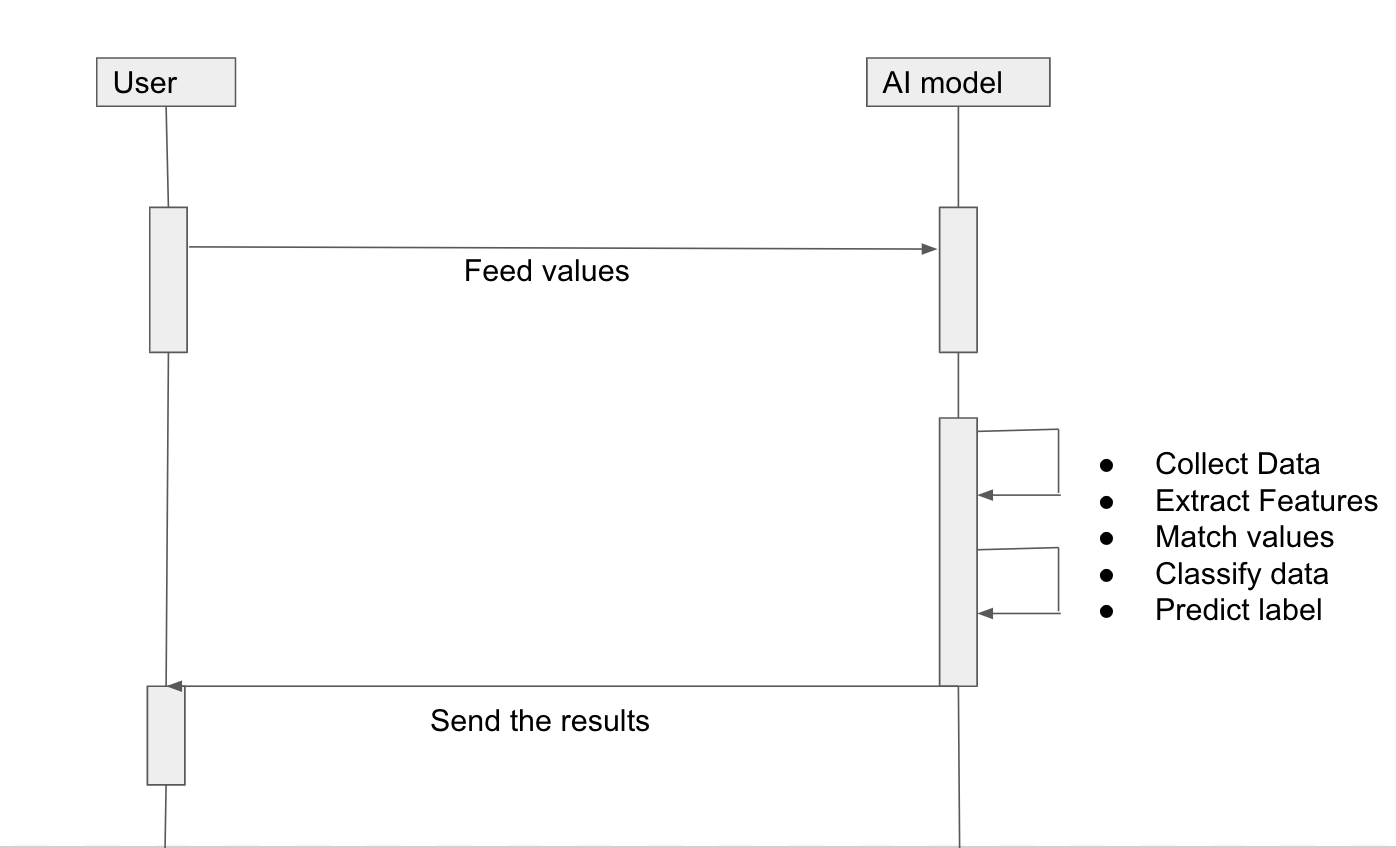
****

Figure 4.6.1: Sequence Diagram

**Chapter 5**

**IMPLEMENTATION**

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed task.

In addition to an informed, working definition of machine learning (ML), we detail the challenges and limitations of getting machines to ‘think,’ some of the issues being tackled today in deep learning (the frontier of machine learning), and key takeaways for developing machine learning applications for business use-cases. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

All of these things mean it's possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results. And by building precise models, an organization has a better chance of identifying profitable opportunities – or avoiding unknown risks.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. The type of algorithm data scientists chooses to use depends on what type of data they want to predict.

We will be using **supervised learning** in our project:

In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified. Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression, Artificial Neural Networks, etc.

