

# Implementation of an Intelligent Online Job Portal Using Machine Learning Algorithms

Zarrin Tasnim<sup>1\*</sup>, F. M. Javed Mehedi Shamrat<sup>1</sup>, Shaikh Muhammad Allayear<sup>2</sup>, Khobayeb Ahmed<sup>3</sup> and Naimul Islam Nobel<sup>1</sup>

**Abstract** Business intelligence and analytics are data management solutions implemented in companies and enterprises to collect historical and present data, while using statistics and software to analyze raw information, and deliver insights for making better future decisions. In the circumstances of today's world, to survive and established own business need an analytical and find an easiest way or intelligence business model. The main objective is to examine the performance of various Machine Learning algorithms in order to perform with the system of an online job portal. This proposed module integrated with three phase such as, the Clusters similar kind of job search phase (CSK) is a way of knowing the demand is to create a visual graph showing clusters of similar kinds of job searched by the job seekers in the website of the job portal, the email notifications send phase (ENS) is responsible to send email notifications to the job seekers when a job circular is posted in the website, extract the job circular phase (EJC) is the way to extract the job circular post from the career section of each of the company's website. The result shows the successful clustering of similar job search, email notification send to specific people and extracts the information from the web.

## 1 Introduction

The intelligent enterprise utilizes connectivity (between things, people, and enterprises), data, cloud applications, algorithms, and advanced analytics (instead of hard-coded rules and rigorous procedures). This enables them to come to the right decisions with minimal human involvement even in turbulent, fast-changing environments. For a machine to do any work as well or even better than human, it need a combination of technologies to collect and process the data (e.g., IoT and cloud applications), artificial intelligence and Machine Learning [1]. We are working on topics that

---

Zarrin Tasnim  
e-mail: zarrint25@gmail.com

F. M. Javed Mehedi Shamrat  
e-mail: javedmehedicom@gmail.com

Shaik Muhammad Allayear  
e-mail: drallayear.swe@diu.edu.bd

Khobayeb Ahmed  
e-mail: khobayeb15-5200@diu.edu.bd

Naimul Islam Nobel  
e-mail: nobel775@diu.edu.bd

<sup>1</sup>Dept. of Software Engineering, Daffodil International University, Dhaka, Bangladesh

<sup>2</sup>Dept. of Multimedia and Creative Technology, Daffodil International University, Dhaka, Bangladesh

<sup>3</sup>Dept. of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

is a business intellectual model that will optimize the process of job seek seeking, skill development for job seekers and for human resource department (HRD) to hire candidates. Within three phase we accomplished the whole process.

It is an essential method for collecting data on, and keeping in touch with the rapidly increasing Internet. This Paper briefly reviews the concepts of web crawler, its architecture and its various types [2].

The paper discusses the Bayesian decision tree algorithm, its structure and computation. In the paper, the authors gave a detailed calculation of the efficiency of the algorithm with the partition space. It describes the structure and formation of the greedy-modal tree (GMT) is given with numerical example [3].

The main focus of the paper is to make clusters of the verses of the Holy Qur'an. In the paper, the authors talked about mining the text from the Holy Qur'an and applying the K-means algorithm to determine the number of steamed and unsteamed words in each cluster. The final visualization shows the different densities in each cluster [4].

The objective of the paper is to give a fast and efficient seeding method for text document clustering using k-means algorithm. The authors suggested vectorizing the text document. After that the initial seed point are select as far away from one another as possible to get the best result. Furthermore they compared the system with Points, K-means++, and KMC2 seeding methods [5].

From the research gap a system is proposed that contains three modules. The web crawler is used to collect data from the web automatically and efficiently. The decision tree algorithm is used to select candidates for sending email notification and finally the K-means clustering algorithm is used to study he job market demand.

## 2 Methodology

### 2.1 Proposed System

For the proposed system a number of machine learning algorithms [6] are implemented. This algorithms calculate the arrangement of the data to exhibit efficient performance. The proposed system incorporates three critical phase that is, automatically extracting and gathering data from the web is optimize and enrich the database automatically for the effectiveness of the machine learning algorithms [7] to be applied. The second phase is sending email notifications to the job seekers based on the data extracted from the first phase comparing it to the data of the job seekers available in the database. Finally the phase where clustering similar kinds of jobs is done using a machine learning algorithm in order to study and illustrate the proportions of demand of job in the job market at present. This proposed system enhance the roundabout time of an organization's site. The outcome demonstrates the fruitful bunching of comparative pursuit of employment, email notice sent to explicit individuals and concentrates the data from the web.

