Personality Detection from Resume

Project Mentor:

PROF. ARNAB CHAKRABORTY

Team Members:

- Sharmistha Das (10800123186)
- Rishika Mishra (10800123164)
- Ragini Gupta (10800123145)
- Shreya Banerjee (10800123192)

CONTENTS

Project Objective and Scope

Objectives:

- Predict candidate personality traits using resume text.
- Provide insights for recruitment and career guidance.
- Apply on the test datasets and compare the differences in the results

Scope:

- Useful for HR teams to shortlist candidates.
- Helps in training and career recommendation.
- Can be applied across industries.

DATA DESCRIPTION

The dataset used for this project consists of structured resume information collected from Kaggle. It contains **~9,500 records** and multiple columns that describe various aspects of a candidate's resume, such as skills, education, experience, and job positions. The target variable is the **personality label (MBTI type)**, which is mapped to resumes based on their features.

➢ **Dataset Characteristics**

- **Total Records**: 9,544
- **Features**: 30+ attributes including career objective, skills, education, job roles, certifications, etc.
- **Target Variable**: MBTI personality type (INTJ, ENTP, INFJ, ESTJ, etc.).
- **Data Types**: Mostly categorical/text, with one numeric column (matched_score).
- **Missing Values**: Present in several columns (e.g., career_objective, experience_requirement, address, etc

Table 1 : Feature Information

Column Name	Non-Null Count	Null Count	Data Type	Description
career_objective	4740	4804	object	Candidate’s professional goal/objective
skills	9488	56	object	Candidate’s listed skills
degree_names	9460	84	object	Academic degree(s) held
educational_institution_name	9460	84	object	Name of university/college
job_position_name	9544	0	object	Current or desired job position
experiencere_requirement	8180	1364	object	Years of required/mentioned work experience
age_requirement	5457	4087	object	Age-related requirements if specified
certifications	2008	7536	object	Certifications acquired by candidate
Responsibilities	9544	0	object	Roles and responsibilities in past jobs
matched_score	9544	0	Float64	Matching score between

Table 1 : Text Feature Length Statistics

Feature	Min Length	25%	Median	75%	Max Length	Description
career_objective	26	144	210	268	1425	Candidate’s objective/summary statement length
skills	2	161	243	504	3104	Length of skills list (characters)
degree_names	6	10	22	39	472	Degree names length (characters)
educational_institut -ution_name	8	27	42	61	212	Institution name length (characters)
job_position_ Name	5	12	22	35	150	Job title length (characters)

➢ Insights from Dataset

1. **Career Objective** is available in ~50% of resumes; it varies from very short (26 characters) to detailed statements (1425 characters).

2. **Skills** are well populated, with a median length of ~243 characters, reflecting a mix of technical and soft skills.

3. **Degree Names and Institutions** are consistent with typical resume formats (median length ~22–42 characters).

