Staff Employment Platform (StEP) Using Job Profiling Analytics

Ezzatul Akmal Kamaru Zaman(✉),
Ahmad Farhan Ahmad Kamal, Azlinah Mohamed, Azlin Ahmad,
and Raja Aisyah Zahira Raja Mohd Zamri

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
Shah Alam, Selangor, Malaysia
{ezzatul,azlinah,azlin}@tmsk.uitm.edu.my

Abstract. Staff Employment Platform (StEP) is a web-based application which employed machine learning engine to monitor Human Resource Management in hiring and talent managing. Instead of using the conventional method of hiring, StEP engine is built using decision tree classification technique to select the most significant skillsets for each job position intelligently, together with classifying the best position. The engine will then rank and predict competent candidate for the selected position with specific criteria. With the ranking method, the weightage of the profile skillset, qualification level and year of experience are summed up. Subsequently, this sum will be resulting in the competency percentage which is calculated by using a Capacity Utilization Rate formula. The proposed formula is designed and tested specifically for this problem. With the accuracy of 63.5% of Decision Tree classification, the integration of machine learning engine and ranking methods using Capacity Utilization Rate in StEP provides a feature to assist the company recruiters in optimizing candidates ranking and review the most competent candidates.

Keywords: Classification · Data analytics and visualization · Data science
Decision tree · Human resources management · SAS Viya · User profiling

1 Introduction

In recent years, job searching applications have brought much ease for job seekers. However, Human Resources (HR) officials face a challenging task in recruiting the most suitable job candidate(s). It is crucial for Human Resources Management (HRM) to hire the right employee because the success of any business depends on the quality of their employees. To achieve the company's goal, HRM needs to find job candidates that fit with the vacant position's qualifications, and it is not an easy task [1]. Besides that, the company's candidate selection strategy model is often changing for every company [2]. Competent talents are vital to business in this borderless global environment [3]. Even for a competent recruiter or interviewer, choosing the right candidate(s) is challenging [4]. In this era of Big Data and advancement in computer technology, the hiring process can be made to be easier and more efficient.

B. W. Yap et al. (Eds.): SCDS 2018, CCIS 937, pp. 387–401, 2019.
https://doi.org/10.1007/978-981-13-3441-2_30

Based on leading social career website, LinkedIn, in the first quarter of 2017, more than 26,000 jobs were offered in Malaysia, where an estimate of 1029 jobs was related to computer sciences field. In April 2017, about 125 job offers were specifically for the data science field. According to the Malaysia Digital Economy Corporation (MDEC) Data Science Competency Checklist 2017, Malaysia targets to produce about16,000 data professionals by the year 2020 including 2,000 trained data scientists.

HR would have to be precise about what criteria they need to evaluate in hiring even though they are not actually working in the field. This can be done by doing a thorough analysis, converted into an analytical trend, or chart from the previous hiring. Hence, this will significantly assist HR in the decision making of job candidates recruitment [5].

In the era of Big Data Analytics, the three main job positions most companies are seeking recently are Data Engineer, Data Analyst and Data Scientist. This study aims to identify the employment criteria the three job position in Data Science by using data analytics and user profiling. We proposed and evaluated a Staff Employment system (StEP) that analyzes the user profiles to select the most suitable candidate(s) for the three data science job position. This system can assist Human Resource Management in finding the best-qualified candidate(s) to be called for interview and to recruit them if they are suitable for the job position. The data is extracted from social career websites. User profiling is being used to determine the pattern of interest and trends where different gender, age and social class have a different interest in a particular market [6]. By incorporating online user profiling in StEP platform, HRM can cost in job advertising and time in finding and recruiting candidates. An employee recruiting system can make it easier for recruiters to match candidates' profiles with the needed skills and qualification for the respective job position [7].

To recruit future job candidates, HRM may have to evaluate user profiles from social career websites like LinkedIn and Jobstreet (in Malaysia). StEP directly use data from these websites and run user profiling to get the required information. This information is then passed to the StEP' classification engine for prediction of the suitability of the candidates based on the criteria required. Based on the user profiles, StEP particularly uses the design of the social website, such as LinkedIn, to evaluate the significance of the user's **skills**, **education or qualification** and **experience** as the three employment criteria.

2 Related Machine Learning Studies

Classification techniques are often used to ease the decision-making process. It is a supervised learning technique to enable class prediction. Classification techniques such as decision tree can identify the essential attributes of a target variable such as Job position, credit risk or churn. Data mining can be used to extract data from the Human Resource databases to transform it into more meaningful and useful information for solving the problem in talent management [8]. Kurniawan et al. used social media dataset and applied Naïve Bayes, Decision Tree and Support Vector Machine (SVM) technique to predict Twitter traffic of word in a real-time pattern. As the result, SVM has the highest classification accuracy for untargeted word, however, Decision tree scores the highest

classification accuracy for the targeted words features [9]. Classification has also been used in classifying the action of online shopping consumer. Decision tree (J48) has the second highest accuracy while the highest accuracy belongs to Decision Table classification algorithm [10]. Xiaowei used Decision Tree classification technique to obtain information about the customer on marketing on the Internet [11]. Sarda et al. also use decision tree in finding the most appropriate candidates for the job [12].