## Implementation of an Intelligent Online Job Portal Using Machine Learning Algorithms

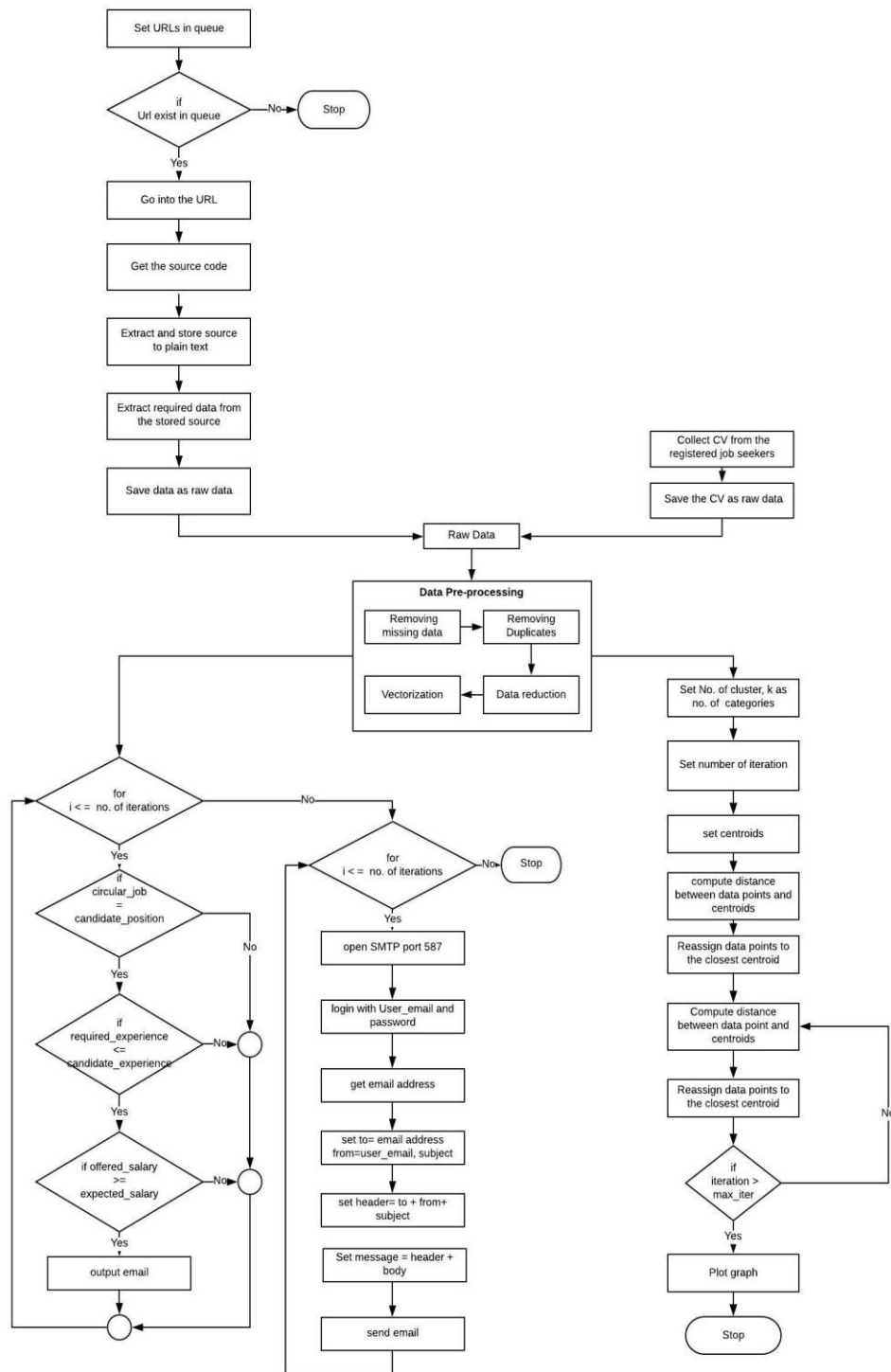


Fig. 1. System Diagram

## 2.2 System Overview

Our proposed system contains three (3) phases. Those are given below:

1. Extract job circular phase (EJC)
2. Cluster similar kind of job search phase (CSK)
3. Email notifications sending phase (ENS)

## 3 Implementation

### 3.1 Extraction of Job Circular Phase (EJC)

In this phase, to extract job circular information from the web, a web crawler is implemented [8]. The URLs of websites are set in a queue that contain job circulars from where the crawlers extract the data and saves in the parameters in the database. The crawler is implemented in order to automatically updating the database and to make to system efficient.