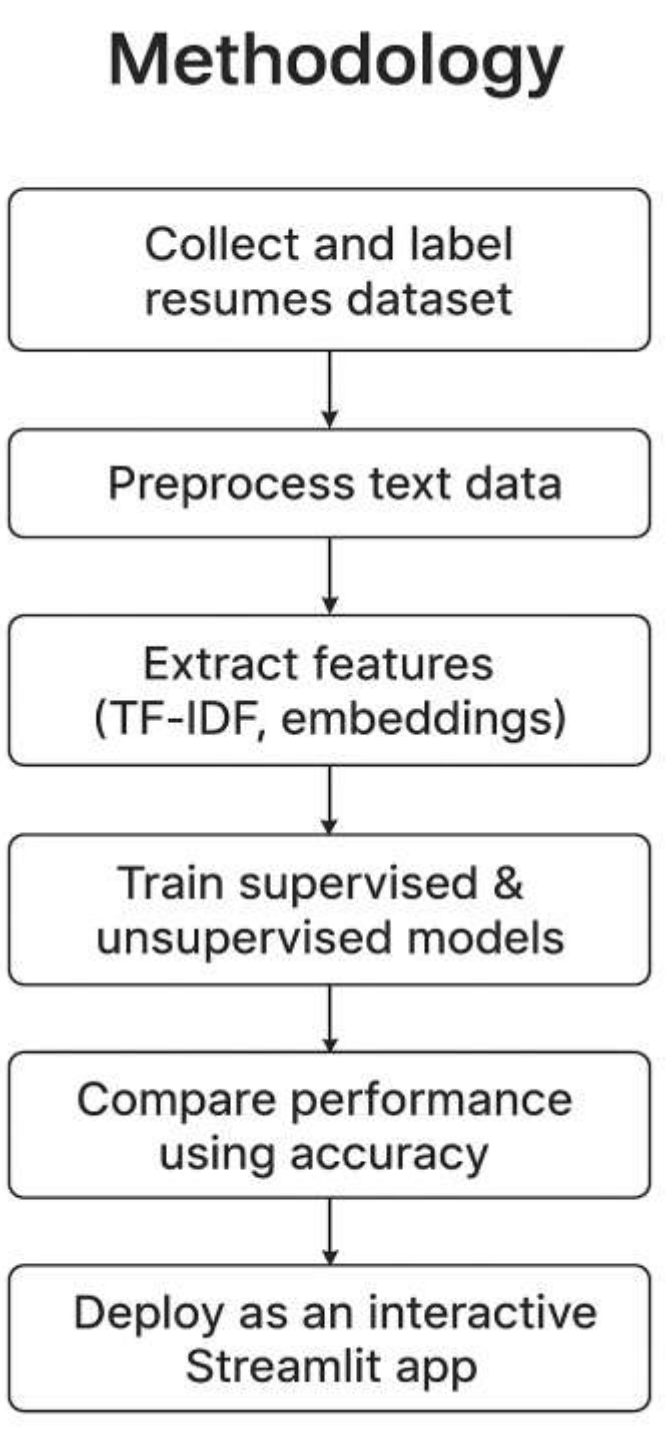
4. **Job Position Names** are always present, with reasonable lengths (median ~22 characters).

5. **Matched Score** is a numeric column (0–0.97), representing how well a resume aligns with a job posting, which can be used as a strong feature for prediction.

6. Several columns like **address, extracurriculars, certifications** have high missing values and may need to be dropped or imputed.

Methodology

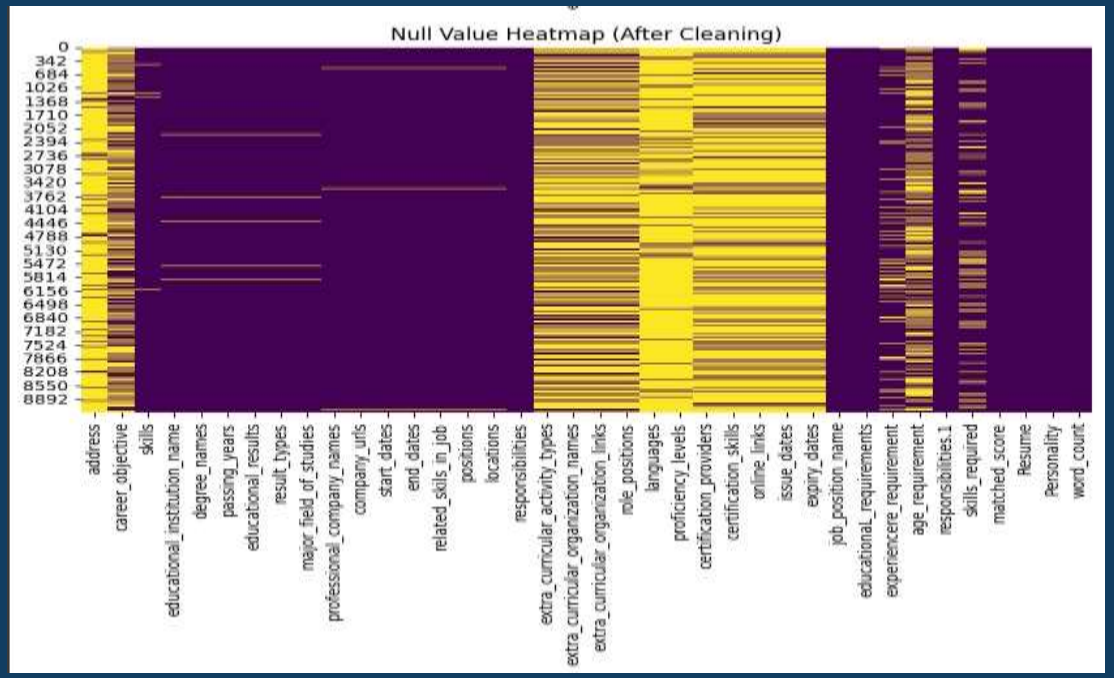
- Collect and label resume dataset.
- Preprocess text (cleaning, normalization).
- Extract features using TF-IDF (unigrams + bigrams).
- Train supervised (Logistic Regression, Random Forest) & unsupervised (K-Means, Hierarchical) models.
- Compare model performance using accuracy.
- Deploy best model in an interactive Streamlit app.



DATA PREPROCESSING

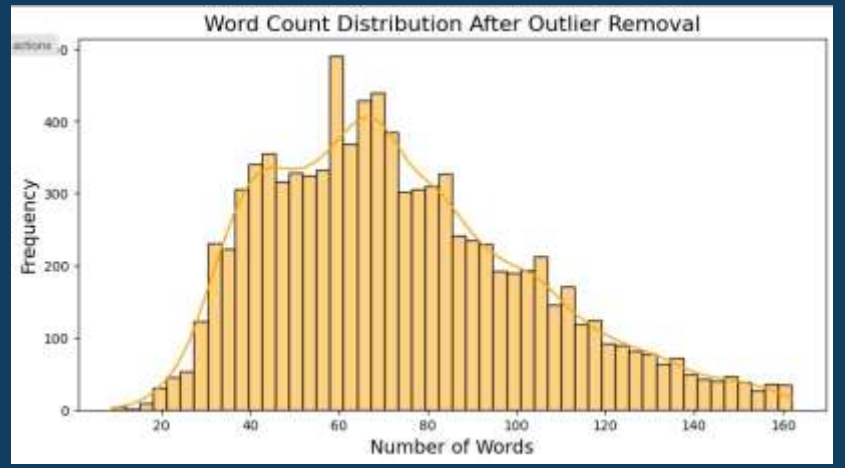
❑Missing Values(Null Values):

Rows with missing text (resume content) or missing labels (personality type) were dropped from the dataset to maintain data integrity.

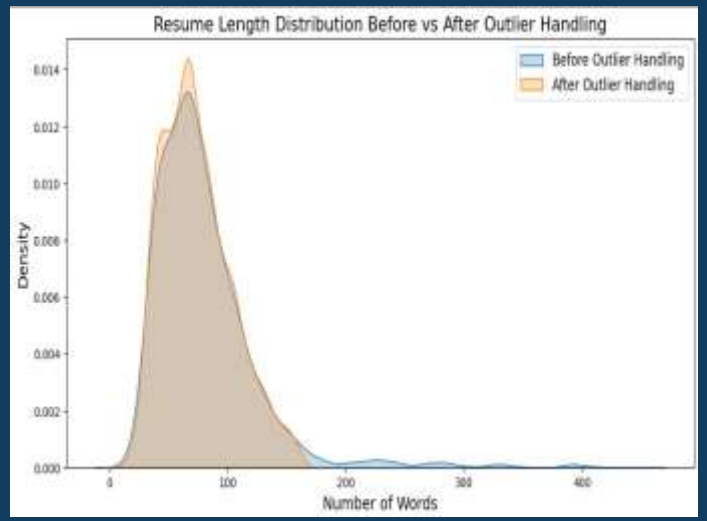


The plot showed that the dataset was complete, with no missing records.

❑Handling outliers:



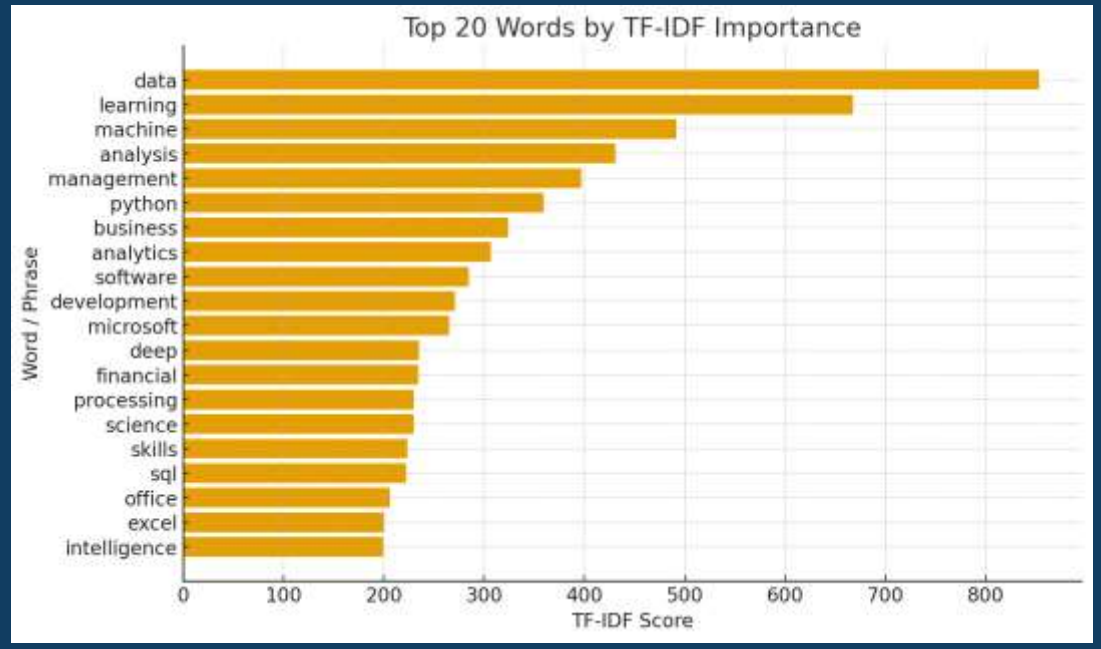
- Word count Distribution after Outlier Removal



- After handling outliers, **3,881 resumes** remained.

❑ Feature Extraction with TF-IDF

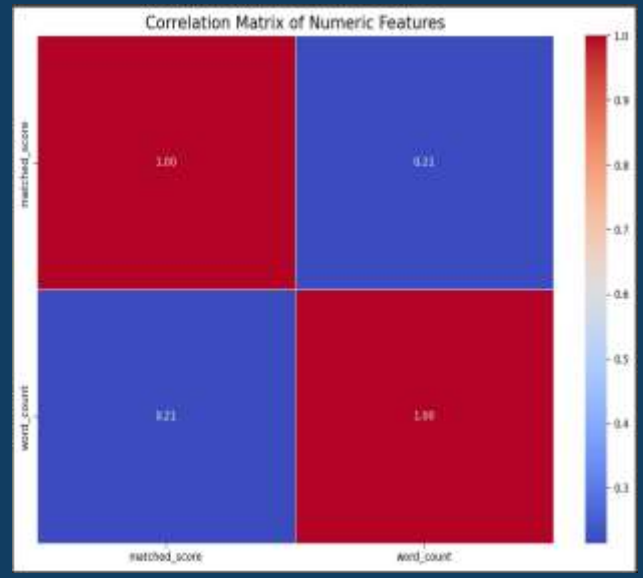
Once the vocabulary was finalized, the resumes were converted into **numerical feature vectors** using **TF-IDF (Term Frequency – Inverse Document Frequency)**.



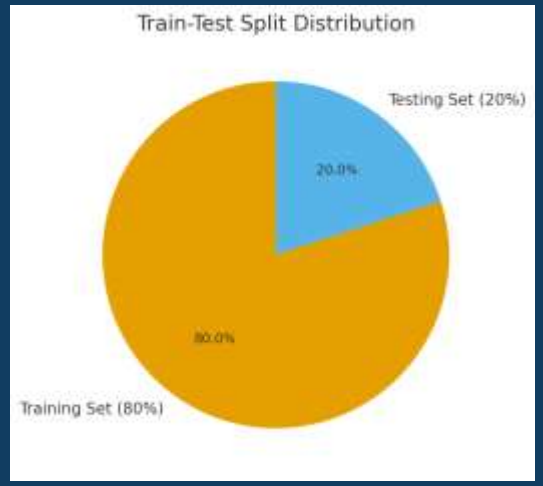
TF-IDF highlights words that are **important in a given resume but less frequent across all resumes**, making it well-suited for personality prediction.

❑ Features by Correlation Matrix

- we computed correlations between features and labels using **Correlation Matrix feature selection**.



❑Train-Test Split



- **Train-Test Split Pie Chart (80% training, 20% testing).**

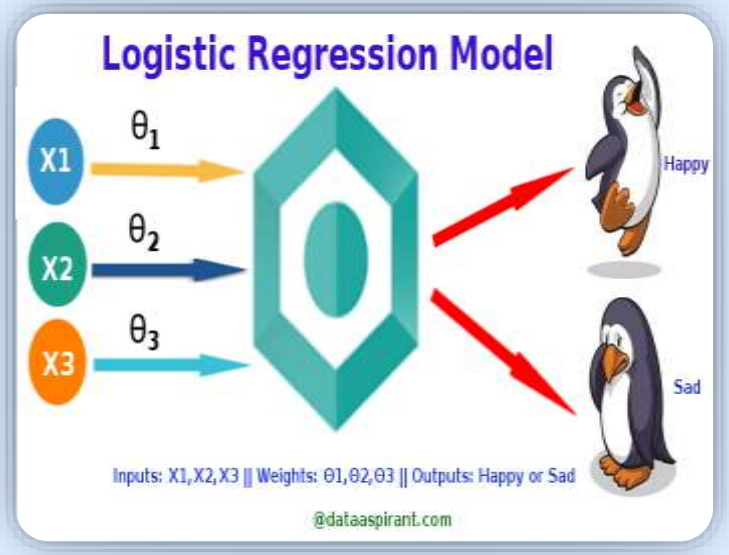
MODELS USED

The Machine Learning models used for this project are:

- Supervised Learning Models
  - Logistic Regression
  - Random Forest
- Unsupervised Learning Models
  - K-Means Clustering
  - Hierarchical Clustering

LOGISTIC REGRESSION(SUPERVISED)

Logistic Regression: is a popular and foundational model used for classification tasks, especially when the outcome variable is binary (i.e., it has two possible classes). Despite its name, logistic regression is not a regression model in the traditional sense (which predicts continuous outcomes); rather, it is used for predicting categorical outcomes.



The Logistic regression equation can be obtained from the equation. We know the equation of the straight line can be written as:

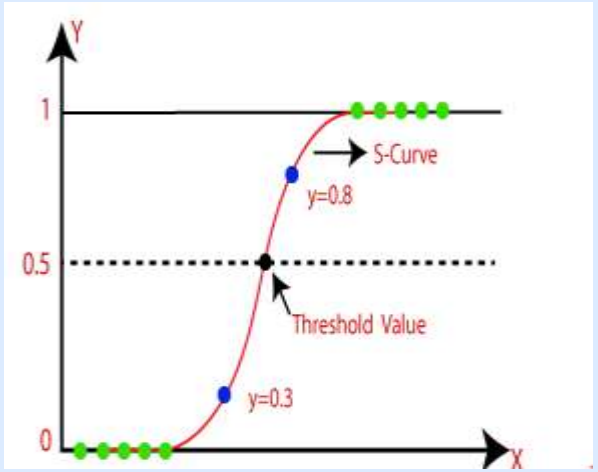
$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In this, y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

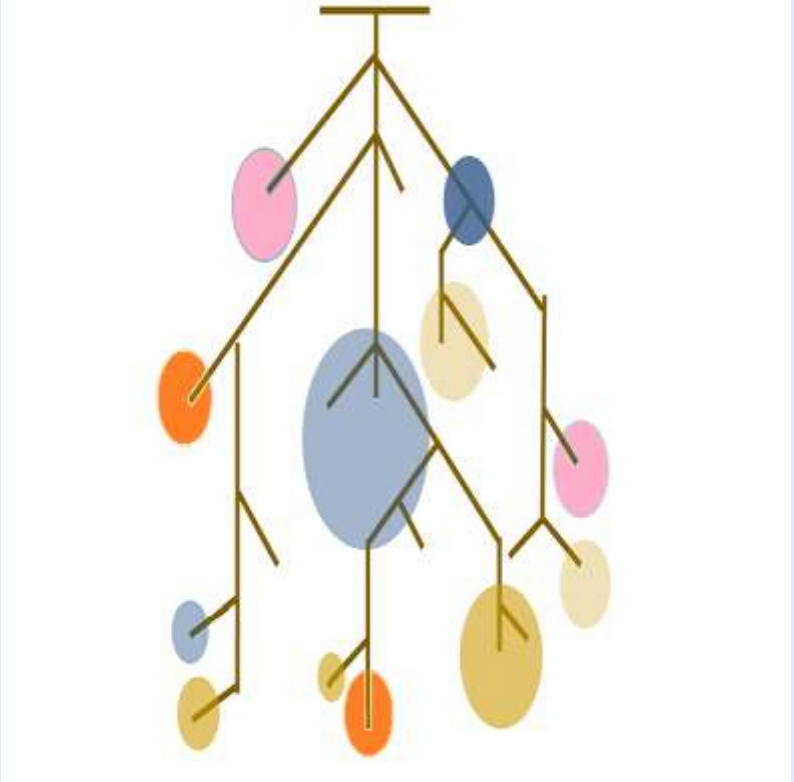
But we need range between -[infinity] to +[infinity], then take logarithm of the equation:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$



RANDOM FOREST(SUPERVISED)

Random Forest is an ensemble method that combines multiple decision trees to improve classification accuracy and robustness. Each tree is trained on a random subset of the data, and the final classification is made by majority vote among the trees.



➢ Why Random Forest?

- Handles **high-dimensional data** like resumes (thousands of words).
- Captures **complex, non-linear patterns** in personality traits.
- More robust and less prone to **overfitting** than single decision trees.

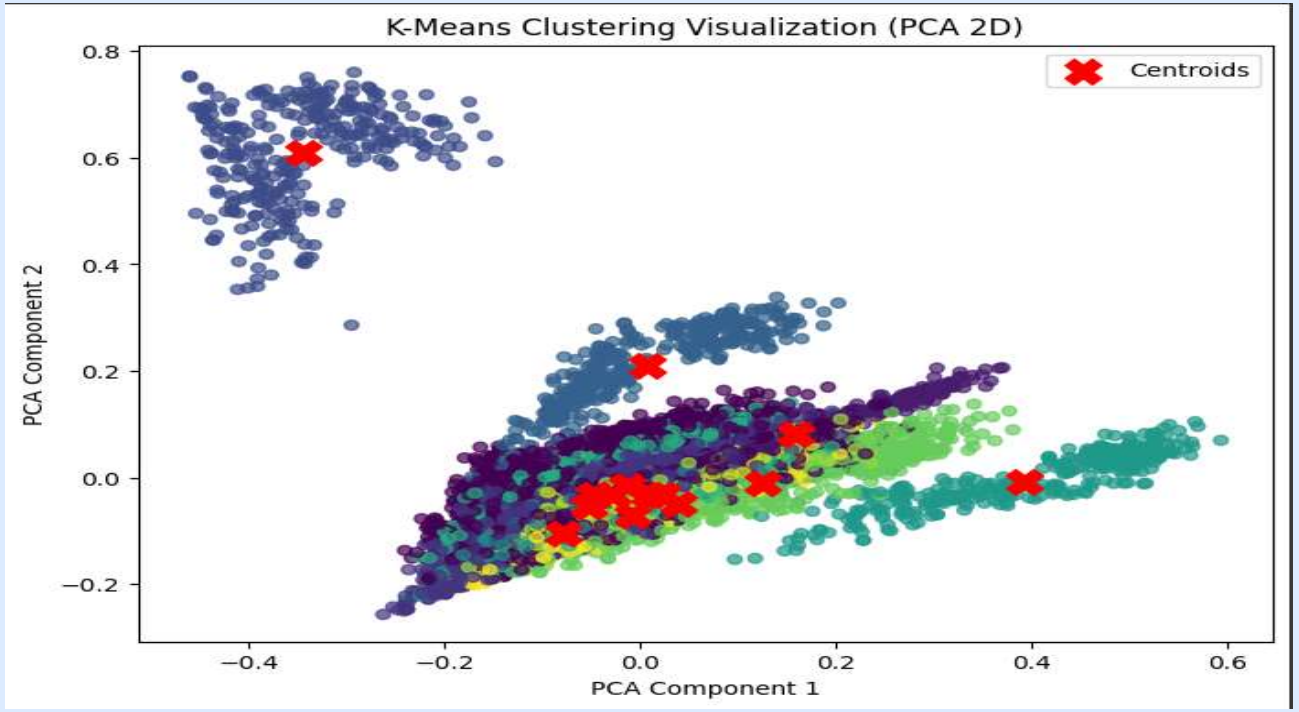
➢ Application in Project

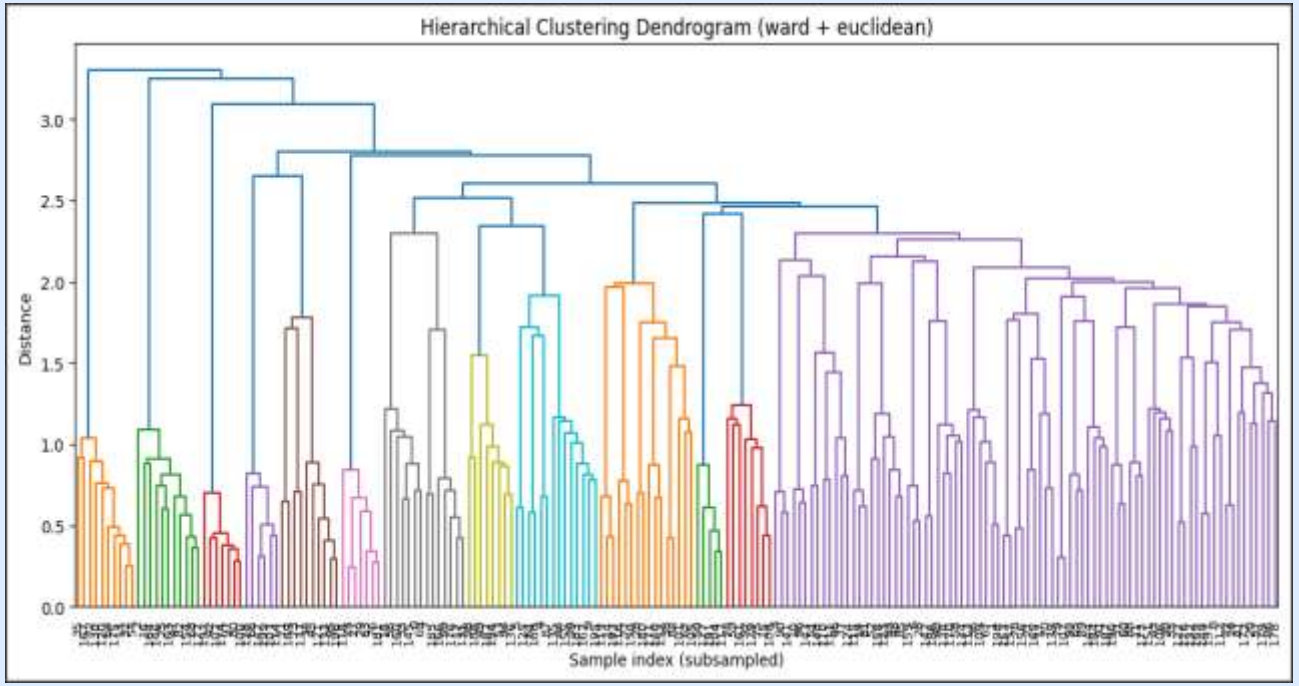
- Trained on resume text features (TF-IDF vectors).
- Compared against Logistic Regression and clustering models.
- Showed **good accuracy**, but more complex than Logistic Regression.

CLUSTERING MODELS

Clustering and models are used in **unsupervised learning** to group similar data points into clusters based on their features. The goal is to find natural groupings within the data without predefined labels.

- **K-Means:** Groups resumes into similar clusters.
- **Hierarchical:** Creates nested clusters for analysis.
- Useful for exploring **hidden personality patterns**, but less accurate.

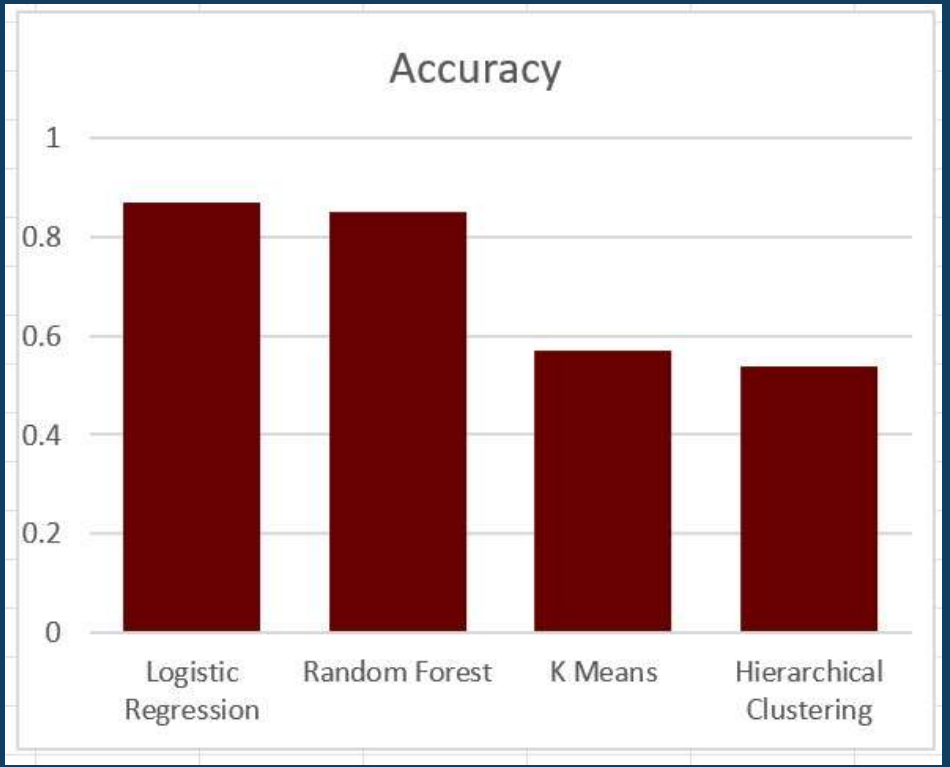




ACCURACY COMPARISON

- Models were evaluated on the resume dataset using accuracy as the metric.
- **Supervised models** outperformed unsupervised models.
- **Random Forest** was close, showing robustness but slightly more complex.
- **Logistic Regression** achieved the **highest accuracy**.
- **Clustering models (K-Means, Hierarchical)** showed weaker alignment with true labels.

Model	Accuracy
Logistic Regression	**0.87**
Random Forest	0.85
K-Means	0.57
Hierarchical Clustering	0.54



INFERENCE

✅**Logistic Regression** provided the **best balance of accuracy (0.87) and simplicity**.**Random Forest** was stable but more complex.

✅**Random Forest** performed well (0.85) but was more complex and harder to interpret.**Supervised ML > Unsupervised ML** for this problem.

⚠️**K-Means (0.57) and Hierarchical Clustering (0.54)** showed **weak alignment** with true personality labels.

⚠️**Supervised Learning models outperformed Unsupervised Learning** for this problem.

✅Personality prediction from resumes is **feasible**, but requires:

- High-quality labeled datasets.
- Better text representation (e.g., embeddings, deep learning).

FUTURE SCOPE OF IMPROVEMENTS

- Expand dataset with more diverse resumes.
- Use advanced NLP models (Word2Vec, BERT).
- Apply deep learning for higher accuracy.
- Develop explainable AI for transparency.
- Deploy as a cloud-based HR tool.
- Integrate with LinkedIn/job portals.

Thank you

Asansol Engineering College

- Sharmistha Das (aec.cse.sharmisthaj904@gmail.com)
- Rishika Mishra (aec.cse.rishikamishra@gmail.com)
- Ragini Gupta (aec.cse.raginigupta555@gmail.com)
- Shreya Banerjee (aec.cse.shreyabanerjee002@gmail.com)