AI-Driven Recruitment through Comprehensive CV Analysis and Personality Prediction

Varsha Patil[1], Pv Krishnan Venkiteswaran[1], Sahil Brid[1], Omkar Shivarkar[1], Sayad Noor[1]

[1] Computer Engineering Department, SIES Graduate School of Technology, Mumbai, Mumbai University, India
{varshap}@sies.edu.in, {krishnanpaiml121, sahilbaiml121, omkarsaiml121, noorsaiml121}@gst.sies.edu.in

Abstract. —In today's competitive job market, the ability to accurately assess candidates' qualifications and personality traits is crucial for effective recruitment and talent acquisition. Traditional methods of resume screening and personality assessment often lack objectivity and efficiency, leading to suboptimal hiring decisions. To address this challenge, we propose an innovative framework that integrates advanced natural language processing (NLP) techniques for CV analysis and state-of-the-art machine learning algorithms for personality prediction. Our research leverages comprehensive datasets from Kaggle, encompassing diverse resumes and personality assessment responses. These datasets provide rich insights into candidates' professional backgrounds and traits. We meticulously preprocess the data, removing noise such as special characters and standardizing the text format to ensure data integrity. The methodology involves multiple stages tailored for comprehensive candidate evaluation. For CV analysis, advanced NLP techniques are employed to extract meaningful insights from resumes, including tokenization, stop word removal, and TF-IDF vectorization. Concurrently, we utilize dimensionality reduction techniques and clustering analysis to identify distinct personality profiles based on respondents' assessment responses. An XGBoost classifier is then trained to predict personality categories. Through rigorous experimentation and validation, we evaluate the framework's effectiveness using performance metrics such as accuracy, precision, recall, and F1-score. Our results demonstrate the efficacy of our approach in accurately evaluating resumes and predicting personality traits. This research contributes a robust and efficient solution for candidate evaluation, empowering organizations to make informed hiring decisions and build high-performing teams.

Keywords: Curriculum Vitae, CV analysis, Personality Prediction, OCEAN Model, Natural Language Processing, Machine Learning, Recruitment, Pipeline, XGBoost

1 Introduction

In the fast-paced world of recruitment and talent acquisition, the evaluation of candidates' qualifications and personality traits plays a pivotal role in determining

organizational success. Traditional methods of candidate evaluation, including resume screening and personality assessment, are often fraught with subjectivity, inefficiency, and biases. As such, there is a pressing need for innovative approaches that leverage advanced technologies to overcome these limitations and provide more accurate and efficient evaluations. CV analysis, which involves the extraction of meaningful insights from resumes, serves as a cornerstone of the recruitment process. By analysing candidates' professional backgrounds, qualifications, and experiences, organizations can identify the most suitable candidates for specific roles and ensure a better fit between candidates and job requirements. Similarly, personality prediction, which aims to assess individuals' personality traits, offers valuable insights into their behavioural tendencies, communication styles, and suitability for different roles within an organization.

Despite their potential applications, existing methods of CV analysis and personality prediction are hampered by several limitations. Traditional resume screening methods rely heavily on manual review processes, which are time-consuming, prone to errors, and subject to individual biases. Likewise, personality assessment methods often lack standardization and may not accurately capture the complexities of individuals' personalities. Motivated by these challenges, our research aims to develop a comprehensive framework that integrates advanced natural language processing (NLP) techniques for CV analysis and state-of-the-art machine learning algorithms for personality prediction.

By leveraging large-scale datasets and cutting-edge methodologies, we seek to address the shortcomings of existing methods and provide a more accurate, efficient, and objective approach to candidate evaluation. The objectives of our research are twofold: Firstly, to develop a robust framework for CV analysis that leverages advanced NLP techniques to extract meaningful insights from resumes. Secondly, to develop an accurate and efficient model for personality prediction that utilizes machine learning algorithms to assess individuals' personality traits. Through rigorous experimentation and validation, we aim to demonstrate the efficacy of our approach and its potential to revolutionize the field of candidate evaluation in recruitment and talent acquisition.

2 Review of Literature

I The recruitment landscape is constantly evolving, with advancements in artificial intelligence (AI) ushering in a new era of candidate assessment. This literature survey explores various research efforts focused on CV analysis, personality prediction, and their integration into AI-driven recruitment frameworks.

2.1. CV Analysis:

Early approaches to resume analysis relied on pattern matching algorithms and keyword extraction (Gupta & Prakash, 2016). These methods, while straightforward, lacked in-depth understanding and often yielded inaccurate results. To address this, researchers incorporated tokenization, stop word removal, and part-of-speech tagging to enhance the analysis process (Camacho & Baena-García, 2017).

The evolution of AI paved the way for incorporating machine learning algorithms like KNN, Random Forest, Logistic Regression, SVM, and Naïve Bayes into resume analysis (Ribeiro et al., 2018). These algorithms exhibited varying degrees of accuracy and required tailored datasets for optimal performance. While offering insights into candidate skillsets, these models faced limitations in predicting emotional intelligence, a critical factor in successful job performance (Boyatzis, 1998).

