# SMART RESUME
# BI TOOL TO SHORTLIST CVs FOR A JOB VACANCY

Y. I. Kodithuwakku
IT 14115776

FINAL REPORT

B.Sc. Special (Honors) Degree in Information Technology Specialization in Information Technology

Department of Information Technology

Sri Lankan Institute of Information Technology

September 2018

# SMART RESUME
# BI TOOL TO SHORTLIST CVs FOR A JOB VACANCY

Y. I. Kodithuwakku
IT 14115776

Dissertation submitted in partial fulfillment of the requirements for the B.Sc. Special Honors Degree in Information Technology.

Department of Information Technology
Sri Lankan Institute of Information Technology

September 2018

# Declaration

I declare that this is our own work and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

| Registration Number | Name | Signature |
|---|---|---|
| IT 14115776 | Y. I. Kodithuwakku | |

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

………………………………… ……………………………

Mr. Lakmal Rupasinghe Date

Project Supervisor

# Abstract

Today in Sri Lanka, most industries follow up one traditional process in hiring new employees. The normal process includes, advertising the vacancy, calling Curriculum Vitaes (CV), short listing them by referring the cvs and interviewing the short listed candidates. Having the right set of cvs is vital since a CV is the representation of the qualifications of an applicant. Also, when it comes to an emergency project, the employer should be able to hire the best employee set within a minimum time period. In this case help of a third-party CV storage which already has a collection of related cvs and has the ability to generate the list of most qualified applicants among them, would be helpful. Already there are many CV storages where people can submit their cvs and afterward they assign the applicants to relevant jobs. But still, there is a shortage of tools that support in selecting the best-qualified set of employees to an employer. The submitted cvs should be read properly and check several attributes such as skills, experiences and some personal information in order to select the best. It is much time consuming for a human to read and draw a mind image of the applicant. Smart Resume is a business intelligence tool for the IT sector, which analyze and classify operational data with classification algorithms to present complex and competitive information to decision makers, in order to dynamically fulfill the business needs. It is built to satisfy the task of generating the list of most suitable candidates. Third party CV storages can use this tool and provide the best solution for any client company. In this paper, we present a combination of desktop and web application that facilitates the task of automating the selection of the most suitable and qualified candidates depending on the attributes given by the user like Age, Gender, Work Experience, Soft skills and Education Qualifications. Depending on the relationship of the attributes (Internal and External) **Smart Resume** will dynamically visualize the most optimal or feasible candidate list.

## Acknowledgement

First and foremost, we would like to thank to our supervisor of this project, Mr. Lakmal Rupasinghe for the valuable guidance and advice. He inspired us greatly to work in this project. His willingness to motivate us contributed tremendously to our project. Besides, we would like to thank the authority of Sri Lanka Institute of Information Technology (SLIIT) for providing us with a good environment and facilities to complete this project. Also, we would like to take this opportunity to thank to the course instructor Mr. Jayantha Amararachchi and Prof. Chandimal Jayawardena, Dean – Faculty of Computing, Department of Information Technology, of Sri Lanka Institute of Information Technology (SLIIT) for offering this Comprehensive Design & Analysis Project (CDAP) subject, Computing Project. Finally, an honorable mention goes to our families and friends for their understandings and supports on us in completing this project. Without helps of the particular that mentioned above, we would face many difficulties while doing this.

## Contents

## List of Tables

## List of Figures

## List of abbreviations

| Abbreviation | Description |
|---|---|
| GUI | Graphical User Interface |
| IT | Information Technology |
| ETL | Extract, Transform, Load |
| BI | Business Intelligence |
| CV | Curriculum Vitae |
| ROC | Receiver Operating Characteristics |
| TP | True Positive – correctly predicted that suitable for job vacancy |
| TN | True Negative – correctly predicted that not suitable for job vacancy |
| FP | False Positive – Incorrectly predicted that suitable for job vacancy |
| FN | False Negative – Incorrectly predicted that not suitable for job vacancy |

# 1. INTRODUCTION

## 1.1. Background

Data in business is very useful only if data is analyzed properly which will aid in strategic business decision making. In the IT industry, workforce management is critical since there is a huge workflow in the industry. Hiring individuals make a big impact on the productivity of the company. To hire best-qualified employees there must be a system, as it is time-consuming to read and study all the curriculum vitae. Standardly curriculum vitae has 2-3 pages sometimes it may have more than 3 pages. So, it may take a long time to go through all the details on it. And also, there is a probability of missing some CVs, as there is no proper method to store them. Therefore, we must find solutions to automatically analyze data, classify and summarize it along with discovering and characterizing trends and flagging anomalies in order to ease the decision-making process effectively in a company. Smart Resume targets at developing a set of tools, technologies and programmed products that are used to collect, integrate and make data available for better, faster decision making.