We assume that T is the set of relevant web pages in test dataset; U is the set of relevant web pages assigned by classifier. Therefore, we define Precision and Recall as follows [9]:

$$Precision = \frac{|U \cap T|}{|U|} \times 100\% \quad (1)$$

$$Recall = \frac{|U \cap T|}{|T|} \times 100\% \quad (2)$$

To implement the proposed phase, algorithm 1 is followed. Setting i number of pages in the web crawler, the process is run until all the expected pages are crawled into and all the required data is gathered.

#### Algorithm 1: Web crawler

**Input:** Number of pages = i, URL link

**Output:** Source, URL link, Job Circular Data

```

Step 1:  Set i;
Step 2:  for iterate till i
Step 3:      Set the URL link to be crawled;
Step 4:      Extract the source of the set URL;
Step 5:      Download and store the source into text format;
Step 6:      Extract data of URL links from the text source;
Step 7:  end for
Step 8:  for all the extracted URLs are crawled;
Step 9:      Crawl in to the extracted links;
Step 10:     Download and store the source of the URL into text format;
Step 11:     Extract required data of job, required experience and offering salary;
Step 12: end for
Step 13: Save the retrieved data into a CSV file;

```

### 3.2 Data Preprocessing

Once the raw data is extracted by the web crawler and stored, the data needs to be preprocessed for further use. Both the extracted data of job circular and he job seekers' data has to be preprocessed before applying them in the decision tree and K-means clustering machine learning algorithm [10]. The preprocessing of the raw data is done following the algorithm 2.

**Algorithm 2: Data preprocessing**

**Input:** Job titles = JT, salary = Sal, experience = Ex and email = Em of job seekers' data;

**Output:** Job Vector = JV, X vector of job title (JT) = XJV and Y vector of job title (JT) = YJV

**Step 1:** Drop null values;  
**Step 2:** Remove duplicate values;  
**Step 3:** Converts string to lower case;  
**Step 4:** for all rows in data frame  
**Step 5:** if JT exists in dataframe then  
**Step 6:** set current JV = previous occurrence;  
**Step 7:** set XJV = previous occurrence of XJV + ( 0.01 X Random number );  
**Step 8:** set YJV = previous occurrence of YJV + ( 0.01 X Random number );  
**Step 9:** end if  
**Step 10:** else  
**Step 11:** set current JV = random generated number;  
**Step 12:** set XJV = Random number + ( 0.01 X Random number );  
**Step 13:** set YJV = Random number + ( 0.01 X Random number );  
**Step 14:** end else  
**Step 15:** end for

**3.3 Email Notification Sending Phase (ENS)**

In this part, a decision tree algorithm is implemented that algorithm gives a list of suitable candidates for a certain job circular based on three attributes i.e. job position, experience and expected salary from the job circular and job seekers' data. The data must be preprocessed such that the categorical variables are based on same degrees. The job position data is considered as the root node of the tree. The tree will provide the emails of the job seekers' once the set conditions are fulfilled. This email addresses will receive to notification in order to be relevant.

Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content.[10]. Entropy is defined as below [11]:

$$E(T) = -\sum_{i=1}^J p_i \log_2 p_i \quad (3)$$

In algorithm 3, the steps of implementation of the decision tree is described where the job circular data is compared with the data of all the job seekers' in order to find a match. The comparison is done based on the conditions set for each case.