According to [13], decision tree also has the balance classification criteria result than other classification technique such as Naïve Bayes, Neural Network, Support Vector Machine. Hence, the result is more reliable and stable in term of classification accuracy and efficiency [14].

Meanwhile, ranking is a method to organize or arrange the result in the order from highest rank to lowest rank (or importance). Luukka and Collan uses the fuzzy method to do ranking in Human Resource selection where the solution is ranked by the weightage assigned for each of it. The best solution is one with the highest weightage. As the result, ranking technique can assist Human Resource Management (selection) in finding the ideal solution (most qualified candidates) for the organization [15]. There also exists a research on Decision Support System (DSS) to be applied in recruiting new employee for the company vacant position. The goal is to decide the best candidates from the calculated weight criteria and criteria itself [16].

In terms of managing interpersonal competition in making decision for a group, Multi Criteria Decision Making(MCDM) is the methods been used to solve it [17]. Recommender system is often being used to assist the user by providing choices of relevant items according to their interest to support the decision making [18].

3 Methods

3.1 Phase I: Data Acquisition by Web Scraping

In this paper, we focus on data science-related jobs, specifically Data Scientists, Data Engineers and Data Analysts. We scraped the profiles of those who are currently in those positions from LinkedIn. We have also specified the users' locations as Malaysia, Singapore, India, Thailand, Vietnam or China. To extract the data from LinkedIn, we performed web scraping using BeautifulSoup in Python. BeautifulSoup is capable of retrieving specific data and then save in any format.

Since the data is highly difficult to be extracted thru online streaming, we saved the offline copy of the data and thoroughly scrapped the details that we need from each page. This way, we can scrape data more cleanly and have control over data saving. The data was saved in CSV format. The raw data extracted were user's name, location, qualification level, skills with endorsements and working experiences measured in years. These are all features available on LinkedIn. A total of 152, 159 and 144 profiles of Data Analyst, Data Engineer and Data Scientist correspondingly have been scrapped. This raw data is saved in. CSV and a sample is shown in Fig. 1 below where each file has a various number of profiles per one run of the Python coding. It also contains some missing values such as profiles without education, experience or skills details. After data extraction, the data was kept in a structured form.

Ahmad I	Data Sci	Kuala Lu	Kuala Lu	Malaysia														
Campus	Graduat	CIM Enç	Research Assistant															
Jun 2017	Dec 201ﾃ	Oct 2016	Feb 2016 – Aug 2016															
Master of Science (MSc)																		
Statistics																		
Bachelor of Sciences with Honors (Industrial Statistics)																		
Statistics																		
3.87 / 4.00																		
Science Stream (Module 1)																		
Pure Sci	Chemis	Physics	Biology)															
4.00 / 4.00																		
Sijil Pelajaran Malaysia																		
Pure Sciences																		
7A+	2A																	
SPSS	Microsc	Statistic	Data An	Researc	Project I	Microsc	RapidMi	SAS Ent	SAS E-M	Statistic	R progra	Data Sci	Data Mi	Binary In	Statistical Pattern Recognition			
3	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1			
NEXT PERSON																		
Ahmad T	Java & p	Big Dat.	Data Sc	Kuala Lu	Malaysia													
LEAD	BI DEVE	LEAD	NBA TE	Senior D	Software Engineer													
Sep 2014	Aug 201	2008 – A	2005 – 2007															
Bachelor of Science (BS)																		
Computer Science																		
Software	Java	Extract	Transfo	Load (E	Netezza	ibm info:	Integrati	C#	Oracle	Hive	Telecom	Requirer	Apache	GWT	XML	MySQL	SQL	Linux
4	15	1	2	2	6	7	5	1	4	1	1	3	1	1	2	3	1	
NEXT PERSON																		
Ahmed Y	Data Sci	Egypt																
Data Sci	Teaching	Data Re	Data An	Business Intelligence Analyst - Internship														
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]													

Fig. 1. Sample of data collected from LinkedIn

3.2 Phase 2: Data Preparation and Pre-processing

Data Pre-processing

The raw data is merged and carefully put into three tables containing the dataset for each job position; Data Scientist, Data Engineer and Data Analyst. From all of the datasets, each of the profiles has a various number of skillset, and some of the skillsets does not apply to the data science field, making it hard to implement in the system for determining which skillset is the most important to the data science field. To solve this problem, a sample of 20 per cent of the total profiles was taken randomly using SAS Viya 'Sampling' feature to identify which skills appeared the most in each of the profiles, such as data analysis, Big Data, Java and more. This process is called features identification.

In this phase, the data itself must be consistent, where situations like if a data scientist lists down his 'Carpenter' skills in his profile, the word Carpenter' has to be removed. Therefore, pruning of features must be done towards the skills stated for each user in the dataset. This is to get a reliable set of features for classification phase later. We upload the data into SAS Viya to produce two graphs to prune the skills feature by removing any skills with ≤ 5 number of profiles. Figure 2 shows the sampled dataset while Figs. 3 and 4 present the data samples before and after the pruning process. The profiles with missing values are removed hence the data is now cleaned and ready for further processing.

#	Name	Position	Location	Experiences (Years)	Education Level	Analytics	Azure	Big data
1	Ali Seyed Shirkhorsh	Sr. Data Scientist	Kuala Lumpur, Malaysia	5.25	Ph.D.	_	_	_
2	Amin Jula	Senior Data Scientist / Train	Kuala Lumpur, Malaysia	9.67	Ph.D.	_	_	
3	Bilal Farooq	Head of Data Science	Selangor, Malaysia	14.82	Master	54	_	_
4	Charles Martin	Data Scientist & Machine Le	San Francisco, United State	16	Ph.D.	36	_	9
5	Matteo Testi	Senior Data Science | Deep L	Rome, Italy	4.67	Master	_	14	_
6	Muhammad Nazmi I	Aspiring data scientist with	Kelantan, Malaysia	1.33	Bachelor's degree	2	_	_
7	Nabilla Farhani	Data Scientist Business Frau	Kuala Lumpur, Malaysia	7.33	Bachelor's degree	2	_	_
8	Poo Kuan Hoong	Data Scientist	Kuala Lumpur, Malaysia	19.33	Ph.D.	_	_	
9	Shahram Sabzevari	Data Scientist | Full Stack De	Kuala Lumpur, Malaysia	17.92	Master	_	_	
10	Wing Yuen Loon	Data Science & Innovation	Kuala Lumpur, Malaysia	29	Master	37	_	_
11	Vincent Granville	Pioneering Data Scientist	Seattle, United States of Am	22.17	Ph.D.	99	_	9
12	Wayne Vovil	Chief Data Scientist / Hadoo	Vietnam	18	Certificate	33	_	9
13	Ronak Talreja	Data Scientist at The Data Te	Mumbai, India	4.92	Bachelor's Degree	20	_	_
14	Daniel Tyszka Junio	Data Scientist at Hariken	Curitiba, Brazil	7.67	Bachelor's Degree	3	2	
15	Rishabh Malhotra	Data Scientist	Pune, India	1.67	Bachelor's Degree	_	_	
16	Sweekar Tanugula	Senior Data Scientist at GE H	Bengaluru, India	12	Bachelor's Degree	16	_	_
17	Subhajit Gupta	Data Scientist at Emirates NB	United Arab Emirates	8.92	Master	22	_	1
18	Khadir LAMRANI	Data Scientist -Senior-	Marrakech, Morocco	6.33	Master	_	_	
19	Neil Eklund	Chief Data Scientist at Schlu	San Francisco, United State	24	Ph.D.	49	_	_
20	Sayali Sonawane	Data Scientist by profession	Maharashtra, India	3.92	Master	_	_	_
Count				20		12	2	1
Total				234.92		373	16	33
AVG				11.746		31.0833	8	33

Fig. 2. Sample of cleaned dataset

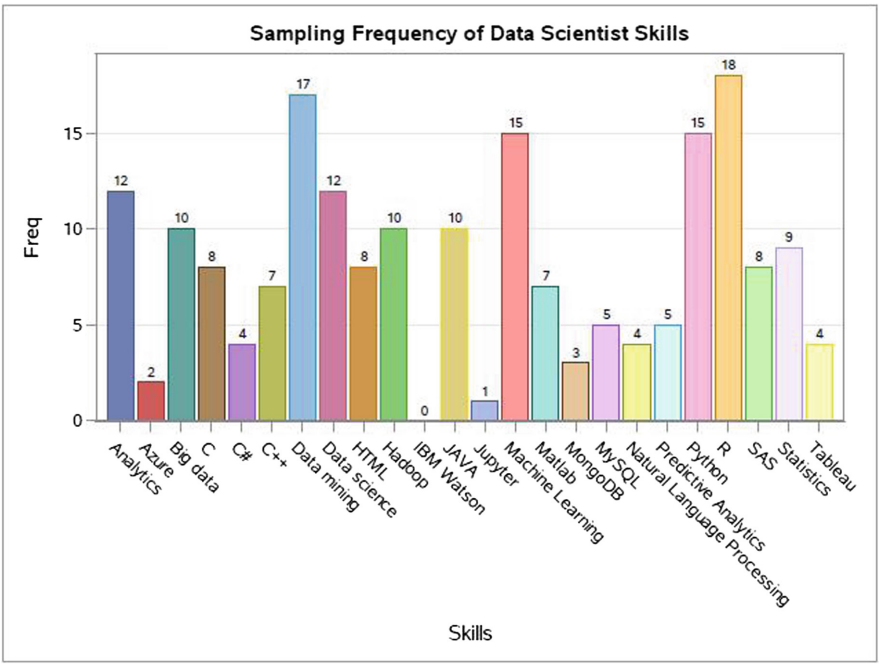


Fig. 3. Sampling frequency of data scientist skills before pruning

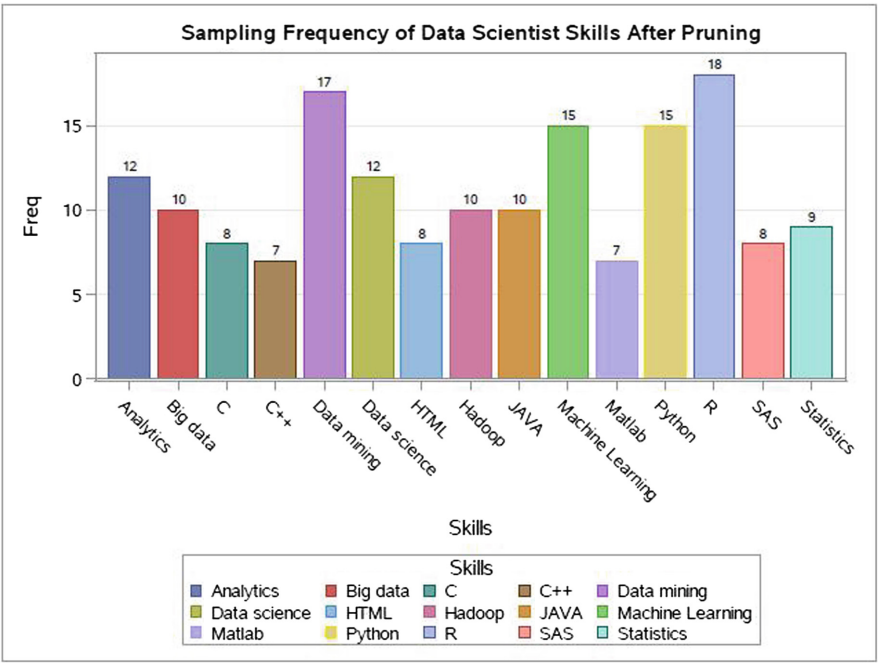


Fig. 4. Sampling frequency of data scientist skills after pruning

Mapping with MDEC Skillsets

Next, we validate the skills that we have identifies by mapping them to the MDEC's Data Science Competency Checklist (DSCC) 2017 skillset groups. This evaluation can confirm the skillsets required in the data science field. As shown in Fig. 5, according to DSCC 2017, there are nine skill sets available; 1. Business Analysis, Approach and Management; 2. Insight, Storytelling and Data Visualization; 3. Programming; 4. Data Wrangling and Database Concepts; 5. Statistics and Data modelling; 6. Machine Learning Algorithms; and 7. Big Data.

Each of the profiles' skillset has endorsement values. The feature is available in every profile of LinkedIn to represent the validation by others for the job candidate's skill as shown in Fig. 6. For this project, we use the endorsement value to determine the competency of the job candidate. This acts as the profile scoring method. This can help us to pinpoint the accurateness of the skillsets that the job candidates claimed they have.

Feature Generation

To properly determine the scores, we have taken the average endorsements from profiles that have endorsements more than 0. Figure 7 shows the average found for Data Scientist skills. To better visualize the scoreboard, the values are summed, and the skills are combined under its main skill group. Data Mining, Excel, R, SAS and statistics skills all belong under 'Statistics and Data Modelling' skill set as shown in Fig. 5. Profiles with endorsement value more than the average will be marked 1 where

Skillsets	MDEC	DS	DE	DA
Business Analysis, Approach and Management	Excel	Excel	Excel	Excel
	Others	Analysis	Analysis	Analysis
Insight, Storytelling and data visualisation	SAS	SAS		
	Others	Analytics		Analytics
Programming	C	C		
	C++	C++		C++
	Java	JAVA	JAVA	JAVA
	Matlab	Matlab		
	Python	Python	Python	Python
	R	R	R	R
	Others	HTML		
Data Wrangling and database concepts	ETL		ETL	
	MongoDB		MongoDB	
	MySQL	MySQL	MySQL	MySQL
	Others	Data Mining	Data Mining	Data Mining
Statistics and data modelling	Data Mining	Data Mining		Data Mining
	Excel	Excel		Excel
	R	R		R
	SAS	SAS		
	SPSS			SPSS
	Others	Statistics		Statistics
Machine learning algorithm	Others	Data mining		
		Data science		
		Machine Learning		
Big Data	Hadoop	Hadoop	Hadoop	
	MongoDB		MongoDB	
	Others	Big data	Big data	
		Data science		

Fig. 5. Skillsets mapped with DSCC skillsets group

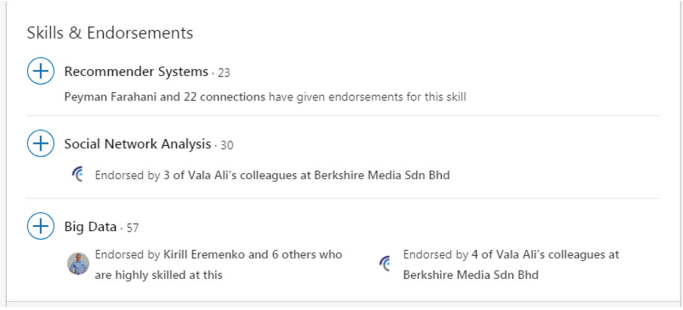


Fig. 6. Endorsement feature in LinkedIn

it means the job candidate is 'Highly skilled'. Profiles with endorsement value below than the average will be marked 0, 'Less skilled'. This binary value is called Skill Level feature that is another feature engineered and generated to enhance classification model. Hence, after data cleaning process is done, the sample sizes are 132, 98 and 99 respectively for Data Analyst, Data Engineer and Data Scientist.

3.3 Phase 3: Predictive Modelling and Ranking

In this phase, we performed two models that are predictive modelling and ranking to gain better accuracy in classifying the best position and ranking the most competent job candidates. The ranking is based on the scoring of job skills where a job candidate with the highest competency percentage is considered most eligible to hold the job position.

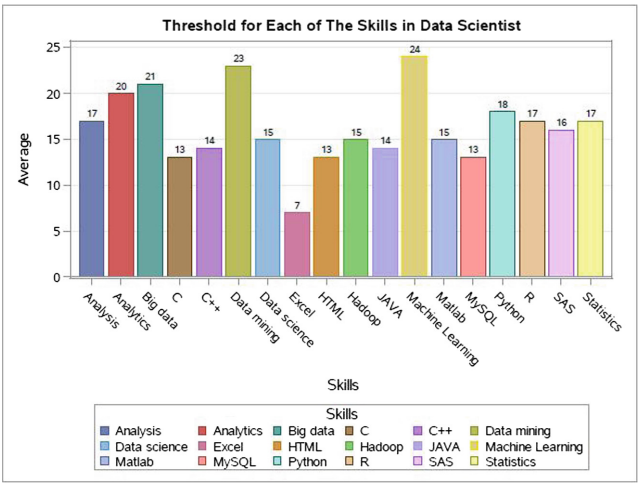


Fig. 7. Bar chart for data scientist skills

First, we calculate the weightage values using Feature Ranking to rank of the skills' score value. Subsequently, predictive modeling is employed where it is done by using classification model adapting decision tree to determine the best job position. Then ranking of job competency is done using Capacity Utilization Rate model. The calculations are explained as follows.

Feature Ranking Process

Each of the skillset group is being identified as the feature of the sample dataset where it comprises a various number of skills. These skills have the value of zero and one indicating the absence or presence of the skill respectively. Each of the skills scores will be summed as the total score for a particular skill set group. The weightage will be determined by the ranking of the skills that are low-rank value for low scores and high-rank value for a higher score. We use decision tree classification to determine the weightage. The skill with the highest importance value is given the highest weightage value. SAS Viya was used to build the decision tree and to determine the weightage for each skill.

We also ranked the qualification level and years of experience. For the qualification level, any job candidate with a professional certificate in data science will most likely be readily accepted in the industry, thus it is weighted as 3. Whereas a postgraduate is weighted 2 and a bachelor's degree graduate is given rank (or weight) of 1. Meanwhile, years of experience of more than 8 years is weighted 3, more than 3 years but less and equal to 8 years is weighted 2 and lastly, less or equal to 3 years is weighted 1.

Predictive Modeling Using Decision Tree

Classification engine is again being used to classify the three data science job position for a candidate. The target variable has three classes: Data Analyst, Data Engineer and Data Scientist. Meanwhile, the input features are the qualification, years of experience, weightage of the seven skillsets group that are 1. Business Analysis, Approach and

Management; 2. Insight, Storytelling and Data Visualization; 3. Programming; 4. Data Wrangling and Database Concepts; 5. Statistics and Data Modelling; 6. Machine Learning Algorithms; and 7. Big Data. This weightage is obtained in feature ranking process above.

Ranking Using Capacity Utilization Rate (CUR)

Then, the ranking is determined by calculating the job candidates' competency regarding their skills, combined with qualification level and years of experience. The job candidates' scores are summed up and calculated using a specific formula that can represent the job candidates' competency percentage. The competency percentage is calculated by using a Capacity Utilization Rate formula as shown in Fig. 8. This formula is adapted and designed specifically for this problem at which the formula is formerly known to be used in industrial and economic purposes.

$$\frac{7(1st\ skillset) + 6(2nd\ skillset) + 5(3rd\ skillset) + \cdots + 1(7th\ skillset)}{7(1st\ skillset\ \max value) + 6(2nd\ skillset\ \max value) + \cdots + 1(7th\ skillset\ \max value)}$$

Fig. 8. Capacity utilization rate

4 Results and Discussion

4.1 Weightage Using Feature Ranking Results

As discussed in 3., the weightage gained from summing the number of people with the skills. It is applied to the skillset groups using feature ranking. Since SAS Visual Analytics uses feature ranking as a measure of decision tree level, an importance rate table is gained, and the weightage is set as per the tree level. The highest importance rate for this data set is Machine Learning Algorithm skillset as shown in Table 1. This will make the weightage for that particular skill is ranked at highest that is 6 followed Big Data is 5, Statistics is 4, Programming is 3, Data Wrangling is 2, and subsequently, Business analysis and, Insight and Data Visualization skillset in which both resulted to 1.

Table 1. Classification weightage for all positions

Variable	Importance
Machine_learning_algorithm	31.7244
Big_Data	28.1628
Statistics_and_data_modelling	7.3953
Programming	3.3391
Data_Wrangling_and_databas_concepts	1.5428
Business_Analysis_Approach_and_Management	0.0000
Insight_Storytelling_and_data_visualization	0.0000

4.2 Classification Results and Discussion

In order to classify into class target that is Data Analyst, Data Engineer and Data Scientist, the data is being fed into SAS Visual Analytics to process the data for the classification engine. Then the tree graph of Decision Tree is produced as represented in Fig. 9. Decision Tree shows the accuracy result that is 63.5%. Figure 10 shows the confusion matrix of Decision Tree engine. Low accuracy of the result may be produced due to the imbalanced cleaned dataset where the sample dataset comprises of a higher number of Data Analyst that is 132 as compared to Data Engineer and Data Scientist with 99 and 98 samples respectively. Not only that, other parameter setting for example different percentage of training and testing set should be considered in order to increase the accuracy result as in this research, 70% training set is used to 30% is used for testing set.

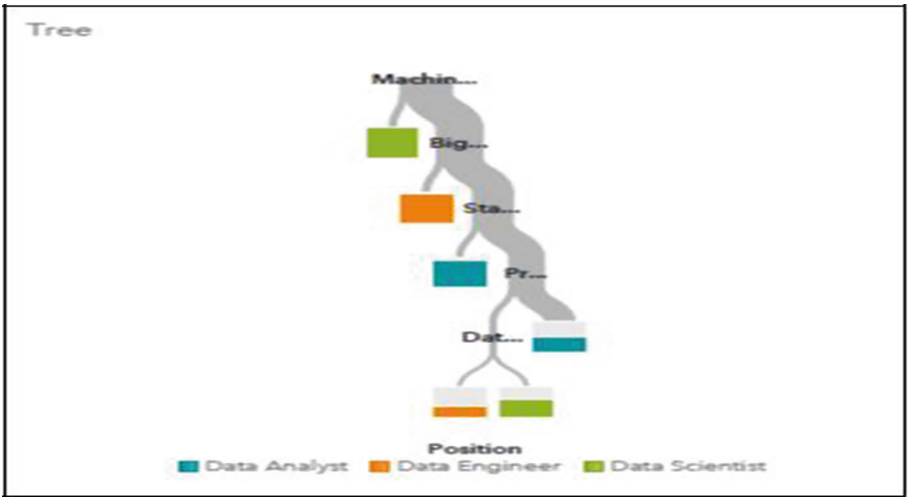


Fig. 9. Decision tree from classification model

4.3 Ranking Using CUR Results

After data position has been determined using classification model, the dataset is summed up for the job candidates' score ranking. Table 2 shows the total of the score for each of the job candidate. Table 3 states the Capacity Utilization Rate calculation and its percentage for each of the job candidates.

After calculating the Capacity Utilization Rate, the percentages are calculated and sorted to determine the best job candidate in the ranking using CUR model. Table 4 represents the ranking of most recommended Data Scientists in Asia. The percentage represents the job candidates' competency for Data Scientist.

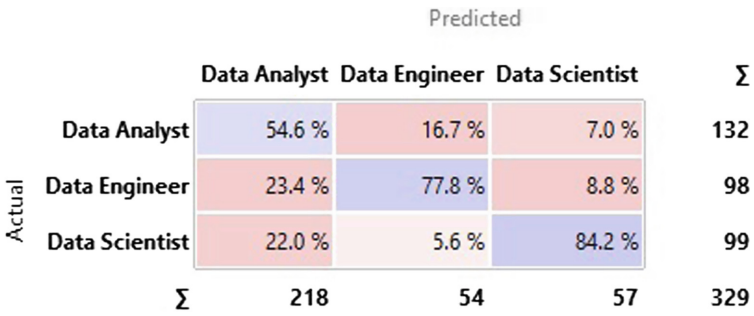


Fig. 10. Confusion matrix of decision tree

Table 2. Sample of data scientist skill score ranking results

#	Name	Exp Weigl	QL Weigh	Programm	Machine l	Statistics a	Business A	Data Wrar	Big Data	Insight, St	TOTAL
1	Ahmad Hakii	1	2	0	0	0	0	0	0	0	3
2	Ahmad Tarme	3	1	1	0	0	0	0	0	0	11
3	Ali Seyed Shi	2	2	2	0	0	0	1	0	0	21
4	Alireza Hoom	1	2	1	1	3	2	0	1	0	41
5	Amin Jula	3	2	5	2	1	1	1	0	0	64
6	Amir Rafieiar	3	2	1	0	0	0	0	0	0	12
7	Asif Muhamn	3	1	0	1	2	1	2	0	2	32
8	Asiya Bhat	2	1	0	0	0	0	0	0	0	3
9	Aswadi A Rah	3	2	0	0	0	0	0	0	0	5
10	Azizi Aziz	3	2	0	0	0	0	0	0	0	5
11	Bharat Bhush	3	2	0	0	0	0	0	0	0	5
12	Bilal Farooq	3	2	1	1	4	1	2	0	2	50
13	Brian Ho	3	2	1	0	0	1	0	0	0	16
14	Buddhika San	3	2	0	0	0	0	0	0	0	5

Table 3. Total score ranking and capacity utilization rate result

#	Name	TOTAL SCORE	TOTAL MAX	C.U.R	%
1	Ahmad Hakiir	3	115	0.026087	2.61
2	Ahmad Tarme	11	115	0.095652	9.57
3	Ali Seyed Shi	21	115	0.182609	18.26
4	Alireza Hoom	41	115	0.356522	35.65
5	Amin Jula	64	115	0.556522	55.65
6	Amir Rafieiar	12	115	0.104348	10.43
7	Asif Muhamn	32	115	0.278261	27.83
8	Asiya Bhat	3	115	0.026087	2.61
9	Aswadi A Rah	5	115	0.043478	4.35
10	Azizi Aziz	5	115	0.043478	4.35
11	Bharat Bhush	5	115	0.043478	4.35
12	Bilal Farooq	50	115	0.434783	43.48
13	Brian Ho	16	115	0.13913	13.91
14	Buddhika San	5	115	0.043478	4.35
15	Caleb Foong	10	115	0.086957	8.70
16	Canh Tran	25	115	0.217391	21.74

Table 4. Ranking of job candidates

#	Name	Country	Experiences (Years)	Qualification Level	TOTAL SCORE	%
73	Sayali Son	India	3.92	Post Graduate	93	80.87
86	Wayne Vo	Vietnam	18	Certificate	67	58.26
29	Eugene Ya	Singapore	18.5	Post Graduate	66	57.39
59	Omar Sale	India	5.83	Bachelors Degree	65	56.52
5	Amin Jula	Malaysia	9.67	Post Graduate	64	55.65
82	Vala Ali R	Malaysia	15.67	Post Graduate	63	54.78
96	Xavier Co	Singapore	18	Post Graduate	62	53.91
57	Ng Kean C	Malaysia	4	Bachelors Degree	55	47.83
69	Ronak Tal	India	4.92	Bachelors Degree	55	47.83
56	Ng Kean C	Malaysia	3.92	Bachelors Degree	53	46.09
81	Sweekar T	India	12	Bachelors Degree	52	45.22
12	Bilal Faroo	Malaysia	14.82	Post Graduate	50	43.48
74	Shahram S	Malaysia	17.92	Post Graduate	50	43.48
17	Chee Bing	Malaysia	4.25	Bachelors Degree	48	41.74
35	Jenna Yan	Malaysia	8.33	Bachelors Degree	48	41.74
92	Weimin W	Singapore	4	Post Graduate	48	41.74
83	Vincent Fi	Singapore	7.83	Post Graduate	47	40.87

4.4 Data Visualization

Viewing the outcome in tables are very hard to understand, especially to the untrained eye. This is where data visualization comes in handy. In this project, SAS Visual Analytics is used again to produce visualization to see a better result of performing machine learning and analytics upon the data.

Figure 11 represents the number of Data Scientists based on their skillsets. This graph uses the same data from Table 1 to produce the weightage for the analytics calculation. It is found that most of the Data Scientists have programming skills with 51 of them having the skills.

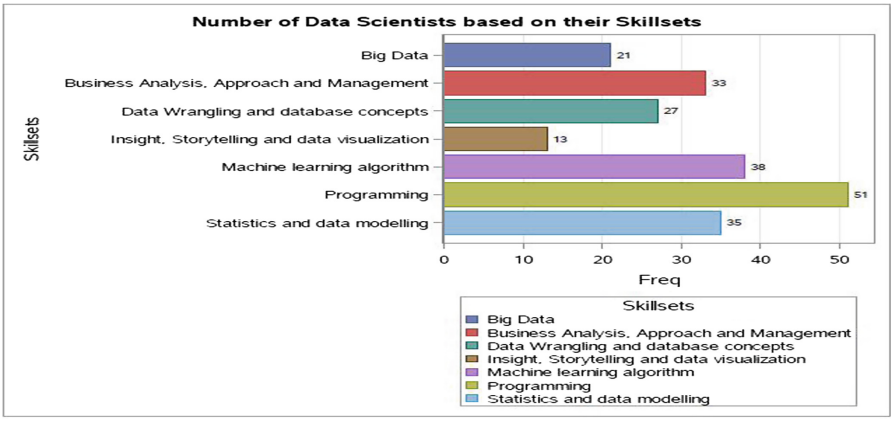
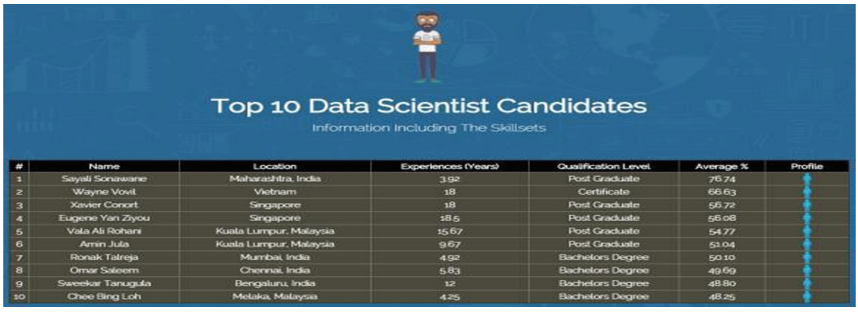


Fig. 11. Number of data scientists based on their skillsets

4.5 Staff Employment Platform (StEP)

The Staff Employment Platform (StEP) is a web-based job candidate searching platform. For this project, this webpage is used to display the top 10 most suitable job candidates in positions, Data Scientist, Data Engineer and Data Analyst, for the company. From the list, company recruiters can search for the job position that they require and view the job candidates recommended for the position.

StEP provides a feature where the company recruiters can view the job candidates' profile to see their details. Figure 12 shows the list of top 10 most recommended job candidates for Data Scientist which is the result of our complex analytics. Furthermore, through these listings, recruiters are also able to view the job candidates' competency according to their job position and display the skills that the candidate has.



#	Name	Location	Experiences (Years)	Qualification Level	Average %	Profile
1	Sayali Sonawane	Maharashtra, India	3.92	Post Graduate	76.74	
2	Wayne Vovil	Vietnam	18	Certificate	66.63	
3	Xavier Conort	Singapore	18	Post Graduate	56.72	
4	Eugene Yan Ziyou	Singapore	18.5	Post Graduate	56.08	
5	Vala Ali Rohani	Kuala Lumpur, Malaysia	15.67	Post Graduate	54.77	
6	Amin Jula	Kuala Lumpur, Malaysia	9.67	Post Graduate	53.04	
7	Ronak Talreja	Mumbai, India	4.92	Bachelors Degree	50.10	
8	Omar Saleem	Chennai, India	5.83	Bachelors Degree	49.69	
9	Sweekar Tanugula	Bengaluru, India	12	Bachelors Degree	48.80	
10	Chee Bing Loh	Melaka, Malaysia	4.25	Bachelors Degree	48.25	

Fig. 12. Top 10 data scientists as job candidates

5 Conclusion

Finding the most suitable employee for a company is a daunting task for the Human Resource Department. Human Resources (HR) have to comb through a lot of information on social-career websites to find the best job candidate to recruit. Staff Employment Platform (StEP) uses SAS Viya in Machine Learning and Visual Analytics to perform job profiling and ranks the competent candidate for data science job positions. Job profiling is better when it is combined with analytics and machine learning, in this case, Classification by using Decision Tree. SAS Viya performs very well in visualizing data and produce clear and understandable charts and graphs as depicted in the sections above. The result is enhanced by using Capacity Utilization Rate formula which is adapted specifically for this problem to do the ranking of competence candidates. As a conclusion, we were able to propose a platform to find the three important criteria needed in the data science field, which are skills, qualification level and years of experience. From job profiles of Data Scientists, Data Engineers and Data Analysts, we were able to perform job profiling and gain thorough analysis of their skills and other important criteria.

Acknowledgement. The authors would like to thank Ministry of Education Malaysia for funding this research project through a Research University Grant; Bestari Perdana 2018 Grant,

project titled "Modified Clustering Algorithm for Analysing and Visualizing the Structured and Unstructured Data" (600-RMI/PERDANA 513 BESTARI(059/2018)). Also appreciation goes to the Research Management Center (RMC) of UiTM for providing an excellent research environment in completing this research work. Thanks to Prof Yap Bee Wah for her time in reviewing and validating the result of this paper.

References

1. Mohammed, M.A., Anad, M.M.: Data warehouse for human resource by Ministry of Higher Education and Scientific Research. In: 2014 International Conference on Computer, Communications, and Control Technology (I4CT), pp. 176–181 (2014)
2. Shehu, M.A., Saeed, F.: An adaptive personnel selection model for recruitment using domain-driven data mining. J. Theor. Appl. Inf. Technol. **91**(1), 117 (2016)
3. Tajuddin, D., Ali, R., Kamaruddin, B.H.: Using talent strategy as a hedging strategy to manage banking talent risks in Malaysia. Int. Bus. Manag. **9**(4), 372–376 (2015)
4. Saat, N.M., Singh, D.: Assessing suitability of candidates for selection using candidates' profiling report. In: Proceedings of the 2011 International Conference on Electrical Engineering and Informatics (2011)
5. Charlwood, A., Stuart, M., Kirkpatrick, I., Lawrence, M.T.: Why HR is set to fail the big data challenge. LSE Bus. Rev. (2016)
6. Farseev, A., Nie, L., Akbari, M., Chua, T.S.: Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 235–242. ACM, Shanghai (2015)
7. Ahmed, F., Anannya, M., Rahman, T., Khan, R.T.: Automated CV processing along with psychometric analysis in job recruiting process. In: 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (2015)
8. Yasodha, S., Prakash, P.S.: Data mining classification technique for talent management using SVM. In: International Conference on Computing, Electronics and Electrical Technologies (ICCEET), pp. 959–963. IEEE (2012)
9. Kurniawan, D.A., Wibirama, S., Setiawan, N.A.: Real-time traffic classification with twitter data mining. In: 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1–5. IEEE (2016)
10. Ahmeda, R.A.E.D., Shehaba, M.E., Morsya, S., Mekawiea, N.: Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining. Paper Presented at the 2015 Fifth International Conference on Communication Systems and Network Technologies (2015)
11. Xiaowei, L.: Application of decision tree classification method based on information entropy to web marketing. Paper Presented at the 2014 Sixth International Conference on Measuring Technology and Mechatronics Automation (2014)
12. Sarda, V., Sakaria, P., Nair, S.: Relevance ranking algorithm for job portals. Int. J. Curr. Eng. Technol. **4**(5), 3157–3160 (2014)
13. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. **160**, 3–24 (2007)
14. Mohamed, W.N.H.W., Salleh, M.N.M., Omar, A.H.: A comparative study of Reduced Error Pruning method in decision tree algorithms. Paper Presented at the 2012 IEEE International Conference on Control System, Computing and Engineering (2012)

15. Luukka, P., Collan, M.: Fuzzy scorecards, FHOWA, and a new fuzzy similarity based ranking method for selection of human resources. Paper Presented at the 2013 IEEE International Conference on Systems, Man, and Cybernetics (2013)
16. Khairina, D.M., Asrian, M.R., Hatta, H.R.: Decision support system for new employee recruitment using weighted product method. Paper Presented at the 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE) (2016)
17. Rosanty, E.S., Dahlan, H.M., Hussin, A.R.C.: Multi-criteria decision making for group decision support system. Paper Presented at the 2012 International Conference on Information Retrieval & Knowledge Management (2012)
18. Najafabadi, M.K., Mohamed, A.H., Mahrin, M.N.R.: A survey on data mining techniques in recommender systems. Soft Comput. (2017)