In Smart Resume, initially, all the CVs that are receiving by the email server in PDF form will be downloaded automatically. After downloading, each and every CV will be fully read from the system itself automatically. Then the downloaded CVs will be converted into CSV (Comma Separated Values) file format and saved. Here the internet BOT technology will be used to implement this functionality. As this is a BOT technology-based system, this will be an end to end system. The system will be worked under one training data set. An Internet BOT simply means a web robot. It is basically a software application which is used to run the automated tasks over the internet. A request to the server will send and the data will be fetched.

There are various attributes help to identify the best candidates for a given vacancy. After the data saved in CSV format, an ETL (Extract, Transform, Load) tool will extract relevant attributes from the Source Files, clean them and save them in the data warehouse for further analyzes.

As the next step, client requirement for the job role is gathered. Then the prediction will be made by considering both client requirements and candidate data. It always depends on the nature of the job role, tasks, and required qualifications. Smart Resume will provide a shortlisted CV collection of candidates according to the user requirements.

Finally, by processing the entire data set based on predefined predictive models, it will generate comprehensive, descriptive and operational dashboards based on fact view data for those responsible for the job vacancies. The optimal or feasible result will be represented as a solution to the task in a user-friendly way. Graphs, charts, hierarchies, and tables will be used to represent those data.

## 1.2. Literature survey

This Research paper presents that selecting the appropriate metric to determine the optimal solution for obtaining an optimized classifier is a decisive step. The correct choice of the metric ensures that generative type classification training classifier is optimal. In this article, it is expected that the reviews of some metrics to recognize the optimum solution will sensitize data mining Researchers on this topic and encourage them to think carefully before choosing and applying metrically suitable to optimize the classification of the training.[1]

In addition, this article also suggests several important aspects in constructing a better metric for recognizing the optimal solution for the generative type of classification algorithms. [1]

In this paper, to classify the engineering materials into different classes for the selection of materials that suit the input design specifications, naïve Bayesian and C4.5 decision tree classifiers(DTC) are applied. Here, classification are analyzed individually and their performance evaluation is analyzed with the predictive parameters of confusion matrix and standard measurements, and the result of the classification is analyzed on different class categories. [2]

## 1.3. Research Gap

Even though there are existing proposed products in the market area, they do not address most of the problems that the proposed system is going to address. The following table shows a comparison of features between the existing products or applications and the proposed solution “**Smart Resume**”.

| Features | **Oracle BI** | **Birst** | **Jobscan** | **Smart Recruit** | **Smart Resume** |
|---|---|---|---|---|---|
| BOT(Automated) | | | | | ✓ |
| ETL Tool | | | | | ✓ |
| Optimal Solution | | | ✓ | | ✓ |
| Feasible Solution | | | | | ✓ |
| All the user does not have to interact with the system | ✓ | ✓ | ✓ | ✓ | ✓ |
| Visualization of data in an abstract way | | ✓ | ✓ | ✓ | ✓ |
| Predictive Analytics | ✓ | ✓ | ✓ | | ✓ |

*Table 1. Comparison with existing system*

## 1.4. Research Problem

The world is a chain of businesses. As the businesses get bigger day by day, the complexity and the competition are highly increasing. New trends are being incorporated into business ecosystems. New technologies are evolving rapidly, and it has significant influence in the organizational processes. So, day by day small to large all companies have to update themselves in terms of resources, manpower, and infrastructures in order to maintain a competent and business system.

It is a known and documented fact that information-driven culture is needed in a company to meet customer needs in today's world. To achieve this, they need a tool or service to discover and prioritize business challenges across their organization with these new one's assessment methods and bring information closer to them, so they can make a smarter decision.

When hiring new employees, a company will have to spend much time, effort and cost on finding suitable candidates among thousands of the educated and qualified ones. As of the recruiting process using in the industry nowadays, companies have to spend a huge cost and time on selecting the perfect ones for the vacant position.