**Algorithm 3: Decision tree**

**Input:** Job position =JP, Job position vector= JPV, expected experience = EEx,

Offering salary = OSal, Job title = JT, Job Vector =JV, experience = Ex, salary= Sal

**Output:** Email

**Step 1:** for all JT in dataset  
**Step 2:** if JP = JT then  
**Step 3:** set JPV as JV;  
**Step 4:** end if  
**Step 5:** else  
**Step 6:** set JPV as random generated number  
**Step 7:** end else  
**Step 8:** end for

```

Step 9:   for all row in dataset
Step 10:  |   if JPV = JV then
Step 11:  |   |   if EEx <= Ex then
Step 12:  |   |   |   if OSal >= Sal then
Step 13:  |   |   |   |   Increment flag;
Step 14:  |   |   |   |   output Email ;
Step 15:  |   |   |   end if
Step 16:  |   |   end if
Step 17:  |   end if
Step 18:  end for
Step 19:  if flag = 0 then
Step 20:  |   output "No candidate";
Step 21:  end if

```

In algorithm 4, the steps taken to send the email notifications to the candidates are given. The email is sent to the candidates those email is received as the output of the decision tree implementation.

#### Algorithm 4: Implementing the decision tree algorithm

**Input:** Job position =JP, Job position vector= JPV, expected experience = EEx,  
Offering salary = OSal, Job title = JT, Job Vector =JV, experience = Ex, salary= Sal

**Output:** Email

```

Step 1:   for all JT in dataset
Step 2:  |   if JP = JT then
Step 3:  |   |   set JPV as JV;
Step 4:  |   end if
Step 5:  |   else
Step 6:  |   |   set JPV as random generated number
Step 7:  |   end else
Step 8:  end for

```

### 3.4 Clustering similar kind of job search phase (CSK)

In the CSK phase, an unsupervised K-means algorithm is implemented that gives the cluster of similar kinds of jobs in the dataset. The algorithm retrieves the data of the job titles from the dataset and categorize them in similar types. This data is then vectorized based on their categories. The vectorized values are then plotted into a scattered graph. The K-means identifies clusters from the scattered graph. Each cluster represent a certain type of job. The density of each cluster can be analyzed to identify the demand of job in the job market.

To calculate the new cluster center the following equation is used [12]:

$$v_i = \frac{1}{c_i} \sum_{j=1}^{c_i} x_i \quad (4)$$

To calculate the distance between the data points,  $p$  and  $q$ , Euclidean distance equation is used. The equation is as follows [13]:

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (5)$$

The implementation of algorithm 5 gives the K-means clustering algorithm. The algorithm gives the cluster of data points as output in a scattered graph which uses the preprocessed data of current job position of job seekers'.

**Algorithm 5: K-means Clustering**

**Input:** X Job Vector = XJV, Y Job Vector = YJV, number of clusters = K, number of centroids = C

- Step 1:** Set K;  
**Step 2:** Set C = K;  
**Step 3:** Plot data points, XJV and YJV;  
**Step 4:** for iterating until no change occurs to the centroids  
**Step 5:**     Compute distance between XJV, YJV and centroids with Euclidean distance [13],  

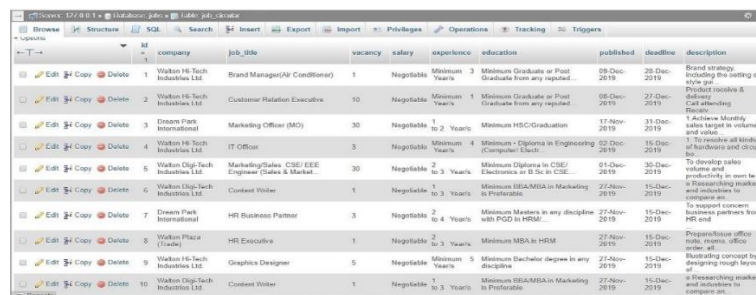
$$dist((x,y), (a,b)) = \sqrt{(x-a)^2 + (y-b)^2}$$
  
**Step 6:**     Assign each data point to the closest cluster;  
**Step 7:**     Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster;  
**Step 8:**     end for

## 4 Result

### 4.1 Extracted data using web crawler

A HTTP link of a website is set in the web crawler using which the web crawler goes in the website and extracts the required data from the website's source, copy it and stores it for further use. Such data is extracted using the implemented web crawler in order to gather data for the system.

In figure 2, the data that is extracted from the website is shown. The web crawler copied the information from the web page [15] and stored it in the database as raw data to enrich the database. This data can be processed for further use in the following phases.



	company	job_title	vacancy	salary	experience	education	published	deadline	description
1	Watson H&Tech Industries Ltd.	Brand Manager(Air Conditioner)	1	Negotiable	Minimum 3 Years	Minimum Graduate or Post Graduate from any reputed...	09-Dec-2019	28-Dec-2019	Brand strategy including the setting of the go...
2	Watson H&Tech Industries Ltd.	Customer Relation Executive	10	Negotiable	Minimum 1 Years	Minimum Graduate or Post Graduate from any reputed...	08-Dec-2019	27-Dec-2019	Product training & Customer support...
3	Dreams Park International	Marketing Officer (MO)	30	Negotiable	1 to 2 Years	Minimum HSC/Graduation	17-Nov-2019	31-Dec-2019	1 Achieve Monthly sales target in volume and value...
4	Watson H&Tech Industries Ltd.	IT Officer	3	Negotiable	Minimum 2 Years	Minimum - Diploma in Engineering Computer Graphi...	02-Dec-2019	15-Dec-2019	7 To resolve all kinds of hardware and software iss...
5	Watson H&Tech Industries Ltd.	Marketing/Sales CSE/EEE Engineer (Sales & Market...	30	Negotiable	2 to 3 Years	Minimum Diploma in CSE/ Electronics or B.Sc in CSE...	01-Dec-2019	30-Dec-2019	7 To develop sales volume and productivity in own te...
6	Watson H&Tech Industries Ltd.	Content Writer	1	Negotiable	1 to 3 Years	Minimum BE/BAMS in Marketing or Professional	27-Nov-2019	15-Dec-2019	9 Researching markets and industries to complete an...
7	Dreams Park International	HR Business Partner	3	Negotiable	2 to 4 Years	Minimum Masters in any discipline with PGD in HRM...	27-Nov-2019	15-Dec-2019	To support current business partners from HR end...
8	Watson Plaza (Dubai)	HR Executive	1	Negotiable	2 to 3 Years	Minimum MBA in HRM	27-Nov-2019	15-Dec-2019	Proactive office work, reports, office admin, all...
9	Watson H&Tech Industries Ltd.	Graphics Designer	5	Negotiable	Minimum 5 Years	Minimum Bachelor degree in any discipline	27-Nov-2019	15-Dec-2019	Revolutionizing concept by designing rough layout of...
10	Watson H&Tech Industries Ltd.	Content Writer	1	Negotiable	1 to 3 Years	Minimum BE/BAMS in Marketing or Professional	27-Nov-2019	15-Dec-2019	9 Researching markets and industries to complete an...

Fig. 2. Extracted data by web crawler

### 4.2 Email Notifications

The phase of sending email notification consists of two parts. First the decision tree algorithm is implemented. The algorithm gives as a decision of the suitability job seekers to a job circular. As output, the algorithm provides the email addresses of the job seekers'. Next an email notification is send out to the address obtained from the result of the decision tree algorithm using an smtp server.

In figure 3, if suitable candidates are found for any job circular, a list of email addresses of those candidates will be provided. This email addresses will only get the notification of that particular circular.

In figure 4, a test email is send to a recipient. This process can be used to send email notification of a job circular to a job seeker who is suitable for the job.

Email addresses :  
 nqqj820u@gmail.com  
 l0wfs71@gmail.com  
 7cxvtrs1@gmail.com  
 6k1m65e1@gmail.com  
 sqhsrd9s@ymail.com  
 tv2edn3m@ymail.com  
 47pljawl@gmail.com  
 28czccsc@yahoo.com  
 nymh2fk6@hotmail.com  
 k030xxx4@gmail.com  
 gu55hbvp@gmail.com  
 pe2781ug@gmail.com  
 hyhmp4jn@gmail.com  
 996r5rec@ymail.com  
 ciiv3f3t@yahoo.com  
 8rqrlwy4@ymail.com  
 jqknuk2n@yahoo.com  
 sqibagfv@ymail.com  
 di86m5ms@gmail.com  
 i9nwafhd@yahoo.com

Fig. 3. Test Email Notification



Fig. 4. Test Email Notification

#### 4.3 Clusters of Similar Kinds of Jobs

Analyzing the clusters, the market demand can be realized as the high density of cluster has higher demand compared to the lower density clusters [14]. The final scatter graph (figure 5) is shown that indicates 28 different clusters in along with a color code legend. Analyzing this clusters and comparing the color legend, we can identify the demand of the jobs.

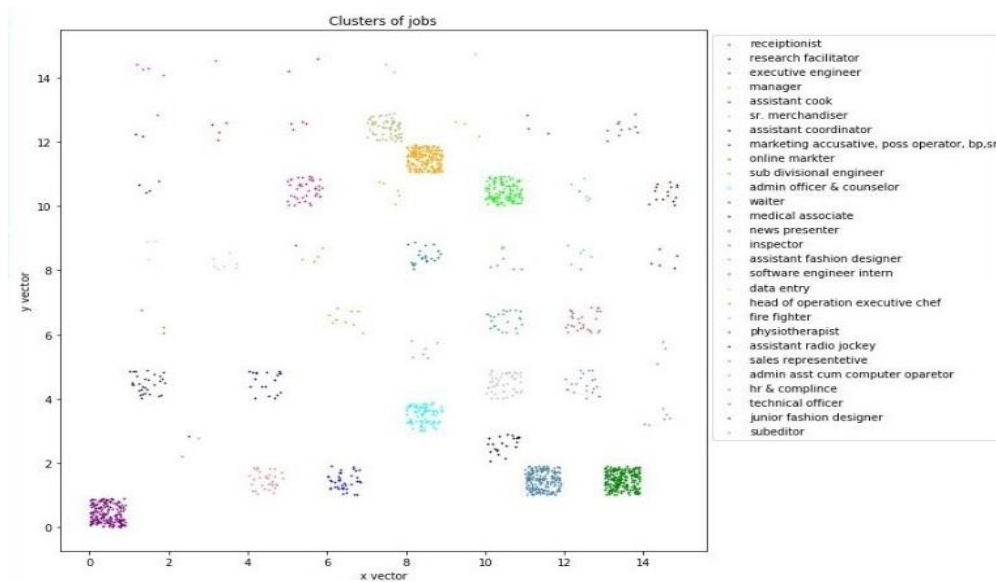


Fig. 5. K-means Cluster of Types of Jobs

#### 4.4 Efficiency Measurement

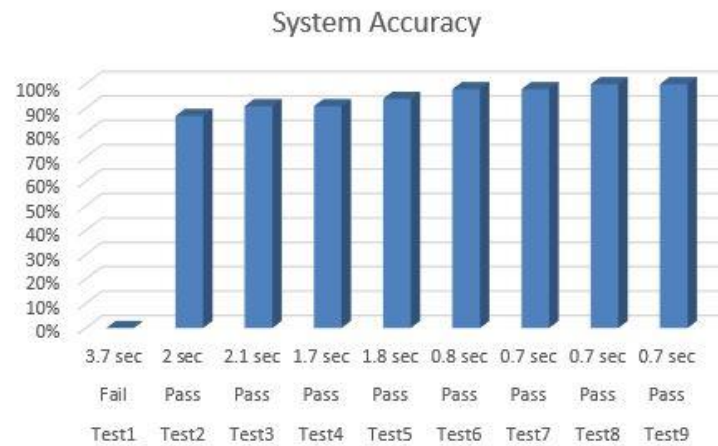
To test the implemented system, a number of test cases were used. In table 1, ten test cases were tested and the result are demonstrated. The test cases contain URLs for web crawler to extract data



and job seekers data for decision tree and clustering algorithms. For Test1, it is seen that the crawler fails to gather any data. From test2 to onwards, the crawler successfully extract data and completed all the processes with different accuracy rate with up to 100% accuracy. A graphical interface is illustrated in figure 6.

**Table 1.** Accuracy measurement for Web crawler

Input (Test Case)	Output	Response Time	System Accuracy
Test1	Fail	3.7 sec	0%
Test2	Pass	2 sec	87%
Test3	Pass	2.1 sec	91%
Test4	Pass	1.7 sec	91%
Test5	Pass	1.8 sec	94%
Test6	Pass	0.8 sec	98%
Test7	Pass	0.7 sec	98%
Test8	Pass	0.7 sec	100%
Test9	Pass	0.7 sec	100%
Test10	Pass	0.7 sec	100%



**Fig.6.** Accuracy graph for Web crawler

## 5 Conclusion

The system is proposed for online job portals. The system contains three different phases. Each phase is responsible to improve the efficiency of the system. First of all, EJC phase is responsible for extracting the job circulars from different company's website automatically using a web crawler. The web crawler extracts the job circular from the company website's career section and store it in the database of skill.jobs. Next is the CSK phase that is tasked to give a graph containing the cluster of same kind of jobs from the data using the K-means algorithm. Analyzing the density of the cluster, skill enhancement training programs can be organized to help jobseekers get better jobs. Finally in ENS phase a decision tree is implemented to make decision about to whom email notification of a certain job circular should be send. This phase uses the data stored in the database, provided by the jobseekers in order to match the job seekers' job position, salary and experience to find a suitable job from the job. When job circular is matched with a suitable job seeker, the job

seeker receives an email allowing him to know about the job circular. This phases together makes an efficient system that automatically gathers job circulars and allow job seekers to know if any job matches their requirement which helps job seekers to apply for jobs more efficiently. At the same time, analyzing the demand job seekers, training programs can be arranged that helps enhance skills to get better jobs.

## References

- [1] Shamrat, F. M. J. M., Raihan, M. A., Rahman, A. K. M. S., Mahmud, I., Akter, R., : An Analysis on Breast Disease Prediction Using Machine Learning Approaches. In. International Journal of Scientific & Technology Research, Volume 9, Issue 02, ISSN: 2277-8616, pp: 2450-2455. (2020)
- [2] Udupure, T. V., Kale, R. D., Dharmik, R. C., : Study of Web Crawler and its Different Types. In. IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI, (2014)
- [3] Nuti, G., Rugama, L. A. J., Cross, : A Bayesian Decision Tree Algorithm A Bayesian Decision Tree Algorithm, in ArXiv (2019)
- [4] Slamet, C., Rahman, A., Ramdhani, M. A., Darmalaksana, W., : Clustering the Verses of the Holy Qur'an using K-Means Algorithm. In. Asian Journal of Information Technology vol.15(24), p: 5159-5162, (2016)
- [5] Sherkat, E., Velcin J., Milios, E. E., : Fast and simple deterministic seeding of Kmeans for text document clustering. In. 9th International conference of the CLEF Association, CLEF, vol.11018 (2018)
- [6] Shamrat, F. M. J. M., Asaduzzaman, M., Rahman, A. K. M. S., Tusher, R. T. H., Tasnim, Z., : A Comparative Analysis Of Parkinson Disease Prediction Using Machine Learning Approaches. In. International Journal of Scientific & Technology Research, Volume 8, Issue 11, ISSN: 2277-8616, pp: 2576-2580, (2019)
- [7] Rahman, A. K. M. S., Shamrat, F. M. J. M., Tasnim, Z., Roy, J., Hossain, S. A., : A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms, In. International Journal of Scientific & Technology Research, Vol. 8, Issue 11, ISSN: 2277-8616, pp: 419-422. (2019 )
- [8] Shamrat, F. M. J. M., Tasnim, Z., Rahman, A. K. M. S., Nobel, N. I., Hossain, S. A., : An Effective Implementation of Web Crawling Technology to Retrieve Data from the World Wide Web (www). In. International Journal of Scientific & Technology Research, Volume 9, Issue 01, ISSN: 2277-8616, pp: 1252-1256. (2020)
- [9] Lu, H., Zhan, D., Zhou, L., He, D., : An Improved Focused Crawler: Using Web Page Classification and Link Priority Evaluation. In: Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2016, Article ID 6406901, 10 pages, (2016) <http://dx.doi.org/10.1155/2016/6406901>
- [10] Decision tree algorithm,  
<https://www.geeksforgeeks.org/decision-tree-introduction-example/>  
last accessed: 12 April 2020
- [11] Decision tree algorithm,  
[https://en.wikipedia.org/wiki/Decision\\_tree\\_learning](https://en.wikipedia.org/wiki/Decision_tree_learning)  
last accessed: 13 April 2020
- [12] K-means algorithm,  
<https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm>  
last accessed: 15 April 2020
- [13] Euclidean distance,  
[https://en.wikipedia.org/wiki/Euclidean\\_distance](https://en.wikipedia.org/wiki/Euclidean_distance), last accessed: 15 April 2020
- [14] Shamrat, F. M. J. M., Tasnim, Z., Mahmud, I., Jahan, M. N., Nobel, N. I., : Application Of K-Means Clustering Algorithm To Determine The Density Of Demand Of Different Kinds Of Jobs. In. International Journal of Scientific & Technology Research, Volume 9, Issue 02, ISSN: 2277-8616, pp: 2550-2557, (2020)
- [15] <https://jobs.waltonbd.com/>  
last accessed: 03 April 2020