# PERSONALITY PREDICTION VIA CV ANALYSIS

## A MINI PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degree of

**Bachelor of Technology**

*in*

COMPUTER SCIENCE AND ENGINEERING
(ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)


**BY**

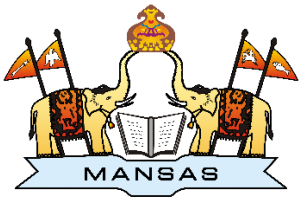
**Neerasa Bhoomika Sai**
**(22331A4241)**

**Dhilli Harshita**
**(22331A4212)**

**Koppoju Chandra Sekhar**
**(23335A4203)**

**Veerla Ramyasree**
**(23335A4207)**

**Under the Supervision of**
**Dr. P. Srinivasa Rao**
**Associate Professor**



**DEPARTMENT OF DATA ENGINEERING**

**MAHARAJ VIJAYARAM GAJAPATHI RAJ COLLEGE OF ENGINEERING**
**(Autonomous)**

**(Approved by AICTE, New Delhi, and permanently affiliated to JNTUGV, Vizianagaram), Listed u/s 2(f) & 12(B) of UGC Act 1956.**

**Vijayaram Nagar Campus, Chintalavalasa,Vizianagaram-535005, Andhra Pradesh**


**APRIL, 2025**

# CERTIFICATE


This is to certify that the project report entitled "**PERSONALITY PREDICTION VIA CV ANALYSIS**" being submitted by N.Bhoomika Sai(22331A4241), D.Harshita(22335A4212), K.Chandra Sekhar(23335A4207), V.Ramyasree(23335A4207) in partial fulfillment for the award of the degree of "**Bachelor of Technology"** in COMPUTER SCIENCE AND ENGINEERING(Artificial intelligence and machine learning) is a record of bonafide work done by them under my supervision during the academic year 2021-2022.

**Dr. P. Srinivasa Rao**
**Associate Professor,**
**Supervisor,**
Department of DE,
MVGR College of Engineering(A),
Vizianagaram.

**Dr. P Satheesh**
**Professor,**
**Head of the Department,**
Department of DE,
MVGR College of Engineering(A),
Vizianagaram.

# DECLARATION

We hereby declare that the work done on the dissertation entitled "**PERSONALITY PREDICTION VIA CV ANALYSIS**" has been carried out by us and submitted in partial fulfilment for the award of credits in Bachelor of Technology in Computer Science and Engineering of MVGR College of Engineering (Autonomous) and affiliated to JNTUGV, Vizianagaram. The various contents incorporated in the dissertation have not been submitted for the award of any degree of any other institution or university.


N.Bhoomika Sai
(22331A4241)

D.Harshita
(22331A4212)

K.Chandra Sekhar
(23335A4203)

V.Ramya Sree
(23335A4207)

# ACKNOWLEDGEMENTS

We express our sincere gratitude to **Dr. P. Srinivasa Rao** for his invaluable guidance and support as our mentor throughout the project. His unwavering commitment to excellence and constructive feedback motivated us to achieve our project goals. We are greatly indebted to him for his exceptional guidance.

Additionally, we extend our thanks to **Prof. P.S. Sitharama Raju (Director)**,**Prof. Ramakrishnan Ramesh (Principal)**, **Dr. P Satheesh (Head of the Department) and Dr. V. Jyothi (Project Coordinator)** for their unwavering support and assistance, which were instrumental in the successful completion of the project. We also acknowledge the dedicated assistance provided by all the staff members in the Department of **Data Engineering.** Finally, we appreciate the contributions of all those who directly or indirectly contributed to the successful execution of this endeavour.

N.Bhoomika Sai (22331A4241)

D.Harshitha (22331A4212)

K.Chandra Sekhar (23335A4203)

V.Ramya Sree (23335A4207)

# LAST MILE EXPERIENCE (LME)

**PROJECT TITLE**

## PERSONALITY PREDICTION VIA CV ANALYSIS

**BATCH NUMBER – 02**

**BATCH SIZE – 4**

**DEPARTMENT OF CSE(AIML)**


**Name:** N.Bhoomika Sai
**Email:**bhoomikaneerasa16@gmail.com
**Contact Number:**7981175287

**Name:** D.Harshita
**Email:**
dhilliharshita@gmail.com
**Contact Number:**6302790262

**Name:**K.Chandra Sekhar
**Email:**chandrasekhar20634@gmail.com
**Contact Number:** 6303163137

**Name:** V.Ramyas Sree
**Email:**ramyasreeveerla11@gmail.com
**Contact Number:**8688065801


**Project Supervisor**


**Name:** Dr. P. Srinivasa Rao
**Designation:** Associate Professor
**Email:** psr.cse@mvgrce.edu.in


### Project Objectives

1. To build an AI-based system that predicts a candidate's personality traits using their CV content.
2. To streamline and enhance the recruitment process by integrating machine learning models for automated personality profiling.

### Project Outcomes

1. Developed a functional web-based tool that analyzes uploaded CVs and predicts Big Five personality traits.
2. Enabled recruiters to make faster, data-driven hiring decisions by matching personality profiles with job roles.

### Domain of Specialisation

Personality Prediction Via CV Analysis uses Artificial Intelligence (AI) and Machine Learning (ML) in Human Resource (HR) Technology. It focuses on Natural Language Processing (NLP) for CV analysis, automated candidate profiling, and AI-driven recruitment processes**.**

### How your solution helping the domains?

Aids HR and recruitment domains by automating personality-based candidate profiling using AI, improving hiring accuracy and reducing manual effort.

### List the Program Outcomes (POs) that are being met by doing the project work

| | |
|---|---|
| **PO9: Individual and teamwork** | Combined individual and team efforts to predict personality traits from CVs, demonstrating effective collaboration in a multidisciplinary setup. |
| **PO10: Communication** | Well-planned and managed with efficient resource use and cost-effective tools, demonstrating leadership and teamwork. |
| **PO11: Project management and finance** | Planned and executed with proper task allocation and timeline management. Ensured efficient use of resources, with cost-effective open-source tools and strong team leadership. |
| **PO12: Life-long learning** | Highlighted the need for continuous learning by exploring new tools and techniques, helping the team adapt to evolving trends in AI and data science. |

### End Users of Your Solution

- Recruitment Teams - Responsible for hiring candidates.
- Software Developers & Researchers - studying e-recruitment processes.
- Job Seekers - Individuals looking to apply for jobs..
- Educational institutions for student placements.

# ABSTRACT

The personality of an individual plays a pivotal role in driving organizational progress and fostering self-development. Identifying a person's personality traits and professional suitability has traditionally been achieved through standard reviews or by analyzing their Curriculum Vitae (CV). However, conventional recruitment methods, which involve manually shortlisting resumes based on specific company requirements, are often time-intensive and prone to bias.

The system is designed as an online platform that facilitates candidate registration and conducts personality assessments using Multiple-Choice Question (MCQ) tests. These personality quizzes aim to provide insights into the behavioral and psychological traits of candidates. Simultaneously, the system evaluates professional eligibility by comparing uploaded CVs against a dataset trained using machine learning algorithms. The core algorithms employed in this process are Logistic Regression,Support Vector Machine,Decision Tree and Random Forest which are chosen for its robust decision-making capabilities and effectiveness in classification tasks. This ensures that recruitment decisions are data-driven and unbiased.

# CONTENTS

# List of Abbreviations

| | | |
|---|---|---|
| **AI** | **–** | Artificial Intelligence |
| **ML** | **–** | Machine Learning |
| **NLP** | **–** | Natural language processing |
| **CV** | **–** | Curriculum Vitae |
| **SVM** | **–** | Support Vector Machine |
| **RF** | **–** | Random Forest |
| **DT** | **–** | Decision Tree |
| **LR** | **–** | Logistic Regression |
| **UI** | **–** | User Interface |
| **TP** | – | True Positives |
| **TN** | – | True Negatives |
| **FP** | – | False Positives |
| **FN** | – | False Negatives |

# List of Figures

# List Of Tables

# INTRODUCTION

Recruitment is a crucial process for organizations to identify and onboard the best talent suited for their needs.As far as employment is considered, selecting the right candidate for the recruitment process from a vast pool of candidates has been a fundamental issue.When there is a requirement in the company, they receive thousands of applications and it is very difficult for the people of the company to go through a lot of CV's.The traditional recruitment involved time-intensive methods such as manually sifting through resumes, conducting multiple rounds of in-person interviews, and relying on subjective evaluations of candidate's qualifications and personalities. These processes were not only labor-intensive but also prone to human biases, inefficiencies, and inaccuracies.

When an organization posts a job opening, it may receive a large number of applications from job portals, advertisements, job consultancies and other sources. The HR department of the company finds it extremely difficult to sort through all of the resumes, address each application and identify the best candidate for the position by following standard processes like interviews, meetings, and specialized testing. Thus, during the actual primary round, they sort through the applicants based on a variety of factors, such as their suitability for the position, their abilities, any poorly chosen CVs and the competitor's skills. Thus, in an effort to reduce the inconvenience associated with the hiring process, we would like to suggest a novel idea in which the standard method of selecting and summarizing applications becomes simpler for the associations by using character expectations through CV examination.

The advent of e-recruitment systems has introduced features like automated resume parsing, online aptitude and personality assessments, and machine learning-driven candidate shortlisting. These systems enable recruiters to handle a larger volume of applicants efficiently while improving the accuracy and fairness of candidate evaluations.

## 1.1 Identification of seriousness of the problem

The recruitment process plays a crucial role in hiring the right candidates for an organization. However, traditional hiring methods have multiple limitations, making recruitment a time-consuming, inefficient, and biased process. Human-based evaluations often lead to inconsistent candidate assessments, as different recruiters may have varying opinions on the same candidate. Existing AI-based hiring models focus mainly on keyword matching, which fails to analyze a candidate's soft skills, personality traits, and behavioral attributes. This

limitation makes it difficult for companies to identify the right candidates beyond just technical skills and qualifications.The lack of an effective and automated approach to personality prediction in recruitment results in inefficient shortlisting, overlooked talent, and potential biases in the hiring process.

### 1.2 Problem Definition

Manual CV screening is time-consuming, prone to bias, and lacks a standardized approach to evaluating personality traits and soft skills .While AI-based hiring systems exist, most rely on keyword matching, which fails to capture behavioral characteristics essential for job suitability. Identifying the right candidates based on both technical skills and personality remains a challenge for recruiters.

### 1.3 Objective

The main objective of this project is to develop an automated personality prediction system using Machine Learning (ML) and Natural Language Processing (NLP) to improve the recruitment process. The system will analyze CV data to extract relevant features and predict personality traits, enabling a more efficient, unbiased, and data-driven candidate shortlisting process.

