Job Trends Analysis Using Rapid Miner

Poorvi Gupta[1], Manu Shrivastava[2], Poovammal E[3*].

Dept. of CSE, SRM Institute of Science and Technology, Kattankulathur, India-603203

[1]poorvigupta0217@gmail.com,[2]shrivastavmanu82@gmail.com, [3]poovamme@srmist.edu.in

Abstract: *Finding jobs online has become common with the increasing number of internet users and websites that provide full time jobs or internships. Due to the Covid-19 pandemic most of the applications and hiring processes are being conducted online through these websites. A person needs to find a suitable job based on their skill sets and interests. Sites like Naukri.com provide such opportunities by enlisting available offers, such that we can apply for a job in a few steps based on our interest. A job requires particular set of skills which also effects the other aspects related to a job such as salary, position, location etc. To analyze these trends many data-mining tools can be used. In this paper we have built a clustering model from the dataset naukri.com. Using the data-mining tool Rapid Miner, the model was developed and we were able to identify the clusters, correlation and other major features for taking decisions.*

Keywords: **Clustering, Data analysis, Mining Tools, Machine learning Algorithms**

1. Introduction

In today's fast-paced world, the companies are accumulating vast amounts of data every day. Data analysis is the process of making this data useful, by extracting and analyzing the data for various purposes. According to Forbes, around 90 % of the data which exists today, was formed and collected in the last two years itself. This happened because of the expeditious growth of internet of things. That said, the need to handle the data has increased. It becomes nearly impossible to clean, transform and evaluate the data manually, therefore numerous software(s), open-source tools exist to make the process easier for us.

In this paper, we have used RapidMiner, which is a data-mining tool and makes the analytics for large datasets a quicker process. The data we used in this paper for analysis is available on Kaggle, a Data-science company. The dataset consists of potential employees with its various attributes which effect their chances of getting hired. It consists of 11 attributes of potential employees and around 29,000 entries.Some important attributes include Job experience in years, the type of industry and role of the employee.

2. Related Works

There are many algorithms which are used for data analysis. In this section some similar works are shared. A similar but slightly different approach called U-k algorithm was proposed in this paper [1] where the variation value was given and clusters were formed. After which extra clusters were automatically removed and the number of clusters were changed.

In another study [2] a classification tool using python was developed in which the implementation of k-nearest neighbor was done for prediction, organization and accuracy of the code. Apart from implementing k-means algorithm, some papers show comparative studies[3].Results of an experiment conducted using neural networks and

*Corresponding author

multiple linear regressions was compared to multivariate KNN algorithm, where KNN showed better results, on a small set of data.

A very similar approach was used in [4], where certain students gave live examination using LMS and the results were analyzed using the Rapid Miner tool which was later used for clustering the same data. The results helped in determining what kind of special need was required by a particular student.

In paper[5], two data mining classification techniques were analyzed,namely SVM(support vector machine) and RF(random forest) using a disease dataset.In [6], data mining techniques were applied using RapidMiner tool to present a new framework on crime prediction.

3. Methodology

Data analysis can be done in various ways, depending upon the data available. The machine learning algorithms are categorized into supervised, unsupervised and semi-supervised algorithms.

In supervised learning, all the available data is labeled and it learns from this data to do the further analysis. Some examples areRain Forest (RF) and SVM. Whereas in unsupervised learning, data is not labeled and is taken as input. For example, k-means and Apriori algorithm. In semi-supervised learning, it is a mixture of known and unknown data,which can be used. One such example is graph-based models.

In the data used in our work, no labels are present therefore the method of clustering has been implemented. The tool chosen was Rapid Miner [7-10]. Clustering refers to the process of fractioning of data into points and then segregating the similar points, so as to identify the common traits between the various attributes.In k-means clustering algorithm, we are required to give in the input for the number of clusters we want.

Under which upper and lower limit have been given too, and is referred to as x-means.Both have their own decision trees along with a single co-relation table. The aim of this paper is to help recruiters in identifying the potential employees faster for their particular positions and requirements.

