IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Resume classification-based on personality using Machine Learning Algorithm

[1]Ms. Praniti Ram Patil

[1]Post Graduate Student

[1]Department of Information Technology,

[1]Thakur College of Engineering and Technology, Kandivali (E), Mumbai, Maharashtra, India, 400101

Abstract: This paper describes personality classification experiment by applying k-means clustering machine learning algorithms. Several previous studies have been attempted to predict personality types of human beings automatically by using various machine learning algorithms. However, only few of them have obtained good accuracy results. To classify a person into personality types. With the onset of the epidemic, everything has gone online, and individuals have been compelled to work from home. There is a need to automate the hiring process in order to enhance efficiency and decrease manual labour that may be done electronically. If resume categorization were done online, it would significantly save paperwork a k and human error. The recruiting process has several steps, but the first is resume categorization and verification. Automating the first stage would greatly assist the interview process in terms of speedy applicant selection. Classification of resumes will be performed using Machine Learning Algorithms Random Forest.

Index Terms - Personality types, machine learning, Jungian, kmeans clustering on personality test, Radom Forest, resume classification Job seekers, Resume recommendation, Job search, Resume matching, Machine learning

I. INTRODUCTION

Interviews are becoming time-consuming affairs. Employees are required to travel to locations and conduct interviews, and it is difficult to manually remember each and every aspect of a candidate or the interview process. In many instances, the artificial intelligence system assists us in simplifying things. Using the conventional method of recruitment, an organization's HR department invites individuals based on their resumes to an interview for a specific position. This HR department manually evaluates a candidate's skills based on their résumé to determine if he or she is qualified for the position or not. HR's conduct interviews, and the panel plays a significant role in determining who is the best applicant for the post. They examine not just the candidate's talents, but also his or her personality is a combination of a person's characteristics and attitudes in dealing with different social situations as in kindergarten, school, university, family, working team, etc. Humans are addicted to biases and prejudices that might affect their judgmental accuracy Personality can be taken as assessment in various fields such as selection of staff, choice of profession, relationship and health counseling. There is a great effect of personality on learning capabilities of humans. For instance, in learning performance we may see significant performance we may see significant differences between extroverts and the ones belonging to introverts.

Maintaining resumes and profiles of all candidates becomes a very tedious job when it comes to big mass recruitment companies because they provide employment in bulk, and thus maintaining or storing data physically is not possible. Machine Learning enables the path through which a computer can be trained to follow specific instructions again and again to make human life easy. The most common usage of machine learning is for the classification of objects. In machine learning, iteration is important because models are exposed to new data and adapt accordingly. Machine learning models learn from previous results and computation to produce correct and reliable decisions. In statistics, classification is a supervised learning concept in which segregation of data.

II. OBJECTIVE

Today the major problem being faced across the industry and candidate is how to acquire the right talent, using minimal resources over the internet and in minimal time. there are three major challenges that are required to be overcome, to bring efficiencies to the complete process find the personality candidate, find the job title for candidate. The solution would help to find the right CV from the large dumps of CVs; would be agnostic to the format in which CV has been created and would give with the list of CVs which are the best match to the job description provided by the recruiter The proposed solution involves supervised learning to classify the resumes into various categories corresponding to the various domains of expertise of the candidates. A multi-pronged approach to classification is proposed

III. METHODOLOGY

Nowadays searching for jobs is a difficult and tedious process for both the employees and the employers. The traditional method for classifying resumes is very time consuming and the concerned authorities need to go through every resume sent by the large number of candidates. This process becomes very complicated because there are millions of engineering graduates passing out every year runs for getting a job. For making the process easier there needs to be match between qualification, experience and many more criteria of candidate and company expectation. In our system, candidates will be sending resume and Classification will be done using Machine Learning such as k-means and random forest

- This project is based on three modules:
- This Module is used to find the Personality of a person.
- This module is used to find the Job title of individual.
- This module is used to classify the resume

IV. DATASET AND MODEL DISCRIPTION

This data was collected (2020-2021) through an interactive on-line personality test. The personality test was constructed with the "Big-Five Factor Markers" from the Participants were informed that their responses would be recorded and used for research at the beginning of the test, and asked to confirm their consent at the end of the test. The following items were presented on one page and each was rated on a five-point scale using radio buttons. The order on page was EXT1, AGR1, CSN1, EST1, OPN1, EXT2, etc. The scale was labelled 1=Disagree, 3=Neutral, 5=Agree Personality test also known as OCEAN Model analyses personality types of individuals based on five dimensions – Openness(O), Conscientiousness (C), Extraversion (E), Agreeableness (A), Neuroticism(N). With each of the dimensions signifying a different personality type. It uses keywords to identify traits and analysed in which personality a person fit. Openness: As the word suggests, This quality features characteristics such as openness and imagination and curiosity. Conscientiousness: Conscientiousness talks about a high amount of thoughtfulness, a goal- oriented attitude and good decision-makers. Extraversion: Extraversion also means extroversion is identified by excitement, talkativeness and assertiveness. Agreeableness: Agreeableness refers to features such as trust, affection and social behavior of an individual. Neuroticism: Neuroticism includes attributes like sadness, moodiness and sudden burst of emotions.

