

CV ANALYSIS AND OPTIMIZING THE RECRUITMENT PROCESS IN THE IT INDUSTRY USING MACHINE LEARNING TECHNIQUES

Project Final (Draft) Report

De Silva M.

(IT20207854)

BSc (Hons) Degree in Information Technology
Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology
Sri Lanka

September 2023

CV ANALYSIS AND OPTIMIZING THE RECRUITMENT PROCESS IN THE IT INDUSTRY USING MACHINE LEARNING TECHNIQUES

De Silva M.

(IT20207854)

The dissertation was submitted in partial fulfillment of the requirements for the BSc (Hons) in
Information Technology Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2023

DECLARATION

I declare that this is my own work, and this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as article or books).

Name	Student ID	Signature
De Silva M.	IT20207854	<i>M. De Silva</i>

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

.....
Signature of the Supervisor
(Dr. Anuradha Karunasena)

.....
Date

ABSTRACT

The success and growth of an organization are heavily dependent on the selection of the right candidates. The conventional approach of relying solely on resumes and academic backgrounds falls short in adequately evaluating a candidate's suitability for a job role.

At the heart of this research lies the identification of prospective employees' personality traits to evaluate their suitability for specific job roles. To accomplish this, the study leverages advanced Machine Learning and Natural Language Processing techniques.

K-means clustering is used to identify the personality clusters a candidate may belong to, while Natural Language Processing techniques are used to extract insights from candidate responses. These results are combined to determine the distribution of the Big Five personality traits among candidates. Importantly, this analysis assesses how closely these traits match the prerequisites of specific job roles. Supervised learning algorithms such as RandomForest, Naïve-Bayes and XGBoost are used to validate the accuracy of personality cluster predictions for candidates, where the Naïve-Bayes model emerges as the top performer.

The insights generated by this study are particularly valuable in the realm of Human Resources technology (HR tech), especially in the IT industry. By incorporating these findings, organizations can enhance their candidate selection processes, ultimately leading to better matches between employees and job roles, and, consequently, contributing to overall organizational success and growth.

Keywords— Recruitment, Big Five personality traits, Natural language processing, Machine Learning

ACKNOWLEDGEMENTS

I would like to offer my sincere gratitude to everyone who supported me in making my final-year research project a success. Firstly, I would like to extend my sincere gratitude to my project supervisor Dr. Anuradha Karunasena, and co-supervisor Dr. Lakmini Abeywardhane, for their guidance and the invaluable knowledge shared throughout the year to make the research successful. The success of this research would not have been possible without their support and constant guidance.

I would also like to express my gratitude to the Sri Lanka Institute of Information Technology and the CDAP team for providing us with the guidance and expertise to finish the project successfully. I also sincerely thank all the project group members, who worked with me and supported me to complete this research successfully.

Maleesha De Silva

Faculty of Computing,

Sri Lanka Institute of Information Technology.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
TABLE OF CONTENTS	v
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF EQUATIONS	ix
LIST OF ABBREVIATIONS	ix
1. INTRODUCTION	1
1.1 Background	1
1.1.1. The Big Five model	3
1.2. Literature Review	5
1.3. Research Gap	9
1.4. Research Problem	10
1.5. Research Objectives	12
2. METHODOLOGY	13
2.1. System Architecture Diagram	13
2.2. Data Collection	14
2.3. Model Training and Keyword Extraction	14
2.3.1. Self-rating responses dataset	15
2.3.2. Open-ended responses dataset	17
2.4. Determining Candidate Big Five Trait Distribution	21
2.5. Determining the expected Personality Requirement for the Job Role	23
2.6. Evaluating candidate personality - job role fit	24
2.7. Frontend Development	25
2.8. Commercialization Aspects of the Product	27
2.9. Testing and Implementation	28
3. RESULTS AND DISCUSSION	32

3.1. Results	32
3.2. Research Findings	34
3.3. Discussion.....	35
4. CONCLUSION.....	37
REFERENCES.....	38
APPENDICES.....	40

LIST OF TABLES

Table 1: Research gap	10
Table 2: Testing the predictions of the K-means clustering model	29
Table 3: Testing the keyword extraction	29
Table 4: Testing the mapping of job role requirements to Big Five trait	30
Table 5: Testing the final score calculation for candidate–job role fit	31
Table 6: Big Five trait distribution for the five personality clusters	32
Table 7: Model performance parameters – Personality prediction	33
Table 8: Comparing different methods of keyword extraction	33

LIST OF FIGURES

Figure 1: The Big Five Personality Traits	3
Figure 2: Personality traits used in peer-reviewed research in 2021	3
Figure 3: Survey results showing the significance of personality-job fit.....	11
Figure 4: Research objective – Personality assessment	12
Figure 5: System architecture diagram.....	13
Figure 6: Implementing the Elbow Visualization Technique	15
Figure 7: Fitting the K-means model	15
Figure 8: Visualization of the six personality clusters	16
Figure 9: Cluster predictions for dataset	16
Figure 10: Big Five trait distributions for the six personality clusters.....	17
Figure 11: Text cleaning of open-ended responses	17
Figure 12: Text tokenization of open-ended responses	18
Figure 13: Stopword removal from open-ended responses.....	18
Figure 14: Lemmatization of open-ended responses.....	18
Figure 15: Set of words from the open-ended responses	19
Figure 16: Keyword extraction using Bag of Words.....	20
Figure 17: Keyword extraction using TF-IDF	20
Figure 18: Keyword extraction using KeyBERT	21
Figure 19: Screenshot of keyword list extracted using TF-IDF approach	21
Figure 20: Traits required for the job role	23
Figure 21: Mapping the personality traits to Big Five traits	23
Figure 22: Obtaining the expected Big Five trait distribution	24
Figure 23: User interface – screenshot 01	25
Figure 24: User interface – screenshot 02.....	25
Figure 25: User interface – screenshot 03	26
Figure 26: Logo – Intellihire	27
Figure 27: Logo - SMMS Software Solutions	27

Figure 28: Elbow Visualization results	32
Figure 29: Radar graph - Candidate vs Expected traits.....	34

LIST OF EQUATIONS

Equation 1: Equation to calculate candidate’s open-ended score	22
Equation 2: Equation to calculate candidate’s self-rating score.....	22
Equation 3: Equation to calculate candidate’s personality score	22
Equation 4: Equation to calculate candidate’s personality-job role fit.....	24
Equation 5: Neuroticism score calculation.....	35

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
FWHR	Facial width-to-height ratio
PPM	Personality Prediction Model
BOW	Bag of Words
BERT	Bidirectional Encoder Representations from Transformers
TF-IDF	Term Frequency – Inverse Document Frequency

1. INTRODUCTION

1.1 Background

Selecting the right candidates for a job is a crucial task for any organization striving for success. However, the traditional recruitment process is often time-consuming, costly and possibly subjected to bias. Moreover, there's often a lack of a clear-cut method for thoroughly assessing vital aspects like a candidate's skills and personality traits and ensuring they align perfectly with the job requirements. This all highlights the need for a more efficient and organized approach to recruitment that takes all these important factors into account, ultimately leading to more effective and streamlined hiring practices.

The Information Technology (IT) industry is a rapidly evolving and promising field, not just in Sri Lanka but globally as well. According to data from the Sri Lanka Information and Communication Technology Agency (ICTA) [1], the country had set an ambitious goal of reaching 200,000 employees within the IT industry by 2022. This highlights the urgent need for recruitment systems that can make the hiring process quicker and less resource-intensive, especially for IT-related roles.

In response to these challenges, forward-thinking organizations have begun embracing automated recruitment systems. These systems utilize advanced technologies like machine learning, data extraction, and natural language processing to thoroughly examine resumes, assess candidates' skills, and even evaluate their personality traits. The ultimate aim of these automated systems is to speed up the recruitment process, boost its accuracy, and make it more cost-effective.

The focus of this research is not only to examine but also to genuinely comprehend the criteria used in candidate selection. It's not just about analysis; it's about creating an advanced automated system that can comprehensively evaluate candidates across all these criteria. The goal here is to identify and hire the best-fit individuals for specific job

positions, thereby contributing to more efficient and effective talent acquisition. This approach isn't just beneficial for organizations; it also serves the broader goal of building a workforce that's well-equipped to meet the ever-evolving demands of the IT industry and other sectors, all while ensuring fairness and inclusivity in the recruitment process.

This research puts forth the idea that using personality assessment can be a highly effective method for identifying the most suitable individuals for specific roles. Swedish psychologist Sofia Sjöberg reinforces this notion by highlighting a crucial point: the significant variations in job performance levels among individuals underscore the importance of personality assessment for organizations [2]. In essence, since everyone exhibits unique traits and characteristics, their job performance also differs. Therefore, utilizing personality assessment tools becomes essential for organizations striving to pinpoint and recruit individuals who possess the personality traits and qualities that align with the demands of a particular role. By doing so, organizations increase their likelihood of hiring individuals who can excel in their assigned tasks, ultimately contributing to higher overall performance within the company.