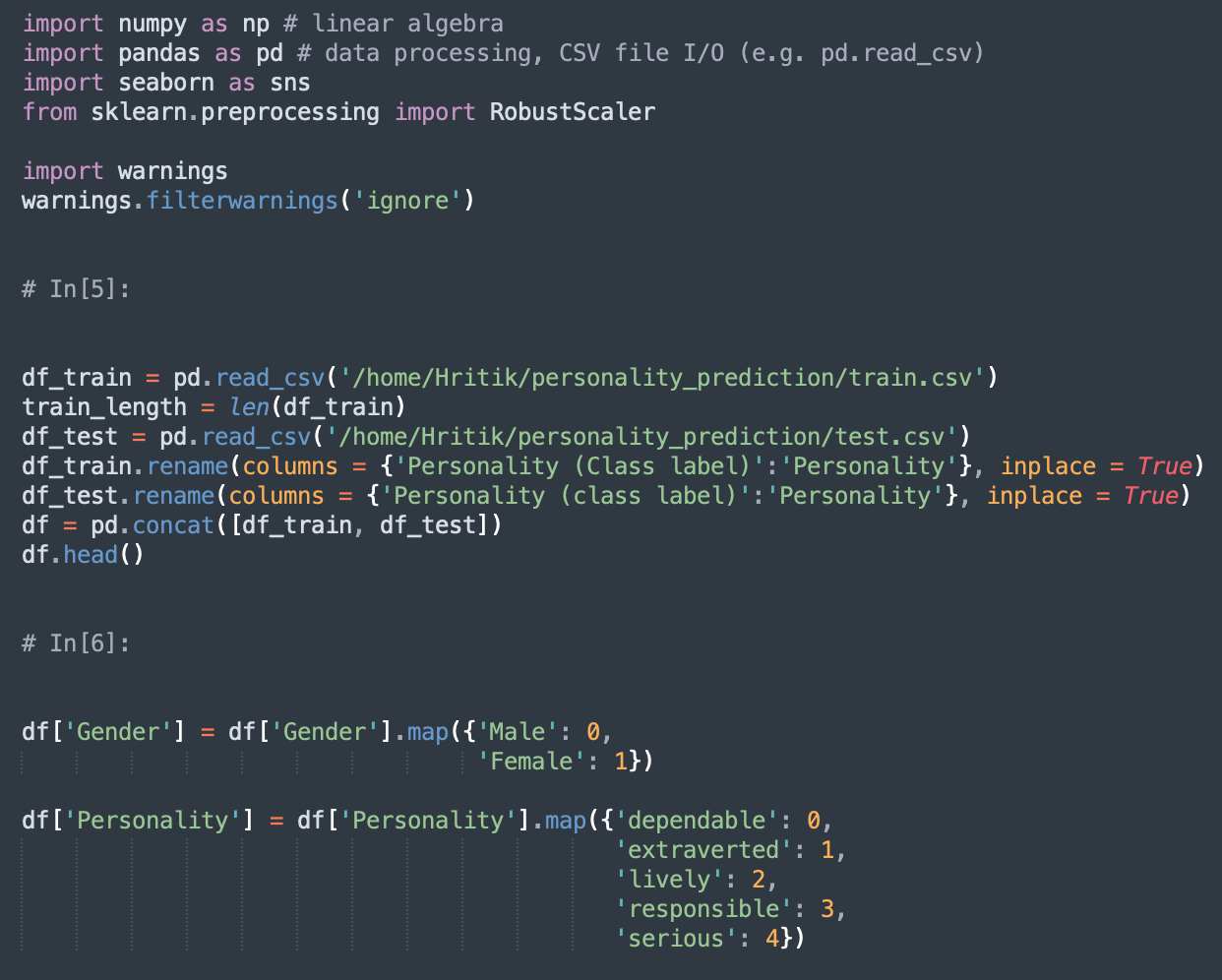
Detailed steps are as follows:

1. Dataset acquisition: To build and develop Machine Learning models, you must first acquire the relevant dataset. This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. Dataset formats differ according to use cases. We use the dataset available on Kaggle.
2. Importing libraries: Since Python is the most extensively used and also the most preferred library by Data Scientists around the world, we’ll show you how to import Python libraries for data preprocessing in Machine Learning. Read more about Python libraries for Data Science here. The predefined Python libraries can perform specific data preprocessing jobs. The three core Python libraries used for this data preprocessing in Machine Learning are NumPy, Pandas, Scikit-learn, TensorFlow, Keras, Streamlit, Matplotlib, and Seaborn.
   1. NumPy: NumPy is one of the most powerful Python libraries. It is used in the industry for array computing. This article will outline the core features of the NumPy library. It will also provide an overview of the common mathematical functions in an easy-to-follow manner. NumPy is gaining popularity and is being used in a number of production systems.
   2. Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.
   3. Scitkit-learn: Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction.
   4. TensorFlow and Keras: Keras is a neural network library while TensorFlow is the open-source library for a number of various tasks in machine learning. TensorFlow provides both high-level and low-level APIs while Keras provides only high-level APIs.
   5. Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, and WxPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebooks, and web application servers also.
   6. Seaborn: Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas’ data structures. Seaborn helps you explore and understand your data.
   7. Streamlit: Streamlit is an open-source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib, etc.
3. Loading the Dataset: In this step, you need to import the dataset/s that you have gathered for the ML project at hand. However, before you can import the dataset, you must set the current directory as a working directory. Once you’ve set the working directory containing the relevant dataset, you can import the dataset using the “read\_csv()” function of the Pandas library. This function can read a CSV file (either locally or through a URL) and also perform various operations on it. The read\_csv() is written as: data\_set = pd.read\_csv(‘Dataset.csv’)
4. Handling missing values: In data preprocessing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data. Needless to say, this will hamper your ML project. In our dataset, we have performed the data imputation and hence there are no missing values.
5. Encode the categorical data: Categorical data refers to the information that has specific categories within the dataset. In the dataset used, there are two categorical variables –Gender and Personality label. Machine Learning models are primarily based on mathematical equations. Thus, you can intuitively understand that keeping the categorical data in the equation will cause certain issues since you would only need numbers in the equations.
6. Feature Scaling: Feature scaling marks the end of the data preprocessing in Machine Learning. It is a method to standardize the independent variables of a dataset within a specific range. In other words, feature scaling limits the range of variables so that you can compare them on common grounds. In our dataset, all values are between 1-and 10, and hence different type of scaling is not required.
7. Extracting important information from data: The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining. In other words, we can say that Data Mining is the process of investigating hidden patterns of information from various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithms, helping decision making and other data requirement to eventually cost-cutting and generating revenue. Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).
8. Python will be used as a development language: Python is a popular object-oriented programming language having the capabilities of a high-level programming language. It’s easy-to-learn syntax and portability capability make it popular these days. It is interpretable, interactive, object-oriented, and beginner-friendly.
9. Model Building and Training:
   1. Algorithm: ANN is used for model building. ANNs are a type of computer program that can be 'taught' to emulate relationships in sets of data. Once the ANN has been 'trained', it can be used to predict the outcome of another new set of input data, e.g., another composite system or a different stress environment.
   2. ANN Tuning: To perform hyperparameter tuning, we want to characterize the hunt space, in other words, which hyperparameters should be advanced and in what range. Here, for this model, there are hyperparameters that can be tuned: dropout rates, number of layers, number of neurons in layers, activation function, optimizer, and learning rate.
   3. Model compilation: Then, at that point, how about we move to display aggregation, where other hyperparameters are likewise present. The gathering step is the place where the streamlining agent alongside the misfortune work and the measurement are characterized. Here, we'll involve absolute entropy as a misfortune capacity and precision as a measurement. For the enhancer, various choices are accessible.
   4. Model training: Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine-learning algorithm to minimize a loss function over the prediction range. The Model is trained for 5000 epochs.
   5. Evaluation: You will have to check the machine created against your evaluation data set that contains inputs that the model does not know and verify the precision of your already trained model. If the accuracy is less than or equal to 50%, that model will not be useful since it would be like tossing a coin to make decisions. If you reach 80% or more, you can have good confidence in the results that the model gives you.
   6. Tuning: If during the evaluation you did not obtain good predictions and your precision is not the minimum desired, it is possible that you have overfitting -or underfitting problems and you must return to the training step before making a new configuration of parameters in your model. You can increase the number of times you iterate your training data- termed epochs. Another important parameter is the one known as the “learning rate”, which is usually a value that multiplies the gradient to gradually bring it closer to the global -or local- minimum to minimize the cost of the function.

Increasing your values by 0.1 units from 0.001 is not the same as this can significantly affect the model execution time. You can also indicate the maximum error allowed for your model. You can go from taking a few minutes to hours, and even days, to train your machine. These parameters are often called Hyperparameters. This “tuning” is still more of an art than a science and will improve as you experiment. There are usually many parameters to adjust and when combined they can trigger all your options. Each algorithm has its own parameters to adjust. To name a few more, in Artificial Neural Networks (ANNs) you must define in its architecture the number of hidden layers it will have and gradually test with more or less and with how many neurons each layer. This will be a work of great effort and patience to give good results.

* 1. Inferences: If the model is giving good accuracy, then you are now ready to use your Machine Learning model to infer results in real-life scenarios.
  2. User interface: In this step, we integrate the final model into the user interface. User interface code is written in python using the Streamlit library.