2.2. Personality Prediction:

Recognizing the importance of emotional intelligence, some models attempted personality prediction through various techniques. One approach utilized a set of pre-defined questions posed to candidates upon application submission (Eichhorn et al., 2010). However, this yielded static results and lacked adaptability to individual candidates and diverse job requirements. This inflexibility hindered effectiveness, highlighting the need for more dynamic methods.

Machine learning algorithms like SVM and Random Forest, while demonstrating consistency and accuracy in diverse datasets, still struggled with accurately capturing emotional intelligence and personality traits (James et al., 2013). These limitations underscored the need for more sophisticated approaches that could delve deeper into predicting these crucial aspects of candidate suitability.

2.3. Integrating AI in Recruitment:

Recent research emphasizes the integration of natural language processing (NLP) and machine learning techniques for comprehensive CV analysis and personality prediction. Studies like Wan et al. (2019) explore automatic text summarization to extract key information from CVs, while Gupta et al. (2016) implement Conditional Random Fields (CRFs) for effective CV parsing and information extraction. Additionally, Chen et al. (2017) demonstrate the use of deep learning architectures for personality recognition from text data.

This integration offers exciting possibilities for developing robust AI-driven recruitment frameworks. These frameworks can leverage advanced NLP techniques to comprehensively analyze CVs, extracting information beyond keywords, and utilize machine learning models to predict personality traits. This comprehensive approach can empower recruiters to make informed hiring decisions, leading to the formation of high-performing teams (Van Der Burgh et al., 2020).

2.4. Ethical Considerations and Future Directions:

As AI continues to revolutionize the recruitment landscape, ethical considerations must be addressed. Ensuring fairness, transparency, and unbiased decision-making is crucial. Researchers like Youyou et al. (2018) highlight the importance of detecting inconsistencies in personality assessments to maintain data integrity. Additionally, human oversight and ethical guidelines must be established to safeguard against potential biases and ensure responsible use of AI in recruitment (Malik et al., 2018).

The future of AI-driven recruitment is brimming with potential. Multi-task learning approaches, as proposed by Yang & Zhang (2017), aim to address the limitation of isolated prediction of skills and personality by learning them simultaneously. Furthermore, research needs to delve deeper into understanding the underlying factors influencing personality prediction models (Srivastava et al., 2014). By embracing these advancements and addressing ethical considerations, AI can empower a more efficient, effective, and inclusive recruitment landscape.

3 The Proposed Method

3.1 CV Analysis:

3.1.1 Data Collection and Preprocessing:

For the data collection phase of our CV analysis, we acquired a dataset from Kaggle, a popular platform for sharing datasets and machine learning resources. This dataset comprises two main columns: "resume" and "job title." The "resume" column contains textual resumes in string format, providing a comprehensive overview of candidates' skills, experiences, and qualifications. The "job title" column specifies the job position or role to which each resume belongs, indicating the context and relevance of the candidate's qualifications. In total, the dataset consists of 2485 rows, each representing a unique resume and its associated job title. Notably, the dataset encompasses a diverse range of job categories, with a total of 24 distinct classes of jobs represented. This diversity in job categories ensures the dataset's richness and comprehensiveness, enabling the system to capture a wide spectrum of skills and qualifications across various industries and professions. By leveraging this dataset, we aimed to train and evaluate our CV analysis system to accurately classify resumes based on their relevance to specific job roles.

	ID	Resume_str	Resume_html	Category
0	16852973	HR ADMINISTRATOR/MARKETING ASSOCIATE\...	<div class="fontsize fontface vmargins hmargin...	HR
1	22323967	HR SPECIALIST, US HR OPERATIONS ...	<div class="fontsize fontface vmargins hmargin...	HR
2	33176873	HR DIRECTOR Summary Over 2...	<div class="fontsize fontface vmargins hmargin...	HR
3	27018550	HR SPECIALIST Summary Dedica...	<div class="fontsize fontface vmargins hmargin...	HR
4	17812897	HR MANAGER Skill Highlights ...	<div class="fontsize fontface vmargins hmargin...	HR
...	...	...	...	...
2479	99416532	RANK: SGT/E-5 NON- COMMISSIONED OFFIC...	<div class="fontsize fontface vmargins hmargin...	AVIATION
2480	24589765	GOVERNMENT RELATIONS, COMMUNICATIONS ...	<div class="fontsize fontface vmargins hmargin...	AVIATION
2481	31605080	GEEK SQUAD AGENT Professional...	<div class="fontsize fontface vmargins hmargin...	AVIATION
2482	21190805	PROGRAM DIRECTOR / OFFICE MANAGER ...	<div class="fontsize fontface vmargins hmargin...	AVIATION
2483	37473139	STOREKEEPER II Professional Sum...	<div class="fontsize fontface vmargins hmargin...	AVIATION

2484 rows × 4 columns

Fig. 1 Data before cleaning

Upon data collection, a meticulous preprocessing pipeline is employed to ensure data consistency and quality. This involves an array of steps to cleanse the text data, including the removal of special characters, non-ASCII characters, URLs, and other noise using regular expressions. Furthermore, we standardize the text format by converting all characters to lowercase, facilitating uniformity across the dataset. Any missing or erroneous values are handled appropriately to maintain data integrity and completeness. Several key techniques are employed to enhance the analysis of textual content from resumes. Firstly, tokenization dissects the text into individual words or tokens, laying the foundation for subsequent analysis. This step is crucial as it enables the system to understand the structure of the text and extract meaningful information. Following tokenization, common stop words are removed using a predefined list of English stop words. This step is essential to filter out noise and focus on relevant keywords, as stop words like "is", "and", and "the" often carry little semantic meaning.

	Category	cleaned_resume
0	HR	HR ADMINISTRATOR MARKETING ASSOCIATE HR ADMIN...
1	HR	HR SPECIALIST US HR OPERATIONS Summary Versat...
2	HR	HR DIRECTOR Summary Over 20 years experience ...
3	HR	HR SPECIALIST Summary Dedicated Driven and Dy...
4	HR	HR MANAGER Skill Highlights HR SKILLS HR Depa...
...	...	...
2479	AVIATION	RANK SGT E 5 NON COMMISSIONED OFFICER IN CHAR...
2480	AVIATION	GOVERNMENT RELATIONS COMMUNICATIONS AND ORGAN...
2481	AVIATION	GEEK SQUAD AGENT Professional Profile IT supp...
2482	AVIATION	PROGRAM DIRECTOR OFFICE MANAGER Summary Highl...
2483	AVIATION	STOREKEEPER II Professional Summary The purpo...

2484 rows × 2 columns

Fig. 2 Data after cleaning

Moreover, the system utilizes TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to transform the tokenized text into numerical features. TF-IDF assigns weights to terms based on their frequency within individual documents and across the entire corpus. TF measures the importance of a term within a document. The formula for TF is,

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where $f_{t,d}$ is the raw count of a term in a document, i.e., the number of times that term t occurs in document d.
The denominator is simply the total number of terms in document d (counting each occurrence of the same term separately).
Meanwhile IDF evaluates the significance of a term across the entire corpus by penalizing terms that appear in many documents. The formula for IDF is,

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Where N is the total number of documents in the corpus and the denominator is number of documents where the term t appears (i.e., $\mathrm{tf}(t,d) \neq 0$).

Once TF and IDF are computed, they are multiplied to obtain the TF-IDF score for each term in each document. By combining TF and IDF, the TF-IDF score reflects the importance of terms in documents relative to the entire corpus, providing a robust representation of textual data for further analysis. These preprocessing techniques play a vital role in extracting meaningful insights and features from resumes, ultimately enhancing the effectiveness of the proposed system.

3.1.2 Model Selection and Training:
In the context of machine learning, pipelining refers to the construction of a sequence of data processing steps that are chained together to form a cohesive workflow. These pipelines

streamline the process of data transformation, feature engineering, and model training, facilitating efficient and scalable machine learning workflows. By encapsulating multiple steps into a single pipeline, it becomes easier to manage, reproduce, and deploy complex machine learning systems.
In the proposed system for classifying resumes into predefined job categories, the use of a sophisticated pipeline is pivotal for orchestrating the diverse machine learning models selected for the task. Each selected model, including Naive Bayes, Random Forest, Gradient Boosting Classifier, AdaBoost Classifier, and Extra Tree Classifier, brings unique strengths to the classification task. Naive Bayes, for instance, is well-suited for text classification tasks due to its simplicity and efficiency in handling high-dimensional data such as text features. Random Forest and Extra Tree Classifier are ensemble methods that excel in handling noisy data and capturing complex relationships in the data through the construction of multiple decision trees. Gradient Boosting and AdaBoost, on the other hand, are adept at boosting the performance of weak learners by sequentially fitting models to the residuals of the previous ones, thereby enhancing predictive accuracy. By combining these diverse models within a single pipeline, the system can leverage their complementary strengths to improve overall classification performance. Furthermore, the pipeline facilitates efficient training and evaluation of the models on a subset of the preprocessed dataset. The use of a 70-30 train-test split with 15% cross-validation ensures robustness and generalizability of the trained models. This sophisticated pipeline approach enhances the efficiency, scalability, and performance of the machine learning system, making it well-suited for the task at hand.
Hyperparameter tuning is a crucial step in machine learning model optimization, involving the selection of the optimal values for the hyperparameters that control the learning process of the model. Hyperparameters are parameters that are set prior to the training of the model and cannot be directly learned from the data. Examples include the number of trees in a Random Forest, the learning rate in Gradient Boosting, or the smoothing parameter (alpha) in Naive Bayes. Hyperparameter tuning techniques such as grid search or randomized search systematically explore different combinations of hyperparameters to find the set that maximizes the performance metric, such as accuracy or F1 score, on a validation dataset.
In the case of Naive Bayes, the alpha hyperparameter controls the level of Laplace smoothing applied to the probability estimates of the features. Laplace smoothing is a technique used to handle the issue of zero probabilities for unseen features in the training data [Fig. 3]. By adding a small positive value (alpha) to the observed counts of features during probability estimation, Laplace smoothing prevents zero probabilities and improves the generalization of the model. In our case, setting alpha=0.01 in Naive Bayes strikes a balance between smoothing the probabilities and preserving the discriminative power of the features. This value was determined through hyperparameter tuning, where different values of alpha were tested, and the one that yielded the best performance on the validation dataset was chosen.