1. Advertise the vacancy
2. Call Curriculum Vitae of the interested candidates.
3. Shortlisting the applied candidate list by referring their Curriculum Vitaes.
4. Interview the shortlisted candidates and recruit the most suitable ones for the position.

But practically, it takes a lot of time and effort for a human to judge an employee's skill and talent just by reading their Curriculum Vitaes. Normally, a CV should contain 2 to 3 pages and all the relevant qualifications should be listed there. Because, according to normal policy, the time dedicated to reading one CV is 6 to 7 seconds. The reader should be able to grab the relevant information within that time period.

But, practically, there may be well qualified, talented candidates, who have a large skill set and a CV extended from 7 to 8 pages since it has to hold each and every qualification they achieved. Sometimes, the required qualifications for the specific position they applied, would be included in the last pages of the CV. In this kind of scenario, the reader would miss the important skills or points because they cannot waste much time on one CV. It is much time consuming for a human reader to read one CV end to end. And also, the most qualified candidates would not be called to the interview just because their CV is too long or not well formatted. It is a huge disadvantage not only to the candidate but also to the company, since the company may lose the best employee to their vacancy.

On the other hand, there may be hundreds of applicants for a vacancy of a large IT industry. In such a scenario, it is very hard and time-consuming to download each and every CV and read them one by one in order to shortlist in human hands and send them to relevant companies who need suitable candidates. And also they need to save those candidates details in their databases for further uses such that another company also want the same set of candidates.

There is no any similar tool to satisfy the exact issue, but there are some similar commercially available BI tools but below are the drawbacks.

1. They are far too expensive, which are being developed by big vendors and often target the big clients.
2. Small and medium scale companies cannot afford a big cost or a time to find the suitable candidates because many of the employees have to engage in multi-tasks inside the company.
3. Cost for advertising for each candidate is very high.

Considering all the above facts, we can come to a point that there is a real need of cost-effective Business Intelligent (BI) tools that can cater the need of recruiting the best employees to a company by sending a set of selected candidates for relevant companies who require them. Therefore, the purpose of this research is to address such mentioned issues.

## 1.5. Research Objectives

### 1.5.1. Main Objectives

Introduce an intelligent system to select the CVs based on the characteristics/attributes of the job vacancy. Depending on that given characteristics an optimal or feasible CVs will be filtered. The optimal CV selection is the one with the lowest probability of unfavorable outcomes an optimal solution is a theoretically proven solution. But this might not be a correct logical solution and we may have to come up with a feasible team. Hence the tool has the option of providing the most feasible (possible and practical) solution as well. The **Smart Resume** enables users to select the most optimal or feasible CVs according to the given attributes for the job vacancy of IT company or industry. The current market of the BI tools has a very complex interface, which requires professional knowledge to perform tasks. Smart Resume will not require technical knowledge or professional expertise to interact and it will be developed in a simple way with fewer controls to increase adaptability and comfort for the user.

### 1.5.2. Specific Objectives

- Download the CVs in the automatically
- To read the downloaded CVs
- Classify the data into relevant columns

- Save the classified data in CSV format
- Mapping CSV data into data warehouse by Extracting most suitable attributes, then cleaning the extracted data into more analytical manner by removing redundant data omitting null values and apply appropriate values. Then map the cleaned and transformed data into the data warehouse created using MySQL for further Processing and analysis.

- ✓ Identify the Attributes with the use of Relevant data
- ✓ Extract the Relevant Data from CSV
- ✓ Cleaning Unorganized, Redundant Data (Data Cleansing)
- ✓ Transform Cleaned Data into Different Format
- ✓ Creating data warehouse using MySQL
- ✓ Load (Map) Cleaned and Transformed Data into data warehouse

- Build a solution to select the most optimal and feasible candidate list.

- ✓ The main objective of this research part is to generate the most suitable candidate list in a faster and more accurate way considering attributes given by the user, related to the IT sector. The prediction is expected to be highly accurate and the final decision is sent to create a graphical view.

  After building the Predictive Model, the predicted details will be stored in a MySQL database. Using that data will find out that that result is the optimal or feasible CV selection according to the given attributes. ROC and classification Metrics will be used for those things. Also, the predicted result will be representing as Graphs, charts, hierarchies, and tables. It will provide a better user interaction through the interactive dashboard.