3.1. Clustering

On Rapid Miner platform, we can perform various functions like predictions, clustering, and deployment. We have created clusters, heat-map, decision tree and a co-relation table.The Rapid Miner user interface consists of four kinds of views on top, which are Designs, Results, Turbo-prep, Auto-model. Under designs, the process model is available, which displays the processes as they occur. We can load data, do outliers casting or forecasting.

There are numerous operators, which we can drag and drop to the process model area, and can be connected to the input/output port, according to our need. In Rapid Miner, the columns are usually referred to as attributes, and the rows as examples. So in the dataset used here, we have 11 attributes and some 29,000 examples.

In our work, we have used the auto-model feature of Rapid Miner, in which clustering is done. For auto modeling:-

1. The selected data is imported to the interface.
2. Selection of task is done. (In this case, clustering).
3. Target is prepared.
4. Attributes like Unique ID, key skills etc were removed for better analysis.

5. Under k-means, the number of clusters was set to 3.
6. Lower and upper limit were set to 3 and 10 respectively. Therefore the variation was from 3 to 10.
7. The extracted features were reduced to 500.
8. The separation of text attributes was done.

The results obtained in as shown in figure 1. The three bars in figure 1 represent the clusters, which are clusters 0, 1 & 2 respectively.

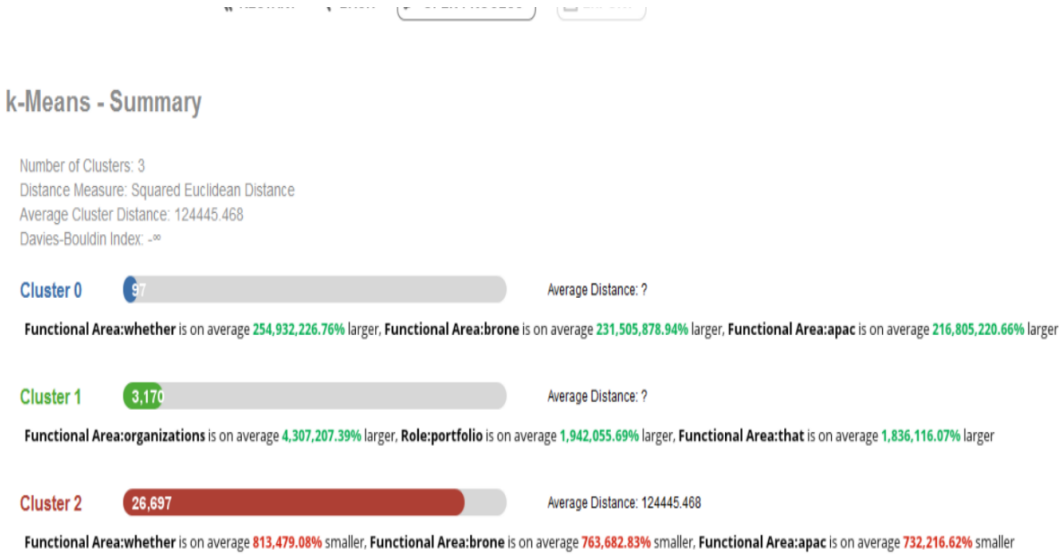

Figure 1. Clustering

Cluster 0 contains 97 out of the 29,000 examples / rows. Cluster 1 contains 3170 out of the 29000 examples, whereas cluster 2 contains 26,697 out of the 29,000 examples. As shown in fig.1, Cluster 3 which has the maximum members has the Functional area as 'brone', and is far greater than the others clusters having functional areas as 'apache hadoop' or 'sales manager product and portfolio'.

3.2. Heat Map

The figure number 2 represents a chart called the heat map. The values which are below the average are displayed in green otherwise it is red for above average values. The darker the color is, higher is the intensity. Therefore, from figure 2, we can see that in cluster 0, there are three high intensity functional areas whereas in cluster 1 & 2 there is no such areas.

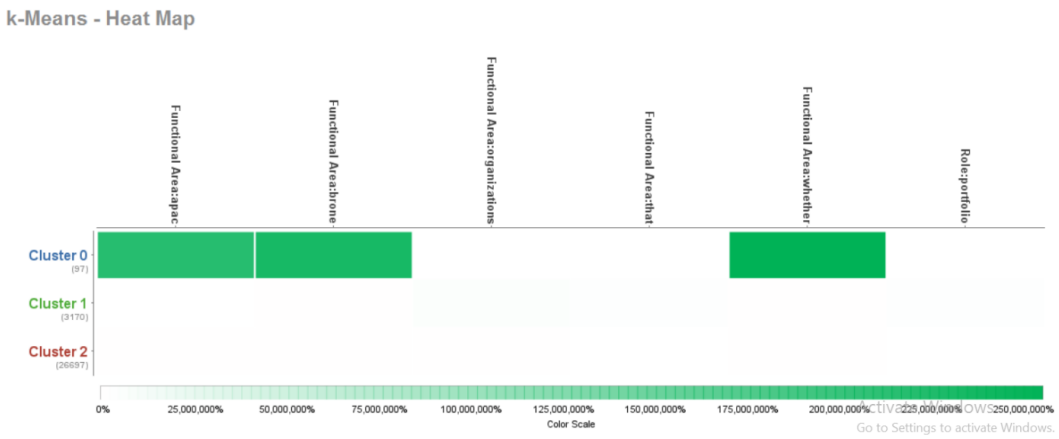

Figure 2. Heat map

3.3. Decision Tree Model

Figure 3 depicts the decision tree which was obtained. Since, the dataset was not having any designated class label, the suitableattribute was chosen. The feature through which the prediction can be done better according to Rapid Miner is the BPO (Business process outsourcing) attribute. Through this attribute we can categorize others. If in BPO the value is greater than 0.160, then it belongs to business category, in which there are further subdivisions in which if the value is greater than 0.08 it will belong to cluster 0, otherwise it will go on according to the tree.

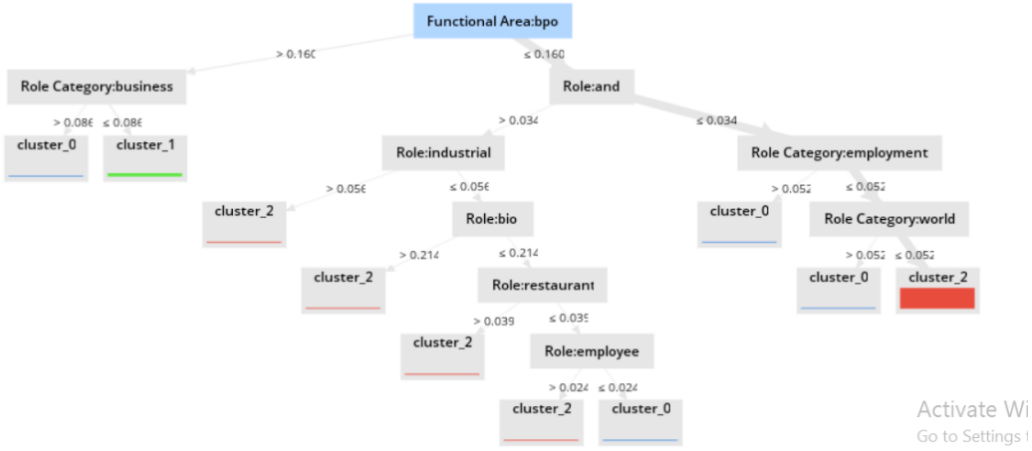

Figure 3.Decision tree model

3.4. Clustering performed by varying the cluster requirement

The 5 clusters were created as shown in figure 4, where in, Cluster 0 contains 22,384 out of the 29,000 examples/rows. Cluster 1 contains 4313 records, Cluster 2 contains 3101, Cluster 3 contains 69 and Cluster 4 contains 97 out of the 29,000 examples / rows. Clearly, cluster 0 has the highest number of similar equipment management roles, whereas cluster 3 consisted of smallest role category- ecommerce.

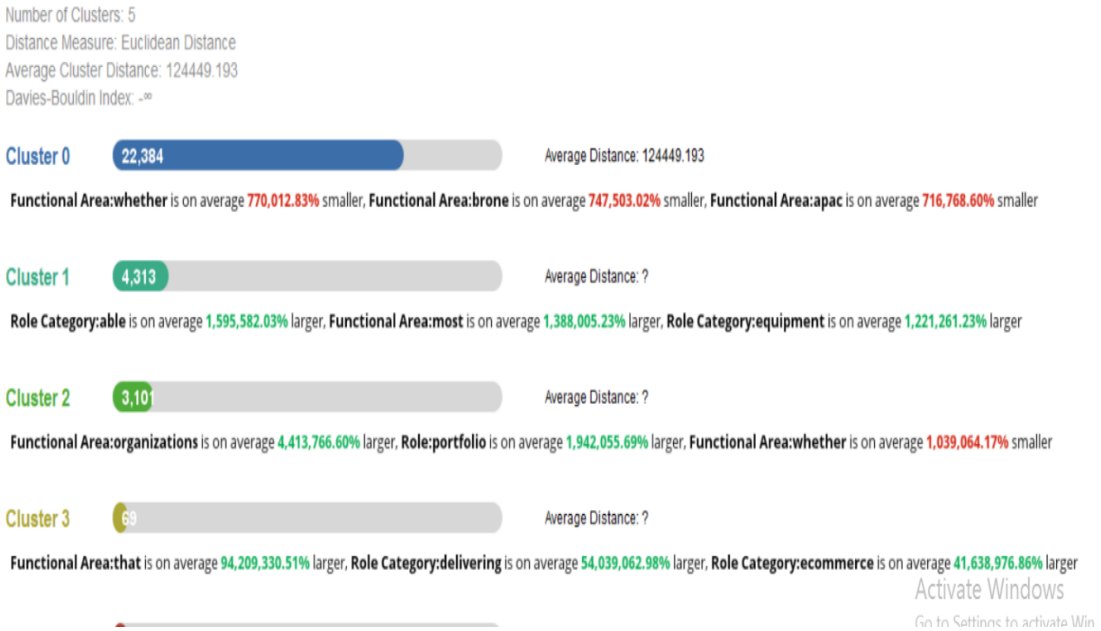

Figure 4. Clusters obtained by choosing the number of cluster as 5.

3.5 Heat map formed by varying the cluster requirement

Here in figure 5 it can be seen that cluster 0 has the highest intensity attributes which are below average whereas cluster 1 has the lowest intensity attributes.

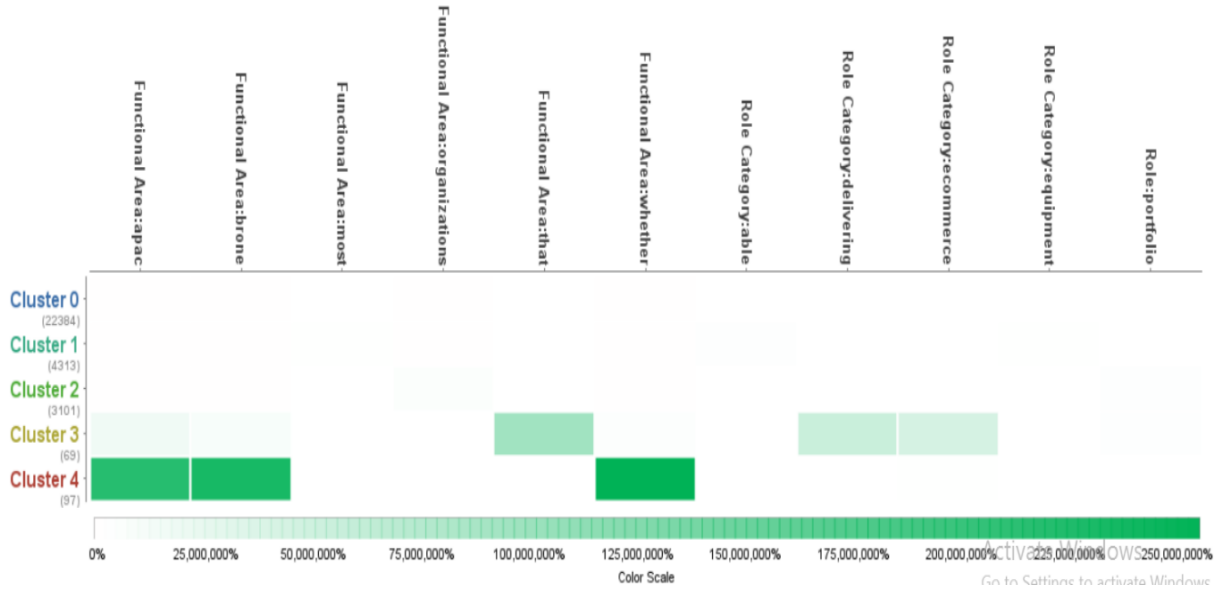

Figure 5. Heat Map using in model 1.

3.6 Decision tree using varied cluster requirements

Just like the previous example, the attribute through which we can start the categorization process is the functional area "business", now if its value is greater than 0.283 or less than that, accordingly the attributes will be distributed in the clusters as shown in fig.6.

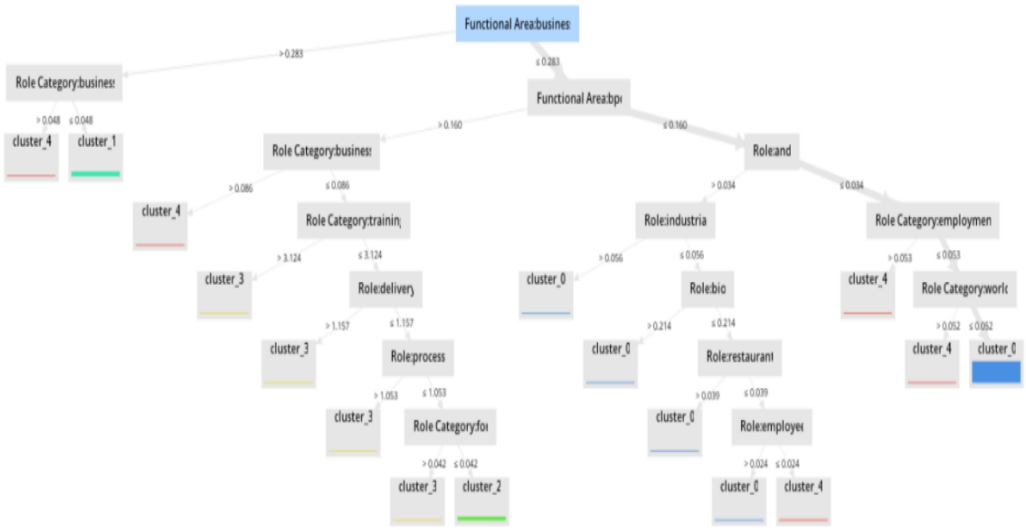

Figure 6. Decision tree

3.7 Co-relation Table

Finally a co-relation table which is common to both the processes is displayed in fig.7, which shows how strongly the attributes are related to each other and can be identified by how dark the box is. A second model was also prepared from the same dataset but with different chosen attributes. The second model was built using attributes key skills, roles, job and job salary.

Correlations

Attribut...	Crawl Ti...	Crawl Ti...	Crawl Ti...	days_di...	Functio...	Functio...	Functio...	Functio...	Functio...	Functio...	Functio...	Functio...	Functio...	Functio...	Fu
Crawl Ti...	1	-0.049	0.002	0.101	0.002	?	-0.001	-0.002	?	0.007	0.008	0.002	0.000	0.006	0.0
Crawl Ti...	-0.049	1	-0.005	-0.009	-0.009	?	0.023	0.005	?	-0.014	-0.010	-0.022	0.003	-0.017	-0.
Crawl Ti...	0.002	-0.005	1	0.009	-0.010	?	-0.010	0.003	?	0.002	-0.001	-0.010	0.022	-0.012	-0.
days_diff...	0.101	-0.009	0.009	1	-0.013	?	0.008	0.016	?	0.022	0.014	0.015	0.024	-0.006	0.0
Function...	0.002	-0.009	-0.010	-0.013	1	?	-0.003	-0.000	?	0.263	-0.003	-0.003	-0.001	0.915	-0.
Function...	?	?	?	?	?	1	?	?	?	?	?	?	?	?	?
Function...	-0.001	0.023	-0.010	0.008	-0.003	?	1	-0.003	?	-0.006	-0.051	-0.039	-0.014	-0.004	-0.
Function...	-0.002	0.005	0.003	0.016	-0.000	?	-0.003	1	?	0.558	-0.003	-0.003	-0.001	-0.000	-0.
Function...	?	?	?	?	?	?	?	?	1	?	?	?	?	?	?
Function...	0.007	-0.014	0.002	0.022	0.263	?	-0.006	0.558	?	1	-0.006	-0.005	-0.002	0.519	0.6

Figure 7. Correlation table in model 1 using algorithm

3.8 Clusters of second model

In this model, 3 clusters were formed based on key skills. As shown in figure 8, Cluster 0 has only 6 points out of 26528 example/rows. The key skills in this cluster are solid works, engineer and sales force. This cluster is also highly scattered. Cluster 1 has 42/26528 examples / rows. Key skills in this cluster include scripts (eg. JavaScript, Test Scripts etc.), equity, managing director. This cluster is also highly scattered as depicted by the size comparison in percentage indicated below the cluster. Cluster 2 has 26480 / 26528 examples / rows. Key skills in this cluster include scripts, solid works and equity. This cluster is highly dense as it has more number of examples in a relatively smaller area. Hence it is safe to conclude that the key skills included in cluster 2 are more in demand and can increase chances of applicant for finding a suitable job.

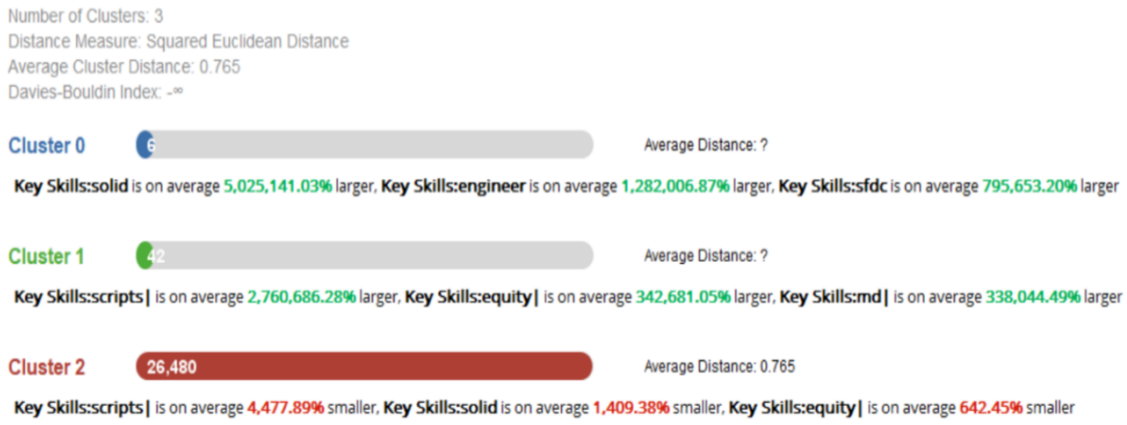

Figure 8. Clusters obtained by using K-Means algorithm in model 2.

3.9 Heat-map

According to the group of key skills present in a cluster the heat map depicts as shown in fig.9, that cluster 0 and cluster 1 have below average values with different functional intensity of each key skill whereas cluster 2 has no values below average.

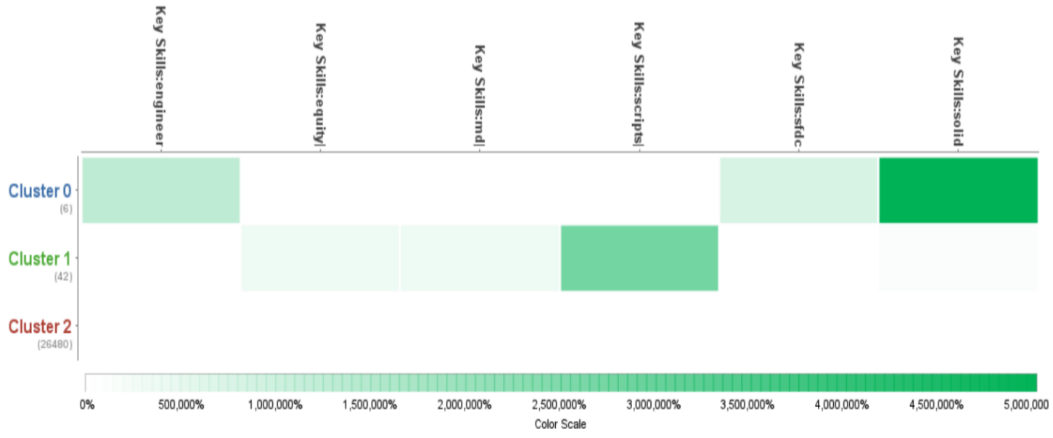

Figure 9. Heat map for the second model.

3.10 Decision tree

The decision tree for the second model depicts (figure 10) the way the examples/rows were distributed into different clusters based on key skills that map the way. For example if the value is greater than 0.256 then it belongs to cluster 1

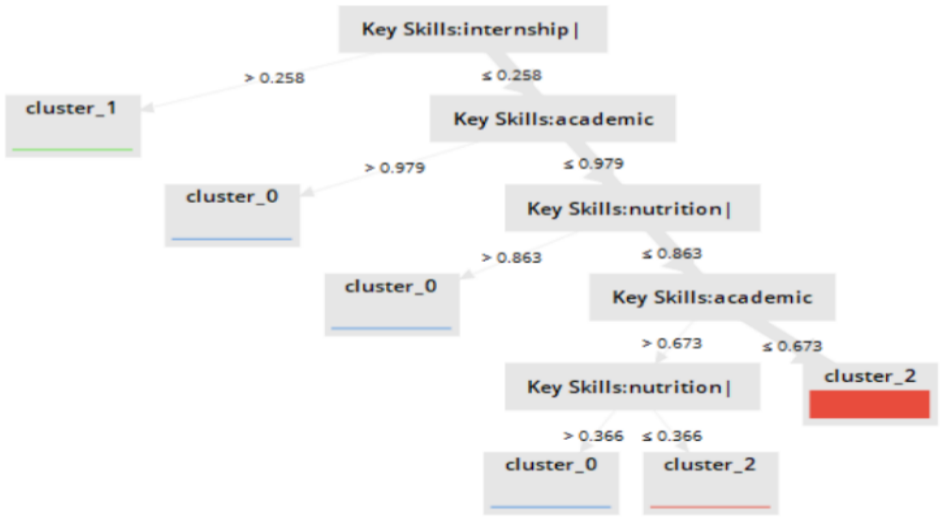

Figure 10. Cluster Tree in model 2.

.3.11 Centroid chart

In this model- 2, a centroid chart was prepared. When using a K-Means algorithm, a cluster is defined by a centroid, which is a point (either imaginary or real) at the center of a cluster. Every point in a data set is part of the cluster whose centroid is most closely located. The graph in fig. 11 has different key skills along the x-axis. This chart helps us to find the key skills that were located closely to the centroid of clusters. For example, the key skill 'react' had a value of more than 3 on y-axis. So the key skill react is more likely to be centroid of a cluster.

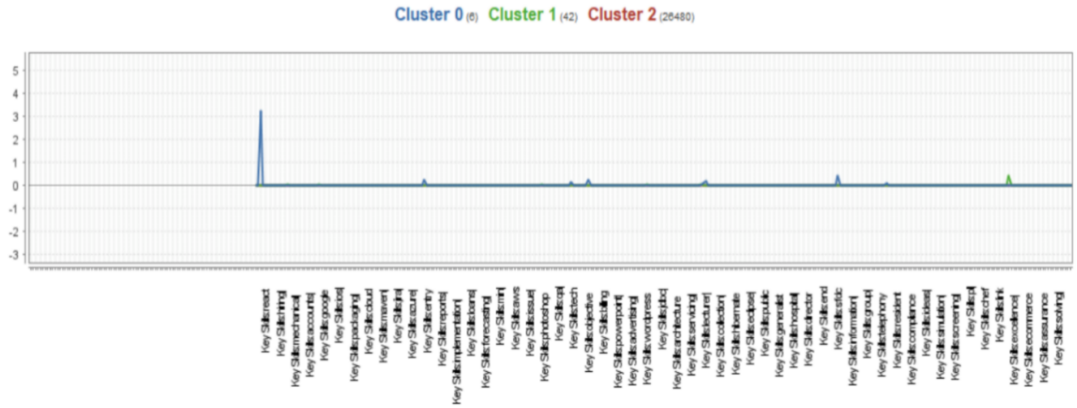

Figure 11. Centroid chart in model 2.

3.12 Centroid Table

The centroid table for this model 2, shown in fig.12, depicts the value for each key skill in the three clusters obtained. K-Means defines clusters by their central data point, i.e. the average of all elements in the cluster. These so called centroids are defined by the centroid table, where each column contains the attribute values of a centroid.

k-Means - Centroid Table

Cluster	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...	Key Skil...
Cluster 0	0.004	0.001	3.229	0.001	0.001	0.002	0.000	-0.001	0.000	0.001	0.001	0.000	-0.000	-0.000
Cluster 1	0.004	0.001	0.002	0.001	0.001	0.002	0.000	-0.001	0.000	0.001	0.001	0.000	0.014	0.040
Cluster 2	0.008	0.004	0.002	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002

Figure 12. Centroid table of model 2 .

4. Conclusion

In this paper we tried to demonstrate how modeling is done on Rapid Miner, using a particular dataset. Rapid Miner provides an easy platform because of integrated data preparation. We were able to identify the grouping of the clusters according to a particular attribute and we were able to form various decision trees, according to which a recruiter should able to easily identify which group the potential employee belongs to. The models also helps the readers to identify the skills of their interest that will help them build a stronger job profile in order to meet the demands of the best recruiters. Furthermore the models provide us with an idea of the path and trends that companies are following in order to maintain optimal taskforce. In the present world with overwhelming competition, this kind of knowledge if provided to students at the beginning stage of career will help them to boost their chances of getting hired.

References

[1] KRISTINA P. SINAGA AND MIIN-SHEN YANG,"unsupervised k-means clustering algorithm", IEEE Access, vol.8, April 2020.

[2] D Pedrozo, F Barajas, A Estupi˜n´an, K L Cristiano, andD A Triana Universidad Aut´onoma de Bucaramanga. " Data analysis for a set of university student lists using the k-NearestNeighbors machine learning method ".Journal of Physics conference series, june 2020 .

[3] BagusPriambodo, SarwatiRahayu, Al HamidyHazidar, Emil Naf'an,Mardhiahmasril, IngeHandriani, ZicoPratama Putra, Asama KudrNseaf ,Deni Setiawan, Yuwan Jumaryad,"Predicting GDP of Indonesia Using K-Nearest Neighbor Regression Predicting GDP of Indonesia Using K-Nearest Neighbor Regression" 2019, Journal of Physics: Conference Series.

[4] Abhinn Pandey, "study and analysis of k-means clustering algorithm using rapidminer , a case study" , Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 12(Part 4), December 2014.

[5] ShakuntalaJata, Vivek Sharma ,"AN ALGORITHM FOR PREDICTIVE D ATA MINING APPROACH IN MEDICAL DIAGNOSIS ",International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018.

[6] Abdo, HananFahmy, Amir Abobaker Shaker,"Data for Crime Prediction", International Journal of Computer Science and Information Security, vol.17(6), pp.56-62, 2019.

[7] Rapidminer, https://rapidminer.com

[8] Rapidminer tutorials, https://www.youtube.com/user/pallabs/playlists

[9] Blog, https://rapidminer.com/blog/
[10] https://docs.rapidminer.com/