1.1.1. The Big Five model

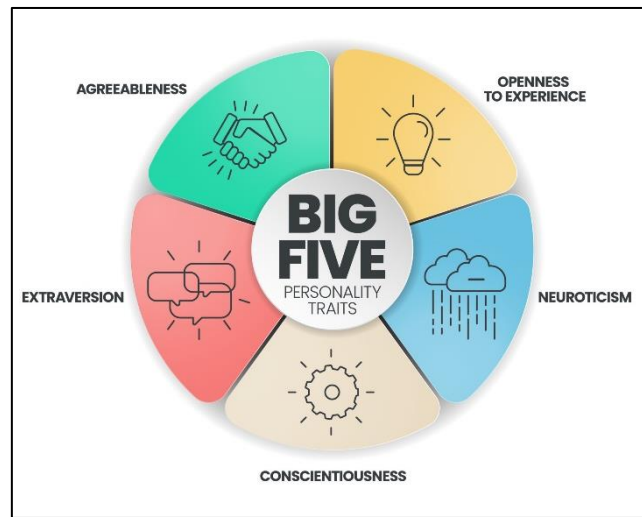
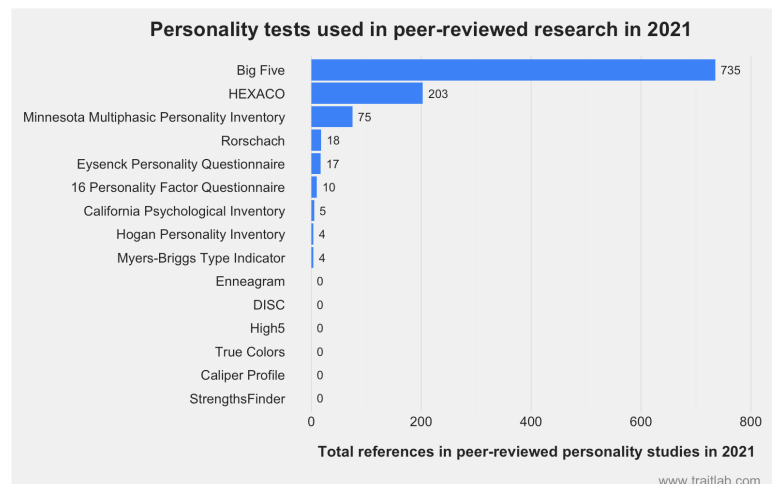


Figure 1: The Big Five Personality Traits

The Big Five Model is the preferred personality assessment model for this component, as it has emerged as the dominant model in modern personality research. Over the past five years, the Big Five Model has been featured in 735 scientific publications, surpassing the combined usage of all other assessment models[3]. This statistical evidence suggests that the Big Five Model is currently the most reliable option for predicting personality traits. Hence, it is considered the leading choice for this component.



Moreover, the Big Five traits are considered to be highly relevant to job performance and organizational success, making it a popular choice for employee selection and development.

1.2 Component Overview

While technical skills, professional expertise, and academic qualifications are undoubtedly crucial for evaluating job performance, they represent just one aspect of the multifaceted puzzle when determining if an individual is a perfect fit for a particular role or work environment. Apart from these qualifications, it's crucial to acknowledge how personality traits profoundly affect job suitability, overall job satisfaction, and overall success, benefiting both the individual and the organization.

Research and studies have shown the significant role that personality plays in predicting professional effectiveness, often even surpassing the importance of prior work experience [4]. In today's world of recruitment, more and more recruiters are using personality assessments as a valuable tool. These assessments help filter through applicants and pinpoint candidates whose inherent qualities align with the organizational requirements and the company's cultural dynamics.

Simply put, this change in perspective acknowledges that finding the perfect candidate is about more than just technical skills and qualifications. It's about understanding that qualities like adaptability, teamwork, leadership potential, and emotional intelligence play a crucial role in a candidate's success in a job and their fit within the organization. So, using personality assessments in hiring is like a thoughtful step toward a more complete and human approach to choosing candidates. This approach benefits both employers and job seekers by making the hiring process more holistic.

The focus of this research component is to determine the personality traits of a prospective employee with the purpose of evaluating their suitability for a particular job position. This

includes determining the existence of the big five traits and aligning the personality traits of the candidate with those anticipated in the job description.

1.2. Literature Review

In light of the growing interest in the science of psychology and its applications in a variety of fields, researchers have been investigating the use of technology and data analysis to predict an individual's personality traits. Over the past few years, machine learning algorithms and predictive models have been developed to analyze various types of data, such as social media activity, facial expressions, and linguistic patterns, to predict personality traits. This literature review aims to provide an overview of the present state of research on predicting personality traits using various types of data and machine learning algorithms.

The paper titled “DevFlair: A Framework to Automate the Pre-screening Process of Software Engineering Job” [5] (Research A) by Jayasekara et al. proposes a system to automate the initial screening of job applicants for software engineering positions. The writers begin by highlighting the difficulties that recruiters encounter when trying to hire software engineers, especially during the initial screening stage. They suggest that conventional screening methods like reviewing resumes and conducting technical interviews are tedious and can be biased. As a solution, the authors introduce ‘DevFlair’, a system that streamlines the screening process by examining a candidate's coding contributions online and utilizing artificial intelligence algorithms to forecast their suitability for the job. In this paper, PPM (Personality Prediction Model) is introduced as a tool for predicting a job applicant's Big Five personality trait distribution based on their LinkedIn profile. PPM retrieves the candidate's profile details from LinkedIn using the ProxyCurl API, pre-processes the data, and uses five different models to predict their Big-Five personality traits. The article describes how probabilities are computed for each trait

and how they are combined to determine a candidate's overall personality. Overall, the paper presents an interesting approach to addressing the challenges faced by recruiters in the software engineering industry and could potentially save recruiters time and resources while improving the accuracy of their pre-screening process.

Another research “Candidate Selection for the Interview using GitHub Profile and User Analysis for the Position of Software Engineer” [6] (Research B), proposes a pre-screening mechanism to minimize the time taken to conduct interviews for a number of candidates. Similar to the previous research, the authors note that traditional hiring processes can be time-consuming and subjective, and argue that recruiters for software engineering positions expect candidates to have a diverse skill set, including the ability to learn independently and solve problems, which cannot be assessed solely through their resume. Therefore, the proposed solution for pre-screening candidates for the position of software engineer is to analyze the candidates using a range of dimensions, including their GitHub account, academic transcript, letters of recommendation, LinkedIn profile, and personality prediction. The research mentioned above involves determining a candidate's personality traits through a phone call interview where the candidate answers open-ended questions. The responses are then transcribed, preprocessed, and analyzed using NLP techniques to predict the distribution of Big Five traits. The algorithms used are Random Forest, SVM, and Logistic Regression, and the study finds that Logistic Regression has the highest accuracy for all five traits. The results are visualized in the form of a radar graph where '1' signifies the presence of a trait, and '0' indicates its absence.

In the research paper titled “Interview Data Analysis using Machine Learning Techniques to Predict Personality Traits” [7] (Research C), the author uses prosodic features such as intensity, pitch and frequency to predict five personality traits; ‘Engaged’, ‘Excited’, ‘Friendly’, ‘Calm’ and ‘Speaking rate’ of interview candidates. The study uses audio-visual recordings from mock interviews conducted with MIT students as the dataset. The results showed that the prosodic features related to intensity played a significant role in predicting traits like "Engaged" and "Excited," while features like pitch and duration of pause were

more relevant in predicting "Friendly." Similarly, the study found that prosodic features related to pitch were important in predicting the trait of "Calm." After selecting the top prosodic features, the researchers applied three different regression models to determine the best method for predicting the personality traits. The study concluded that Decision Tree was the best choice for predicting the selected traits using the chosen prosodic features.

The research paper titled ‘A Multimodal Interviewee Evaluation Approach for Candidates Facing Video Interviews’ [8] uses The ‘First Impressions Challenge’ dataset, consisting of ten thousand clips extracted from three thousand YouTube videos, served as the foundation for this study. Its primary objective was to discern apparent personality traits from video recordings where individuals spoke directly to the camera. The methodology was multifaceted: visual features were extracted by capturing 15 random frames from each video, resizing them, and scaling them down; audio features were extracted from the audio waveforms of these videos. The model was a fusion of audio and visual processing. It initially used a bimodal time-distributed approach to independently analyze audio and visual traits. These were merged, and a stacked LSTM model with dropout layers was used to capture the temporal dimension of the videos. The final layer produced predictions for five personality traits with sigmoid activation. The integration of audio and visual data yielded higher predictive accuracy compared to previous studies that focused solely on visual information, marking a notable advancement in personality assessment from multimedia content.

Research done in 2022 titled “Personality Prediction for Online Interview” by S. K. Nivetha et al [9] (Research D) explores the application of machine learning algorithms for predicting the personality of job candidates during online interviews. The study focuses on the use of the ‘Big Five Personality Traits’ as the basis for personality prediction. The authors note that while online interviews have become a popular method for recruitment, they often lack the personal touch of face-to-face interviews, which can make it difficult to accurately assess a candidate's personality. The paper discusses the use of different

methods such as text, image and video analytics to forecast personality traits by applying technologies such as support vector machines (SVM) and neural networks. The study uses two approaches to identify the Big Five personality traits of candidates participating in online interviews. The first approach utilizes the facial height-to-width ratio (FWHR) of the candidates captured in their images, and a CNN-based model is trained to determine the traits. The second approach involves using a questionnaire consisting of situational questions to assess how a candidate might behave or respond in specific situations, and the responses are analyzed through K-means clustering to predict the presence of Big Five traits. According to the findings of this research, the K-Means clustering algorithm generated a result with 90% precision. Additionally, the use of FWHR and CNN to analyze a person's personality resulted in an accuracy of approximately 92%.

The paper titled “Predicting Personality Using Answers to Open-Ended Interview Questions” by M. Jayaratne and B. Jayatilleke [10] (Research E) aims to investigate the possibility of using open-ended interview questions to predict the personality traits of individuals. This research uses NLP and machine learning techniques to predict the personality characteristics of job applicants. The data for the study is gathered from an interview conducted through online chat. This research suggests that algorithms can determine a candidate's personality objectively by analyzing their answers, thus eliminating any subjective biases that arise from human interviewer assessments. The paper emphasizes the importance of job-personality suitability as it greatly affects job performance, job satisfaction, and duration of employment. According to the research, individuals tend to experience greater job satisfaction when their personality aligns with their chosen profession. However, the research also highlights that candidates may find a personality test cumbersome, and it adds expense to the hiring process. Therefore, the authors propose that interview responses are a more appropriate way of predicting a candidate's personality. The approach utilized in this study involved developing a regression model to estimate a score for each of the six characteristics in the HEXACO model by analyzing the written responses provided by the candidates in the form of open-ended questions. In the HEXACO-based self-rating questions, the candidate rates

themselves on a 5-point scale. 9 The final scores for the individual traits are calculated but taking the average over all the responses for each of the 6 traits.

In conclusion, this literature review suggests that several methods, such as social media platforms, FWHR, prosodic features, and responses to open-ended questions, have been utilized for candidate personality prediction. However, further improvements are needed to make these methods more reliable and ensure more accurate results.

1.3. Research Gap

Based on the literature survey done above, the following were identified as research gaps,

- Existing research has primarily concentrated on predicting the distribution of Big Five Personality traits among job candidates. However, there is a research gap in evaluating a candidate's fitness for a specific job role based on their personality traits. This study aims to address this gap by not only predicting candidates' Big Five Personality trait distribution but also assessing their compatibility with the job role they are applying for, with a focus on making hiring decisions more informed and human-centered.
- Majority of the studies have generally taken a broad approach, without specific industry focus. This research component focuses on examining the hiring of candidates in the IT industry, with a particular emphasis on identifying the essential personality traits required for success in different job roles within the IT sector.

	Uses the Big Five Model	Focused on the entire IT industry	Determines the presence and level of each trait	Assesses a candidate's suitability for a job based on their personality traits
Research A	✓	✗	✓	✗
Research B	✓	✗	✗	✗
Research C	✗	✗	✓	✗
Research D	✓	✗	✓	✗
Research E	✗	✗	✓	✗
Proposed system	✓	✓	✓	✓

Table 1: Research gap

1.4. Research Problem

Organizations are increasingly using personality prediction methods to improve their recruitment and selection processes. These methods can help them gain valuable insights into a candidate's personality traits, which can be useful in finding the right fit for a job. However, there is still a need to investigate how to effectively apply these methods to achieve desired outcomes. Organizations need to find accurate and reliable assessment methods that align with their specific needs and goals.

In addition to identifying accurate assessment methods, it is also crucial for organizations to understand how to assess the fit between a candidate's personality and a job role. A good personality-job fit can have a significant impact on an employee's performance, job satisfaction, and retention rate. Therefore, organizations need to develop strategies for assessing personality-job fit to ensure that new hires can succeed in their roles. Given the

above-mentioned considerations, this research problem aims to study how personality prediction methods can enhance recruitment and selection processes. The study will identify the best assessment methods for predicting candidate personality traits and explore strategies for evaluating personality-job fit. The results will be useful for organizations looking to improve their recruitment and selection processes and increase the success of their workforce.

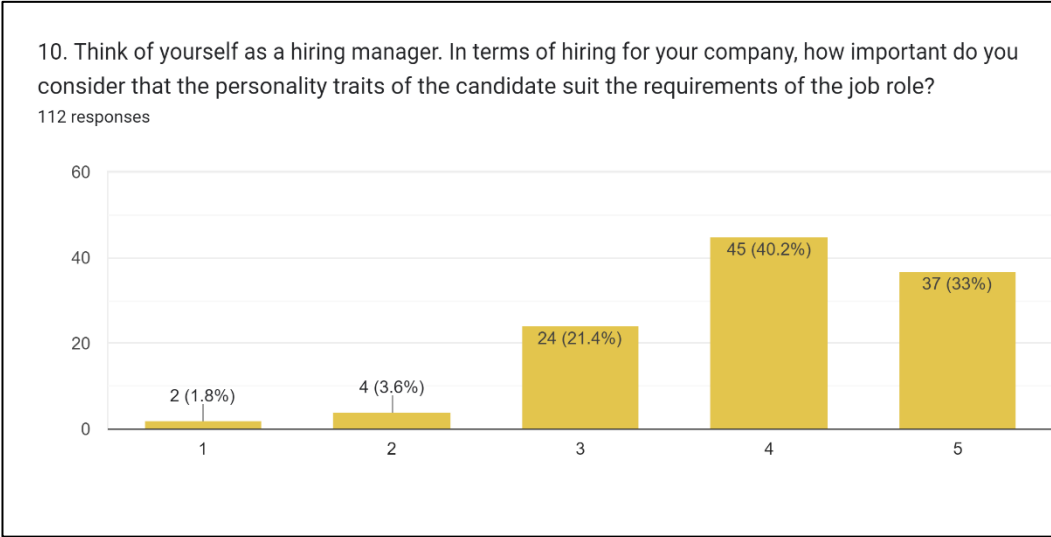


Figure 3: Survey results showing the significance of personality-job fit

Figure 3 depicted above illustrates the significance of personality traits in the workplace, as revealed by a survey conducted to assess their relevance in the professional environment. According to the results, around 80% of the respondents indicated their willingness to take into account a candidate's personality traits during the recruitment process.

1.5. Research Objectives

Main Objective

The primary objective of this research is to use machine learning and natural language processing (NLP) to gain comprehensive insights into a candidate's personality and evaluate how well their qualities align with the specific job they're seeking.

Specific Objectives

- **Develop an algorithm that identifies the presence of the Big Five Personality Traits of candidates.**

This objective aims to determine the presence and degree of big five traits in a prospective employee allowing the hiring manager to gain insight into the candidate's personality.

- **Develop an algorithm to assess how well the personality traits of the candidate aligns to suit the job role.**

This objective aims at determining the personality-job fit of a prospective employee. The candidate's personality traits will be evaluated against the personality traits deemed necessary by the hiring manager for the particular job position to generate a personality score that shows how compatible their personality traits are with the requirements of the job role.

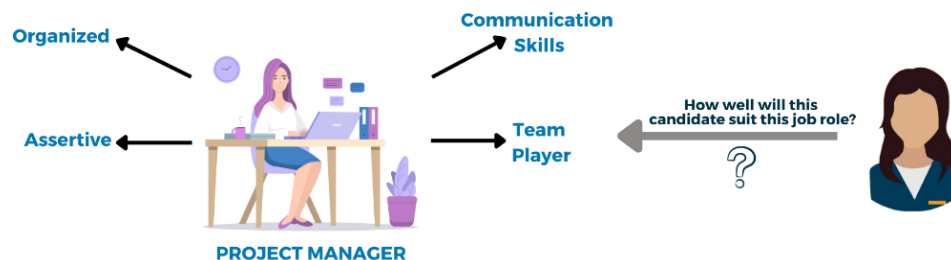


Figure 4: Research objective – Personality assessment

2. METHODOLOGY

2.1. System Architecture Diagram

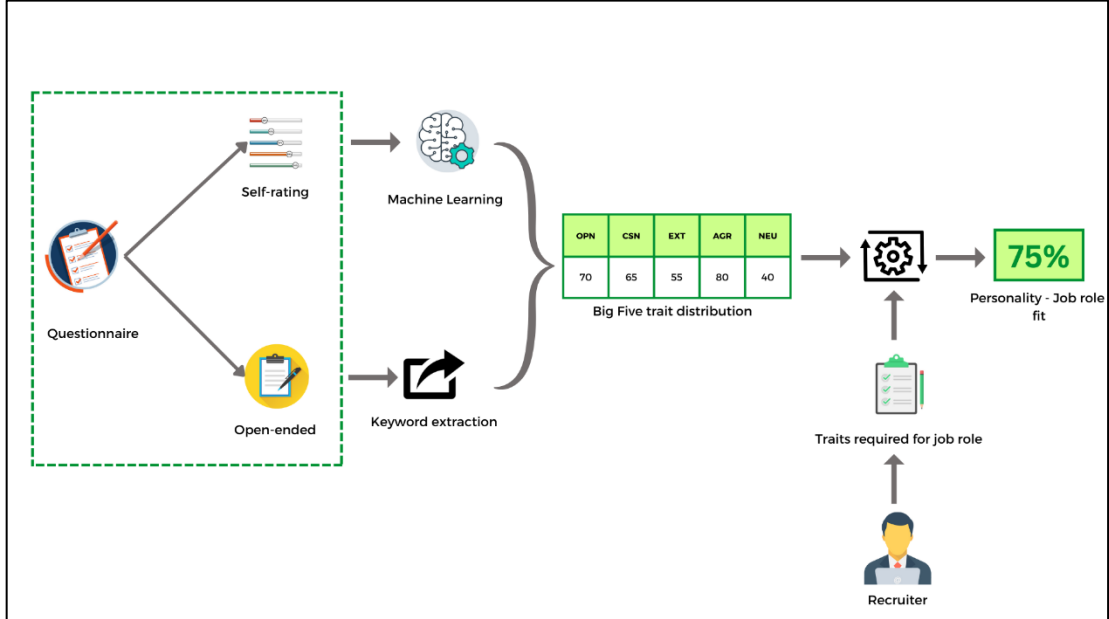


Figure 5: System architecture diagram

Figure 5 shown above presents the architecture diagram outlining the proposed methodology for personality assessment. Candidates will be assessed based on their responses to a questionnaire comprising two question types:

- Self-rating questions, where candidates rate their responses on a scale ranging from 1 (strongly disagree) to 5 (strongly agree).
- Open-ended questions, allowing candidates to freely express their thoughts and ideas.

This questionnaire serves as the foundation for determining the distribution of the Big Five personality traits in each candidate.

Simultaneously, the recruiter will outline the specific personality criteria essential for the job position that the candidate is seeking. These criteria will cover ten distinct personality

traits, which will be mapped to the Big Five personality traits to determine the desired Big Five trait distribution for the job role.

Ultimately, the candidate's distribution of Big Five traits will be compared against the expected trait distribution, resulting in a score that reflects the alignment. to the candidate indicating the degree to which the candidate suits the job role.

2.2. Data Collection

Two datasets were primarily utilized for this research:

- Self-rated questionnaire responses

This dataset was obtained from a publicly available dataset collection on 'Kaggle.' It comprises responses to twenty-five questions, with ten questions corresponding to each of the traits in the Big Five Model. The candidates' responses are rated on a scale ranging from 1 to 5.

Link to dataset: <https://www.kaggle.com/datasets/tunguz/big-five-personality-test>

- Open-ended questionnaire responses

This dataset was created by collecting responses through a Google Form survey. The survey consists of five questions, each designed to assess one of the Big Five traits.

2.3. Model Training and Keyword Extraction

2.3.1. Self-rating responses dataset

The dataset, which includes responses to self-rating questions, was used to determine the potential personality clusters to which a candidate might belong to. Initially, the dataset underwent preprocessing, followed by exploratory data analysis to gain insights into its characteristics.

To identify personality clusters within the dataset, the K-means clustering algorithm was used. K-means is an unsupervised machine-learning technique that groups data points into clusters based on similarity. In this context, each respondent's self-rating responses are used as data points, and the algorithm aims to group them into clusters of similar responses.

Before applying the K-means algorithm, it is essential to determine the optimal number of clusters, often referred to as the 'k-value'. This is done to ensure that the clustering results are meaningful. The 'Elbow Visualization' technique was used for this purpose. It involves running the K-means algorithm with different values of k (from 2 to 15 in this case) and plotting the variance explained by each number of clusters. The 'elbow' point on the graph is typically chosen as the optimal k-value, as it represents a balance between maximizing within-cluster similarity and minimizing the number of clusters.

```
[ ] # Visualize the elbow
    from sklearn.cluster import KMeans
    from yellowbrick.cluster import KElbowVisualizer

    kmeans = KMeans()
    visualizer = KElbowVisualizer(kmeans, k=(2,15))
    visualizer.fit(df_sample)
    visualizer.poof()
```

Figure 6: Implementing the Elbow Visualization Technique

```
from sklearn.cluster import KMeans

df_model = data_df.drop('country', axis=1)

# define 6 clusters and fit the model
kmeans = KMeans(n_clusters=6)
```

Figure 7: Fitting the K-means model

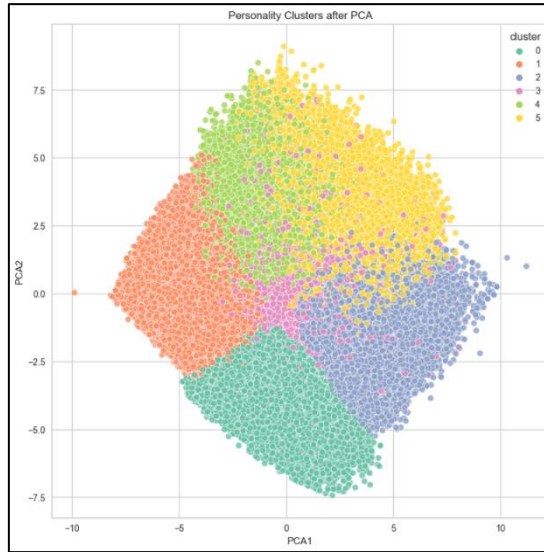


Figure 8: Visualization of the six personality clusters

Once the K-means model was trained, it was used to predict clusters for each response in the dataset. These cluster predictions underwent validation through the application of supervised learning algorithms like Random Forest, Naïve-Bayes, and XGBoost.

	EXT1	EXT2	EXT3	EXT4	EXT5	...	OPN2	OPN3	OPN4	OPN5	cluster
0	4.0	1.0	5.0	2.0	5.0	...	1.0	4.0	1.0	4.0	1
1	3.0	5.0	3.0	4.0	3.0	...	2.0	4.0	2.0	3.0	0
2	2.0	3.0	4.0	4.0	3.0	...	1.0	2.0	1.0	4.0	0
3	2.0	2.0	2.0	3.0	4.0	...	2.0	5.0	2.0	3.0	4
4	3.0	3.0	3.0	3.0	5.0	...	1.0	5.0	1.0	5.0	1
5	3.0	3.0	4.0	2.0	4.0	...	1.0	5.0	1.0	3.0	1
6	4.0	3.0	4.0	3.0	3.0	...	2.0	4.0	3.0	4.0	0
7	3.0	1.0	5.0	2.0	5.0	...	1.0	3.0	1.0	5.0	1
8	2.0	2.0	3.0	3.0	4.0	...	1.0	5.0	1.0	4.0	0
9	1.0	5.0	3.0	5.0	2.0	...	1.0	3.0	1.0	3.0	5

Figure 9: Cluster predictions for dataset

The average answer to each question group is obtained to identify the Big Five trait distribution for each cluster. This will be used to identify the Big Five trait distribution of candidates belonging to each cluster.

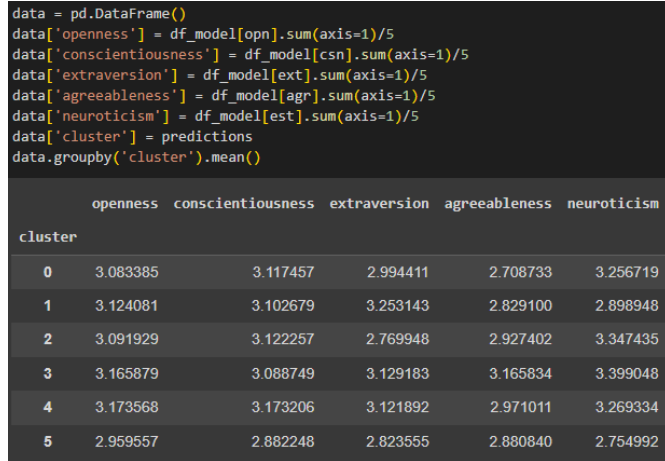


Figure 10: Big Five trait distributions for the six personality clusters

2.3.2. Open-ended responses dataset

The open-ended responses are processed using Natural Language Processing techniques to identify the keywords that can be used to determine the presence of each trait from candidate responses.

The first step in the process of keyword extraction is text cleaning which includes converting the text to lowercase and removing non-word and whitespace characters as well as digits.

```
# convert to lowercase
responses_df = responses_df.applymap(lambda x: x.lower() if isinstance(x, str) else x)

# remove non-word and non-whitespace characters
responses_df = responses_df.replace(to_replace=r'^\w\s', value='', regex=True)

# remove digits
responses_df = responses_df.replace(to_replace=r'\d', value='', regex=True)
```

Figure 11: Text cleaning of open-ended responses

Next, the responses undergo tokenization, where they are split into individual words using the ‘word_tokenize’ function.

```
[ ] # Tokenize the text data
responses_df['Openness'] = responses_df['Openness'].apply(word_tokenize)
responses_df['Conscientiousness'] = responses_df['Conscientiousness'].apply(word_tokenize)
responses_df['Extraversion'] = responses_df['Extraversion'].apply(word_tokenize)
responses_df['Agreeableness'] = responses_df['Agreeableness'].apply(word_tokenize)
responses_df['Neuroticism'] = responses_df['Neuroticism'].apply(word_tokenize)
```

Figure 12: Text tokenization of open-ended responses

Next, stopwords are removed from the tokenized words to remove commonly used words such as ‘the’, ‘an’, and ‘a’ that do not provide meaning.

```
[ ] # Remove stop words
stop_words = set(stopwords.words('english'))

responses_df['Openness'] = responses_df['Openness'].apply(lambda x: [word for word in x if word not in stop_words])
responses_df['Conscientiousness'] = responses_df['Conscientiousness'].apply(lambda x: [word for word in x if word not in stop_words])
responses_df['Extraversion'] = responses_df['Extraversion'].apply(lambda x: [word for word in x if word not in stop_words])
responses_df['Agreeableness'] = responses_df['Agreeableness'].apply(lambda x: [word for word in x if word not in stop_words])
responses_df['Neuroticism'] = responses_df['Neuroticism'].apply(lambda x: [word for word in x if word not in stop_words])
```

Figure 13: Stopword removal from open-ended responses

The next step of extracting keywords from the dataset is Lemmatization. Lemmatization helps in improving the accuracy of text analysis by reducing words to their base or dictionary form [11]. This makes it easier to identify and analyze words that have similar meanings. The ‘WordNetLemmatizer’ function of the nltk library is used for this purpose.

```
import nltk
from nltk.stem import WordNetLemmatizer
from nltk.corpus import wordnet
import pandas as pd

# initialize lemmatizer
lemmatizer = WordNetLemmatizer()

# define function to lemmatize tokens
def lemmatize_tokens(tokens):
    # convert POS tag to wordnet format
    def get_wordnet_pos(word):
        tag = nltk.pos_tag([word])[0][1][0].upper()
        tag_dict = {"J": wordnet.ADJ,
                    "N": wordnet.NOUN,
                    "V": wordnet.VERB,
                    "R": wordnet.ADV}
        return tag_dict.get(tag, wordnet.NOUN)

    # lemmatize tokens
    lemmas = [lemmatizer.lemmatize(token, get_wordnet_pos(token)) for token in tokens]

    # return lemmatized tokens as a list
    return lemmas

# apply lemmatization function to column of dataframe
responses_df['Openness_lemmatized'] = responses_df['Openness'].apply(lemmatize_tokens)
```

Figure 14: Lemmatization of open-ended responses

After completing the above steps, the resulting data consists of a set of words in each response.

Openness_lemmatized	Conscientiousness_lemmatized	Extraversion_lemmatized	Agreeableness_lemmatized	Neuroticism_lemmatized
[strongly, believe, seek, new, experience, tak...	[prioritize, work, accord, deadline, level, im...	[im, quite, comfortable, social, situation, al...	[try, view, feedback, criticism, opportunity, ...	[face, difficult, situation, try, remain, calm...
[someone, thrives, novelty, excitement, make, ...	[believe, accuracy, attention, detail, critica...	[im, outgo, person, nature, social, situation,...	[initially, tough, receive, criticism, negativ...	[experience, one, effective, way, handle, stre...
[naturally, inclined, seek, new, experience, t...	[manage, time, effectively, planning, ahead, a...	[someone, value, personal, relationship, im, a...	[handle, feedback, stay, openminded, actively,...	[deal, stressful, situation, like, take, proac...
[personal, professional, growth, two, top, pri...	[ensure, work, accurate, complete, time, follo...	[wouldnt, necessarily, describe, extroverted, ...	[take, feedback, seriously, also, try, take, p...	[find, deal, stressful, situation, communicati...
[natural, risktaker, recognize, value, seek, n...	[believe, set, realistic, goal, key, complete,...	[someone, naturally, extroverted, thrive, soci...	[view, feedback, opportunity, strengthen, rela...	[experience, effective, way, handle, stressful...
[someone, enjoys, take, risk, seek, new...	[accuracy, timeliness, essential,	[social, situation, always,	[take, proactive, approach,	[face, difficult, situation,

Figure 15: Set of words from the open-ended responses

The final step is to extract the keywords for each trait from the above data. These keywords will be the indicators of the presence of each trait. Keyword extraction was done using 3 methods. The top 40 keywords with the highest word frequencies were considered.

a) Keyword extraction using the Bag of Words (BOW) method

This method considers the term frequency of each word occurring in the dataset. The 40 words with the highest term frequency will be returned as the keywords.

```

# OPENNESS

# Create a bag of words representation of the text data
vectorizer = CountVectorizer()
bag_of_words = vectorizer.fit_transform(responses_df['Openness_lemmatized']).apply(lambda x: ' '.join(x))

# Get the sum of the counts of each word in the corpus
sum_words = bag_of_words.sum(axis=0)

# Get the frequency of each word in the corpus
word_freq = [(word, sum_words[0, idx]) for word, idx in vectorizer.vocabulary_.items()]

# Sort the list of words by frequency in descending order
word_freq = sorted(word_freq, key=lambda x: x[1], reverse=True)

# Print the top 40 most common words
print('Bag of words representation for Openness trait:\n')

Openness_top40 = []

for word, freq in word_freq[:40]:
    print(word, freq)
    Openness_top40.append(word)

```

Figure 16: Keyword extraction using Bag of Words

b) Keyword extraction using the TF-IDF method

This method considers both term frequency and inverse document frequency to identify keywords.

```

# create a TfidfVectorizer
tfidf = TfidfVectorizer()

# fit and transform the text column
tfidf_matrix = tfidf.fit_transform(df['Openness'])

# get the sum of the TF-IDF scores for each word
tfidf_sum = tfidf_matrix.sum(axis=0)

# convert the sum to a list of tuples (word, score)
word_scores = [(word, tfidf_sum[0, idx]) for word, idx in tfidf.vocabulary_.items()]

# sort the list by score (in descending order)
word_scores.sort(key=lambda x: x[1], reverse=True)

# print the top 40 most frequent words
print('Most frequently used words for Openness trait:\n')

Openness_tokens = []

for word, score in word_scores[:40]:
    print(f'{word}: {score}')
    Openness_tokens.append(word)

```

Figure 17: Keyword extraction using TF-IDF

c) Keyword extraction using the KeyBERT method

KeyBERT is a keyword extraction technique that uses BERT embeddings to identify the keywords that best capture the essence of the given text document. [12].

```
from keybert import KeyBERT

def extract_keywords_from_dataframe(df, text_column, num_keywords=5):
    # Initialize the KeyBERT model
    model = KeyBERT()

    # Concatenate all text phrases in the column into a single string
    text = ' '.join(df[text_column])

    # Extract keywords from the concatenated text
    keywords = model.extract_keywords(text, keyphrase_ngram_range=(1, 1), stop_words='english', top_n=num_keywords)

    # Return the extracted keywords
    return [keyword for keyword, _ in keywords]
```

Figure 18: Keyword extraction using KeyBERT

2.4.Determining Candidate Big Five Trait Distribution

After extracting the keywords, a keyword list was prepared for each keyword extraction method. These lists consist of 5 columns, each consisting of keywords that will be used to identify each of the Big Five traits.

	Openness_tokens	Conscientiousness_tokens	Extraversion_tokens	Agreeableness_tokens	Neuroticism_tokens
0	new	work	new	feedback	stay
1	risk	time	connection	try	situation
2	take	ensure	people	take	try
3	life	task	make	improve	help
4	im	complete	try	make	take
5	professional	project	others	believe	problem
6	personal	accurately	relationship	use	make
7	try	team	social	receive	find
8	believe	include	meeting	stay	stress
9	learn	also	situation	approach	solution
10	opportunity	make	able	understand	manage
11	potential	deadline	building	ask	difficult

Figure 19: Screenshot of keyword list extracted using TF-IDF approach

The candidate responses to the open-ended questions were subjected to the process of text cleaning, tokenization, stopword removal, and lemmatization. Afterward, keywords were extracted from the candidate responses using the three methods mentioned above.

The keywords extracted from the candidate responses are compared against the prepared keyword lists to identify common words. The score for each trait will be calculated as follows,

$$Openness_score(open_ended) = (number\ of\ common\ words / 40) * 5$$

Equation 1: Equation to calculate candidate's open-ended score

A score was calculated for each of the Big Five traits for the keywords extracted using all 3 keyword extraction methods. By looking at the scores, it was determined that TF-IDF keywords extraction method was the most successful in identifying the keywords that can be used to identify the presence of Big Five traits in candidates.

The score from the self-rating responses will be calculated by taking the average value for each question group as follows,

$$Openness_score(self_rating) = (OPN1 + OPN2 + OPN3 + OPN4 + OPN5)/5$$

Equation 2: Equation to calculate candidate's self-rating score

Finally, a single score is obtained by considering both the self-rating and open-ended scores of the candidate. The self-rating score is given a weight of 40% to the final score, whereas the open-ended score is given a weight of 60%.

$$Candidate_score(openness) = (self_rating(openness) * 0.4) + (open_ended(openness) * 0.6)$$

Equation 3: Equation to calculate candidate's personality score

2.5. Determining the expected Personality Requirement for the Job Role

The expected personality traits are taken as a list of ten traits needed in the workplace. These traits are rated on a scale of 1 to 5 by the recruiter to indicate the importance of each trait for a specific job role (Figure 20). Then, these ten traits are mapped to the Big Five traits [13] to obtain the expected Big Five trait distribution.

```

Innovative 5
Fast learner 4
Organization skills 3
Attention to detail 4
Assertiveness 2
Leadership skills 2
Team Player 5
Communication skills 3
Confidence 3
Adaptability to changes 4
Name: 0, dtype: object

```

Figure 20: Traits required for the job role

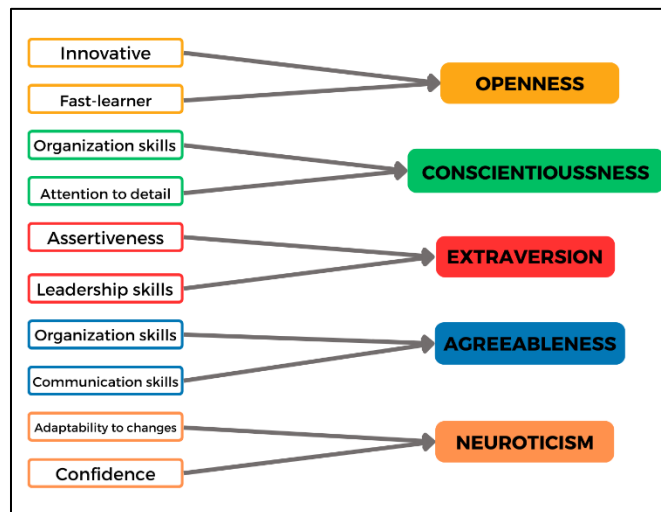


Figure 21: Mapping the personality traits to Big Five traits

```

# Map the personality requirements with Big Five Traits

expected['openness'] = [(data['Fast learner'] + data['Innovative'])/2]
expected['conscientiousness'] = [(data['Attention to detail'] + data['Organization skills'])/2]
expected['extraversion'] = [(data['Assertiveness'] + data['Leadership skills'])/2]
expected['agreeableness'] = [(data['Team Player'] + data['Communication skills'])/2]

# expected (anti-)neuroticism score
expected['neuroticism'] = [(data['Confidence'] + data['Adaptability to changes'])/2]

# Show the updated DataFrame
print(expected)

```

	openness	conscientiousness	extraversion	agreeableness	neuroticism
0	4.5	3.5	2.0	4.0	3.5

Figure 22: Obtaining the expected Big Five trait distribution

Figure 22 above depicts how the expected Big Five trait distribution was obtained. This distribution will be compared against the candidate's Big Five trait distribution to evaluate how closely the candidate's personality aligns with the requirements of the job role.

2.6. Evaluating candidate personality - job role fit

The final personality score for the candidate is calculated by taking into consideration the candidate's Big Five Trait distribution and the expected Big Five trait distribution obtained in the sections discussed above. The final score is calculated as follows,

$$\text{Score} = S \left(\frac{\text{trait}(\text{candidate})}{\text{trait}(\text{expected})} \right) * \frac{100}{5}$$

Equation 4: Equation to calculate candidate's personality-job role fit

2.7. Frontend Development

The frontend development utilized HTML, CSS, and JavaScript technologies, while Flask served as the backend engine to facilitate the connection between the frontend and the Machine Learning and NLP algorithms.

The primary objective in the development of user interfaces was to create user-friendly interfaces that enable recruiters to quickly gain comprehensive insights into a candidate's personality.

Candidate Name: Manushika Maldeniya					
	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Final Scores	4.49	2.89	3.43	3.62	1.96
Expected Scores	4.25	4.0	2.25	4.0	3.5
Candidate personality - Job role match: 79.75%					
Show details					

Figure 23: User interface – screenshot 01

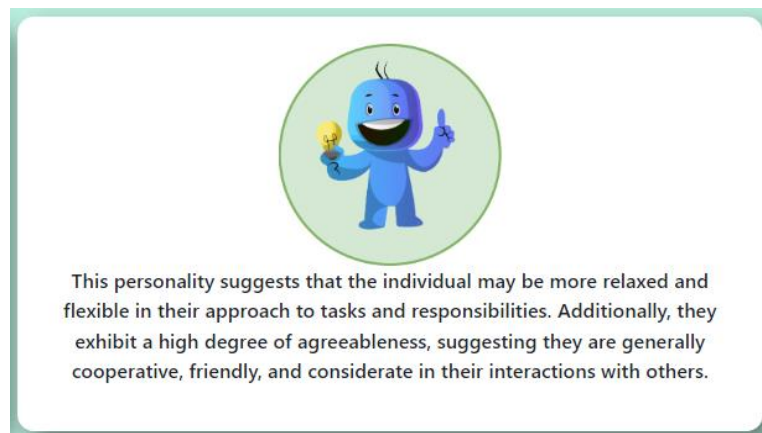


Figure 24: User interface – screenshot 02

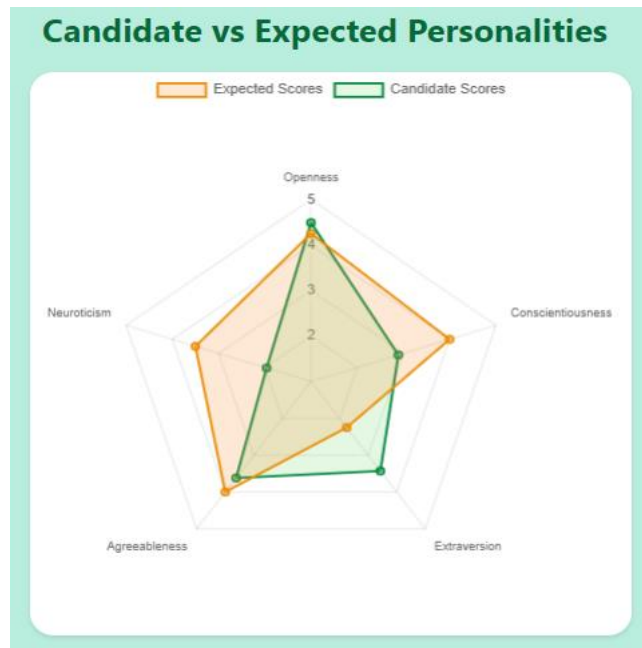


Figure 25: User interface – screenshot 03

2.8. Commercialization Aspects of the Product

‘Intellihire’ is a recruitment software created by ‘SMMS Software Solutions’ to simplify the hiring process in IT companies. Intellihire helps to streamline the recruitment process by using advanced machine learning algorithms and natural language processing techniques to analyze candidate profiles, job requirements, and other relevant data to identify the most suitable candidates for a particular job.

The personality prediction feature of Intellihire minimizes the effort required by hiring managers to evaluate a candidate's personality through informal conversations or HR interviews. This can save time in the hiring process by quickly providing insights into a candidate's personality traits and how they may fit into a particular role or team. Additionally, by ensuring that candidates are a good fit for the role and the company culture, a candidate personality prediction system can help reduce turnover rates and increase employee retention improving the hiring process of the organization and ultimately leading to a more productive and satisfied workforce.

Intellihire also features an intuitive user interface that makes it easy for hiring managers to navigate and manage the recruitment process. Additionally, the system can be customized to meet the specific needs of different organizations.



Figure 26: Logo – Intellihire



Figure 27: Logo - SMMS Software Solutions

2.9. Testing and Implementation

Testing was done to verify the functionality of all aspects within the personality prediction component.

This included:

1. Assessing the K-means clustering model's predictive capabilities for determining a candidate's personality cluster.
2. Validating the correct extraction of keywords from open-ended responses.
3. Confirming the accurate alignment of job role requirements with the Big Five traits.
4. Evaluating the precision of candidate personality-job role fit calculations.

Test case Number	01
Description	Testing the predictions of the K-means clustering model – when the candidate provides answers to the self-rating questions(on a scale of 1 to 5) the personality cluster the candidate belongs to should be displayed.
Input	3,2,1,1,2,1,2,2,1,4,2,5,4,4,4,5,5,4,4,4,5,3,3,4,5
Expected Output	A number from [0] to [5] (to represent the six clusters)
Actual Output	[3]

Test Result	Pass
-------------	------

Table 2: Testing the predictions of the K-means clustering model

Test case Number	02
Description	Testing the keyword extraction from open-ended responses
Input	“I am someone who enjoys taking on new challenges and seeking out new experiences.”
Expected Output	'someone', 'enjoy', 'take', 'new', 'challenge', 'seek', 'new', 'experiences.'
Actual Output	<code>['someone', 'enjoy', 'take', 'new', 'challenge', 'seek', 'new', 'experiences.']</code>
Test Result	Pass

Table 3: Testing the keyword extraction

Test case Number	03
Description	Testing the mapping of job role requirements to Big Five traits – when the recruiter provides the requirement for the job role it should be mapped to the Big Five traits

Input	Innovative 5 Fast learner 4 Organization skills 3 Attention to detail 4 Assertiveness 2 Leadership skills 2 Team Player 5 Communication skills 3 Confidence 3 Adaptability to changes 4
Expected Output	Openness 3.5 Conscientiousness 4.4 Extraversion 2.0 Agreeableness 4.0 Neuroticism 3.5
Actual Output	<div> openness conscientiousness extraversion agreeableness neuroticism 4.5 3.5 2.0 4.0 3.5 </div>
Test Result	Pass

Table 4: Testing the mapping of job role requirements to Big Five trait

Test case Number	04	
Description	Testing the final score calculation for candidate–job role fit	
Input	Expected scores	Candidate scores
	Openness 3.5 Conscientiousness 4.4 Extraversion 2.0 Agreeableness 4.0 Neuroticism 3.5	Openness 3.31 Conscientiousness 3.26 Extraversion 2.04 Agreeableness 3.86 Neuroticism 3.04
Expected Output	88.3	
Actual Output	Candidate personality score: 88.3	

Test Result	Pass
-------------	------

Table 5: Testing the final score calculation for candidate–job role fit

3. RESULTS AND DISCUSSION

3.1. Results

The Elbow Visualization technique revealed a k-value of 6 indicating that the candidates can be grouped into six distinct personality clusters.

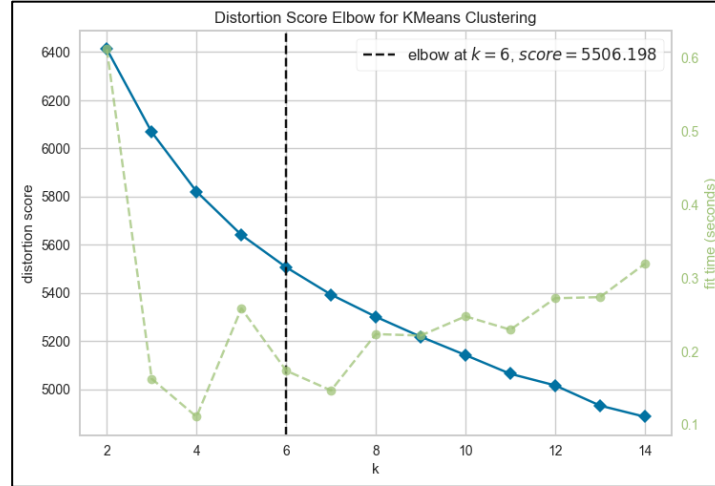


Figure 28: Elbow Visualization results

The Big Five trait distribution for the six clusters was obtained as shown below.

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Cluster 0	3.08	3.11	2.99	2.70	3.25
Cluster 1	3.12	3.10	3.25	2.83	2.89
Cluster 2	3.09	3.12	2.77	2.93	2.34
Cluster 3	3.16	3.08	3.12	3.17	2.39
Cluster 4	3.17	3.17	3.12	2.97	3.26
Cluster 5	2.96	2.88	2.82	2.88	2.75

Table 6: Big Five trait distribution for the five personality clusters

Identifying which cluster, a candidate belongs to would help recruiters to determine what kind of personality a candidate has.

The cluster predictions were validated using using RandomForest, Naïve-Bayes, and XGBoost algorithms. The performance parameters of the algorithms are shown below.

Model	Accuracy	Precision	Recall	F1 score
RandomForest	0.78	0.78	0.77	0.77
Naive-Bayes	0.78	0.78	0.78	0.78
XGBoost	0.76	0.75	0.75	0.75

Table 7: Model performance parameters – Personality prediction

Three different methods were utilized for keyword extraction and the extracted keywords were subsequently compared to the candidate responses to identify the most effective method. Table 8 below displays the scores obtained (rated on a scale of 1 to 5) for each keyword extraction method.

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Bag of Words	2.7	2.67	3.4	2.65	2.9
TF-IDF	3.31	2.8	3.2	3.5	2.6
KeyBERT	2.4	1.66	3.0	2.5	1.4

Table 8: Comparing different methods of keyword extraction

A radar graph is utilized to visually represent and compare the candidate's Big Five trait distribution with the expected Big Five trait distribution for the job role. This visualization provides recruiters with a quick and intuitive understanding of how well the candidate's personality aligns with the requirements of the job role.

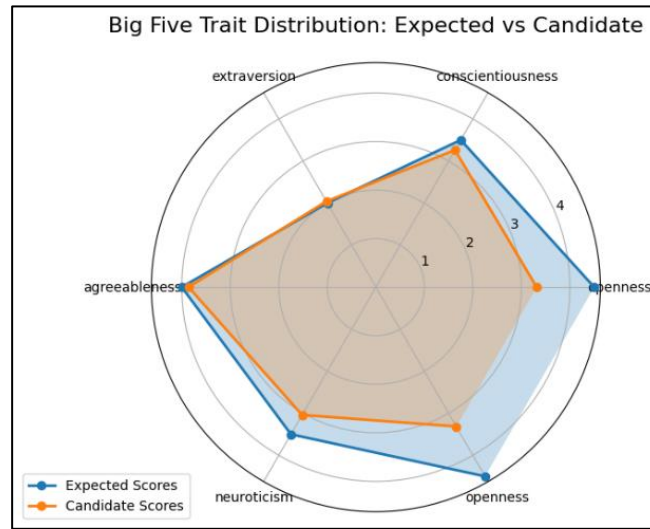


Figure 29: Radar graph - Candidate vs Expected traits

3.2. Research Findings

Based on the results shown in Table 7, the Naive-Bayes algorithm showed the most promising results in validating the cluster predictions.

The scores presented in Table 8 illustrate the performance of three different methods—Bag of Words, TF-IDF, and KeyBERT—in extracting keywords, with each trait rated on a scale of 1 to 5. Among these methods, TF-IDF demonstrated the most favorable outcomes, leading to its selection as the preferred method for generating keyword lists and extracting keywords from the candidate responses..

The sum of ratios between the candidate and the expected scores for each of the Big Five traits is taken as a percentage. As neuroticism is negatively correlated with job performance [5], the probability that the candidate does not possess the neuroticism personality trait is considered.

$$Candidate_score = OPN + CSN + EXT + AGR + (1 - NEU)$$

Equation 5: Neuroticism score calculation

The observation from the study revealed that every candidate falls into one of six distinct clusters, and within each of these clusters, the personality traits exhibit variations. These variations indicate that candidates within each cluster possess somewhat different personality profiles, which can make them more suitable for specific job positions or roles.

To visually represent and summarize these findings effectively, the study uses radar graphs. Radar graphs offer a clear and concise way to present complex information about the alignment between personality traits and job suitability. Each cluster's personality traits can be graphically displayed on a radar chart, allowing for an easy, at-a-glance assessment of how well candidates within that cluster match the requirements of various job roles. This visual representation simplifies the process of understanding the compatibility between a candidate's personality and specific job positions, providing valuable insights for decision-makers in the recruitment process.

3.3. Discussion

Initially, it was planned to include only open-ended questions in the questionnaire. However, it was realized that this approach might not be equitable for candidates who struggle with expressing themselves in English. It could put them at a disadvantage in the recruitment process. As a result, it was decided to make a slight adjustment by incorporating self-rating questions into the questionnaire. This modification ensures that recruiters have the chance to obtain a deeper and more accurate understanding of each candidate's personality.

The most closely related research conducted previously [5] employed 5 distinct machine learning models to identify the Big Five traits. However, the accuracy achieved by each of these models was lower compared to the accuracy attained by the Naïve-Bayes model discussed in this paper, thus demonstrating significant improvement over previous research.

Furthermore, this research introduces a novel element by evaluating the alignment between a candidate's personality and the job role. This represents a substantial advancement beyond the scope of prior related studies. In essence, it goes beyond merely predicting personality traits; we also assess how well a candidate's responses in the questionnaire correspond to their actual personality. This level of analysis adds depth and practicality to the assessment process, as it helps in understanding not only what personality traits a candidate possesses but also how well they suit a job position.

4. CONCLUSION

The central goal of this research is to revolutionize the recruitment process within the IT industry by embracing the transition towards automated techniques. This research component proposes a novel approach that centers on the assessment of an individual's personality traits. By doing so, recruiters can obtain a holistic and in-depth understanding of a candidate's character and how it aligns with the specific requirements of the job role they are seeking.

Traditional recruitment methods often rely on qualifications and experience alone, which can overlook crucial aspects of a candidate's suitability for a role, such as their work style, communication skills, and adaptability. The proposed innovative method seeks to bridge this gap by delving into the realm of personality traits.

Using personality trait assessments can provide us with valuable insights into a candidate's natural behaviors, how they approach problem-solving, their teamwork skills, and much more. This information helps us assess how well their personality traits match the requirements of the job. Essentially, it allows us to look beyond the surface and understand a candidate's potential for success and their ability to fit in with our organization in the long run.

The importance of this approach is that it has the potential to greatly improve the accuracy of the hiring process. Instead of relying solely on resumes and interviews, which may not always give a complete picture of a candidate's skills and abilities, we want to use data to make our evaluations. This data-driven method can help us make more informed and fair hiring choices, which can lower the chances of candidates not fitting well into their job roles. This research aims to bring about a fresh approach to hiring in the IT industry.

In future work, the inclusion of techniques to validate the information provided in the personality assessment questionnaire responses will significantly improve the accuracy of the proposed solution. This enhancement will prove advantageous for organizations seeking to identify the most fitting candidates for their workforce.

REFERENCES

- [1] “Sri Lanka aiming 200,000 ICT workforce by 2022,” *ICTA*, Aug. 08, 2019. <https://www.icta.lk/news/sri-lankaaiming-200000-ict-workforce-by-2022/>
- [2] S. Sjöberg, *Utilizing research in the practice of personnel selection: general mental ability, personality, and job performance*. Stockholm: Department of Psychology, Stockholm University, 2014.
- [3] “15 Personality Tests Compared for 2022.”
- [4] A. Furnham, K. V. Petrides, C. J. Jackson, and T. Cotter, “Do personality factors predict job satisfaction?,” *Personality and Individual Differences*, vol. 33, no. 8, pp. 1325–1342, Dec. 2002, doi: 10.1016/S0191-8869(02)00016-8.
- [5] R. T. R. Jayasekara, K. A. N. D. Kudarachchi, K. G. S. S. K. Kariyawasam, D. Rajapaksha, S. L. Jayasinghe, and S. Thelijjagoda, “DevFlair: A Framework to Automate the Pre-screening Process of Software Engineering Job Candidates,” in *2022 4th International Conference on Advancements in Computing (ICAC)*, Colombo, Sri Lanka: IEEE, Dec. 2022, pp. 288–293. doi: 10.1109/ICAC57685.2022.10025337.
- [6] R. G. U. S. Gajanayake, M. H. M. Hiras, P. I. N. Gunathunga, E. G. Janith Supun, A. Karunasenna, and P. Bandara, “Candidate Selection for the Interview using GitHub Profile and User Analysis for the Position of Software Engineer,” in *2020 2nd International Conference on Advancements in Computing (ICAC)*, Malabe, Sri Lanka: IEEE, Dec. 2020, pp. 168–173. doi: 10.1109/ICAC51239.2020.9357279.
- [7] S. Chopra and S. Urolagin, “Interview Data Analysis using Machine Learning Techniques to Predict Personality Traits,” in *2020 Seventh International Conference on Information Technology Trends (ITT)*, Abu Dhabi, United Arab Emirates: IEEE, Nov. 2020, pp. 48–53. doi: 10.1109/ITT51279.2020.9320879.
- [8] W. G. Y. Randika, M. T. A. R, K. L. O. G. Liyanage, A. Karunasena, and K. M. L. P. Weerasinghe, “A Multimodal Interviewee Evaluation Approach for Candidates Facing Video Interviews,” in *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India: IEEE, Oct. 2022, pp. 1–7. doi: 10.1109/ICCCNT54827.2022.9984585.

- [9] S. K. Nivetha, M. Geetha, R. S. Latha, K. Sneha, S. Sobika, and C. Yamuna, "Personality Prediction for Online Interview," in *2022 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India: IEEE, Jan. 2022, pp. 1–4. doi: 10.1109/ICCCI54379.2022.9740980.
- [10] M. Jayaratne and B. Jayatilleke, "Predicting Personality Using Answers to Open-Ended Interview Questions," *IEEE Access*, vol. 8, pp. 115345–115355, 2020, doi: 10.1109/ACCESS.2020.3004002.
- [11] Y. R, "Python | Lemmatization with NLTK." <https://www.geeksforgeeks.org/python-lemmatization-with-nltk/>
- [12] M. Q. Khan *et al.*, "Impact analysis of keyword extraction using contextual word embedding," *PeerJ Computer Science*, vol. 8, p. e967, May 2022, doi: 10.7717/peerj-cs.967.
- [13] J. Darby, "HR Blog," *thomas.co*, May 16, 2023. <https://www.thomas.co/resource-type/hr-blog>

APPENDICES

Appendix A: Turnitin Report

Turnitin Originality Report

Document Viewer

Processed on: 12-Sep-2023 17:16 +0530
ID: 2056674043
Word Count: 7317
Submitted: 10

Final Report (Draft) | IT20207854 By Maleesha De Silva

Similarity Index

11%

Similarity by Source

Internet Sources: 7%
Publications: 3%
Student Papers: 7%

include quoted

include bibliography

exclude small matches

mode: quickview (classic) report

print

download

2% match (student papers from 13-Oct-2021)
[Submitted to Sri Lanka Institute of Information Technology on 2021-10-13](#)

1% match (student papers from 12-Oct-2021)
[Submitted to Sri Lanka Institute of Information Technology on 2021-10-12](#)

1% match (R.T.R Jayasekara, K.A.N.D Kudarachchi, K.G.S.S.K Kariyawasam, Dilini Rajapaksha, S.L Jayasinghe, Samantha Thellijagoda. "DevFlair: A Framework to Automate the Pre-screening Process of Software Engineering Job Candidates", 2022 4th International Conference on Advancements in Computing (ICAC), 2022)
[R.T.R Jayasekara, K.A.N.D Kudarachchi, K.G.S.S.K Kariyawasam, Dilini Rajapaksha, S.L Jayasinghe, Samantha Thellijagoda. "DevFlair: A Framework to Automate the Pre-screening Process of Software Engineering Job Candidates", 2022 4th International Conference on Advancements in Computing \(ICAC\), 2022](#)

1% match (student papers from 15-Dec-2022)
[Submitted to Liverpool John Moores University on 2022-12-15](#)

<1% match (student papers from 11-Oct-2021)
[Submitted to Sri Lanka Institute of Information Technology on 2021-10-11](#)

<1% match (student papers from 03-Sep-2023)
[Submitted to Sri Lanka Institute of Information Technology on 2023-09-03](#)

<1% match (student papers from 21-Mar-2021)
[Submitted to Sri Lanka Institute of Information Technology on 2021-03-21](#)

<1% match (student papers from 24-Mar-2021)
[Submitted to Sri Lanka Institute of Information Technology on 2021-03-24](#)

Appendix B: Survey Questionnaire

10. Think of yourself as a hiring manager. In terms of hiring for your company, how important do you consider that the personality traits of the candidate suit the requirements of the job role? *

1

2

3

4

5

Not important

☐

☐

☐

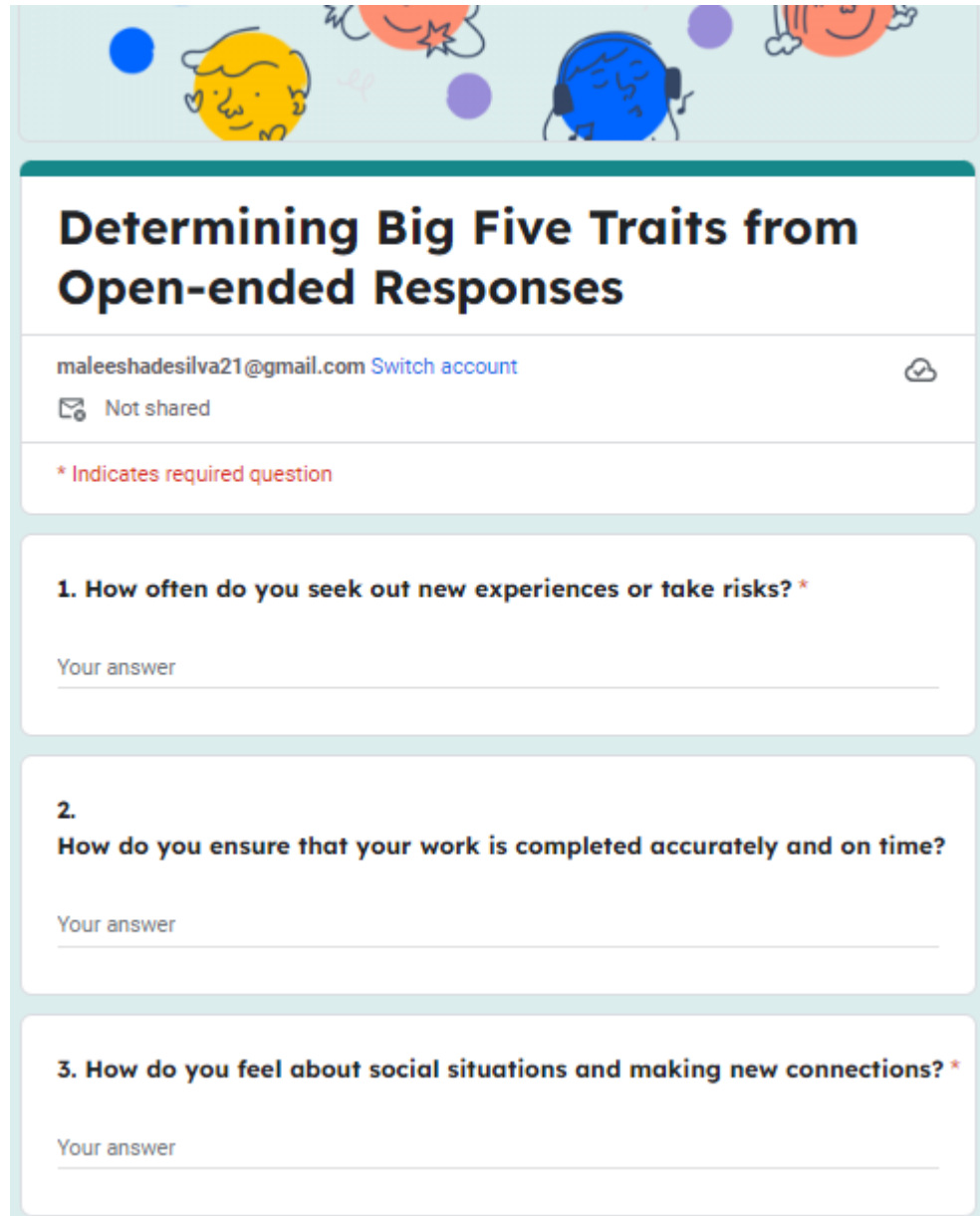
☐

☐

Very important

40

Appendix C: Google form used to prepare open-ended response dataset



The image shows a Google Form titled "Determining Big Five Traits from Open-ended Responses". The form is created by maleeshadesilva21@gmail.com and is not shared. It contains three required questions, each with a text input field for the answer.

Determining Big Five Traits from Open-ended Responses

maleeshadesilva21@gmail.com [Switch account](#)

Not shared

* Indicates required question

1. How often do you seek out new experiences or take risks? *

Your answer

2. How do you ensure that your work is completed accurately and on time?

Your answer

3. How do you feel about social situations and making new connections? *

Your answer

4. How do you handle feedback or criticism from others? *

Your answer

5. Can you tell me about a time when you had to *
deal with a difficult or stressful situation.
How did you handle it?

Your answer

Submit

Clear form