# 2. METHODOLOGY

This section includes detailed descriptions about the techniques and mechanism employed to make **Smart Resume** a reality. The descriptions include how software implementation of our project is carried out, what are the materials and data needed, and how they will be collected. It also includes time frames and schedules that are required in achieving its objectives. In addition to them, the research areas that we have identified in order to carry out this project are explained rationally.

### 2.1.1. Model Evaluation & Dashboard Simulation

#### 2.1.1.1 Model Evaluation

The evaluation of the predictive model is carried out using three methods.

- ✓ Classification Table (Confusion Metrix)
- ✓ ROC (Receiver Operating Characteristics)
- ✓ Model Accuracy

The classification table shows the number of correct and incorrect forecasts made in comparison with the actual results (objective value) to the data.
The ROC curve is a graphical mark that illustrates the operation of a binary classifier system, such as it's the threshold of discrimination is different. A curve is created by tracking the true positive rate against false positive speed in various threshold parameters.

The precision of the model is measured by the proportion of correct predictions to the total number of cases evaluated.

After the Model creation it is very important to test the accuracy of that model. I have used two main model evaluation techniques; Confusion Matrix and Cross Validation to test the overall accuracy.

- **Confusion Matrix**

  A classification is generally evaluated by a confusion matrix. Figure.1, illustrates that TN denotes the successfully classified negative examples count. FN denotes the incorrectly classified positive examples count. FP denotes the incorrectly classified negative examples count and TP denotes the successfully classified positive examples count.

| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | TN | FP |
| Actual Positive | FN | TP |

*Figure 1 Confusion Matrix*

**Overall Accuracy = TP+FN/ (TN+FP+FN+TP)**

When is each candidate eligible for a job vacancy, does the classifier accurately predict it? This measurement is called “Recall”. When a classifier predicts a candidate will suitable, how often does that candidate actually suitable? This measurement is called "precision".

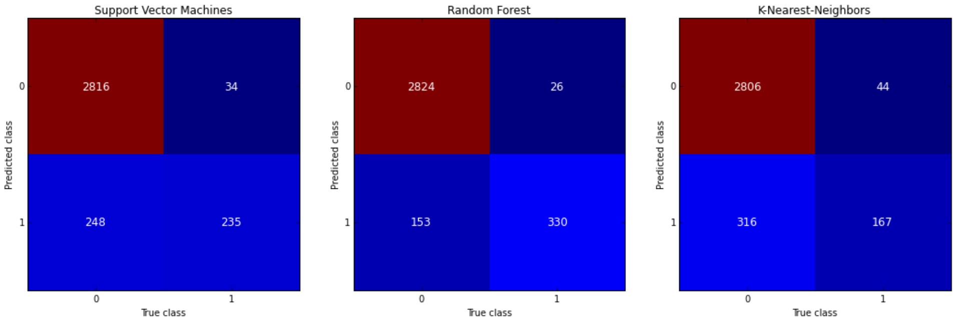


*Figure 2 Confusion Matrix for each Algorithm*

As mentioned above Support Vector Machine has the highest accuracy, precision, and recall from the algorithms I trained. But this will not always be the same. This will heavily depend on the dataset. Since different job vacancies have different datasets, different patterns, and attributes; we used all three algorithms and will show the accuracies of each in the interface. So that the company can decide on which algorithm to select according to their need or they can allow the system to select the best model automatically by changing the feature given in system settings of the dashboard.

- **Cross Validation**

Cross validation attempts to avoid overfitting (training on and predicting the same data point) while still producing a prediction for each observation dataset. This is trained in a variety of datasets while training a systematic set of models. After training, according to the hypothesis that has hide from each model, many train tests are successively subdivided. When done correctly, each observation has a ‘fair’ correspondence. Here, I used the Python Scikit Learn for cross validation, it calculates the Model Accuracy is measured by the correct predictions ratio to the total number of cases evaluated.

#### 2.1.1.2 Dashboard Simulation

After analyzing the filtered result data will present it in a format that will make user perfectly understands the difference between raw data and predictable data that would be able to represent the number of candidates got selected for the interview. These data representations will help the Evaluators/Interviewers take the decisions on selecting most suitable professional candidates according to the company requirement. The dashboard will represent the data in an interactive way using graphs, charts, hierarchies, and tables. The dashboard will provide you Summary of results and key points of analyzed data.

This approach enables the user to identify optimal or feasible CVs according to the specified attributes (requirement) given by the company. So that, the applicants who haven’t submitted a good quality CV, also got the opportunity. It is not an only a static panel, where the user presents a set of predefined data each time they are loaded. This system is expected that it will provide a better user interaction through the interactive dashboard.