### 1.4 Existing Models & Limitations

Several AI-based recruitment models have been developed to automate hiring, but they primarily rely on resume parsing, keyword-based screening, and chatbot-based assessments. Resume parsing systems extract structured data from CVs but do not analyze personality traits or behavioral attributes. Keyword-based screening matches resumes with job descriptions but lacks contextual understanding, often filtering candidates based on keyword presence rather than actual relevance. Chatbot-based hiring assistants assist in initial screening; however, they rely on predefined responses and fail to provide deep personality insights.

# LITERATURE SURVEY

Xinyu Shui.,et al., aimed to determine the individual's Big-Five personality traits by physiological signals. A bracelet was attached to the people which tracks their heart rate and their daily activities were observed. They used linear regression analysis to predict an individual scores. Their findings shows that individuals show a good response in multiple situations where they are dependent on people rather than in single faced situation.

Nandani Agarwal.,et al., developed an automated system that predicts personality traits from resumes using various ML algorithms like Decision Tree, Random Forest, k nearest neighbor and Naive Bayes for effectively predicting personality through CV (Curriculum Vitae) analysis. Big five personality models are used to analyze personality.

Jayashree Rout.,et al., introduced a model which checks the candidate's personality and aptitude scores. They used Five Factor Model to predict the personality and TF-IDF to identify the significant text from a document. This helped the recruiters to make fair decisions based on the test scores.

Lakshyajit Thapa.,et al. addressed a system for predicting the personality which helped the automated system. This framework collected the data of the user from various platforms like CV's, social media and their responses to the tests conducted and predicted their traits by using Big Five, Myers Briggs and HEXACO personality models. They also compared various ML algorithms to develop models which can predict and help the recruiters to select the candidate similar to their job specifications.

G.Sudha.,Sasipriya. et al], developed an online platform for registering the candidate details and analysing their personality through an online test containing MCQ type personality quizzes. Then their CV is matched to the job requirements by using a trained dataset which is trained by Logistic Regression. This platform helped with better hiring satisfaction for the recruiters.

Atharva Kulkarni.,et al. predicted the personality based on the score obtained from labels of Big Five model. This approach parsed the entire CV and looks for texts matching to OCEAN model. In order to identify the text NLP. NLTK, SpaCy were used and for predicting KNN, SVM, Naïve Bayes and Random Forest algorithms were used which resulted in low accuracy.

Muthu Selvi M., Angeline Ranjitha Mani.,et al. designed a framework which reduces the workload on humans. This framework helped the recruiters to identify suitable candidate for a required role. The administrator can easily select a candidate based on the test results by using various clustering techniques without even an option to go through the CV.

Amal Khalifa Al Aamer.,et al , study found that AI technology can efficiently improve the process of hiring with many advantages considering limited costs, enhanced accuracy and reduced human work for scrutinizing the candidates.

Anum Jaffar.,et al., proposed system that determines the traits using non verbal movements like facial expressions, head and body postures. Few questions with expert analysis and a dataset which recognizes the facial movements is trained using CNN.

Madhura Jayaratne, Buddhi Jayatilleke.,et al. gathered data from 46,000 applicants who attempted an online interview through chat which also included personality questions based on HEXACO model. Using this HEXACO and Random Forest, they built a model that predicts the personality.

# 3. THEORETICAL BACKGROUND

## 3.1 Personality Prediction in Recruitment

Personality plays a crucial role in determining a candidate's suitability for a job. Beyond technical skills and experience, employers seek individuals who align with their company culture and job requirements. Traditional hiring methods rely on interviews and psychometric tests to assess personality traits, but these approaches can be subjective, time-consuming, and inconsistent.

### 3.1.1 Importactance Of Personality Prediction In Hiring

Personality traits significantly influence an individual's job performance, teamwork, and adaptability. A strong personality-job fit leads to higher productivity, job satisfaction, and employee retention. Traditional hiring methods often fail to objectively evaluate soft skills, leadership potential, and problem-solving abilities. Automated personality prediction helps overcome these limitations by providing consistent, unbiased, and scalable personality assessments. It enables recruiters to identify candidates who not only meet technical qualifications but also align with the organization's values and culture, ensuring a more effective hiring process.

### 3.1.2 Traditional vs AI-Based Personality Assessment

**Traditional Personality Assessment**

Traditional hiring methods rely on interviews, psychometric tests, and behavioral questionnaires to evaluate a candidate's personality. These assessments are time-consuming, subjective, and prone to bias, as different interviewers may interpret responses differently. Additionally, manual personality evaluations often focus on limited traits, overlooking deeper behavioral insights that influence job performance.

**AI-Based Personality Assessment**

AI-based personality assessment leverages Machine Learning (ML) and Natural Language Processing (NLP) to analyze CVs, written responses, and text patterns to predict personality traits. Unlike traditional methods, AI-driven models provide objective, data-driven, and scalable personality evaluations. These systems assess behavioral indicators, communication style, and sentiment analysis from CV text, ensuring consistency and efficiency in hiring. AI-based methods reduce human bias, automate candidate shortlisting, and improve hiring accuracy, making recruitment more efficient and data-driven.

### 3.1.3 Role Of Curriculum Vitea(CV) in Personality Prediction

A Curriculum Vitae (CV) is not just a document listing a candidate's qualifications and experience; it also provides valuable insights into their personality, communication style, and work approach. The way a CV is structured, the choice of words, and the descriptions of achievements can reflect key personality traits such as openness, conscientiousness, extraversion, agreeableness, and emotional stability.

## 3.2 Machine Learning in Personality Prediction

Machine Learning (ML) plays a crucial role in automating and improving personality prediction from various data sources, such as CVs, social media profiles, and psychometric tests. Traditional hiring methods rely on human judgment, which can be subjective and inconsistent. In contrast, ML models analyze patterns in text, language, and past hiring data to provide objective and data-driven personality assessments.

Using Natural Language Processing (NLP), classification algorithms, and deep learning models, ML can identify key personality traits based on established frameworks like the Big Five Personality Traits (OCEAN model). These models process textual data from CVs and extract behavioral indicators, helping recruiters assess candidates beyond just skills and experience. By leveraging ML, companies can achieve faster, unbiased, and more reliable personality predictions, enhancing the overall hiring process.

### 3.2.1 What is Machine Learning?

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that enables computers to learn from data and make predictions or decisions without being explicitly programmed. Instead of following predefined rules, ML models identify patterns in data and improve their performance over time.In the context of personality prediction, ML algorithms analyze text data from CVs to detect patterns that correlate with different personality traits. By using Natural Language Processing (NLP) and classification techniques, ML helps automate and enhance the accuracy of personality assessments in recruitment.

### 3.2.2 Why Machine Learning For Personality Prediction?

Traditional personality assessment methods, such as manual CV screening and psychometric tests, are time-consuming, subjective, and prone to human bias. Machine Learning (ML) provides an automated, data-driven, and objective approach to personality prediction, making the hiring process more efficient and reliable.

ML models analyze CV text, writing style, and word choices using Natural Language Processing (NLP) to detect behavioral patterns and personality traits. Unlike traditional methods, ML can process large volumes of CVs quickly, ensuring consistent and unbiased

evaluations. By leveraging ML, recruiters can identify the best candidates based on both technical skills and personality compatibility, improving overall hiring decisions.

### 3.2.3 Machine Learning vs Traditional Psychometric Tests

**Traditional Psychometric Tests**

Traditional psychometric tests assess personality using questionnaires and self-reported answers, measuring traits based on established models like the Big Five Personality Traits (OCEAN model). While effective, these tests are time-consuming, prone to human bias, and can be manipulated by candidates to give socially desirable responses. Additionally, administering and evaluating these tests require significant effort and resources.

**Machine Learning-Based Personality Prediction**

Machine Learning (ML) automates personality assessment by analyzing CV text, writing patterns, and word usage using Natural Language Processing (NLP). Unlike psychometric tests, ML-based models can extract hidden personality traits without requiring direct user input. They are faster, unbiased, and scalable, allowing recruiters to assess thousands of candidates efficiently. ML models improve over time with more data, making them more adaptive and accurate compared to traditional psychometric approaches.

## 3.3 Natural Language Processing (NLP) for CV Analysis

NLP plays a crucial role in extracting valuable insights from unstructured text data, helping in personality prediction and candidate evaluation. Using NLP techniques, CVs can be analyzed for word choice, sentence structure, and writing style, which reflect a candidate's personality traits, communication skills, and professional experience.

### 3.3.1 What is NLP?

Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that focuses on enabling computers to understand, interpret, and process human language. It combines linguistics and machine learning to analyze text and speech, allowing machines to extract meaningful insights from unstructured data.

In the context of CV analysis, NLP helps in extracting skills, experience, and personality traits by analyzing word choice, sentence structure, and writing patterns. This makes recruitment more efficient, objective, and data-driven by automating the evaluation of candidates based on their CV content.

### 3.3.2 NLP Techniques for Resume Parsing

Resume parsing involves extracting and structuring information from CVs using Natural Language Processing (NLP) techniques. These techniques help in analyzing text data to identify key details such as skills, experience, education, and personality traits.

**Key NLP Techniques for Resume Parsing**:

**Tokenization** – Splitting text into words or sentences for analysis.

**Stopword Removal** – Eliminating common words (e.g., "the," "and") to focus on meaningful terms

**Named Entity Recognition (NER**) – Identifying names, job titles, company names, and skills.

**TF-IDF (Term Frequency-Inverse Document Frequency)** – Identifying important words in a resume based on their frequency.

### 3.3.3 Feature Extraction from CVs using NLP

Feature extraction is a critical step in Natural Language Processing (NLP) that converts unstructured text data from CVs into structured information for analysis. By examining a candidate's resume, NLP techniques can identify key attributes such as skills, job roles, educational background, and personality traits. One of the most common methods involves keyword extraction, where relevant terms related to a candidate's expertise and experience are identified. Named Entity Recognition (NER) helps detect important details such as company names, job positions, and technical skills, making the recruitment process more efficient.

Another important technique is Term Frequency-Inverse Document Frequency (TF-IDF), which helps rank words based on their significance in the document, ensuring that important qualifications stand out. Additionally, sentiment analysis can assess a candidate's self-description and communication style, providing insights into their confidence and personality traits.

## 3.4 Machine Learning Models Used

### 3.4.1 Logistic Regression for Personality Prediction

Logistic Regression is a widely used classification algorithm in Machine Learning that predicts categorical outcomes based on input features. In personality prediction, Logistic Regression helps classify candidates into different personality traits by analyzing their CV text data. It is particularly useful when predicting binary or multi-class personality traits, such as whether a candidate is highly conscientious or not. Its simplicity, efficiency, and interpretability make it a strong choice for text-based classification tasks. However, since

personality traits are often complex and nonlinear, Logistic Regression may be combined with other advanced models for improved accuracy.

### 3.4.2 Decision Tree Classification

A Decision Tree is a supervised learning algorithm that classifies data by splitting it into branches based on feature importance. In personality prediction, it analyzes CV text features such as word choice, sentence length, and key phrases to make classification decisions. Each split in the tree is based on a condition, leading to different personality categories. Decision Trees are easy to interpret and work well with structured data but may suffer from overfitting, making them less reliable with complex text data.

### 3.4.3 Random Forest for Personality Prediction

Random Forest is an ensemble learning method that combines multiple Decision Trees to improve classification accuracy. Instead of relying on a single tree, it builds several trees using random subsets of data and averages their predictions. This reduces overfitting and enhances generalization, making Random Forest more effective than a single Decision Tree. In personality prediction, Random Forest can process large amounts of CV data, improving classification accuracy by analyzing multiple linguistic and behavioral patterns.

### 3.4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful classification algorithm used in text-based prediction tasks, including personality prediction from CVs. SVM works by finding the optimal hyperplane that best separates different personality classes based on extracted text features. It is particularly effective in handling high-dimensional data, making it well-suited for Natural Language Processing (NLP) applications.

## 3.5. Personality Traits and Theories

Personality traits define an individual's behavior, thinking patterns, and interactions with others. Various psychological theories have been developed to categorize personality traits and understand human behavior. In personality prediction, these theories provide a structured framework for analyzing and classifying individuals based on linguistic patterns and textual data extracted from CVs.

One of the most widely accepted models in personality research is the Big Five Personality Traits (OCEAN model), which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This model is commonly used in AI-driven personality prediction, as it provides measurable traits that can be linked to job performance and workplace behavior.

# 4. APPROACH DESCRIPTION

## 4.1. Approach Flow

The personality prediction system follows a structured flow, starting from data collection to personality classification and candidate shortlisting. The process begins with CV submission, where users upload their resumes to the system. The CV text is then extracted and preprocessed using Natural Language Processing (NLP) techniques such as tokenization, stopword removal, and named entity recognition (NER).

Once the data is cleaned, feature extraction is performed using methods like TF-IDF and word embeddings, converting text into numerical representations for Machine Learning (ML) models. The processed data is then fed into classification models such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) to predict personality traits based on established theories like the Big Five Personality Model.

Finally, the predicted personality traits are matched with job requirements, and the system shortlists the most suitable candidates. The results are displayed through a user-friendly interface, allowing recruiters to make informed hiring decisions based on both technical skills and personality compatibility.

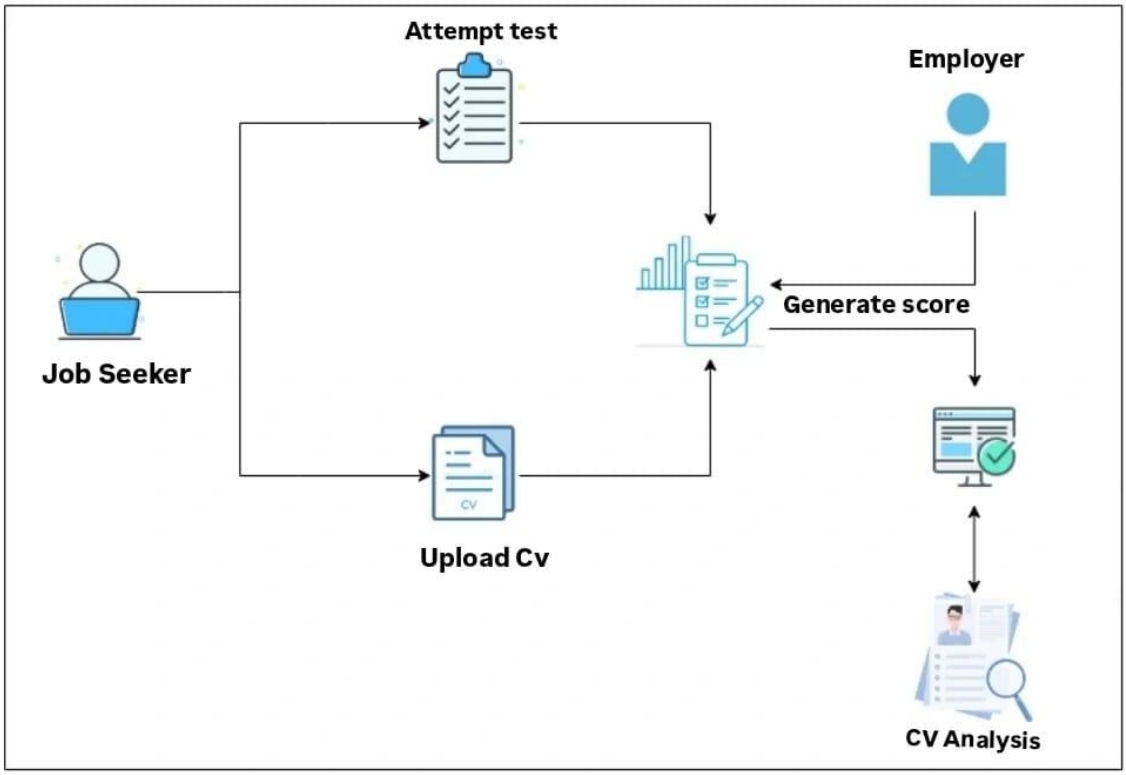



**Figure 4.1. System Model**

### 4.2 Admin Module

The Admin Module manages the recruitment system, oversees personality prediction, and ensures efficient hiring. With secure access, the admin can create and manage aptitude and personality test questions, define job requirements, and review candidate assessments. Using AI-driven recommendations, the admin shortlists candidates, reducing bias and improving decision-making through Machine Learning (ML) and Natural Language Processing (NLP).

1. System Oversight: Manages recruitment and ensures smooth operations.
2. Test Management: Creates and maintains aptitude and personality assessments
3. Job & Skills Management: Defines job roles, required skills, and personality traits
4. Candidate Evaluation: Reviews test results, personality predictions, and CVs.

### 4.3 User Module

The User Module allows candidates to interact with the system by submitting their CVs and taking aptitude and personality tests. Users begin by registering and logging into the system, after which they can upload their CVs for analysis.

Once the CV is submitted, the system processes the text using Natural Language Processing (NLP) to extract relevant features such as skills, experience, and writing patterns. Users are also required to complete aptitude and personality assessments, which help improve the accuracy of personality prediction.

### 4.4 Unified Modelling Language (UML) Diagrams

4.4.1 Class Diagram

A class diagram is an illustration of the relationships and the source code dependencies among classes in the UML.

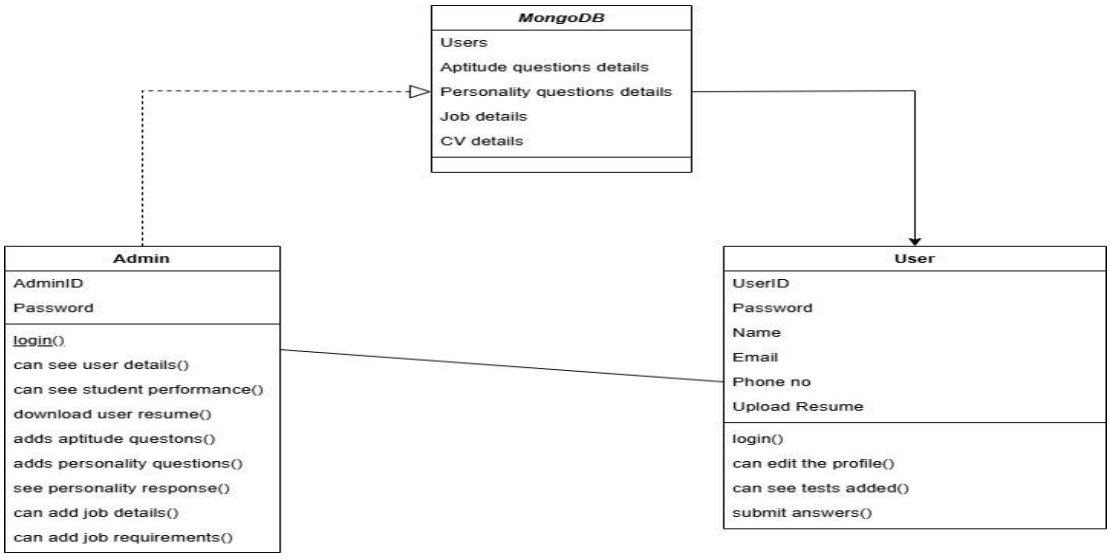
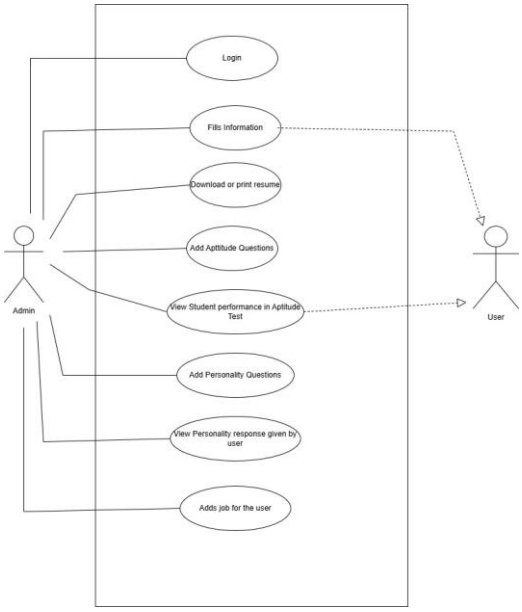


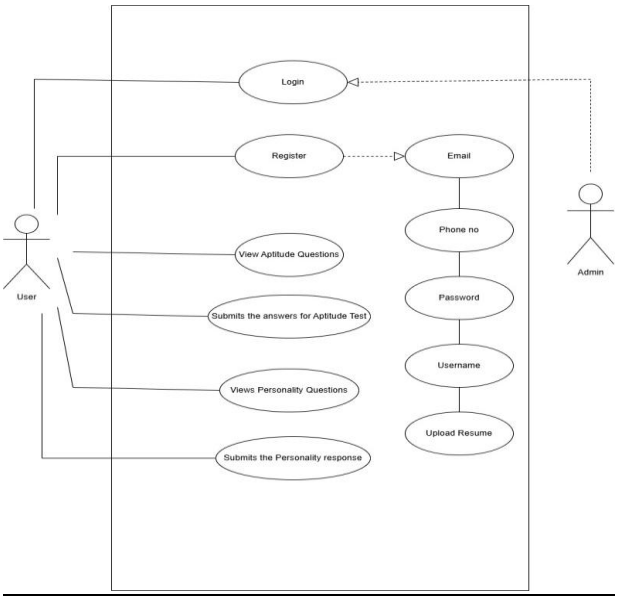

**Figure 4.4.1. System Class Diagram**

### 4.4.2 Use Case Diagram

Use case diagram are behaviour diagrams used to describe a set of actions that some application should or can perform in collaboration with one or more external users of the application(Actors




**4.4.2.1. Admin Use Case Diagram**




**4.4.2.2 . User Use Case Diagram**

### 4.4.3 Activity Diagram

Activity diagram describes the dynamic aspects of the application. It is essentially an advanced version of flowchart modelling the flow from one activity to another activity. It describes how activities are coordinated to provide a service which can be at different levels of abstraction

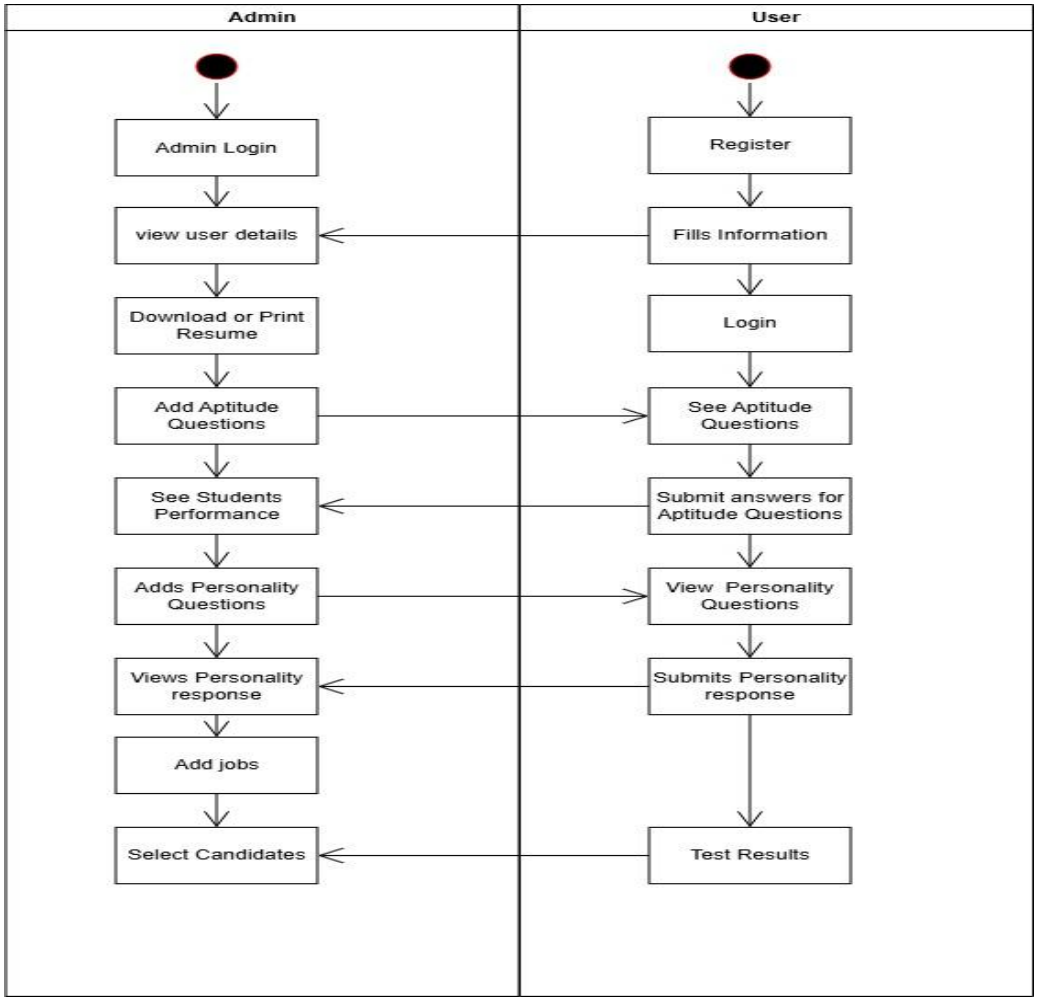



**Figure 4.4.3. System Activity Diagram**

## 4.5 CV Analysis & Personality Prediction

The CV Analysis & Personality Prediction module is responsible for extracting meaningful insights from a candidate's resume using Natural Language Processing (NLP) and Machine Learning (ML) techniques. Once a user uploads their CV, the system processes the text by performing tokenization, stopword removal, named entity recognition (NER), and feature extraction using methods like TF-IDF .

After preprocessing, the extracted features are fed into ML models such as Logistic Regression, Decision Tree, Random Forest, and SVM, which classify the candidate based on

established personality theories like the Big Five Personality Traits (OCEAN Model). The system predicts traits such as openness, conscientiousness, extraversion, agreeableness, and neuroticism, helping recruiters assess the candidate's suitability for a job.

This module enhances the recruitment process by providing an automated, unbiased, and data-driven approach to personality assessment, allowing companies to make better hiring decisions.

**4.6 Shortlisting & Result Display**

The Shortlisting & Result Display module is responsible for identifying the most suitable candidates based on CV analysis, personality prediction, and test scores. Once the system processes the CV and personality traits, it matches the extracted features with job requirements to determine candidate suitability.

Candidates who meet the required criteria are automatically shortlisted, reducing manual effort and bias in the hiring process. The results, including personality traits, test scores, and skill compatibility, are displayed in a structured format for the admin to review. This module ensures an efficient and objective candidate selection process, helping recruiters make data-driven hiring decisions quickly.

**4.7 Deployment & Evaluation**

The Deployment & Evaluation module ensures the implementation and performance assessment of the personality prediction system. The trained Machine Learning (ML) models are deployed using Flask, allowing real-time CV analysis and personality prediction through a user-friendly interface.

For evaluation, the system's accuracy is measured using key performance metrics, including Accuracy, Precision, Recall, and F1 Score. These metrics help assess how well the model predicts personality traits based on CV data. Additionally, the performance of different ML algorithms such as Logistic Regression, Decision Tree, Random Forest, and SVM is compared to select the most effective model. This module ensures that the system remains reliable, scalable, and optimized for accurate personality prediction and hiring decisions.

# 5. DATA EXPLORATION

## 5.1 Data Collection

The data collection process involves gathering CVs from various sources to build a dataset for training and testing the personality prediction system. CVs can be obtained from public datasets, job portals, recruitment agencies, or manually collected resumes.

The collected CVs contain structured (name, education, skills, experience) and unstructured (summary, achievements, personal statements) data.

## 5.2 Data Preprocessing

Data preprocessing is a crucial step in preparing the CV dataset for analysis and personality prediction. It involves cleaning and transforming raw CV text into a structured format suitable for Natural Language Processing (NLP) and Machine Learning (ML) models.

The preprocessing steps include text extraction from different file formats (PDF, DOCX), tokenization to break text into words, stopword removal to eliminate common words, and named entity recognition (NER) to identify key details like skills, job roles, and organizations.

## 5.3 Text Extraction from CVs

Text extraction from CVs is a fundamental step in automated resume analysis, enabling efficient processing of structured and unstructured data. The process begins with file handling, where resumes in various formats such as PDF, DOCX, and TXT are loaded using libraries like PyPDF2 and python-docx. Once the text is extracted, preprocessing techniques such as tokenization, stopword removal, and stemming or lemmatization are applied to clean and normalize the data. Identifying key sections such as the candidate's name, contact details, education, work experience, skills, and certifications is done using Regular Expressions (RegEx) and Named Entity Recognition (NER) models from libraries like spaCy and NLTK. To enhance the analysis, keyword extraction techniques such as TF-IDF.

Additionally, semantic analysis techniques help in understanding the context of extracted words and phrases, improving accuracy in classification. Finally, the processed data is stored in a structured format like JSON or CSV, making it ready for further personality prediction and job-matching analysis.

# 6. DATA ANALYSIS

## 6.1. Exploratory data analysis (EDA) and its types

Exploratory Data Analysis (EDA) is an essential step in data preprocessing that involves analyzing datasets to uncover patterns, detect anomalies, and summarize key characteristics. It helps in understanding the structure of the data, identifying missing values, spotting outliers, and exploring relationships between variables before applying machine learning models. EDA can be categorized into different types based on the nature of analysis. Univariate analysis focuses on examining a single variable using statistical measures such as mean, median, variance, skewness, and kurtosis, often visualized through histograms and box plots. Performing EDA ensures better data quality, improves feature selection, and enhances model performance, making it a crucial step in applications like resume screening and personality prediction from CV analysis.

## 6.2. Importance of Data Analysis in Personality Prediction

Data analysis plays a crucial role in personality prediction as it helps in extracting meaningful insights from structured and unstructured data, such as resumes and textual responses. By applying various analytical techniques, patterns and correlations between a candidate's skills, experience, language usage, and personality traits can be identified. Feature extraction methods like TF-IDF and word embeddings allow for the numerical representation of text, making it easier for machine learning models to process and classify personality characteristics based on frameworks like the Big Five Personality Model. Additionally, Exploratory Data Analysis (EDA) aids in detecting inconsistencies, handling missing values, and understanding data distribution, which ensures the reliability of predictions. Advanced techniques such as Natural Language Processing (NLP) and Named Entity Recognition (NER) further enhance the extraction of relevant attributes from resumes, improving the accuracy of personality classification. By leveraging data analysis, recruiters and HR professionals can make informed, data-driven hiring decisions, reducing bias and enhancing candidate-job matching efficiency. This automation streamlines the recruitment process, making it more objective and effective in evaluating candidates based on both professional qualifications and personality traits.

## 6.3. EDA Inferences

### 6.3.1. Word Frequency Analysis

Word Frequency Analysis is a fundamental technique in text processing that involves identifying and counting the occurrence of words in a given dataset. It helps in understanding

the most commonly used terms, extracting key information, and analyzing text patterns. In personality prediction via CV analysis, word frequency analysis can highlight frequently mentioned skills, job titles, and personality-related keywords, aiding in candidate profiling. This technique is typically implemented using tokenization and stopword removal to filter out irrelevant words and focus on meaningful terms. Methods such as Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF) can be applied to measure word importance in a document or across multiple CVs. Visualization tools like word clouds, bar charts, and histograms can further represent word frequency trends effectively.

### 6.3.2. Personality Traits Vs Word Usage

The relationship between personality traits and word usage is a key aspect of text-based personality prediction, as linguistic patterns often reflect an individual's psychological characteristics. The Big Five Personality Model—which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN)—is commonly used to analyze how word choices correlate with personality traits.

Individuals with high Openness tend to use imaginative, abstract, and intellectual words, while those with low Openness may prefer conventional and straightforward language. People high in Conscientiousness often use structured, goal-oriented words, whereas those low in this trait might have more disorganized or impulsive writing patterns. Extraverts frequently use words related to social interactions and emotions, such as "party," "friends," and "exciting," while introverts may use introspective and solitary-related words. Agreeable individuals favor positive and cooperative language, whereas less agreeable people may use more direct or even argumentative words. Lastly, those with high Neuroticism often express more negative emotions through words like "stress," "anxiety," and "worried," while emotionally stable individuals use more neutral or positive language.

By analyzing word usage in resumes or written text, machine learning models and Natural Language Processing (NLP) techniques can classify personality traits, helping recruiters gain deeper insights into a candidate’s behavioral tendencies and cultural fit for a role.

### 6.3.3. Skill Set Vs Job Role

The relationship between skill set and job role is crucial in determining a candidate’s suitability for a particular position. Different job roles require specific technical and soft skills, which can be extracted from resumes using Natural Language Processing (NLP) and Machine Learning techniques. A well-defined skill set ensures that a candidate is equipped with the necessary knowledge and competencies to perform effectively in a given role.

For example, technical roles like Software Developer or Data Scientist require expertise in programming languages (Python, Java, SQL), data analysis, and problem-solving skills. On the other hand, management roles such as Project Manager or HR Executive emphasize leadership, communication, and decision-making skills.

By analyzing the correlation between skills and job roles, AI-driven recruitment systems can automate the job-matching process, ensuring that candidates are aligned with the right positions.

# 7. MODELLING

## 7.1. Model Development

### 7.1.1. Multi-class classification for Personality Prediction

Multi-class classification for personality prediction is a machine learning approach that categorizes individuals into multiple personality trait classes based on textual data extracted from resumes, cover letters, or assessments. This method is commonly used in Natural Language Processing (NLP) to analyze linguistic patterns and predict personality traits based on the Big Five Personality Model (OCEAN), which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Machine learning models like Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forestare used for classification. After model training, evaluation metrics such as accuracy, precision, recall, and F1-score help measure the model's effectiveness in predicting personality traits.

### 7.1.2 Machine Learning Models Used

#### 7.1.2.1 Logistic Regression

Logistic Regression works by first transforming text data into numerical features using methods like TF-IDF. These features are then fed into the logistic regression model, which applies a linear function to compute a weighted sum of input features and passes it through the sigmoid activation function, producing probability scores. If the probability exceeds a predefined threshold (e.g., 0.5 in binary classification), the model assigns a class label accordingly.

#### 7.1.2.2. Decision Tree

A decision tree first converts text into numerical features using techniques like TF-IDF. The algorithm then selects the most informative features using measures like Gini Impurity or Entropy (used in Information Gain) to determine the best splits. Each decision node represents a condition on a feature, leading to branches that further divide the data until reaching leaf nodes, which represent the final class labels.

#### 7.1.2.3. Random Forest

Random Forest is an ensemble learning algorithm that improves classification accuracy by combining multiple Decision Trees. It is widely used in text-based classification tasks such as resume screening, sentiment analysis, personality prediction, and spam detection due to its robustness and ability to handle high-dimensional data.The algorithm works by constructing

multiple decision trees during training, where each tree is trained on a random subset of the dataset using a technique called bagging (Bootstrap Aggregation).

**7.1.2.4. Support Vector Machine (SVM)**

Support Vector Machine maps these features into a higher-dimensional space and identifies the decision boundary that maximizes the margin between different classes. This is done using support vectors, which are the data points closest to the hyperplane and crucial in defining the classification boundary.

SVM can handle linear and non-linear classification using different kernel functions like Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid kernels.

**7.2. Model Evaluation**

Model evaluation is a crucial step in machine learning to assess how well a trained model performs on unseen data. In text-based classification tasks like resume screening, personality prediction, and sentiment analysis, various metrics are used to measure the model's accuracy, reliability, and effectiveness.

The evaluation process begins by splitting the dataset into training, validation, and test sets to ensure unbiased performance measurement. Commonly used metrics for classification models include accuracy, precision, recall, F1-score, confusion matrix.

1.Accuracy : It measures the overall correctness of the model and it is given b

$$\text{Accuracy} = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ \ Positives + False\ Negatives}$$

2.Precision : It measures how many of the predicted positive instances were actually correct.

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

.3.Recall : It measures how many actual positive instances were correctly identified.

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

4.F1 Score : It is the harmonic mean of precision and recall balancing both the metrics.

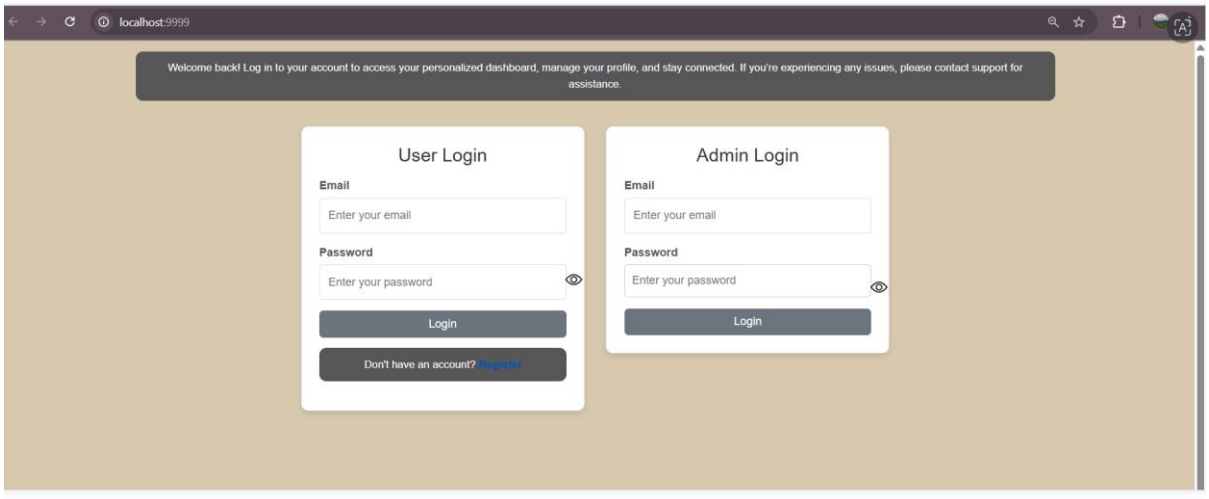
$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

# 8. RESULTS AND CONCLUSIONS

## 8.1. Results

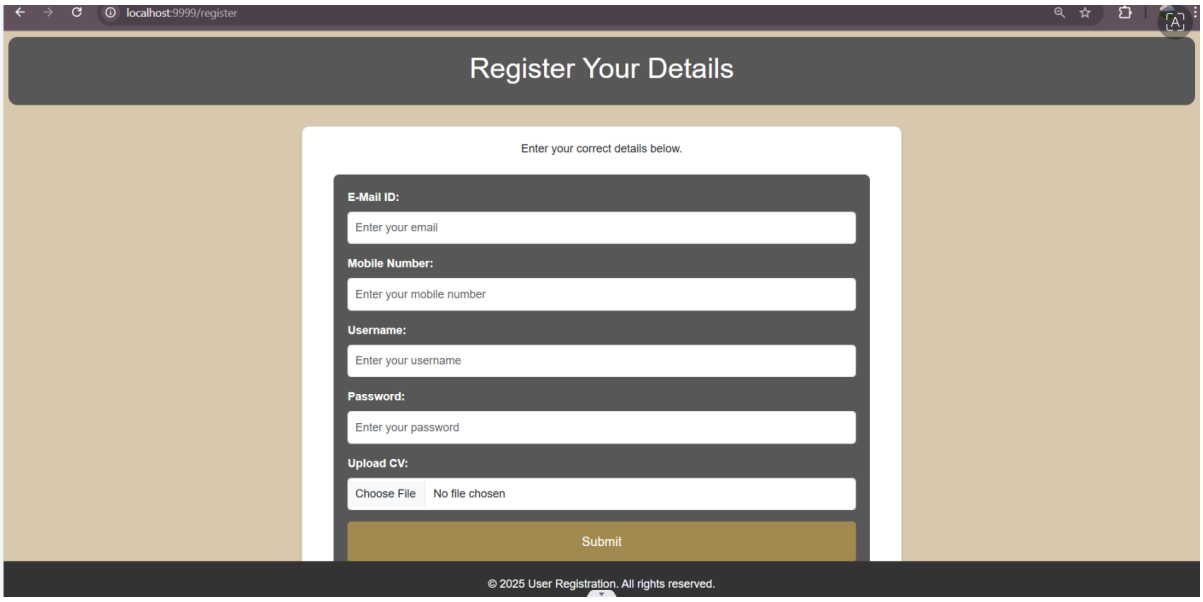
The results of the personality prediction via CV analysis demonstrate the effectiveness of machine learning models in identifying personality traits based on textual information extracted from resumes. Various classification algorithms such as Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) were evaluated using different feature extraction techniques, including TF-IDF.

**Figure 8.1.1. Login page**




Here new user will register.

**Figure 8.1.2. New User Registration**



Then the details of the candidate will be displayed in this way.

**Figure 8.1.2.1. User Dashboard**

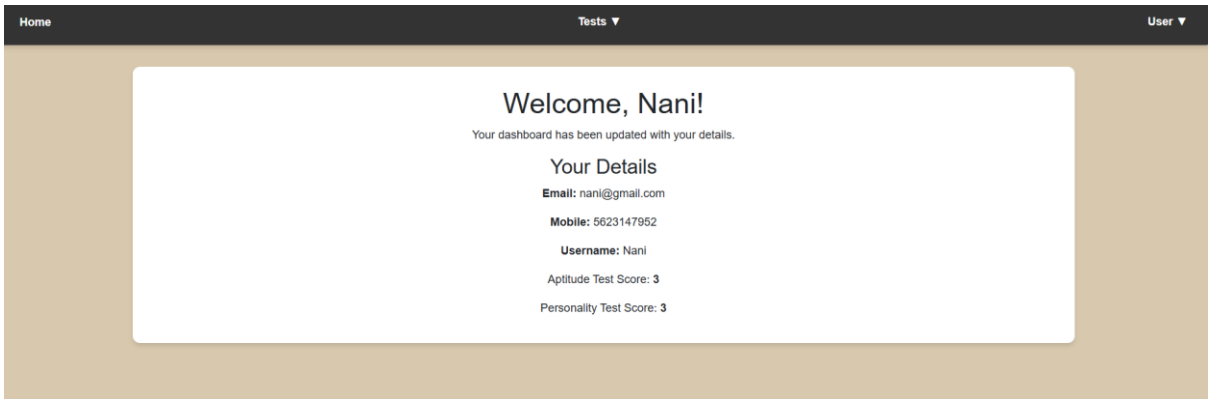
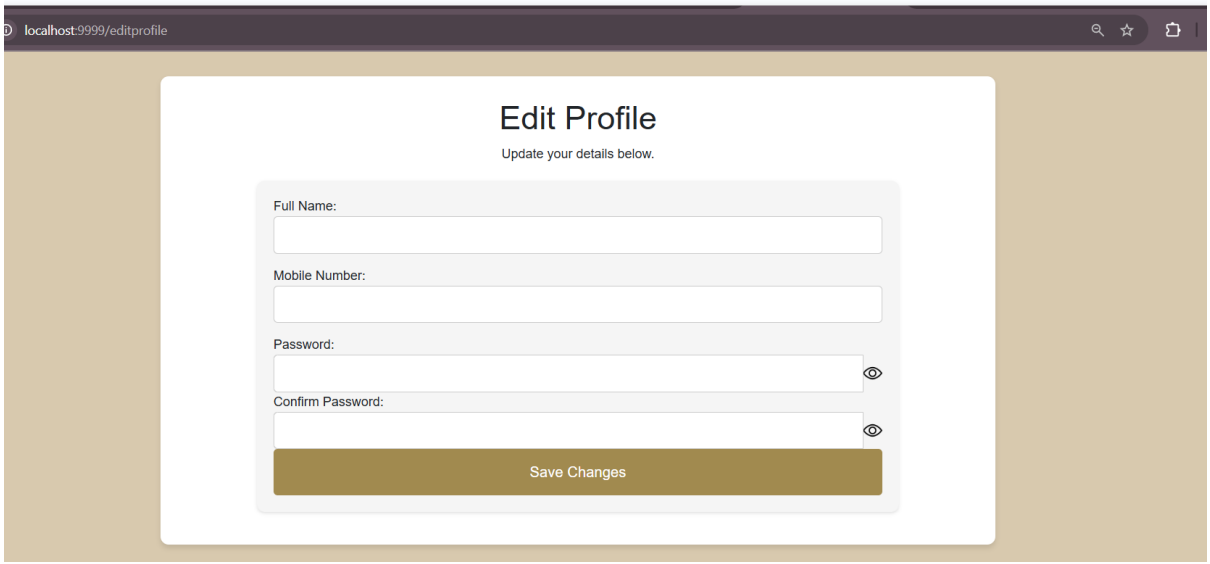
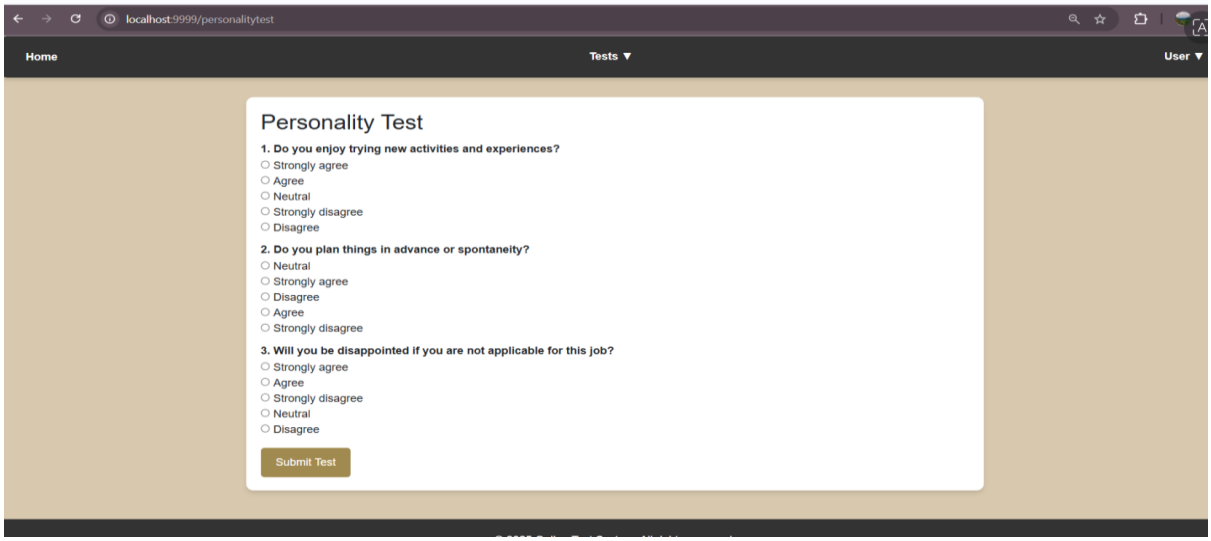



**Figure 8.1.2.2. Edit Profile**



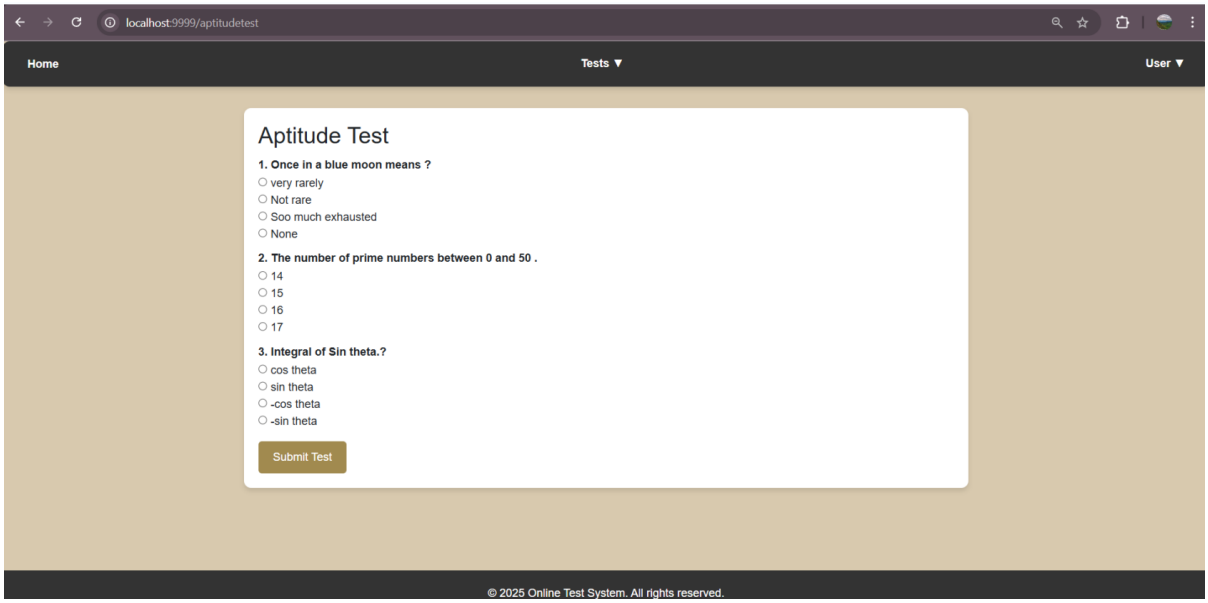

Here the candidates have to attempt the personality test.

**Figure 8.1.2.3. Personality Test**



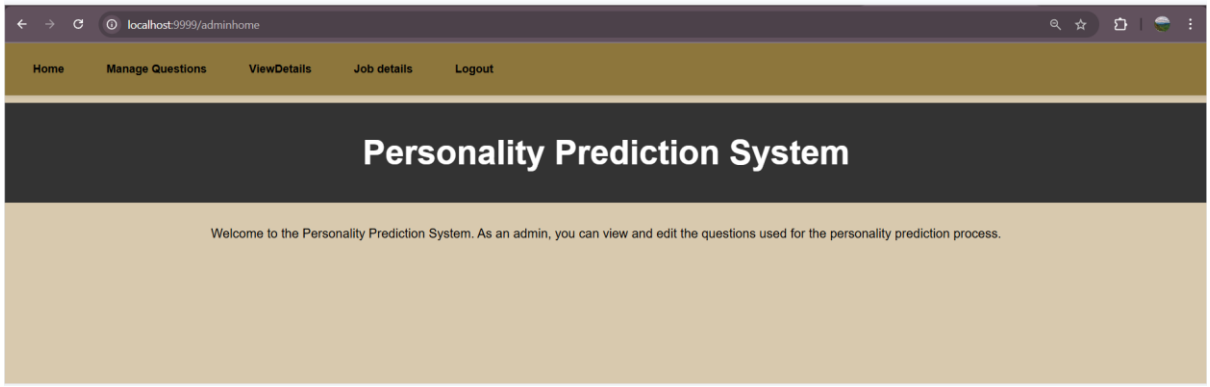
Here the candidates have to attempt the aptitude test.

**Figure 8.1.2.4. Aptitude Test**



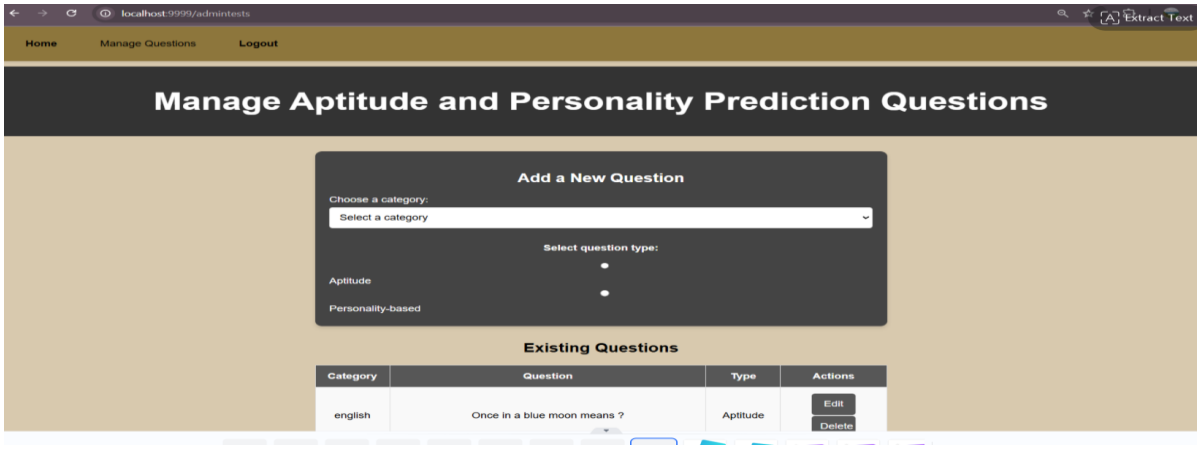

The admin dashboard of the Personality Prediction System, allowing the admin to manage questions and view details.

**Figure 8.1.3. Admin Dashboard**



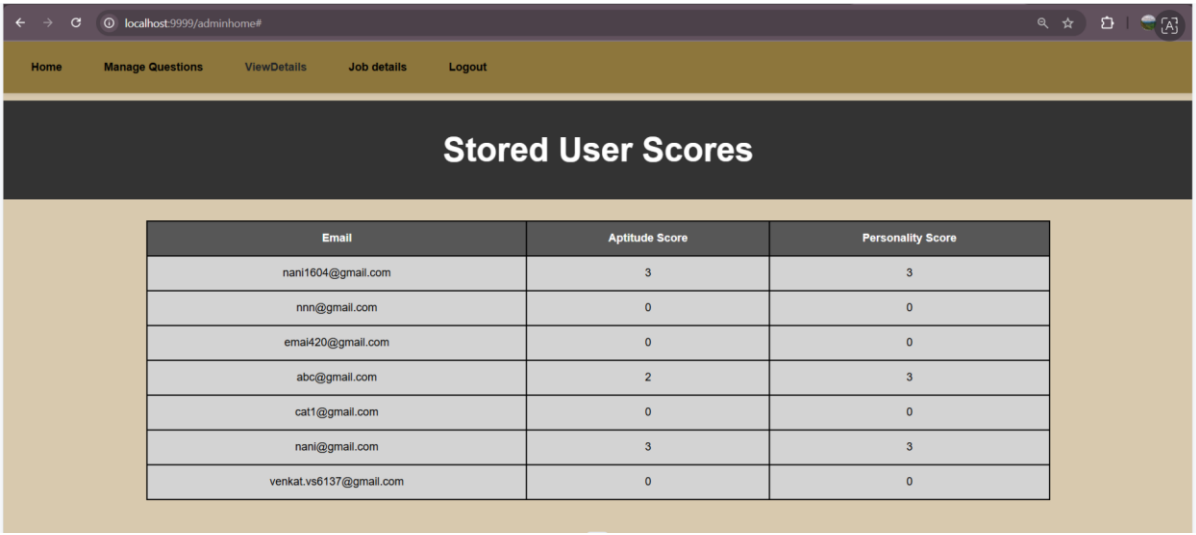

The recruiter can make their own questions to test aptitude and the personality via MCQ's.

**Figure 8.1.3.1. Manage Questions**



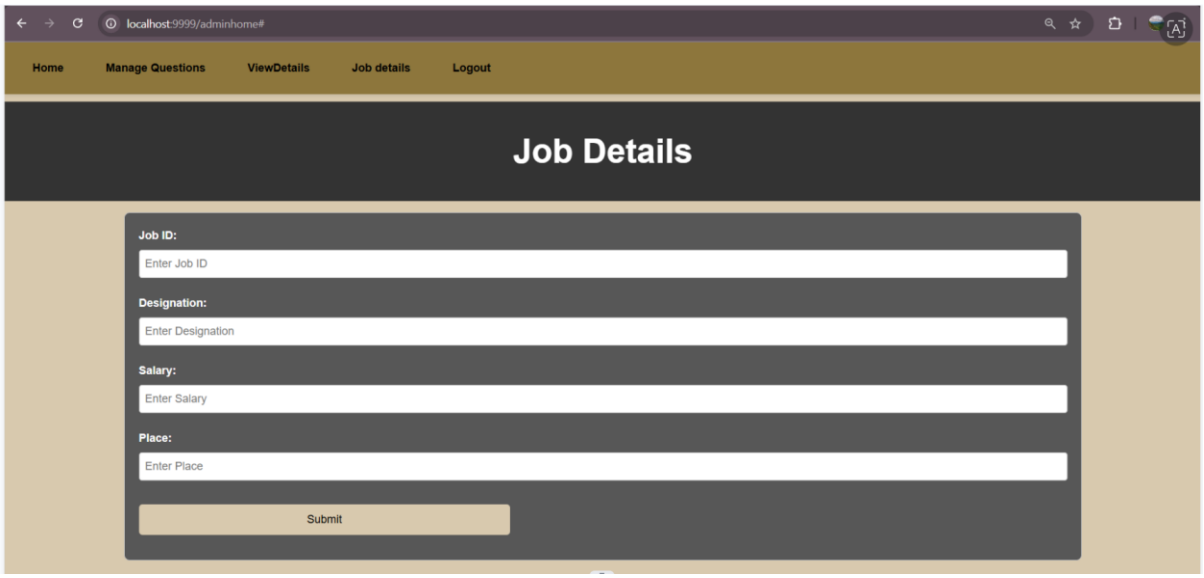
Details of the candidates those who were registered into the system will be displayed to the recruiter.

**Figure 8.1.3.2. User Scores**




Admin enters the job details and requirements for a specified role.

**Figure 8.1.3.3. Job Details**

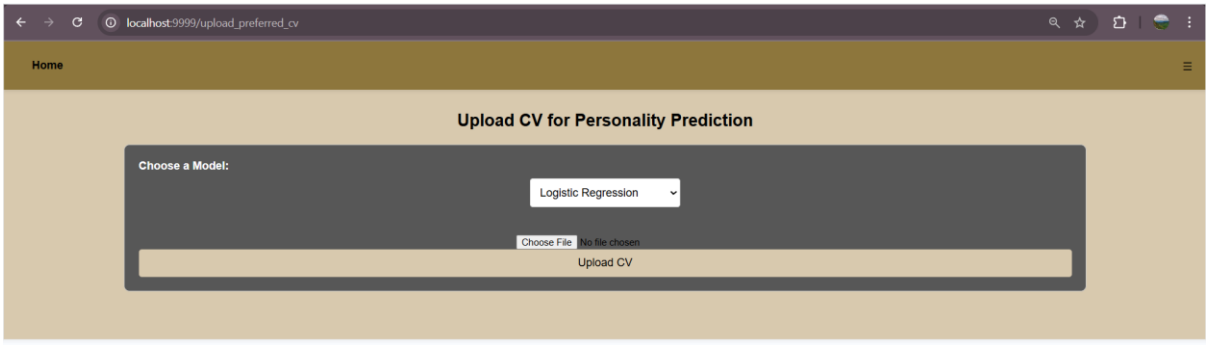



The candidate's CV to download is shown in Shortlist CVs page when he passes 75% in the test.

**Figure 8.1.3.4. Shortlist Cvs**

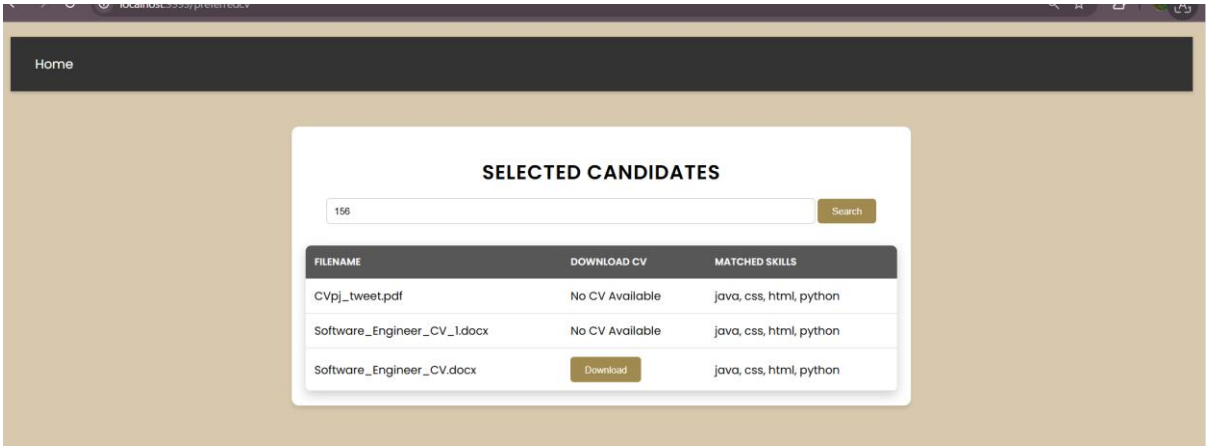
The downloaded CV is uploaded for predicting the personality and extracting the skills.

**Figure 8.1.3.5. Upload CV**



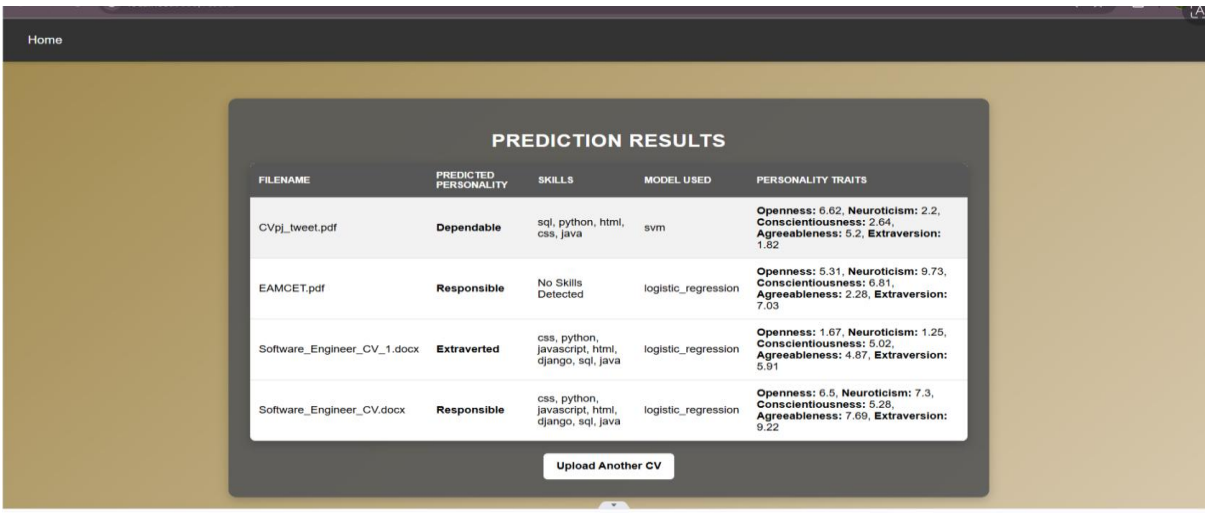

Selected Candidates who are close to the requirements of the that particular job role will be displayed.

**Figure 8.1.3.6. Selected Candidates**




This is the overall output showing the model used, predicted personality and the skills extracted.

**Figure 8.1.3.7. Predicted Results**

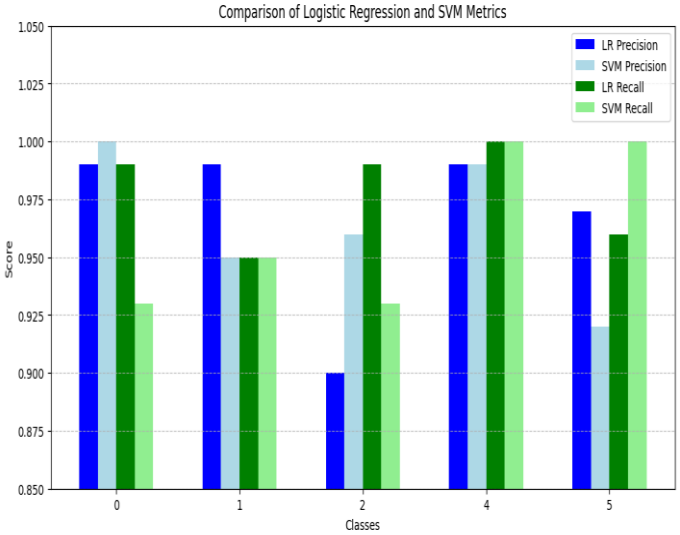


### 8.2 Comparison Between Machine Learning Models

After training the model and uploading of CV,we observed that Logistic Regression and Random Forest have the nearest and same accuracy.We handpicked the best two models which are Logistic Regression and Support Vector Machine.They are compared as shown.

Table 1 : Performance Metrics

| Model | Accuracy | Precision | F1-score | Recall |
|---|---|---|---|---|
| Logistic Regression | 0.9636 | 0.97 | 0.96 | 0.96 |
| Support Vector Machine | 0.9608 | 0.96 | 0.96 | 0.96 |



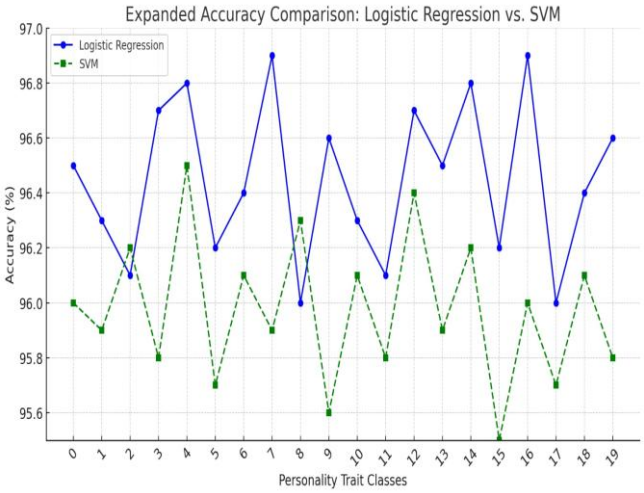




**Figure 8.2.1. Bar Graph of Logistic and SVM** **Figure 8.2.2. Line Chart of Both Models**

Figure 8.2.1. shows the bar graph obtained after comparing both Logistic Regression and SVM models with precision and recall scores highlighting the best model. Figure 8.2.2 represents a line chart which presents the overall comparison of accuracies of Logistic Regression and SVM models considering 20 samples showing a variation in accuracy between the two classifiers indicating Logistic has a highest accuracy when compared.

### 8.3 Conclusion

This project successfully demonstrated the working of an E-Recruitment System in helping the recruiters to shortlist the candidates required.It has concluded that ML models can effectively predict the personality from CV text. These models can also be useful in automated recruitment, candidate profiling and personality assessment, reducing bias and improving hiring efficiency. It also requires shortspan of hiring process reducing the manual

work. The integration of an admin-user interface and secure authentication mechanisms further enhances the system's usability and reliability, streamlining the recruitment workflow in a more structured and unbiased manner.

### 8.4 Future Scope

In future enhancements, the system can be extended to analyze data from social media, emails, and cover letters to strengthen personality prediction with multi-modal data. Incorporating advanced NLP models like BERT or GPT for contextual understanding of CVs can significantly improve the precision of text analysis. Additionally, expanding support for multilingual CVs will enable a more inclusive and globalized solution. The integration of real-time chatbots for candidate interaction, emotion recognition, and interview transcript analysis can further enrich the profiling process. Moreover, building an adaptive model that evolves with feedback from HR professionals can continuously improve the relevance and accuracy of predictions, making it a robust tool for modern, AI-driven recruitment.

# REFERENCES

[1]. Xinyu Shui.,Yiling Chen.,Xin Hu., "Personality in Daily Life: Multi-Situational Physiological Signals Reflect Big-Five Personality Traits" , IEEE Journal of Biomedical and Health Informatics , Volume: 27, Issue: 6 ,2023.

[2]. Nandini Agarwal.,Justina.,M.J.Akshaya.,Nihar Shetty.,"Machine Learning driven Personality Prediction System using the concept of CV Analysis" ., International Conference on Electrical Electronics and Computing Technologies (ICEECT) ,Conference Paper.,2024.

[3]. Lakshyajit Thapa; Aniket Pandey; Deepanshu Gupta; Ayush Deep; Rakesh Garg.,A Framework for Personality Prediction for E-Recruitment Using Machine Learning Algorithms., 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence).,21 March 2024.

[4]. Jayashree Rout.,Sudhir Bagade.,Pooja Yede.,Nirmiti., "Personality Evaluation and CV Analysis using Machine Learning Algorithm"., International Journal Of Computer Science and Engineering.,Volume.-7., Issue-5., May 2019.

[5]. G.Sudha.,Sasipriya., Sri Janani S.,Nivethitha D.,Saranya S., Karthick Thyagesh G., "Personality Prediction Through CV Analysis using Machine Learning Algorithms for Automated E-Recruitment Process"., International Conference on Computing and Communications Technologies (ICCCT)., Conference Paper 2021.

[6]. Atharva Kulkarni, Tanuj Shankarwar, Siddharth Thorat.,"Personality Prediction Via CV Analysis using Machine Learning"., International Journal of Engineering Research & Technology (IJERT)., Volume 10, Issue 09.,September 2021.

[7]. Muthu selvi M., Angeline Ranjitha Mani.,"Personality Prediction System Through CV Analysis"., International Journal on Cybernetics & Informatics ., Vol 10(02) ,2020.

[8]. Amal Khalifa Al Aamer, Allam Hamdan & Zaher Abusaq.,"The Impact of Artificial Intelligence on the Human Resource Industry and the Process of Recruitment and Selection International Conference on Business and Technology"., Conference paper., March 2023., pp 622–630.

[9]. Anum Jaffar; Sara Ali; Khawaja Fahad Iqbal; Yasar Ayaz; Ali R. Ansari; Muhammad A. B. Fayyaz.,"A Comprehensive Multimodal Humanoid System for Personality Assessment Based on the Big Five Model".,IEEE Access.,Journal Article.,Volume: 12,2024.

[10]. Madhura Jayaratne; Buddhi Jayatilleke.,"Predicting Personality Using Answers to Open-Ended Interview Questions".,IEEE Access.,Journal Article.,Volume: 8,2020.

## Appendix: A- Packages, Tools used & Working Process

### Python Programming language

Python is a high-level Interpreter based programming language used especially for general-purpose programming. Python features a dynamic type of system and supports automatic memory management.

It supports multiple programming paradigms, including object-oriented, functional and Procedural and also has its a large and comprehensive standard library. Python is of two versions. They are Python 2 and Python 3.

This project uses the latest version of Python, i.e., Python 3. This python language uses different types of memory management techniques such as reference counting and a cycle-detecting garbage collector for memory management. One of its features is late binding (dynamic name resolution), which binds method and variable names during program execution.

Python's offers a design that supports some of things that are used for functional programming in the Lisp tradition. It has vast usage of functions for faster results such as filter, map, split, list comprehensions, dictionaries, sets and expressions. The standard library of python language has two modules like itertools and functools that implement functional tools taken from Standard machine learning.

### Libraries

### NumPy

Numpy is the basic package for scientific calculations and computations used along with Python. NumPy was created in 2005 by Travis Oliphant. It is open source so can be used freely. NumPy stands for Numerical Python. And it is used for working with arrays and mathematical computations.

Using NumPy in Python gives you much more functional behavior comparable to MATLAB because they both are interpreted, and they both allows the users to quickly write fast programs as far as most of the operations work on arrays, matrices instead of scalars. Numpy is a library consisting of array objects and a collection of those routines for processing those arrays.

Numpy has also functions that mostly works upon linear algebra, Fourier transform, arrays and matrices. In general scenario the working of NumPy in the code involves searching, join, split, reshaping etc. operations using NumPy.

The syntax for importing the NumPy package is → import NumPy as np indicates NumPy is imported alias np.

**Pandas**

Pandas is used whenever working with matrix data, time series data and mostly on tabular data. Pandas is also open-source library which provides high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

This helps extremely in handling large amounts of data with help of data structures like Series, Data Frames etc. It has inbuilt methods for manipulating data in different formats like csv, html etc.,

Simply we can define pandas is used for data analysis and data manipulation and extremely works with data frames objects in our project, where data frame is dedicated structure for two-dimensional data, and it consists of rows and columns similar to database tables and excel spreadsheets.

In our code we firstly import pandas package alias pd and use pd in order to read the csv file and assign to data frame and in the further steps. We work on the data frames by manipulating them and we perform data cleaning by using functions on the data frames such as df.isna().sum(). So, finally the whole code depends on the data frames which are to be acquired by coordinating with pandas. So, this package plays a key role in our project.

**Matplotlib**

Matplotlib is a library used for plotting in the Python programming language and it is a numerical mathematical extension of NumPy. Matplotlib is most commonly used for visualization and data exploration in a way that statistics can be known clearly using different visual structures by creating the basic graphs like bar plots, scatter plots, histograms etc.

Matplotlib is a foundation for every visualizing library and the library also offers a great flexibility with regards to formatting and styling plots. We can choose freely certain assumptions like ways to display labels, grids, legends etc.

In our code firstly we import the matplotlib.pyplot alias plt, This plt comes into picture in the exploratory data analysis part to analyze and summarize datasets into visual methods, we use plt to add some characteristics to figures such as title, legends, labels on x and y axis as said earlier ,to understand more clearly we can also use different plots.

**Seaborn**

Seaborn is used for drawing attractive statistical graphics with just a few lines of code. In other words, we can say seaborn is a data visualization library based on the matplotlib and closely combined with Pandas data structures in Python. Visualization is the central theme of Seaborn which helps in exploration and understanding of data.

Plots are used for visualizing the relationship between variables. Those variables can be numerical or categorical.

Using Seaborn, we can also plot wide varieties of plots like Distribution plots, Pie chart and bar chart, Scatter plots, Pair plots, Heat maps.

In our code we use seaborn library in EDA where sns is used to create countplot between skin disease and count of target values and we use facetgrid with respect to sns for looking out distribution of age based on the diseases and sns with respect to the countplot is also used to find perspective of data analysis i.e., is the disease due to family genes i.e., family history vs count.

**Scikit-learn**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. Scikit learn is an efficient, and beginner, user friendly tool for predictive data analysis and it provides a selection of tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library is built upon different libraries such as NumPy, SciPy and Matplotlib.

Scikit learn is used when identifying to which category an object likely belongs, predicting continuous values and grouping of similar objects into clusters.

When coming to our code, sklearn plays an important with respect to classification algorithm, the final result and performance. Accuracy of the algorithm can be determined using sklearn. The different modules imported from sklearn library is train_test_split, GaussianNB, one vs rest classifier and all the metrics. When going much detail into the modules and packages, train_test_split means it splits the data into random training and testing subsets. We used gaussian naive bayes classification algorithm in order to classify the values of the model. On taking the syntax as example from sklearn. metrics, import * describes to import all the metrics required for doing some kind of mathematical, or evidential calculations. Similarly from sklearn.model_selection, import train_test_split described above and there are few preprocessing steps such as from sklearn. preprocessing import Label Encoder where labelencoder encode labels with a value between 0 and n_classes-1 where n is the number of distinct labels, and other step is from sklearn.preprocessing import label_binarize where label

binarize is used to convert multi-class labels to binary labels (belong or does not belong to the class) and we use multiclass classification in our which will be explained in detail in the above document, and when coming to metrics we use confusion matrix in order to calculate the performance of classification model by using certain measures like precision,recall,f1 sore and threshold value.

NLP tools : NLTK (Natural Language Toolkit), spaCy ,TF-IDF

Framework : Flask

Frontend Interface : HTML,CSS,JavaScript

Tools used VS Code