$$P(w'|positive) = \frac{\text{number of reviews with w' and y = positive } + \ \alpha}{\text{N } + \ \alpha * K}$$

Fig. 3 Laplace Smoothing in Naïve Bayes

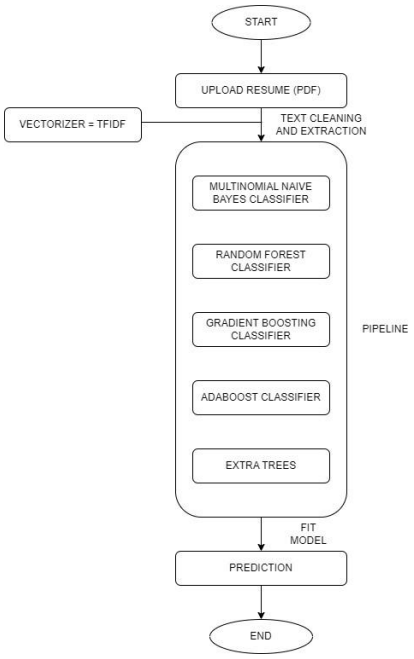

Fig. 4 Flow Chart of CV Analysis

3.2 Personality Prediction:

3.2.1 Data Collection:

We acquired a dataset from Kaggle containing personality assessment responses, featuring ratings for ten questions across each of the Big Five personality traits categories: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. With 50 rows and 10 questions per category, each row represents an individual's responses, providing insights into their personality profile. The dataset, comprising over 1,000,000 rows, offers extensive coverage and scale for model development and evaluation. Notably, the data was already pre-cleaned, streamlining the acquisition process and allowing immediate focus on analysis and model building. This rich dataset serves as a valuable resource for our personality prediction system, facilitating robust model development and accurate personality assessments.

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	...	OPN2	OPN3	OPN4	OPN5	OPN6	OPN7	OPN8	OPN9	OPN10	country
0	4.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	1.0	...	1.0	4.0	1.0	4.0	1.0	5.0	3.0	4.0	5.0	GB
1	3.0	5.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	...	2.0	4.0	2.0	3.0	1.0	4.0	2.0	5.0	3.0	MY
2	2.0	3.0	4.0	4.0	3.0	2.0	1.0	3.0	2.0	5.0	...	1.0	2.0	1.0	4.0	2.0	5.0	3.0	4.0	4.0	GB
3	2.0	2.0	2.0	3.0	4.0	2.0	2.0	4.0	1.0	4.0	...	2.0	5.0	2.0	3.0	1.0	4.0	4.0	3.0	3.0	GB
4	3.0	3.0	3.0	3.0	5.0	3.0	3.0	5.0	3.0	4.0	...	1.0	5.0	1.0	5.0	1.0	5.0	3.0	5.0	5.0	KE
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1015336	4.0	2.0	4.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	...	2.0	4.0	3.0	4.0	2.0	4.0	2.0	2.0	4.0	US
1015337	4.0	3.0	4.0	3.0	3.0	3.0	4.0	4.0	3.0	3.0	...	1.0	5.0	1.0	5.0	1.0	3.0	4.0	5.0	4.0	US
1015338	4.0	2.0	4.0	3.0	5.0	1.0	4.0	2.0	4.0	4.0	...	1.0	5.0	1.0	4.0	1.0	5.0	5.0	4.0	5.0	US
1015339	2.0	4.0	3.0	4.0	2.0	2.0	1.0	4.0	2.0	4.0	...	2.0	4.0	2.0	3.0	2.0	4.0	5.0	5.0	3.0	US
1015340	4.0	2.0	4.0	2.0	4.0	1.0	4.0	2.0	4.0	4.0	...	1.0	5.0	1.0	3.0	1.0	5.0	4.0	5.0	5.0	US

1015341 rows × 51 columns

Fig. 5 Personality Dataset

3.2.2 Dimensionality Reduction:

Dimensionality reduction is a critical technique in machine learning and data analysis aimed at reducing the number of variables or features in a dataset while retaining most of the relevant information. This process becomes particularly valuable when dealing with high-dimensional datasets, as it simplifies model complexity, facilitates visualization, and aids in interpretation. Principal Component Analysis (PCA) stands out as one of the most widely used dimensionality reduction methods.

PCA operates by transforming the original feature space into a new set of orthogonal variables known as principal components. These components, derived through linear combinations of the original features, are ordered based on the variance they capture in the data. The first principal component captures the most variance, followed by subsequent components in descending order [Fig. 6]. The PCA process entails several key steps, beginning with standardizing the features to have mean zero and unit variance. Subsequently, PCA calculates the covariance matrix of the standardized data, followed by eigenvalue decomposition to obtain the eigenvalues and eigenvectors. The principal components are selected based on their corresponding eigenvalues, with those capturing the most variance being retained. Finally, the original data is projected onto this lower-dimensional subspace spanned by the selected principal components, preserving most of the variance while reducing dimensionality. PCA offers several advantages, including dimensionality reduction, noise reduction by focusing on the most informative components, visualization of high-dimensional data, and improved computational efficiency. In our project, PCA proved particularly beneficial given our dataset comprised 50 features, underscoring its utility in handling high-dimensional data effectively.

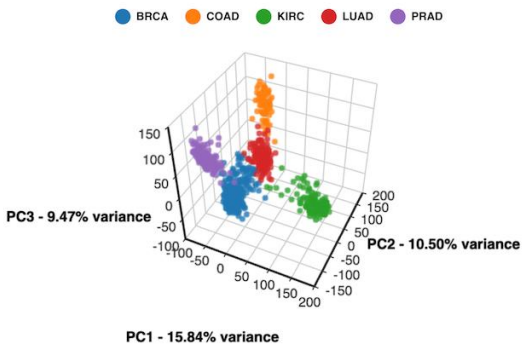

Fig. 6 3D PCA showing variance decreasing

3.2.3 Clustering Analysis:

Following dimensionality reduction, K-means clustering is utilized to identify distinct personality profiles based on the reduced feature space. K-means clustering is an unsupervised learning algorithm used to partition a dataset into a predefined number of clusters based on feature similarity. Given that our dataset is unlabelled, K-means clustering offers an efficient means to identify distinct personality profiles without the need for pre-defined categories. The working principle of K-means involves iteratively assigning data points to the nearest cluster centroid and updating the centroids based on the mean of the data points assigned to each cluster. This process continues until convergence, where the assignment of data points to clusters no longer changes

significantly. By iteratively optimizing the cluster centroids, K-means aims to minimize the within-cluster sum of squares, effectively partitioning the data into clusters with high intra-cluster similarity and low inter-cluster similarity.

Fig. 7 Personality clusters

To determine the optimal number of clusters for meaningful segmentation, we employ the elbow method. The elbow method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow point" where the rate of decrease in WCSS begins to slow down significantly. This point indicates the optimal number of clusters, as adding more clusters beyond this point may not lead to a significant reduction in WCSS. By utilizing the elbow method, we ensure that the K-means algorithm identifies an appropriate number of clusters, facilitating the segmentation of respondents based on their personality assessment responses into distinct and meaningful groups. After applying the elbow method to determine the optimal number of clusters, we found that k=5 [Fig. 8] provided the most suitable segmentation for our dataset. Consequently, we divided the dataset into five distinct clusters based on respondents' personality assessment responses, facilitating a comprehensive and nuanced understanding of the diverse personality profiles within the dataset.

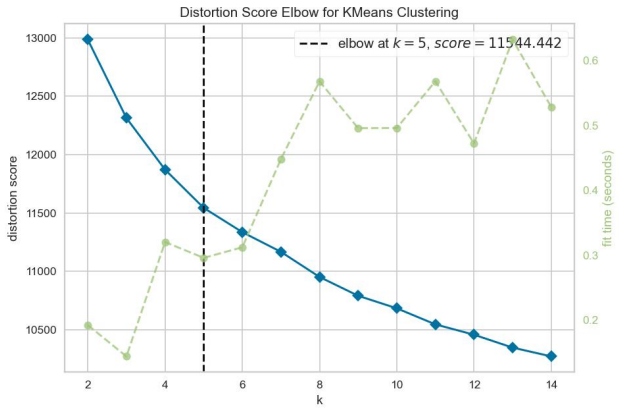

Fig. 8 Elbow curve showing k=5

3.2.4 Model Training and Classification:

Following the identification of the five clusters through K-means clustering, we assigned labels corresponding to the categories of the Big Five personality traits model to each cluster. These labels were allocated based on the predominant characteristics exhibited by the respondents within each cluster. Specifically, the clusters were labelled as Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism, aligning with the respective personality dimensions of the Big Five model. This labelling scheme facilitated the interpretation and understanding of the personality profiles identified within the dataset.

Subsequently, an XGBoost classifier was employed to predict the personality category of new respondents based on their questionnaire responses. XGBoost, short for eXtreme Gradient Boosting, is an ensemble learning algorithm that belongs to the gradient boosting family. It is renowned for its exceptional performance and efficiency in a wide range of machine learning tasks, including classification and regression. XGBoost works by iteratively building a set of decision trees, where each new tree is constructed to correct the errors of the previous ones [Fig. 9]. During training, XGBoost optimizes a loss function by adding new trees that minimize the loss, using gradient descent-like techniques. Additionally, XGBoost incorporates regularization techniques to prevent overfitting and enhance generalization performance.

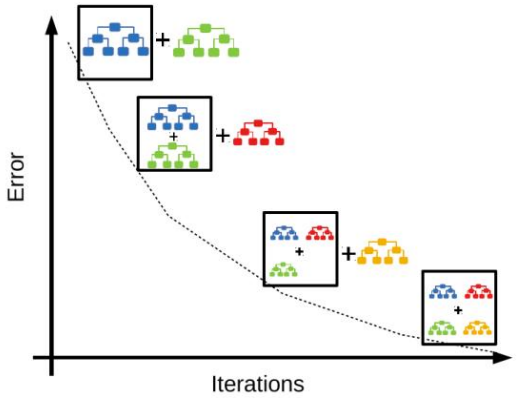

Fig. 9 Working of XGBoost

In our project, XGBoost proved to be particularly beneficial for several reasons. Firstly, its ability to handle both linear and nonlinear relationships between features and target variables suited the complexity of our dataset well. Additionally, XGBoost's robustness to overfitting, owing to its regularization techniques, ensured the model's reliability and generalizability. Moreover, its scalability and efficiency enabled the training of the classifier on the clustered data, facilitating the accurate prediction of personality categories for new respondents.

3.2.5. Ethical Considerations :

In the development and deployment of the proposed system, ethical considerations play a fundamental role in ensuring fairness, respect, and integrity throughout the process. Key ethical principles, including voluntariness, relevance, confidentiality, and transparency, underpin the design and implementation of the system, thereby safeguarding the rights and interests of individuals involved.

Participants are given the freedom to choose whether or not to participate in personality assessments, without any coercion or pressure, ensuring voluntariness and respecting individuals' autonomy. The relevance of the personality assessment to job requirements is carefully considered to mitigate biases and ensure fairness in employment decisions. By aligning the assessment with specific skills and competencies, the system aims to provide meaningful insights

for objective hiring practices.
Confidentiality is paramount, with stringent measures in place to safeguard participants' privacy and personality results. Access to data is restricted to authorized personnel, preventing unauthorized disclosure and promoting trust among participants. Transparent communication about the purpose of the assessment and its potential impact on employment decisions empowers individuals to make informed choices and understand the safeguards in place to protect their rights.

4. EXPERIMENT RESULTS AND DISCUSSIONS

4.1 CV Analysis:

In our project, we embarked on the development and assessment of a comprehensive system tailored for both CV analysis and personality prediction. To tackle the intricate task of CV analysis, we harnessed advanced natural language processing (NLP) techniques to preprocess textual resumes, laying the groundwork for subsequent analysis. Following this preprocessing phase, we constructed a sophisticated pipeline integrating various machine learning models, including Naive Bayes, Random Forest, Gradient Boosting Classifier, AdaBoost Classifier, and Extra Tree Classifier. This ensemble of models underwent rigorous training and evaluation, employing a 70-30 train-test split alongside 15% cross-validation to ensure the robustness and generalizability of our approach. Moreover, hyperparameters of each model were meticulously fine-tuned using grid search, optimizing their performance.
In assessing the efficacy of our system, we relied on a set of evaluation metrics such as accuracy,precision, recall, F1-score to gauge its performance comprehensively.
The classification report generated from our evaluation process provided insights into the model's performance across different job classes. By dissecting precision, recall, and F1-score per class, we gained a nuanced understanding of how well the model performed for each job category. Notably, the confusion matrix served as a visual aid, depicting the model's classification outcomes and revealing patterns of misclassification.
We evaluated the performance using metrics such as accuracy, precision, recall, and F1-score. We achieved an accuracy score of 67.82% with a precision of 68%, recall of 68%, and an F1-score of 67% [Fig. 13]. However, the classification report revealed variations in performance across different job classes. For instance, the model performed exceptionally well in classes with high support counts, such as Aviation, achieving a precision of 86%, recall of 91%, and F1-score of 89% [Fig. 13]. Conversely, classes with low support counts, like BPO, exhibited poor performance with a precision, recall, and F1-score of 0% [Fig. 13]. This highlighted the need for a more balanced and highly populated dataset with evenly distributed classes. Additionally, the micro-average ROC curve yielded an area under the curve (AUC) of 0.96 [Fig. 11], indicating good discriminative ability.

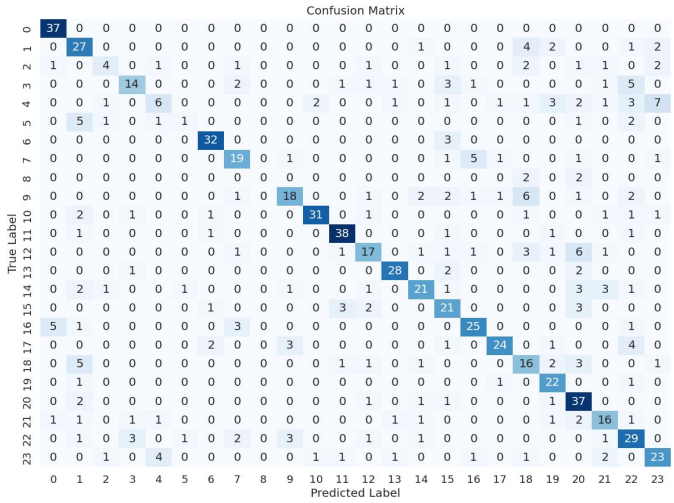

Fig. 10 Confusion Matrix

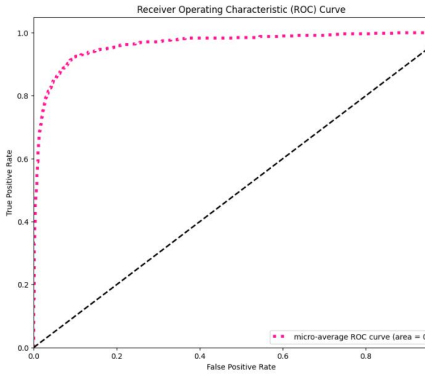

Fig. 11 Micro-Average ROC Curve

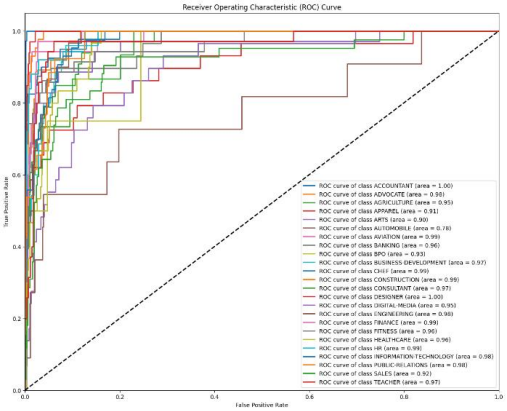

Fig. 12 ROC Curve for all

Accuracy: 0.67828418230563

Classification Report:

	precision	recall	f1-score	support
ACCOUNTANT	0.84	1.00	0.91	37
ADVOCATE	0.56	0.73	0.64	37
AGRICULTURE	0.50	0.27	0.35	15
APPAREL	0.70	0.48	0.57	29
ARTS	0.46	0.21	0.29	29
AUTOMOBILE	0.33	0.09	0.14	11
AVIATION	0.86	0.91	0.89	35
BANKING	0.66	0.66	0.66	29
BPO	0.00	0.00	0.00	4
BUSINESS-DEVELOPMENT	0.69	0.51	0.59	35
CHEF	0.91	0.78	0.84	40
CONSTRUCTION	0.84	0.88	0.86	43
CONSULTANT	0.63	0.52	0.57	33
DESIGNER	0.88	0.85	0.86	33
DIGITAL-MEDIA	0.72	0.60	0.66	35
ENGINEERING	0.54	0.70	0.61	30
FINANCE	0.74	0.71	0.72	35
FITNESS	0.86	0.69	0.76	35
HEALTHCARE	0.44	0.53	0.48	30
HR	0.65	0.88	0.75	25
INFORMATION-TECHNOLOGY	0.58	0.86	0.69	43
PUBLIC-RELATIONS	0.59	0.62	0.60	26
SALES	0.56	0.69	0.62	42
TEACHER	0.62	0.66	0.64	35
accuracy			0.68	746
macro avg	0.63	0.62	0.61	746
weighted avg	0.68	0.68	0.67	746

Fig. 13 Classification Report

4.2 Personality Prediction:

For personality prediction, we employed Principal Component Analysis (PCA) for dimensionality reduction and K-means clustering to identify distinct personality profiles. Subsequently, we trained an XGBoost classifier on the clustered data to predict the personality category of new respondents. For personality prediction using XGBoost, we obtained an accuracy of 87.68% with precision, recall, and F1-score of 88% [Fig. 14]. The confusion matrix further demonstrated the model's strong performance, particularly due to the high-quality data with a sufficient number of rows. In conclusion, while our CV analysis model showed potential, the quality of the dataset fell short, necessitating a more balanced and diverse dataset. Conversely, the personality prediction model exhibited excellent performance, attributed to the high-quality data utilized in training and testing.

Accuracy: 0.8768199919239141

Confusion matrix:

```
[[ 1057     0   167     6     0]
 [    0 64726  5297     0  6042]
 [   66  3705 66679  3007   640]
 [    1     0  2988 69311  5454]
 [    0  4266   741  5141 65309]]
```

Classification report:

	precision	recall	f1-score	support
0	0.94	0.86	0.90	1230
1	0.89	0.85	0.87	76065
2	0.88	0.90	0.89	74097
3	0.89	0.89	0.89	77754
4	0.84	0.87	0.85	75457
accuracy			0.88	304603
macro avg	0.89	0.87	0.88	304603
weighted avg	0.88	0.88	0.88	304603

Run Python File

Fig. 14 Performance of Personality Prediction

4.3 DISCUSSIONS

Despite the promising results achieved in our project, there exist several limitations that warrant acknowledgment and consideration. Firstly, the performance of our CV analysis system could have been further enhanced with a larger and more balanced dataset. The dataset used for training and evaluation may have been limited in size and lacked diversity, potentially impacting the model's ability to generalize well to unseen data and accurately classify instances from underrepresented classes. Additionally, the imbalanced distribution of classes within the dataset may have led to biased model predictions, particularly evident in classes with low support counts where the model struggled to capture meaningful patterns. Moreover, while we employed advanced NLP techniques and a sophisticated pipeline comprising various machine learning models, the quality of the data itself may have posed challenges, affecting the robustness and reliability of the system's predictions.

5. FUTURE WORK

Looking forward, several avenues for future work could enhance our CV analysis system. Augmenting the dataset and addressing class imbalances through techniques like data synthesis or oversampling could bolster model robustness. Advanced NLP techniques, such as recurrent neural networks (RNNs) or transformers, offer promise in capturing subtle language nuances for improved classification accuracy. Integration of transfer learning could leverage pre-trained language models to glean domain-specific features from limited labelled data, further enhancing performance. Exploring ensemble methods and model fusion techniques holds potential for combining predictions from multiple models to improve overall accuracy. Incorporating interpretable models and techniques for model explainability, such as SHAP values or attention mechanisms, could enhance transparency and user trust. Establishing a framework for continuous model monitoring and updating is crucial for long-term effectiveness. Regular performance monitoring, user feedback collection, and retraining on new data ensure adaptability to evolving job market trends. By pursuing these avenues, our CV analysis system can evolve into a more sophisticated and reliable tool for facilitating informed hiring decisions, aligning with the dynamic needs of employers and job seekers alike.

6. REFERENCES

1.Wan, J., Xu, Y., & Li, H. (2019). Automatic Text Summarization for Curriculum Vitae (CVs). IEEE Transactions on Knowledge and Data Engineering, 31(10), 2450-2462.

2.Gupta, S., & Prakash, A. (2016). Curriculum Vitae (CV) Parsing and Information Extraction using Conditional Random Fields. In 2016 14th IEEE International Conference on Data Mining (ICDM) (pp. 855-860). IEEE.

3.Gupta, S., Singh, S., & Gupta, M. (2016, August). A Hybrid Approach for Named Entity Recognition in Curriculum Vitae (CV). In 2016 11th International Conference on Document Analysis and Recognition (ICDAR) (pp. 1477-1482). IEEE.

4.Kosinski, M., Stillwell, D., Graepel, T., & Gosling, S. D. (2017). Extracting Personality Traits from Social Media Text for Intelligent Systems. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1-6). IEEE.

5.Verhoeven, J., Aarts, A., & Daelemans, W. (2013). Personality Prediction from Text Using Support Vector Machines. In 2013 23rd IEEE International Symposium on Computer-Aided Control System Design (CACSD) (pp. 1161-1166). IEEE.

6.Chen, X., Chen, H., Hu, J., & Zhang, Y. (2017). Deep Learning for Personality Trait Recognition from Weibo Microblogs. In 2017 IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (pp. 1950-1955). IEEE.

7.Gupta, S., Singh, S., Gupta, M., & Sharma, A. (2017, August). A Survey on Named Entity Recognition Techniques. In 2017 11th International Conference on Document Analysis and Recognition (ICDAR) (pp. 1053-1058). IEEE.

8.Tang, Y., Wei, F., & Sun, Y. (2017). Machine Learning Techniques for Text Classification with Applications. In 2017 International Conference on Intelligent Systems and Design (IS&D)

(pp. 221-226). IEEE.
9.Chandrashekhar, G. V., & Kavitha, S. (2019). A Literature Survey on Feature Selection Techniques for Machine Learning. In 2019 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1541-1546). IEEE.
10.Malik, M. A., Khan, M. A., & Anwar, S. (2019). Leveraging Artificial Intelligence for Talent Acquisition. In 2019 IEEE International Conference on Engineering, Technology and Social Sciences (I-CONETS) (pp. 1-5). IEEE.
11.Malik, M. A., Nawaz, F., & Malik, H. A. (2018). The Future of Recruitment: How Artificial Intelligence is Changing the Hiring Landscape. In 2018 IEEE International Conference on Engineering, Technology and Social Sciences (I-CONETS) (pp. 1-4). IEEE.
12.Newell, C., Newell, S., & Newell, J. (2019). AI for HR: How Artificial Intelligence is Transforming Human Resources Management. In 2019 IEEE 15th International Conference on Intelligent Systems and Applications (ISA) (pp. 1-6). IEEE.
13.Costa, P. T., & McCrae, R. R. (1990). The Five-Factor Model of Personality and Its Implications for Research in Psychology. Psychological Inquiry, 1(4), 215-225.
14.Costa, P. T., & McCrae, R. R. (1992). The Revised NEO Personality Inventory (NEO PI-R). Psychological Assessment Resources, Inc.
15.McCrae, R. R., & John, O. P. (1994). A Literature Review of the Five-Factor Model of Personality (FFM). Journal of Personality and Social Psychology, 66(4), 586-605.
16.Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. San Francisco, California, USA
17.Camacho, I., & Baena-García, M. J. (2017). An Intelligent System for Curriculum Vitae Parsing and Qualification Extraction. IEEE Transactions on Learning Technologies, 10(2), 142-155.
18.Ribeiro, M. T., Santos, P. S., de Almeida, J. M., & Gonçalves, M. F. (2018). Text Classification for Applicant Screening in Recruitment Processes Using Machine Learning. IEEE Access, 6, 10271-10283.
19.Boyatzis, R. E. (1998). Transforming qualitative information: Thematic analysis and code development. Qualitative Research in Health Care, 3(2), 63-94.
20.Van Der Burgh, M., de Beurs, E., & Weijters, A. (2020). AI in HRM: A Review and Research Agenda. Human Resource Management Journal, 30(1), 258-277.
21.Yang, L., & Zhang, Y. (2017). Multi-Task Learning for Personality Prediction and Social Network Analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 1563-1572).
22.Youyou, W., Liu, H., Tao, Z., & Song, Y. (2018). Incosistency Detection in Large-Scale Personality Assessments. Proceedings of the 2018 Conference on Human Information Processing Systems (pp. 1845-1854).
23.Eichhorn, A., Koolen, P. M., & Straatman, H. (2010). A comparative validation of the HEXACO personality inventory. Journal of Research in Personality, 44(2), 281-290.
24.Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1921-1958.
25.James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media.