## 2.1. Testing and Implementation

| | |
|---|---|
| Test Case ID | TC1 |
| Use case description | Validate valid User Credentials |
| Pre –Condition | Login interface is loaded |
| Test Procedure | Type username and password<br>Click login button |
| Test Input | User name = admin<br>Password = admin123 |

| | |
|---|---|
| Expected Output | User should be able to successfully log in to the system |
| Actual Output | User logs into the system |

*Table 2 Test case 1*

| | |
|---|---|
| Test Case ID | TC2 |
| Use case description | View Predictive model evaluation results |
| Pre –Condition | Should be logged into the system |
| Test Procedure | Type username and password<br>Click login button<br>Click model evaluation results button |
| Test Input | User name = admin<br>Password = admin123 |
| Expected Output | User should be able to successfully log in to the system |
| Actual Output | User logs into the system |

*Table 3 Test case 2*

## Research Findings

The Smart Resume system is specified for the ETL Tool, Prediction, model accuracy and Dashboard. Model accuracy is calculated using confusion matrix, cross validation for overall model. And ROC curve displays the specific model’s accuracy.

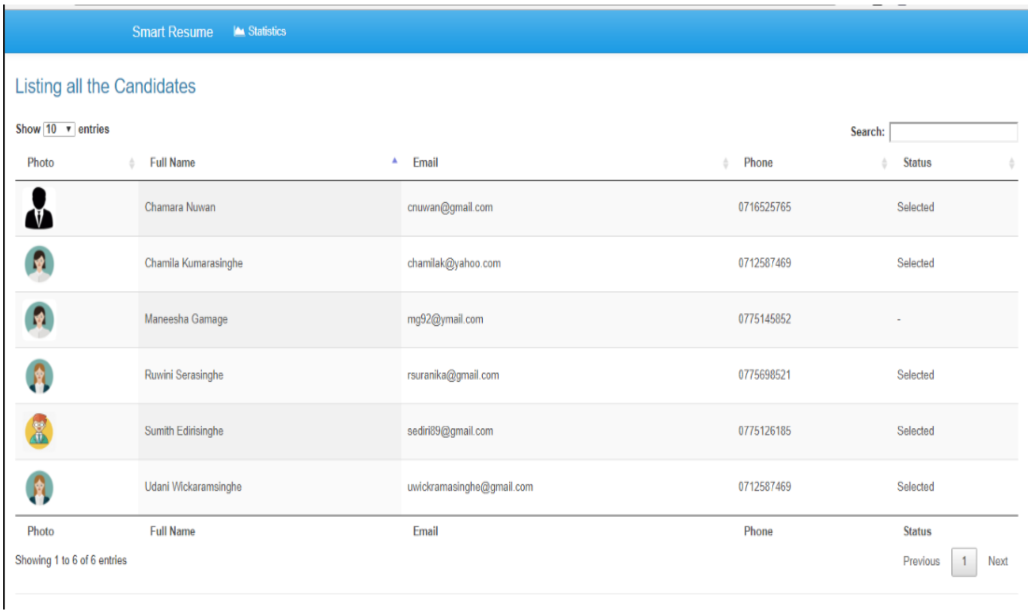
In order to build the Model evaluation scikit-learn a machine learning library in Python is used. Dashboard is build using php for the interfaces and for the function python is used.

# 3.RESULT AND DISCUSSION

This section conveys about the results of the research. The purpose of this topic is to give the reader to compare and contrast the research topic and the implemented system. After reading this section Reader can understand the research completely and the implemented system correctly.

## 3.1. Result

The final output of the system is an attractive web-based dashboard which includes detailed graphs, charts, tables and hierarchies. According to the selection made by the Predictive Model, selected candidates will be display in a table with their required personal details. This would be a great help for the company to easily and efficiently select some candidates. That table can sort, search the results. Every candidate is shown a summary of his/her chosen attributes.