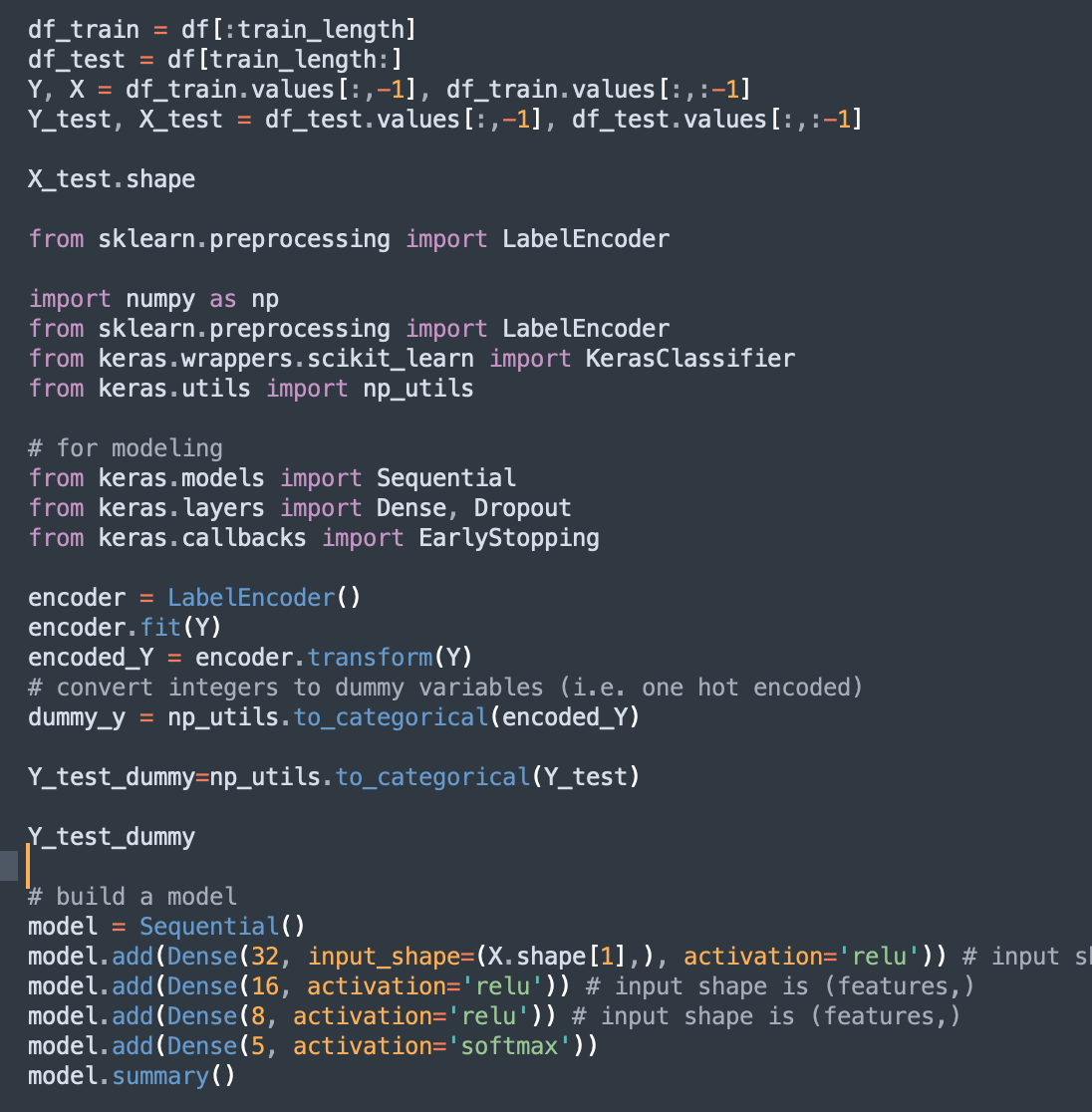
**Sample code for ANN model building:**

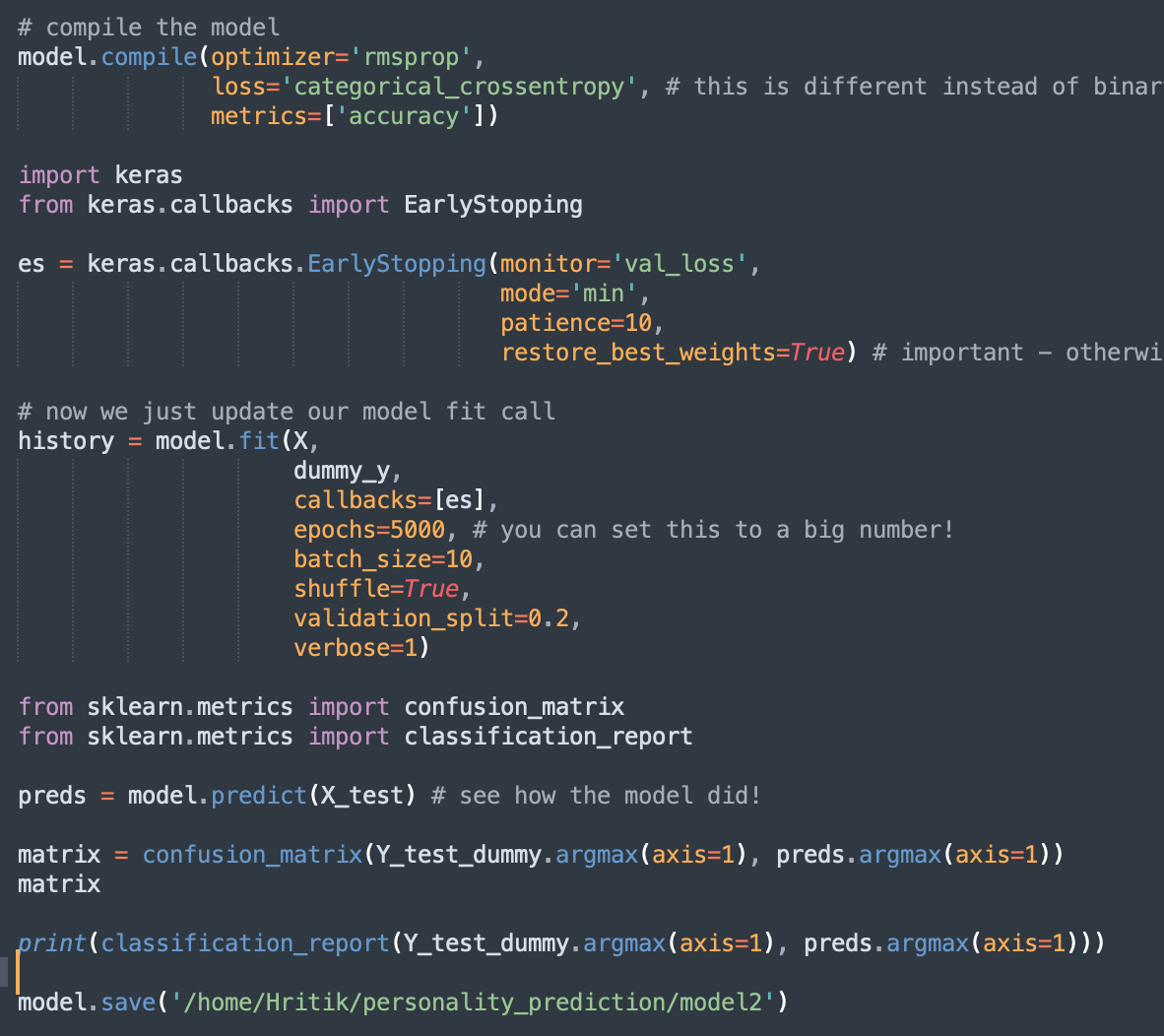
****

****

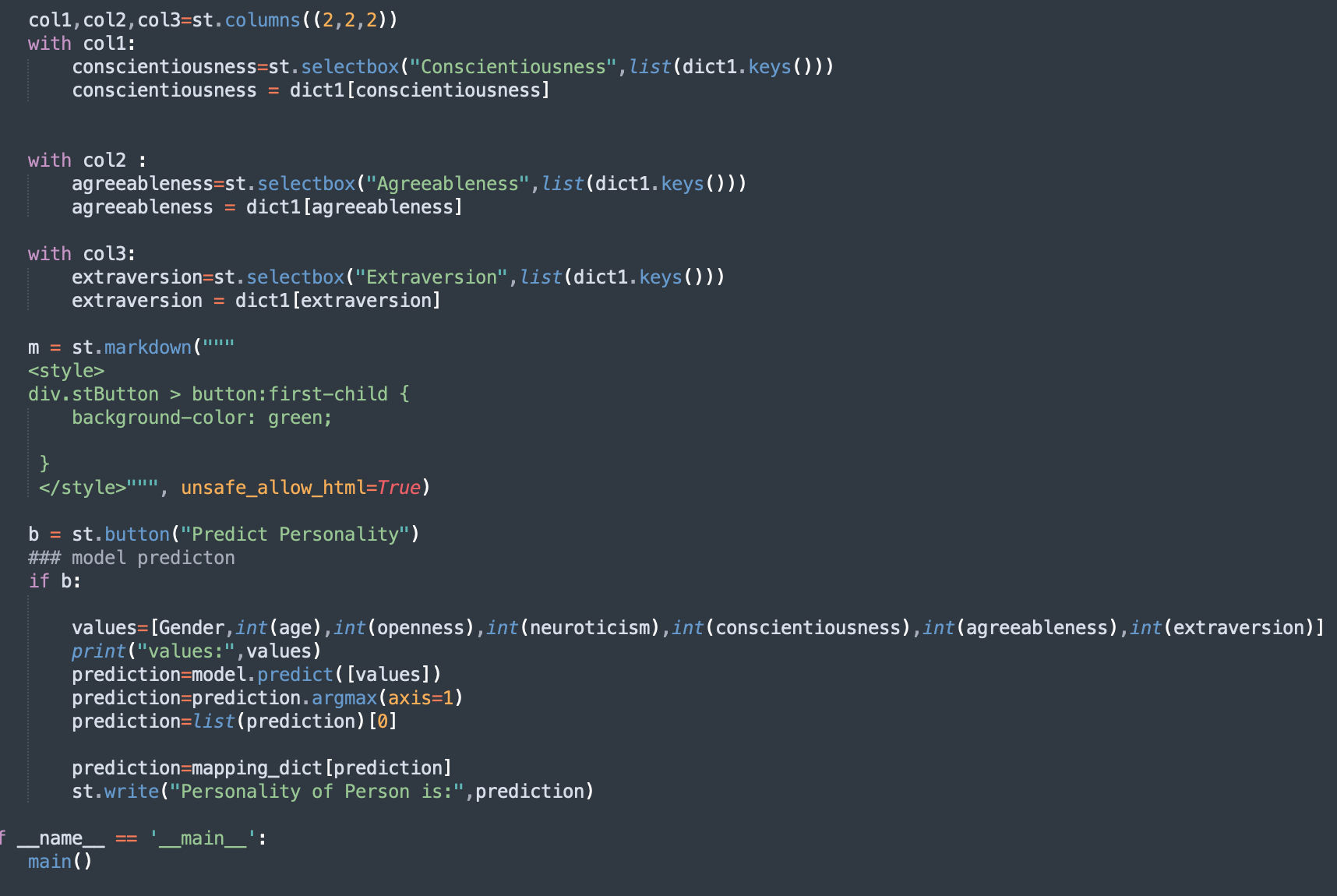
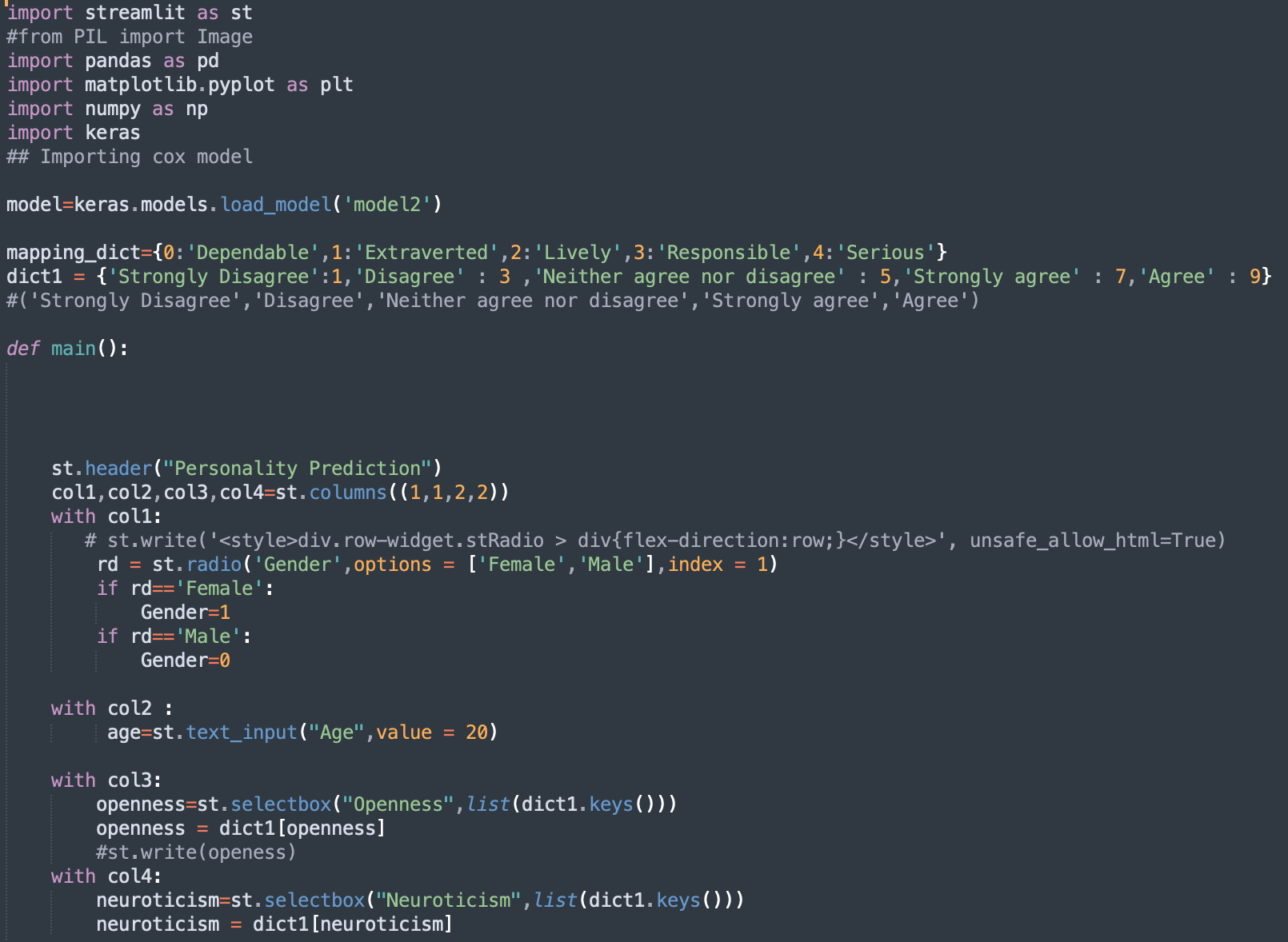
****

****

****

****

**Sample code for the User interface:**

****

**Chapter 6**

**Result and Evaluation**

After Model tuning and refinement, we get the final model that can be used for predictions on unseen data. The accuracy for the final model reaches up to 86%.

Confusion Matrix of final AI model based on ANN algorithm:

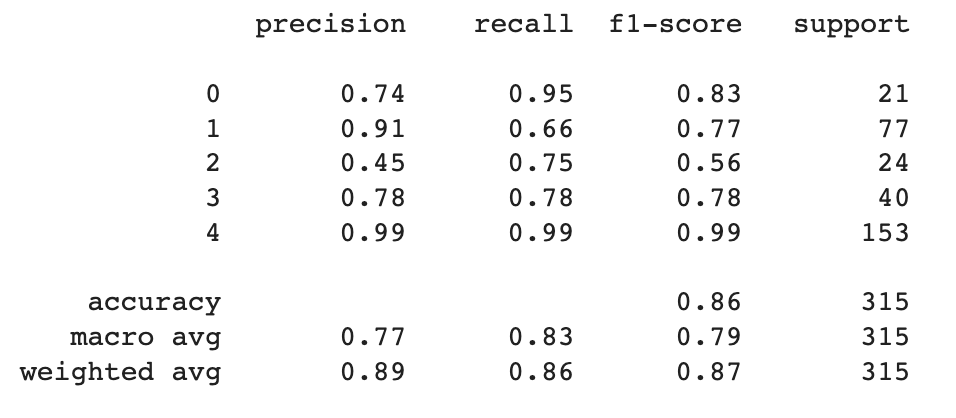


Figure 6.1 Confusion Matrix of ANN Model

The user interface was developed using Streamlit which looks as follows:

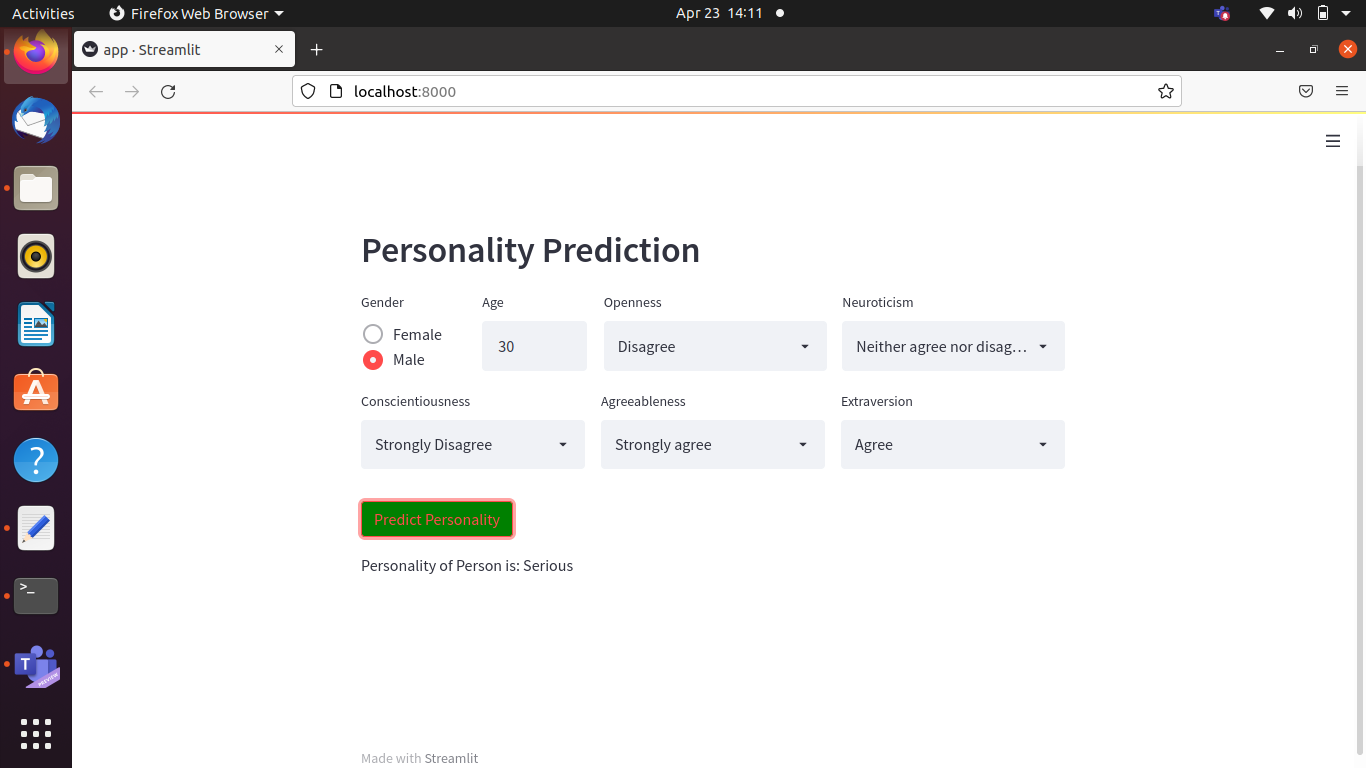


Figure 6.2 Final User Interface

**Chapter 7**

**CONCLUSION**

The system will help the HR department to select the right candidate for a job. This in turn will provide expert workforce for the organization. The recruiter need not sort through thousands of resumes. The recruiter enters a score and the personality trait output is the basis of selecting or rejecting a resume. We will make use of deep learning-based ANN algorithms to train our AI model and they have high time complexity. The system can be integrated with the recruiting company’s existing website. Thus, the genuineness of the employer is guaranteed as the system is part of the company's official website. The system would be used in many business sectors that will require expert candidates, thus reducing the workload on the human resource department.

**Future Scope**

In this project, we discuss how personality is identified using different classification algorithms. Here we study the relationship between the user and his/her personality. In this, we used ANN because it gives the best accuracy around 86% while compare to other algorithms that are used previously like naive Bayes, SVM, Logistic regression is fast and give accurate results compared to other algorithms. Personality system is used in E-commerce sites, Competitive exams, Psychometric tests, matrimonial sites, and in Government sectors like the army, navy, and Air force. Thus, the personality is automatically classified by the system after the user attempts the survey by the data set provided in the back end. Personality analysis and prediction are more in recent times so further in the future more personality traits can be added. Further, any improvement can be done using the data set and algorithms to improve the accuracy and can be helpful for the career guidance module, if the user has good speaking and convincing skills. This project discusses personality Analysis and Prediction.

**References**

1) Harari, G.M., Lane, N.D., Wang, R., Crosier, B.S., Campbell, A.T. and Gosling, S.D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science. Perspectives on Psychological Science, 11(6), pp.838–854.

2) Gosling, S.D. and Mason, W. (2015). Internet Research in Psychology. Annual Review of Psychology, 66(1), pp.877–902.

3) Kunte, A.V. and Panicker, S. (2019). Using textual data for Personality Prediction:A Machine Learning Approach. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/abstract/document/9036220 [Accessed 26 Apr. 2022].

4) Pratama, B.Y. and Sarno, R. (2015). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. 2015 International Conference on Data and Software Engineering (ICoDSE).

5) van der Linden, D., te Nijenhuis, J. and Bakker, A.B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. Journal of Research in Personality, [online] 44(3), pp.315–327.

6) Kalghatgi, M.P., Ramannavar, M.M. and Sidnal, D.N. (2015). A Neural Network Approach to Personality Prediction based on the Big-Five Model. [online]

7) Faliagka, E., Tsakalidis, A. and Tzimas, G. (2012). An integrated e‐recruitment system for automated personality mining and applicant ranking. Internet Research, 22(5), pp.551–568.

8) Golbeck, J., Robles, C., Edmondson, M. and Turner, K. (2011). Predicting Personality from Twitter. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/document/6113107>.

9) Komisin, M. and Guinn, C. (2012). Identifying Personality Types Using Document Classification Methods. [online] Semantic Scholar.

10) Wan, D., Zhang, C., Wu, M. and An, Z. (2014). Personality Prediction Based on All Characters of User Social Media Information. Communications in Computer and Information Science, pp.220–230.

11) Li, C., Wan, J. and Wang, B. (2017). Personality Prediction of Social Network Users. [online] IEEE Xplore.

12) Tandera, T., Hendro, Suhartono, D., Wongso, R. and Prasetio, Y.L. (2017). Personality Prediction System from Facebook Users. Procedia Computer Science, [online] 116, pp.604–611.