Fig.1. personality question

V. SYSTEM ARCHITECTURE

First the dataset needs to be scraped which can be done using various websites. Once the dataset has been scrapped, pre-processing of data is to be done by doing proper stemming and lemmatization, removing stop words and filler words store the relevant words in a separate column for further process. After the pre-processing of data is done, making sure that the relevant data have the words which have occurred the highest number of times need to be put in a term-frequency document using TFIDF Vectorization. Once the previous steps are done, the cleaned data is now ready to test and form classification models using machine learning algorithms. Analyzing the model by their accuracy and the forming confusion matrix for the same is required for better understanding of the results.

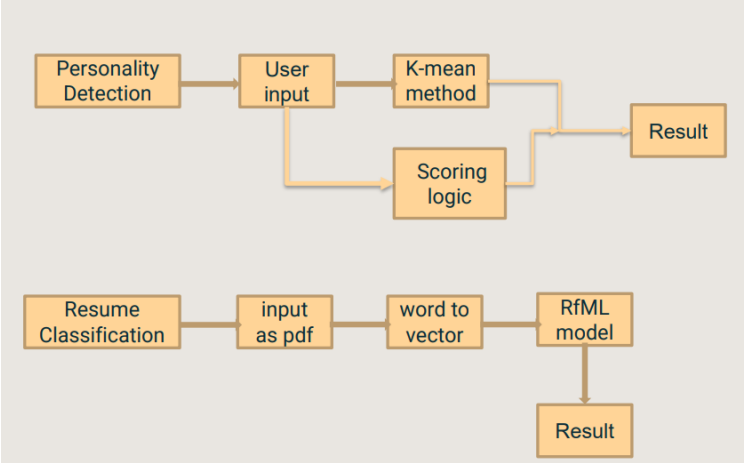


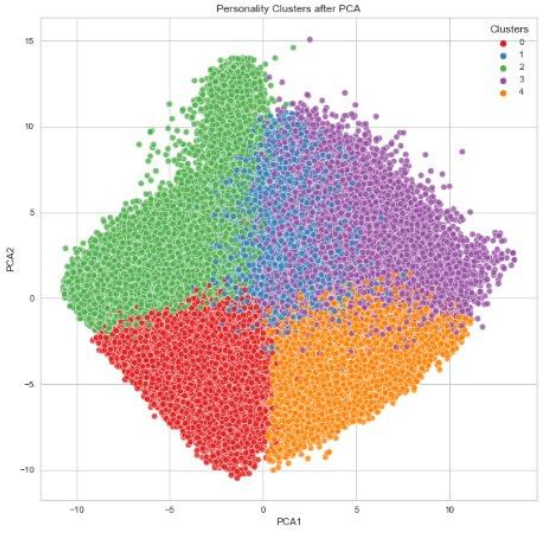
Fig.2 System Architecture

VI ALGORITHM

7.1 k-means Clustering Algorithm

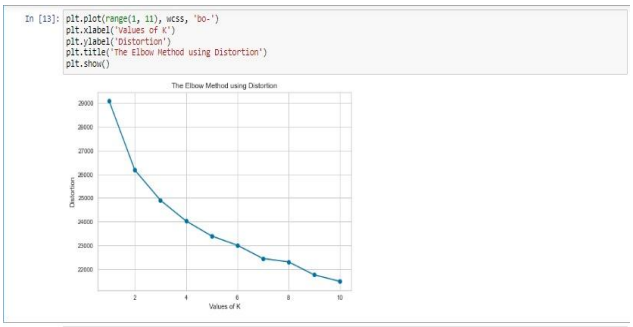
The k-means method is the most popular clustering method. It refers to unsupervised learning part in machine learning field. The most well-known algorithm utilizes an iterative refinement technique. By cause of pervasiveness, it is regularly called the k-means algorithm. The algorithm continues by switching back and forth between two stages Assignment step: Assign every observation to the cluster with mean (average of a set of values), which has the slightest squared Euclidean distance which shows how close or far away two observations from each other [36]. We calculated the distance as follows: where x and y are samples in n-dimensional feature space. Update step: Calculation of values of new means is made of clusters that will become the centroids of observations in the new clusters. If the assignments do not change, the algorithm converges. Generally, we define randomly k clusters in the plane, then we calculate distances of each data point to these k clusters and assign the observation to the closest centroid, after that we move centroids to the mean of assigned values to them. We repeat this process until convergence, when the values of centroids do not change after iteration. Using this algorithm, we cannot be sure that the most optimal result will be found. We can stop the algorithm from converging using other distance functions as Manhattan distance functions for instance. There are other k-means transformations, especially the spherical k-means and k-medoids.

$$similarity = \cos(\theta) = \frac{A \,.\, B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i\, B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\,.\sqrt{\sum_{i=1}^{n} B_i^2}}$$



the error of each data point

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} || \, x_i^j - C_j||^2$$



7.2 Random Forest Algorithm

Random Forest is a classification algorithm that works on the principle of decision trees. It takes in input of many decision trees and gives the best majority output from all the inputted decision trees. On the RF Classifier the training data is fitted. Then, in thevalidation dataset labels are being predicted.

Step 1 − First, start with the selection of random samples from a given dataset.

Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
Decision tree

$$fi_j = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

fi sub(i)= the importance of feature i
ni sub(j)= the importance of node j
Step 3 − In this step, voting will be performed for every predicted result.

Normalized to a value between 0 and 1

$$normfi_j = \frac{fi_j}{\sum_{j \in all\ nodes} fi_j}$$

Step 4 − At last, select the most voted prediction result as the final prediction result.

$$RFfi_i = \frac{\sum_{j \in all\ trees} norm\ fi_{ij}}{T}$$

RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
normfi sub(ij)= the normalized feature importance for i in tree j
T = total number of trees

VII. RESULTS AND DISCUSSION

The results obtained from our model are summarized in the following table:

Model Evaluation

Algorithm name	Accuracy
Random Forest	
Decision Tree	

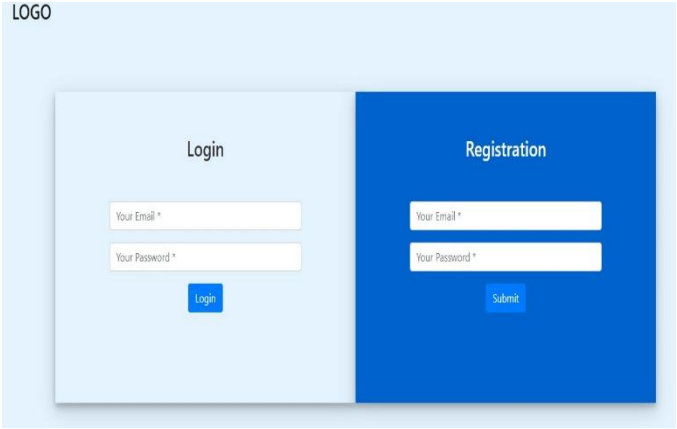


Fig. 3 Login Page

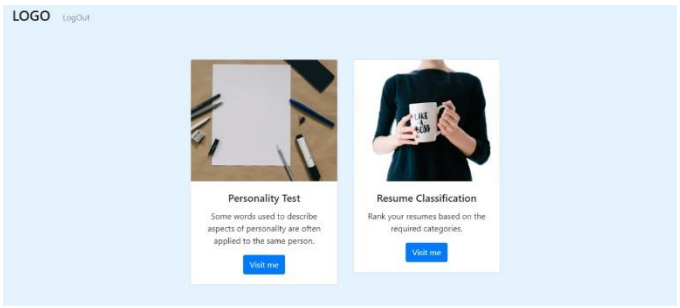


Fig. 4 Home Page

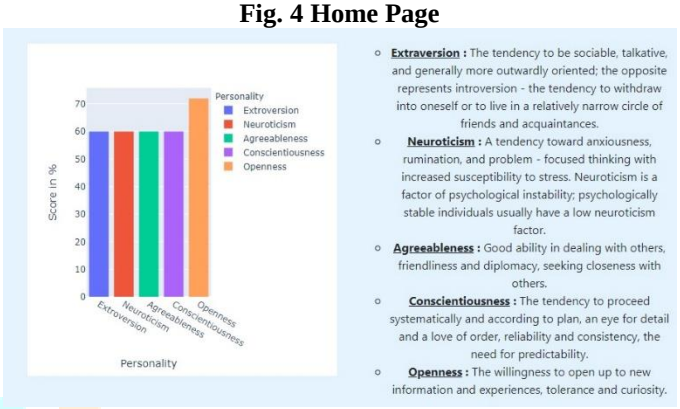


Fig. 5. Personality Types

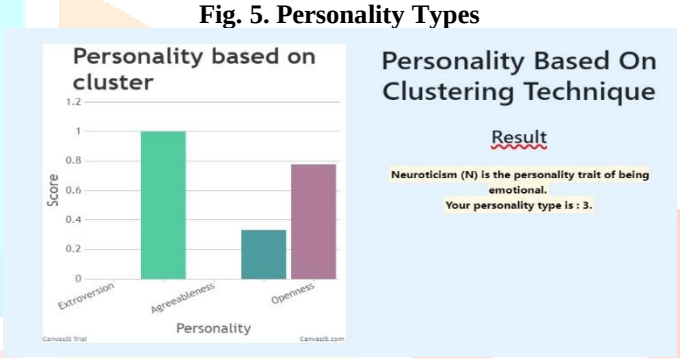
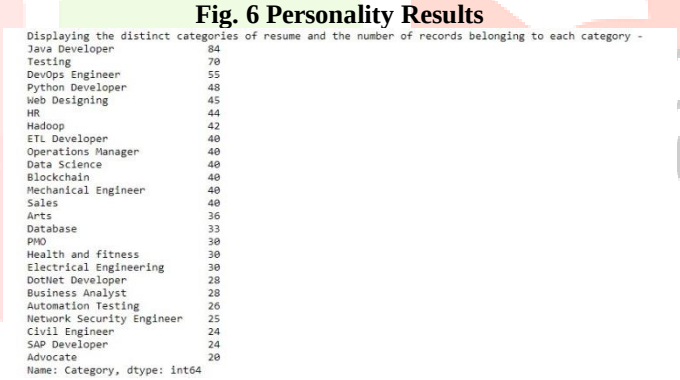


Fig. 6 Personality Results



```
Displaying the distinct categories of resume and the number of records belonging to each category -
Java Developer                84
Testing                       70
DevOps Engineer               55
Python Developer              48
Web Designing                 45
HR                            44
Hadoop                        42
ETL Developer                 40
Operations Manager            40
Data Science                  40
Blockchain                    40
Mechanical Engineer           40
Sales                         40
Arts                          36
Database                      33
PMO                           30
Health and fitness            30
Electrical Engineering        30
DotNet Developer              28
Business Analyst              28
Automation Testing            26
Network Security Engineer     25
Civil Engineer                24
SAP Developer                 24
Advocate                      20
Name: Category, dtype: int64
```

Fig.7. Job Title

Fig. 8 Resume Classification

VIII. CONCLUSION

Two models have been built on the cleansed data: i) Classification - Based on the resume and category the model has been designed to categories the resume in the right category and ii) describes personality classification by applying k-means clustering. We provided explanations of advantages and disadvantages of k-means clustering.

REFERENCES

[1] Iqbal H. Sarker, Machine Learning:Algorithms, Real world Applications and Research Directions, Springer Nature Computer Science(2021) 2:160, https://doi .org/10.1007/s42979-021-00592-x.

[2] Mr. Ramraj S, Dr.V. Sivakumar, Kaushik Ramnath G, Real-Time Resume Classification System Using LinkedIn Profile Descriptions, IEEE International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE-2020), July 29-31, 2020, Odisha, India.

[3] David C. Funder. Personality. Annual Review of Psychology, 52(1):197–221, 2001.

[4] Jieun Kim, Ahreum Lee, Hokyoung Ryu. Personality and its effects on learning performance: Design guidelines for an adaptive e-learning system based on user model In: 2013 Elsevier. International Journal of Industrial Ergonomics. 2013 p. 1–12. https://doi. org/10.1016/j.ergon.2013.03.001

[5].Ajmal M., Ashraf M.H., Shakir M., Abbas Y., Shah F.A. (2012) Video Summarization: Techniques and Classification. In: Bolc L., Tadeusiewicz R., Chmielewski L.J., Wojciechowski K. (eds) Computer Vision and Graphics. ICCVG 2012. Lecture Notes in Computer Science, vol 7594. Springer, Berlin, Heidelberg.https://doi.org/10.1007/978-3-642-33564-8_1.

[6]. H. Raksha, G. Namitha and N. Sejal, "Action based Video Summarization," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 457-462, doi: 10.1109/TENCON.2019.8929597. REFERENCES

[7]. S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hommeaux and R. Ptucha, "Semantic Text Summarization of Long Videos," 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, 2017, pp. 989-997, doi: 10.1109/WACV.2017.115.

[8] Das, A.S., Datar, M., Garg, A., Rajaram, S., 2007. Google news personalization: scalable online collaborative filtering, in: Proceedings of the 16th international conference on World Wide Web, ACM. pp. 271–280.

[9] Diao, Q., Qiu, M., Wu, C.Y., Smola, A.J., Jiang, J., Wang, C., 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars), in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 193–202.

[10] Farber, F., Weitzel, T., Keim, T., 2003. An automated recommendation approach to selection in personnel recruitment. AMCIS 2003 proceed- ¨ ings , 302.

[11] Golec, A., Kahya, E., 2007. A fuzzy model for competency-based employee evaluation and selection. Computers & Industrial Engineering 52, 143–161.

[12] Gunaseelan B, Supriya Mandal, Rajagopalan V , Automatic Extraction of Segments from Resumes using Machine Learning , 2020 IEEE 17th Indian council International conference.