Listing all the Candidates

| Photo | Full Name | Email | Phone | Status |
|---|---|---|---|---|
| | Chamara Nuwan | cnuwan@gmail.com | 0716525765 | Selected |
| | Chamila Kumarasinghe | chamilak@yahoo.com | 0712587469 | Selected |
| | Maneesha Gamage | mg92@ymail.com | 0775145852 | - |
| | Ruwini Serasinghe | rsuranika@gmail.com | 0775698521 | Selected |
| | Sumith Edirisinghe | sediri89@gmail.com | 0775126185 | Selected |
| | Udani Wickaramsinghe | uwickramasinghe@gmail.com | 0712587469 | Selected |
| Photo | Full Name | Email | Phone | Status |

Showing 1 to 6 of 6 entries

*Figure 3 Listing all selected candidates*

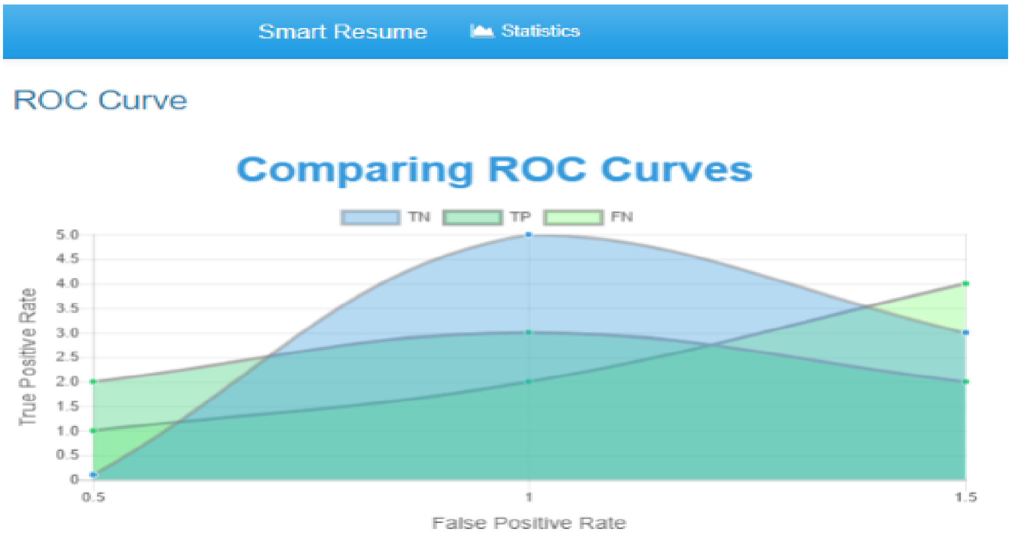


*Figure 4 ROC Curve for the predictive Model*

## 3.2. Discussion

With hard working throughout the year, we have successfully completed the **Smart Resume** BI Tool for selecting CVs for IT industry. Helpful advices and guidance by our supervisor Mr. Lakmal Rupasinghe and my team members supported to achieve this goal. Improvements to the proposed research have made plans to reliably restart internationally through localization of major languages around the world. Smart Resume is Due to the time constraint, we wouldn't able to finish some feature which is useful for future researchers and it will discuss in the next chapter.

# 4.CONCLUSION

Today for most companies like IT, receive huge number of CVs for a vacancy as there are lot of graduates coming out from a university within a year. The quality of the company depends with the capabilities of the recruiters. Therefore, getting the best qualified people from incoming applications is very difficult for them. Currently the selections of CVs are done manually. As the huge number of CVs are summarized by manually, qualified CVs can be skipped by CV errors and human errors. That is where Smart Resume come in to play by reducing all the difficult tasks of selecting optimal set of candidates in place of referring CVs. Smart Resume contains a Desktop application with ETL (Extract, Transform, Load) Tool to transform data into a meaningful order and Smart Resume has a Dashboard as a Web application in order to predict best candidates for a given job opportunity by analyzing the CV data. As the initial step Smart Resume analyze and extract the professional skills, personal skills, personal details and etc. from the CV according to the job vacancy. Finally, Smart Resume display the results in an attractive proper manner in a dashboard with simple graphs and charts which shows the user friendliness of the system.

# 5.REFERENCES

[1] H. M. and S. M.N., &quot;A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS&quot;, Pdfs.semanticscholar.org, 2015. [Online]. Available:https://pdfs.semanticscholar.org/6174/3124c2a4b4e550731ac39508c7d18e520979.pdf. [Accessed: 01- Apr- 2018].

[2] H. K.S., "Performance Evaluation of Predictive Classifiers For Knowledge Discovery From Engineering Materials Data Sets", Arxiv.org, 2008. [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1209/1209.2501.pdf. [Accessed: 03- Apr- 2018].

# 6. APPENDICES

## Appendix A: