Skill Evaluation and Career Mapping

BY

Shourav Roy Badhon
ID: 172-15-9654

Nazmul Hasan
ID: 172-15-9762

Nadira Zaman
ID:172-15-9666

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori
Associate professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised By

Aniruddha Rakshit
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

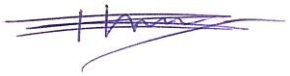
DHAKA, BANGLADESH

MAY 2021

APPROVAL

This Project titled "**Skill Evaluation and Career Mapping**", submitted by Shourav Roy Badhon, ID No:172-15-9654, Nazmul Hasan, ID No:172-15-9762 and Nadira Zaman, ID No:172-15-9666 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation was held on 31 May, 2021.

BOARD OF EXAMINERS



Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Nazmun Nessa Moon
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Aniruddha Rakshit

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin

Professor

Department of Computer Science and Engineering

Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr**. **Sheak Rashed Haider Noori, Associate professor & Associate Head, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

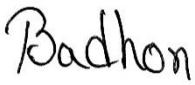
Supervised by:

Dr. Sheak Rashed Haider Noori
Associate professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:

Aniruddha Rakshit
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Shourav Roy Badhon

ID: 172-15-9654

Department of CSE

Daffodil International University

Nazmul Hasan

ID: 172-15-9762

Department of CSE

Daffodil International University

Nadira Zaman

ID: 172-15-9666

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First we would like to thank almighty God for His divine blessing that has made it possible for us to complete the final year project successfully.

We would like to express our gratitude to our Supervisor **Dr. Sheak Rashed Haider Noori**, Associate professor & Associate Head, Department of CSE Daffodil International University, Dhaka. The guidance and support of our supervisor has enabled us to carry out this project. His knowledge in the field of *Data Mining & Machine Learning* has helped us tremendously. He has shown us the way out when we were stuck with a problem. This work would not have been possible without the vision of our former department head, **Dr. Syed Akhter Hossain.** We would also like to give our heartiest appreciation to **Aniruddha Rakshit** and **Professor Dr. Touhid Bhuiyan**, Head**,** Department of CSE, for his kind help to finish our project.

We would like to thank our friends who supported and helped us in any way in this project. Finally, we thank our parents for their support and believe in us that has enabled us to complete this project successfully.

ABSTRACT

Software development is an emerging sector with a very big potential in Bangladesh. This is a sector where there is a need for professionals with knowledge of various methodologies, tools, and techniques. To have a knowledge of the requirements for the software industry is essential for job seekers in this sector. It is also required for universities and training centers for setting their curriculum. However, determining the necessary skills for a job and associating them with a job seeker is a difficult challenge. There is no easy way to find out requirements for specific types of job within the software development industry. For that reason, we have gone through 708 job ads from the Bangladesh software industry and made a dataset. Only technical skills were used to create the dataset. Soft skills such as teamwork, communication and customer orientation were excluded. Machine learning classifiers were used in the dataset to correlate different technical skills with different types of jobs. We have reported findings of various performance evaluation metrics that were used to evaluate this work. We found the best result from the Random Forest classifier. We have then used the best classifier to make a model. Using the model, we have developed an application where someone can check which type of job is suited for their skills. The purpose of this study is to help job seekers better understand the requirements of the Bangladesh software industry.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Software development requires highly skilled people with knowledge of various tools and methods carrying out different tasks at the same time in software projects. It is essential for job seekers who are usually fresh graduates to have a decent knowledge of the skills desired by the industry. Skills could be both technical and non-technical but technical skills are the main requirements. Understanding the skills desired by the software industry is also essential for designing curriculum for universities, training centers and online courses. But there is a gap between industry requirements and skills of job seekers. There have been some studies [8,9] done to address the gap using the traditional approach of interviewing and surveying. There have also been some studies [1,2] done using data mining. However, there is variation in the requirements of the software industry in different countries. The skill requirements also change quite quickly in the software development industry.

For that reason, we have gone through 708 job ads of software development companies to make a dataset of the requirements of the software development industry of Bangladesh. The dataset contains the required technical skills for different types of jobs of software development. Data mining techniques have been used on the dataset to retrieve information from the data. Various machine learning classifiers were used to determine the best one for making a model. Using the model, we have demonstrated how someone can find their suitable job according to their skills.

The target of this study is to determine the skills required by the software development industry of Bangladesh. It also finds the correlation between different technical skills in different types of jobs. The result of the study will help individuals to make decisions about finding jobs suitable to their skills and which skills are more necessary for their desired jobs.

1.2 Motivation

Bangladesh has a rapidly developing software development sector. As the sector gets more matured the number of jobs this is producing is also growing exponentially. According to a BASIS survey [5] of 2018 there are more than 4500 IT companies that are employing 300000 professionals approximately. Demand for software in the local market is about US $1.18 billion. Revenue from the software and ITES industry is approximately US $800 million in 2017. This is a lucrative sector for fresh graduates. So, a significant number of people are choosing this as their desired career.

However, there is a gap between the requirements of the software industry and individual skills. The university curriculum is not designed according to market demand. But the main reason is due to lack of knowledge of the industry requirements at individual level. As a result, it is hard for job seekers to find appropriate jobs and also determine which skills are necessary for their desired jobs. There is no comprehensive guideline from where they can find the necessary information. There are some studies done in different countries in order to address this problem. But those are region specific. This study is intended to analyze the skills required by the software industry of Bangladesh. There are also certain motives in our research:

1)To make a dataset of job requirements of the Bangladesh software industry.
2)Using data mining extracting useful information from the dataset
3)To describe the finding from the data
4)Creating a model from the dataset using machine learning
5)To make a tool where someone can easily determine best jobs according to their skill

1.3 Rationale of the Study

Bangladesh has a software industry that is growing exponentially every year. Software industry generates a substantial amount of revenue. The demand in the local market is also very high. This industry employs a significant number of the population currently and the potential for employment in future is great. Skilled professionals are needed for the growth of the industry. But there is a lack of knowledge about the industry requirement among job seekers. Which is detrimental to the growth of skilled professionals. This study addresses the problem of the knowledge gap of industry required skills among job seekers. Information retrieved from this research can be used by universities and training centers for designing their course curriculum. The main benefit of the study will be at individual level for job seekers in the software development industry. As the web application which will be running a machine learning model created from the data will enable anyone to check their skill with respect to their desired jobs in software development.

1.4 Research Questions

1. What technical skills (experience on languages, libraries, frameworks and tools) are in most demand in Bangladesh industry?
2. Which skills are required for a certain type of development?
3. How can we classify job ads?
4. What information is extracted from the data?
5. How can we use the information?
6. How can we make a model from the data?
7. How to utilize the model in a practical way?

1.5 Expected Outcome

The principal expected outcome of this project is to show the correlation between skills and industry requirements in the software development industry of Bangladesh. There has been no study done on the skill requirements of the software industry of Bangladesh. There is study done [1] that uses personal and academic factors but not skill required. There have been some studies done in other countries. But those are done for their respective industry requirements. So, our dataset will be from job ads which are specific to Bangladesh. The dataset can be used to make a machine learning model. We will use our model in the backend of a web application to show the correlation between skills and requirements by different jobs. This is a research project, so naturally our concern is to publish a research paper of our findings. Our dataset will be available for others to add to as the more a machine can learn the better prediction it can deliver. In this research, we will make a dataset, analyze the findings, make a machine learning model from the dataset to show the correlation between different job requirements and skills and show the necessary steps to do that.

1.6 Project Management and Finance

We have done the project in a relatively short scope. The tools used in the project are found online free of charge. We are a small group of 3 people. The web application running the model is also running on a local server. So there has not been any funding needed thus far.

But to widen the scope and to host the web application where it can handle a decent traffic load funding will be necessary. We are optimistic about finding necessary funding to expand the project.

1.7 Report Layout

This report comprises 6 chapters all of which have subsections of their own.

Chapter 1, Introductions gives a basic idea about the research. The 7 sections in this chapter are 1.1 Introduction, 1.2 Motivation, Rationale, 1.3 Research Questions, 1.4 Expected Outcome, 1.5 Finance and 1.6 Report Layout, respectively.

Chapter 2, Background has 4 sections that are 2.1Introduction, 2.2Related works, 2.3 Comparative Analysis and Summary, and 2.4 Scope of the problem. This chapter is about previous work done on the research topic.

Chapter 3, Research Methodology has 8 sections. Those are 3.1 Introduction, 3.2 Research Subject and Instrumentation, 2.2 Data collection, 2.4 Data Labeling, 3.5 Data Preprocessing, 3.6 Statistical analysis of dataset, 3.7 Implementation Requirement, and 3.8 Implementation Procedure in that order. In this chapter the whole process of this research work is summarized. Implementation Requirement is divided into two sub sections, Problem discussion and Changing Data format.

Chapter 4, Experimental Result and Discussion. 4.1 Introduction, 4.2 Experimental Results and 4.3 Summary are 3 sections of this chapter. However, Experimental Result is divided into 3 sub sections. All the findings from this research are shown in this part.

Chapter 5, Impact on Society, Environment and Sustainability has 4 sections which are 5.1 Impact on Society, 5.2 Impact on Environment, 5.3 Ethical Aspects and 5.4Sustainability plan.

Chapter 6, Conclusion and Future work is the last chapter. This chapter is divided into 4 sections and those are 6.1 Summary of the Study, 6.2 Conclusions, 6.3 Recommendation and 6.4 Implication for Further Study. This is the description of the report layout.

CHAPTER 2
BACKGROUND

2.1 Introduction

Software development industry is a huge industry which is employing millions of people across the world. Understanding the requirements of the industry is important for many stakeholders in this industry. For that reason, there have been several studies done on the subject. Different studies approached it in different ways and some studies are based on countries to better understand the requirements of the industry. These studies try to give comprehensive analysis of the industry. Points out the correlation of individual skills with respect to industry requirements. Understanding the knowledge and skills required by the software development industry will help institutions such as universities and training centers better construct their course curriculum. But mostly the knowledge will help individuals self-develop and better understand the suitable job according to their skills and knowledge.

2.2 Related Works

There have been previous studies conducted regarding this field in various ways using different methods. Some studies [7,8] were done using traditional methods such as surveying, interviewing and some were done using data mining. There are also some studies [9,10,12,13] that are specifically targeted towards a country. Some major research will be discussed in this section.

A. A. Biswas et al. [1] have used data mining for career predictions and analyzed the influencing factors. They have used survey data of 494 individual records. The survey contained 13 questions of personal and academic factors. Prediction was done to determine whether their career would be private job, government job or business. They have included description of their data and results of 6 different classifiers used in the study. They found the best result using the Part classifier.

Yeasin et al. [2] used data mining to predict students' career including student’s strength and weakness. The dataset used in this study is collected from 506 former CS students of

13 universities of Bangladesh. They have analyzed the results of 5 classifiers used in the data.

Roy et al. [3] showed career prediction of CS domain candidates using advanced machine learning techniques. They have collected nearly 20 thousand records from various sources for their dataset. SVM performed best among other machine learning classifiers so they made a web application using it as a background algorithm.

Al-Saiyd and Al-Takrouri introduced [4] to investigate how IT jobs prediction changes concerning graduate students' knowledge, skills and experience using back propagation artificial neural networks. They have used data from 50 graduate students arranged in a format of 35 input factors.

Gorad et al. [5] showed a web application that would help high school students in selecting course which are best for their career. The recommendation is done based on personality traits, interest and the capacity of taking courses of the students.

Panda et al [6] showed a Random Forest based approach to predict career based on students interest and skills. Their study mainly focused on Indian education system. They also did a comparative study of 4 machine learning classifiers.

Moreno et al [7] conducts the study to help academic institutions and software industry to have a better understanding of proficiencies of fresh graduates of undergraduates and graduate software engineering programs. This study helps academic institutions determine which skills from their curriculum are needed by the industry. This study was conducted in 2012.

Akman et al [8] explains employer's expectation in technical, personal and educational competencies among IT graduates. This study was done by surveying.

Hiranrat et al [9] aims to determine technical knowledge and soft skills required by the software industry in Thailand. Text mining was used in this study. The result of this study is intended to be used for designing training courses which will help to better prepare students for employment. It was published in 2018.

Aken et al [10] used data mining to conduct the study that will help software engineering professionals to update their skills according to the demand of computing jobs. This study

used a huge dataset consisting almost quarter of a million unique IT jobs description from various search engines. The jobs are from all across the US.

The objective of the study [12] by Giray and Görkem is to understand the desired skills in Turkish software industry. This study to help universities curriculum making, online course design and for software developers for self-development. This study analysed 1597 job ads from Turkish software development industry.

Matturro and Gerardo in their study [13] identifies soft skill or non-technical skills required by software companies in Uruguay. For this study they have analyzed 678 job ads.

Ahmed et al [14] finds out similarity of soft skill requirements by employers across different cultures in various roles of software development. 500 job ads posted online from North America, Europe, Asia and Australia are analyzed in the study.

Bailey et al [15] reports their study finding the knowledge, skills and abilities needed by computer programmers. This study was conducted by site interviews, web-based surveys and web research.

Papoutsoglou et al [16] proposes a framework which collects IT job ads from web sources and extracts required skills and competences from raw text data. They used Stack Overflow as the web source for collecting data.

In this research work we have collected data from online job ads. Then we used data mining to extract useful information from the data. Machine learning algorithms were used to build a model for web application.

Katore et al. [17] used the C4.5 algorithm in their research for career prediction and recommendation based on personality traits. Their dataset was made from records of 200 students that have 12 attributes and got 86% accuracy.

Lou, Yu, Ran Ren, and Yiyang Zhao [18] modeled a career path based on the sequence of someone's education or experience. They used 67000 profiles from LinkedIn as data to apply the K-means clustering algorithm.

2.3 Comparative Analysis and Summary

There have been various studies done to determine the gap between industry requirements and skills of graduates. Different studies were done in different regions and different time frames. For example, a research [9] was conducted to determine the technical knowledge and soft skill or non-technical skill required by the software industry in Thailand. Similarly, another research [10] finds out desired by Turkish software Industry. Boths [9,10] were done using text mining. Various methods were also used in other similar studies to address the gap between industry requirement and graduate skill. But these studies do not necessarily fully reflect the requirements of the Bangladesh software industry. The studies [1,2] conducted in Bangladesh are not strictly based on skills required by the software development industry. Our research addresses this problem by using job ads from software companies of Bangladesh. The dataset is used to make a model that runs in the backend of a web application. This will help job seekers to find appropriate jobs according to their skills as well as skills required for their desired jobs.

2.4 Challenges

There was no dataset of job requirements in the Bangladesh software industry available at the time we started our research. Our first challenge was to make a dataset from job ads. But the job portals show the job ads for a certain period. So, finding enough data from only one place was not feasible. There was also the problem of imbalanced data because certain types of jobs are more in demand than others. Labelling 708 data by going through each one manually was quite a challenge. After the data collection was complete the next challenge was to transform the data into a form where using ML a model can be created to be used in the backend of a web application. There was no such work done in making a model from the job requirement dataset so there was nothing to take guidance from. The previous research works were about analysis of collected data.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the entire methodology of the research work. Each research approaches a problem in a unique way. We have described in detail the approach we have taken in the methodology part. In our research we have used 708 job ads from online sources to make the dataset. The process of collecting data from the ads have been described in detail. We have labeled the data then. After that the dataset was preprocessed. Statistical analysis was done. The data is in text form. So the data was processed again to convert into a form in order to make a machine learning model. The model was then used in the backend of a web application. The application was made using Django. Every step is detailed in the methodology part individually. This is done to better explain the research work. As better explanation increases the efficacy of work and give the nobility. Graphics representation with their description are used to better demonstrate the work. This makes understanding the work easier. A good explanation of the methodology helps in further research.

Every step of the methodology of our research work has been presented in this chapter. Sub section of the main sections makes understanding the summary of work much easier. It also helps to make sense of the purpose of each step. Figure 3.1.1 which is given below displays the research method that we used in this study.

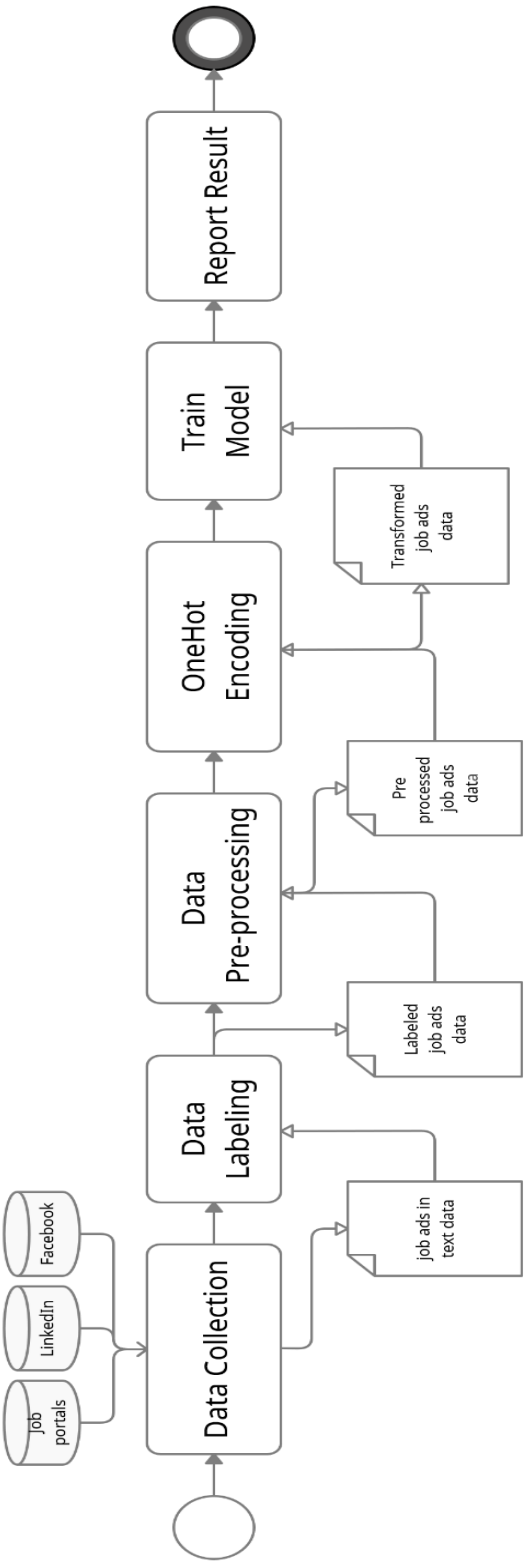


Figure 3.1.1 The research methodology used in this study

3.2 Research Subject and Instrumentation

The name of our research work is “Skill evaluation and Career Mapping”. This is an important research in the software development industry of Bangladesh. We have discussed the process of making a model for job prediction with the conceptual and theoretical process. A machine learning model needs a decently configured computer with high processing power. There is a list given below of the required instrument for this work.

Hardware and Software:

- Intel Core i3 8th generation processor with 12 GB RAM.
- 240 GB SSD
- Google Collab with 12 GB RAM
- Microsoft Visual Studio

Development Tools:

- Windows 10
- Python 3.8.8
- Pandas
- Numpy
- Django

3.3 Data collection

The data used in this project is our own collected data. The data is collected from job ads posted online between January 2019 and May 2021. The sources are Bdjobs, Skill.jobs, Facebook and LinkedIn. The job ads are from 427 different companies and each ad is unique. As the data was collected from different sources and some posts were in graphical form instead of text, we had to manually input every data. We have discarded irrelevant job ads. There were some job ads that are not related to software development, such as those seeking networking experts, IT technicians, etc. Apart from that, we have also ditched job ads whose working places are out of Bangladesh. We have only used job ads that

required explicit technical skill and knowledge in software development such as programming languages or frameworks. After collecting only relevant job ads the end result was a dataset of 708 job ads. We only considered languages, frameworks, libraries, tools, databases and did not add any soft skills such as teamwork, communication, customer orientation to the dataset. Figure 3.3.1 shown below is the initial dataset collected from ads.

HTML, CSS, PHP	Laravel, Bootstrap	jQuery	MVC, Database
C#, JavaScript, SQL	ASP.NET, .NET	jQuery	OOP, MVC, Ajax
HTML, CSS, JavaScript, SQL	.NET, AngularJs	jQuery	OOP, JSON, Node.js, Ajax, API
JavaScript, SQL	Not Required	React	Database, REST
HTML, CSS, JavaScript	Bootstrap	jQuery, React	Node.js
C#, SQL, Java	ASP.NET, .NET	Not Required	API, REST
JavaScript, PHP	Not Required	jQuery	OOP
C#, JavaScript, C++, SQL	ASP.NET, AngularJs	React	OOP, MVC
JavaScript, PHP, Java, SQL	Laravel, Bootstrap, Vue.js, CodeIgniter	jQuery, React	JSON, API, Jira
HTML, CSS, JavaScript	Not Required	jQuery	UI, JSON
HTML, C#, PHP, SQL	AngularJs, Laravel, Bootstrap	jQuery, React	OOP, MVC, JSON, Ajax, Database
PHP, SQL	Bootstrap, Vue.js	jQuery	JSON, Ajax, API, Database
HTML, CSS, PHP, SQL	Not Required	jQuery	Wordpress
JavaScript, C++, PHP, SQL, Java	.NET, Laravel, CodeIgniter	jQuery	OOP
HTML, CSS, JavaScript, PHP	Laravel, Vue.js	React	MVC, Database
HTML, CSS, PHP	Laravel, Bootstrap	React	API
HTML, CSS, JavaScript, SQL	Laravel, Bootstrap, Vue.js, CodeIgniter	jQuery, React	OOP, MVC
Java, Kotlin,Swift	Not Required	Not Required	OOP, JSON, API
JavaScript, PHP	Laravel	jQuery	Node.js, Database, REST
HTML, C#, JavaScript, Python	Django	React	UI, MVC, API, REST

Figure 3.3.1 Initial dataset collected from 708 job ads

3.4 Data Labeling

Machine learning algorithms would be used in the dataset to build a model. In order to apply ML algorithms, labeling the data was important. We labeled the data by using the category of the job. Web, Mobile and Others were used as labels. We labeled data manually by going through each job. The title of job and description was used to determine the category of the job.

3.5 Data Preprocessing

At first manual cleaning was done on the data. It was done using the steps described below:

1. Inserting only relevant data such as languages that are required.
2. Removing redundant jobs with the same company and qualification.
3. Inserting only the data that is relevant to software development.
4. Removing irrelevant data such as company name, job title etc.

After manual cleaning the data was further cleaned using python to remove duplicate values, unnecessary spaces and characters and make the spelling of all values the same.

3.6 Statistical Analysis

The total amount of data is 708 and has 110 explicit technical skills specified in [10,16] and others. The technical skills are programming languages such as Java, PHP, Python, Scripting languages, Markup languages. Technical skills also consist of frameworks such as Flutter, Laravel, Selenium etc. and library, tools and more. For some cases such as databases or API, several things were generalized to keep the data less complicated. After labeling the occurrence and frequency of the label is shown on figure 3.6.1 below.

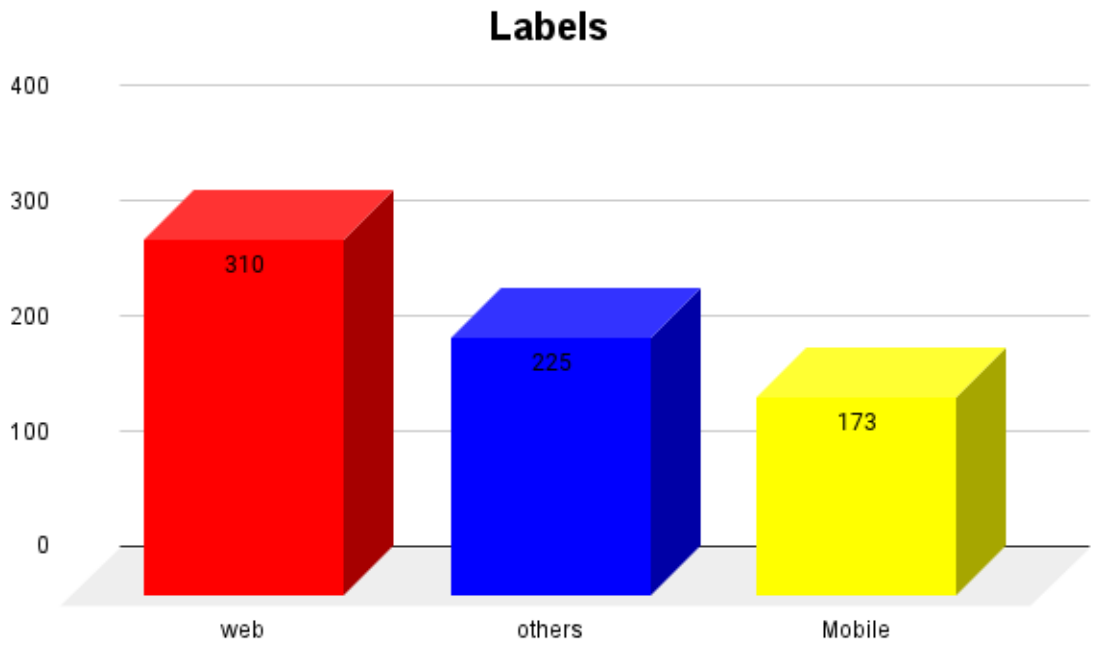


Figure 3.6.1 Occurrence of label

3.7 Implementation Requirements

3.7.a. Problem Discussion

The dataset consists of text data of different attributes such as language framework and so on. But we have to use the data to make a ML model that can be used in the backend of a web app. To do that we had to change the format of the dataset. As Machine Learning classifiers were not producing any meaningful result in the current form of data.

3.7.b. Changing Data format

Converting text data to numerical data was done using One Hot Encoding. One Hot Encoding [19] converts categorical values in data to numerical formats. It creates additional features based on the number of unique values in the categorical feature. A new feature will be added for every unique value. One Hot encoding converts the data into a form which

can be easily inputted into machine learning algorithms. In our dataset it was necessary as each column consisted of many unique values. For machine learning algorithms to work better, transformation of the data was necessary. Language column in our dataset contained 20 unique values. After the transformation 20 columns were added to the data. Each column represents a unique value of languages such as C, PHP, Python etc. If a job requires PHP and does not want C the corresponding value of the job in the PHP column will be 1 and in C column it will be 0. OneHot encoding was applied in every column except the label. After it was done the data consisted of 110 columns. Each column except label contained two integer values either 1 or 0. Suppose there is Java in a line but PHP is not present so the program will put a 1 in the Java column and 0 in php column. After changing from text to numerical the data is displayed below in figure 3.7.b.1.

	Category	CSS	Golang	JavaScript	C#	Objective-C	TypeScript	C	C++	Kotlin	Ruby	PHP	Swift
0	Mobile	0	0	0	0	0	0	0	0	1	0	0	0
1	Web	1	0	1	0	0	0	0	0	0	0	0	0
2	Mobile	0	0	0	0	0	0	0	0	1	0	0	0
3	Mobile	0	0	0	0	0	0	0	0	1	0	0	0
4	Web	1	0	1	0	0	1	0	0	0	0	0	0

Figure 3.7.b.1 Dataset after One Hot Encoding

3.8 Implementation procedure

After the dataset was in appropriate format the data was cleaned using python. It was analyzed. Finally, various Machine Learning algorithms were applied to it. A model was chosen from the algorithms with the best result. We then saved and loaded the machine learning model in python. Scikit-learn [20] was used to save and load the model. Pickle was the specific module used to save the model. Pickle operation is used to serialize ML algorithms and then to save the serialized format to a file. After that the file is loaded to deserialize the model and use it for making new predictions. We used the model in the backend of our web application. The web application was built with Django.

This section describes the various classifier algorithms used in our dataset to predict career in the software development industry. We have discussed the difference in performance of the models used for prediction. We have used 80% of our dataset for training and 20% of the dataset was used for testing the classifiers. Scikit Learn library was used for implementing data mining algorithms on the dataset.

Random Forest

Random forest is a favored supervised machine learning algorithm. This algorithm consists of several decision trees. Value of random vector samples from each tree combined affects tree predictors of Random Forest. Random Forest performed best in our dataset. The parameter values set for Random Forest are given below:

- n_estimators' = 100 (100 trees in the forest)
- min_samples split = 15 (Minimum 25 samples to split on)
- min_samples_leaf = 1 (Minimum 1 samples to at leaf node)
- bootstrap = True (Bootstrap aggregating was used)

The confusion matrix generated from the Random Forest classifier is shown on figure 3.8.1. The plotting of the confusion matrix was done using python library Matplotlib:

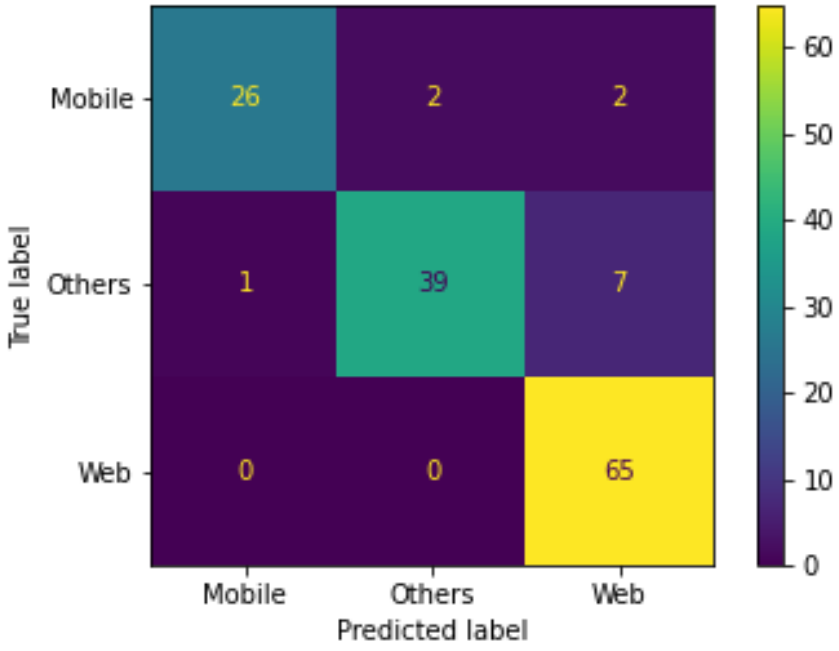


Figure 3.8.1 Confusion Matrix -Random Forest

K-Nearest Neighbor

KNN is regarded among the simplest supervised machine learning algorithms. KNN works by putting new data into a category based on the similarity with the categories available. Parameter values set for KNN are given below.

- leaf_size = 15 (Size of leaf in tree. It affects the speed of the construction and query)
- p = 2 (Distance metric used for tree is Euclidean))
- n_neighbors = 7 (No. used for neighbors queries is 5)
- weights = uniform (Uniform denoted all points in each neighborhood are weighted equally.)

The confusion matrix generated from KNN classifier is shown on figure 3.8.2 below:

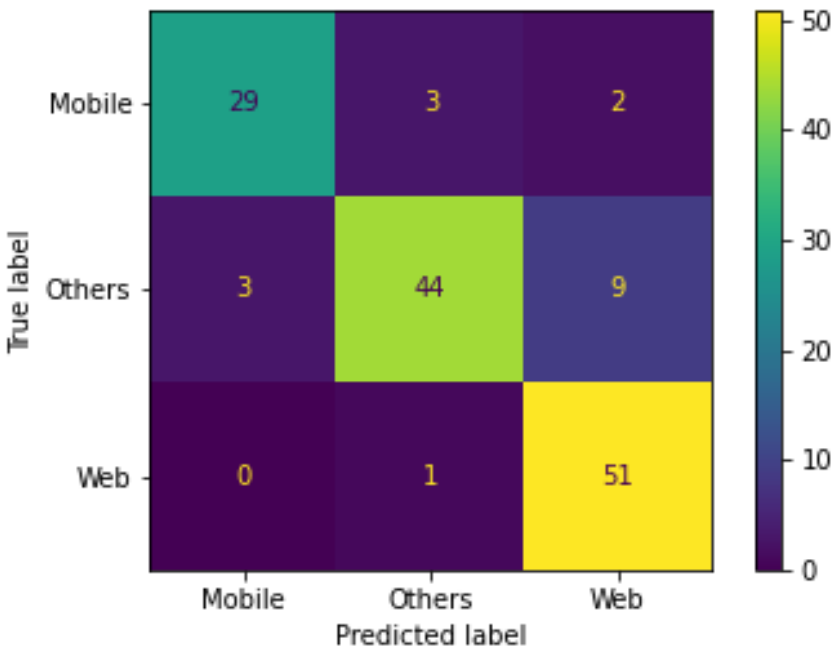


Figure 3.8.2 Confusion Matrix - KNN

Decision Tree

Decision tree is a supervised algorithm which is preferred for classification problems like ours. Decision tree builds a model to predict class using training data. The model is then used for prediction of classes. Parameter values set for Decision Tree are given below:

- Criterion = gini (The function for measuring quality of split. Here gini impurity is the splitting criterion)
- max_depth = 1o(Maximum depth of the tree is 10)
- min_samples_split = 3 (Minimum 3 sample required for splitting)
- splitter = best (It denotes selection of best split at each node)

The confusion matrix generated from KNN classifier is shown on Figure 3.8.3 below:

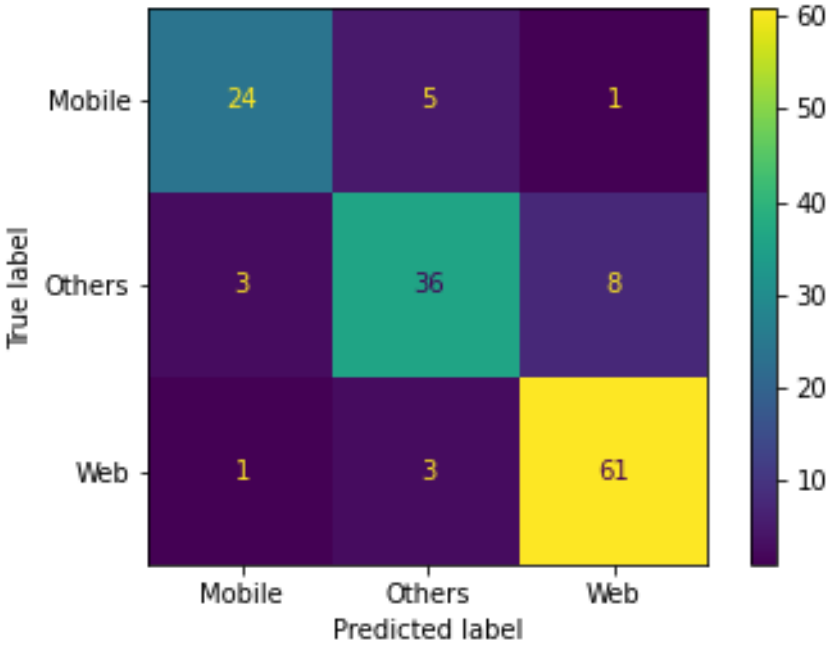


Figure 3.8.3 Confusion Matrix – Decision Tree

Multinomial Naïve Bayes

Naïve Bayes is based on Bayes theorem. It is supervised classifier that works with the assumption of independence between features of a class. We have used Multinomial NB on our dataset. It implements the algorithm for data that is multinomial distributed. Parameter values set for Multinomial NB Tree are given below:

- alpha = 1.0 (Additive smoothing parameter is 1)
- class_prior = None (It denoted no specific priors adjusted)
- fit_prior = True (Learning from prior probabilities)

The confusion matrix generated from Multinomial NB classifier is shown on figure 3.8.4 below:

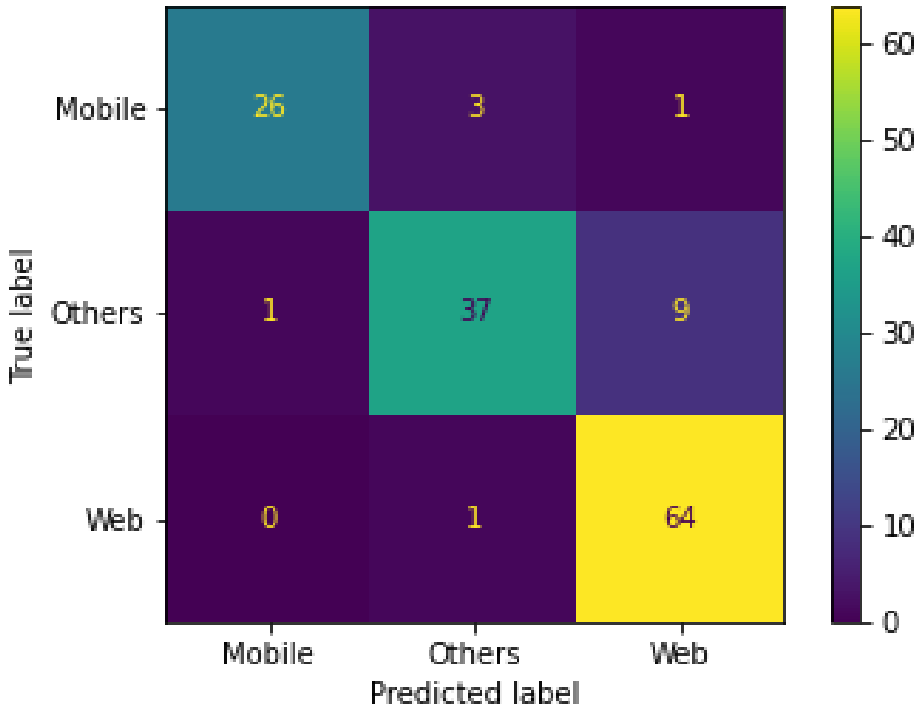


Figure 3.8.4 Confusion Matrix – Multinomial NB

Support Vector Machine

Our dataset has more than two classes, so we needed to use multi-class classification. That is why we used the 'One vs Rest' method for the Support Vector classifier. Parameter values set for SVM are given below:

- c =10 (Regularization Parameter)
- Kernel = rbf (Kernel type is Radial Basis Function)
- decision_function_shape = ovr (Returns one-vs-rest decision function shape)
- gamma = 0.01 (It uses 1 / (n_features * X.var()) as value of gamma,)

The confusion matrix generated from SVM classifier is shown on figure 3.8.5 below:

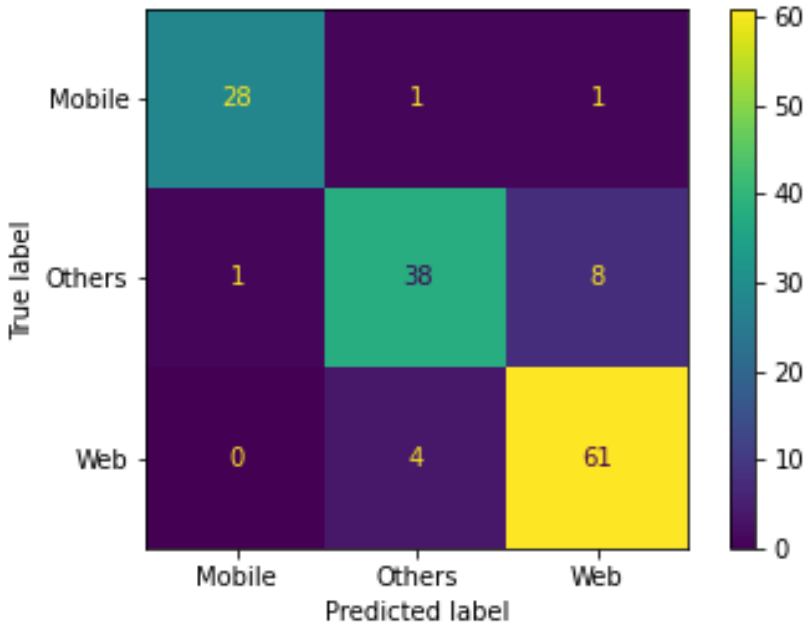


Figure 3.8.5 Confusion Matrix – SVM

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

To properly understand the outcome of this research work and utilize it, a clear understanding of the result is necessary. One of the motivations for this study is to clearly understand the required skills of the Bangladesh software industry. It will help universities and training centers better design their course curriculum. To that extent we have reported the summary of our findings from the data. Most required skills are mentioned with proper details. The main goal of the study however was to make it easier for job seekers to better match their skills to jobs in the software industry. A machine learning model was made to do so. The model is selected by comparing the results of several classifiers. The results of those classifiers are also shown. That is why we have outlined the features of our web application that is running a machine learning model made from the dataset of job ads. We will demonstrate the features of the application.

In this chapter the findings from the research work is described. A proper description about the dataset is given in Experimental Results. After that the implementation of the machine learning model is outlined. The descriptive analysis of the finding is given. Finally, a summary of the whole thing ends the chapter.

4.2 Experimental Results and Implementation

4.2.a Result Analysis

We have done a detailed analysis of the difference in performance of the classifiers used in our research in this section. We have used 4 performance metrics based on the confusion matrix of the classifiers. Those performance matrix calculation was done using following Eqs. (1-4)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (2)$$

$$Percision = \frac{TP}{TP+FP} \times 100\% \quad (3)$$

$$F1\ \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (4)$$

Here, TP, TN, FP, FN are True Positive, True Negative, False Positive, False Negative respectively.

Fig. 7 shows the accuracy of the different classifiers used in this research. We have found the best accuracy 92% from Random Forest classifiers. Accuracy of Multinomial NB and SVM both is 89%. Accuracy of KNN and Decision Tree are 87% and 85% respectively.

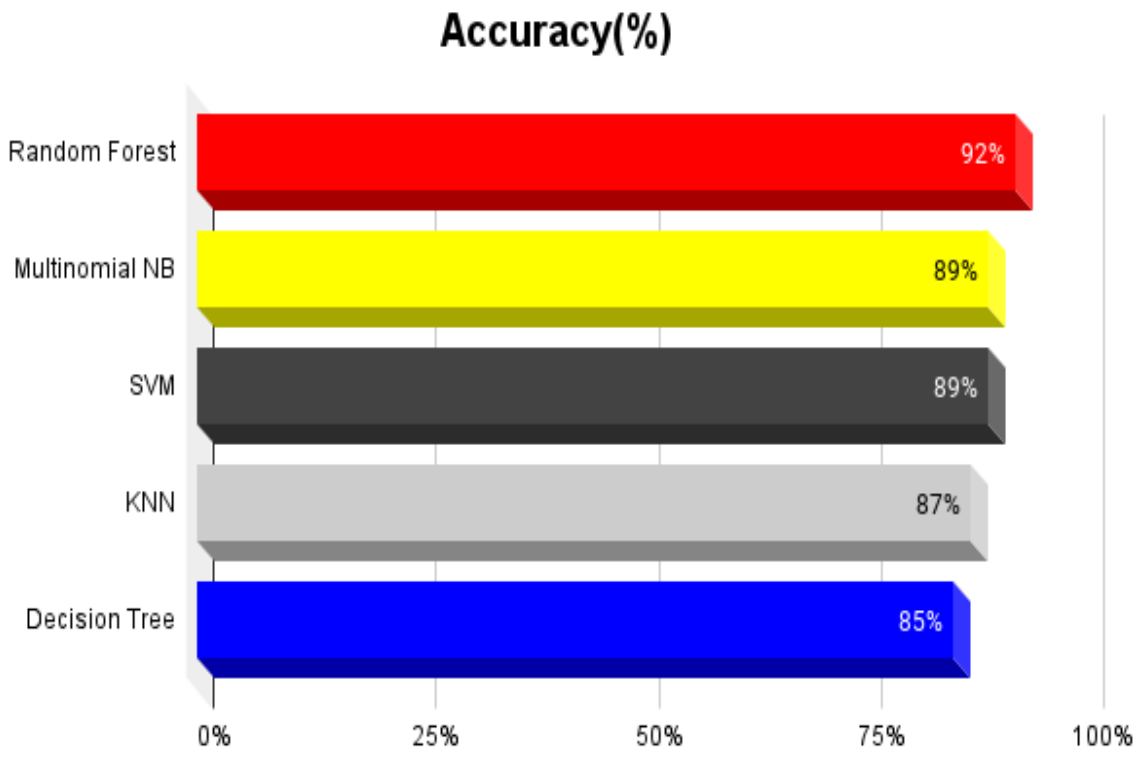


Figure 4.2.a.1 Accuracy in percentage

Precision, Recall and F1 Score of the classifiers are given in the Table 2 below. As shown by all the data of classifiers it is clear that the Random Forest model performs the best. So, we have made a web application which is running the model in the backend. Anyone can go and check which career of software development is better suited according to their skills.

Table 4.2.a.2 Comparison between classifiers performance

Classifier	Class Name (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	Mobile	96	87	91
	Others	95	83	89
	Web	88	100	94
Multinomial NB	Mobile	96	87	95
	Others	90	79	84
	Web	86	98	92
SVM	Mobile	97	93	95
	Others	88	81	84
	Web	87	94	90
KNN	Mobile	91	85	88
	Others	92	79	85
	Web	82	98	89
Decision Tree	Mobile	86	80	83
	Others	82	77	79
	Web	87	94	90

4.2.b Machine learning model implementation

Using the best classifier which in our case Random Forest we a model was made. We then saved and loaded the machine learning model and used Scikit-learn [20]. Pickle was the specific module used to save the model. We used the model in the backend of our web application. The web application was built with Django. The implementation of the model

running in the backend of the web application. The web application is shown below. Figure 4.2.b.1shows the homepage of the web application. Figure 4.2.b.2 shows the page where someone can select their technical skills. Figure 4.2.b.3 shows the result page where prediction is shown after the model has done its job.

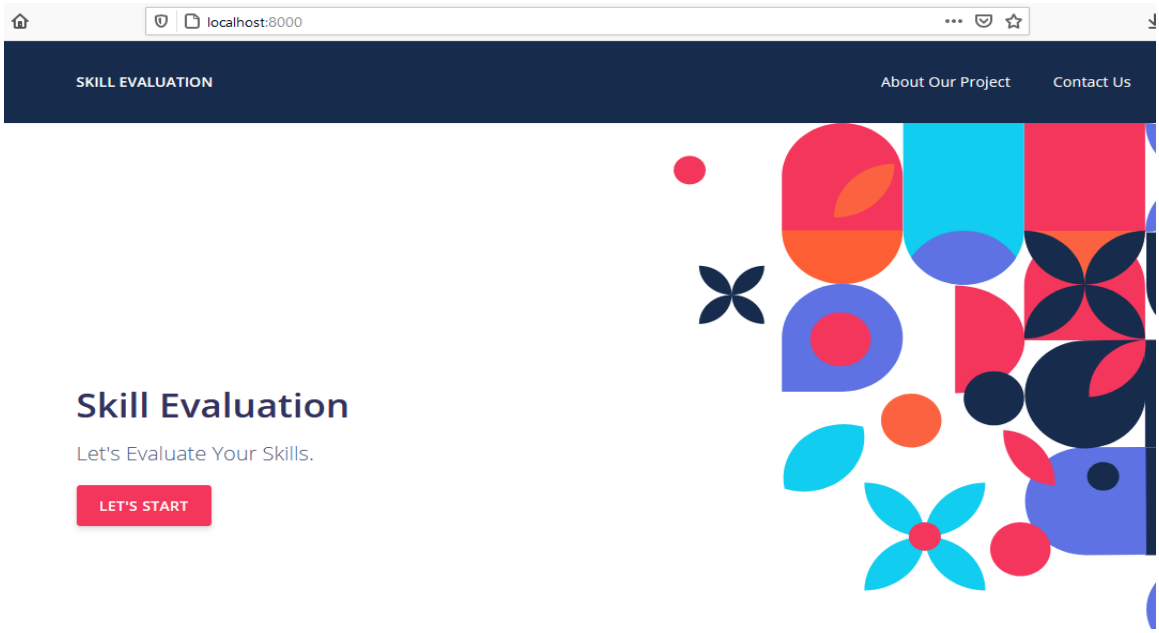


Figure 4.2.b.1 Home page of the web application

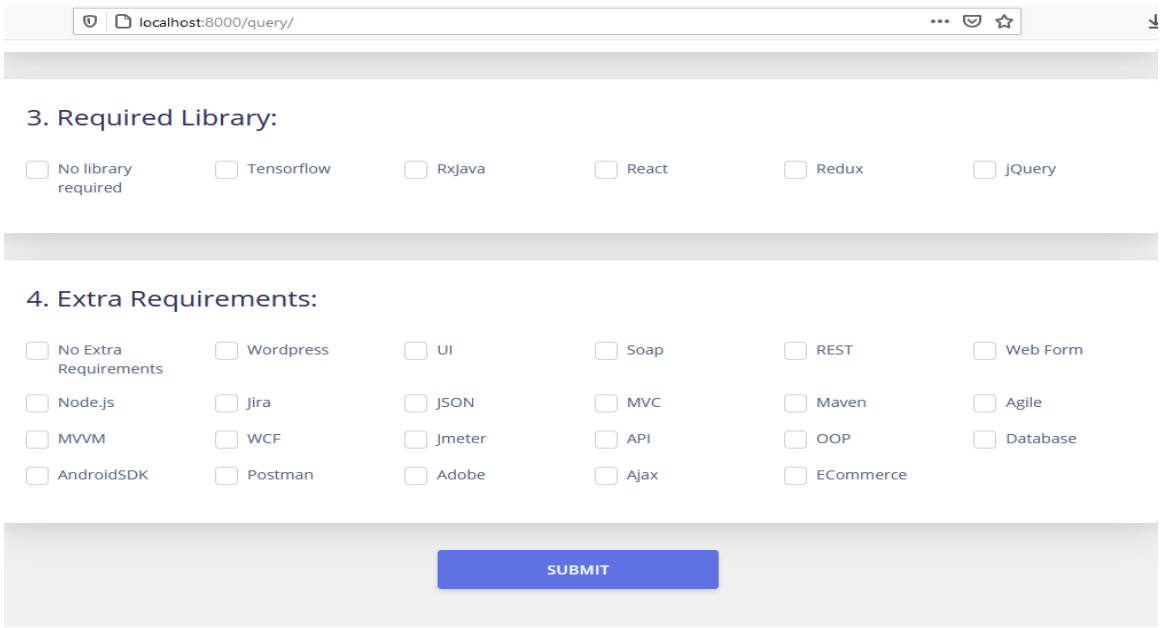


Figure 4.2.b.2 Page to select skills in the web application

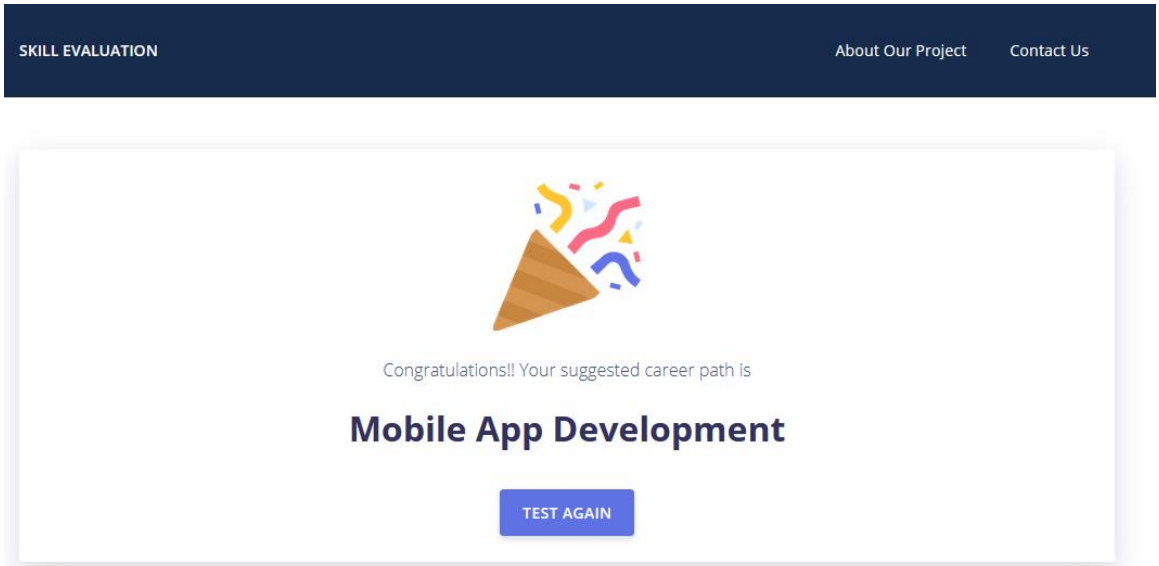


Figure 4.2.b.3 Result page of the web application

4.3 Summary

We collected a dataset of job ads that was used to determine the most required technical skills and knowledge in the software industry of Bangladesh. From the dataset we found out the most desired languages, frameworks, libraries and other requirements. We have made a model that can predict the suitable job type for someone according to their skills. To make the model we had to check the accuracy of different classifiers to pick the best one. We then implemented the model in the backend of a web application. The operation of the web application using a machine learning model was shown.

Chapter-05

Impact on Society, Environment and Sustainability

5.1 Impact on Society:

Software development industry of Bangladesh is rapidly developing but to maintain this pace there is need for skilled personnel. Our research work can help job seekers in better preparing for the software development industry. It will be possible for anyone interested in software development to find appropriate jobs for appropriate skills. For software development industry to flourish it will need a significant work force all of whom may not have any institutional qualification. So, they do not have any good way of knowing about the requirements of the industry. Our work will have a huge positive impact for these individuals as well as others.

5.2 Impact on Environment:

Our work does not have any negative impact on the environment. Some positives can be derived through. As this will help in employment of people it will help in GDP of the country which will help in revenue. That it can be used for improving the environment.

5.3 Ethical Aspects

The data used in this study is publicly available data. The web application will be free for use by anyone. It will be helpful for anyone without any condition and discrimination. So the research work does not violate any ethical aspects, on the contrary it inspires others to help everyone.

5.4 Sustainability Plan

The web application is relatively lightweight so maintaining it will be quite easy. It will be able to handle decent web traffic and the chance of a huge number of people trying to access the web site is slim. So, the web allocation can perform well on any decent server. But there is no functionality added for the model to learn to add this new work has to be done. Adding other functionality will mean better optimization will be needed. For the future there would be a need for funding.

CHAPTER 6

Conclusion and Future Work

6.1 Summary of the Study

Our project is about helping bridge the gap between software industry requirements and knowledge of the jobseekers. To do so we made a web application with a machine learning model that will help job seekers to make better decisions. The model was created with a dataset of 708 job ads of 427 different companies. We have completed the project in 5 months. The entire summary of the work is given below in steps.

Step 1: Data collection
Step 2: Data cleaning
Step 3: Data Labeling
Step 4: Data format conversion
Step 5: Data summarization
Step 6: Using machine learning classifiers
Step 7: Training Model
Step 8: Implementing the model in web application
Step 9: Reporting result

Our dataset will help other people in the research work of the software industry. Finding from the study can be used by academic institutions. The web application can be used by anyone to find appropriate jobs according to their skills. In the following section conclusion and future work of this research will be discussed..

6.2 Conclusion

The primary objective of our research is to help jobseekers better prepare for the software development industry of Bangladesh. In order to do that we build a web application using a machine learning model that will help them in their prospect. The model running in the backend of the application was built from a dataset. Information from the dataset will also help academic institutions to better design their course offering according to the requirements of the software industry of Bangladesh. Making the first dataset of required skills of the software development industry is an achievement. But out biggest accomplishment is making an application which anyone can use to get a better idea of suitable jobs according to their skills.

6.3 Recommendations

The next logical step in our work is to improve the model performance and to do thar the size of the dataset. The model will perform better if it gets to train on more data. The web application can also be improved. It is currently suggesting the best type of jobs according to the user's skill. But it can be programmed to give suggestions on what more should the user learn. Some recommendations are given below:

- Increasing the size of the dataset.
- Increase the number of labels.
- Trying Neural Network classifier in bigger dataset.
- Making the model better.
- Increasing the functionality of the web application.

6.4 Implication for Further Study

There are some things that could be improved to make the work much better. Increasing the size of the dataset will be the first step. A bigger dataset can better train the model to give better predictions. Tweaking the algorithm of the model can also make a better model. The functionality of the web application can be increased by making automatic suggestions available for the user. The suggestion will be on which skills users add to their inventory to better prepare for their desired job. This research work is only for the software development industry but helping other job seekers can also be possible by making dataset and models for other kinds of jobs. Our dataset and implementation has shown what is possible now it needs to grow and branch out more.

REFERENCES

[1]Biswas, Al Amin, Anup Majumder, Md Jueal Mia, Rabeya Basri, and Md Sabab Zulfiker. "Career Prediction with Analysis of Influential Factors Using Data Mining in the Context of Bangladesh." In Proceedings of International Conference on Trends in Computational and Cognitive Engineering, pp. 441-451. Springer, Singapore, 2021.

[2]Arafath, Md Yeasin, Mohd Saifuzzaman, Sumaiya Ahmed, and Syed Akhter Hossain. "Predicting career using data mining." In 2018 International Conference on Computing, Power and Communication Technologies (GUCON), pp. 889-894. IEEE, 2018.

[3]Sripath Roy, K., K. Roopkanth, V. Uday Teja, V. Bhavana, and J. Priyanka. "Student career prediction using advanced machine learning techniques." Int J Eng Technol 7, no. 2.20 (2018): 26.

[4]Al-Saiyd, Nedhal A., and Amjad S. Al-Takrouri. "Prediction of IT jobs using neural network technique." learning 3 (2015): 4.

[5]Gorad, Nikita, Ishani Zalte, Aishwarya Nandi, and Deepali Nayak. "Career counselling using data mining." Int. J. Eng. Sci. Comput.(IJESC) 7, no. 4 (2017): 10271-10274.

[6]Panda, Subhalaxmi, Priyadarshini Adyasha Pattanaik, and Tripti Swarnkar. "A Higher Education Predictive Model Using Data Mining Techniques." In DIAS/EDUDM@ ISEC. 2017.

[7]Moreno, Ana M., Maria-Isabel Sanchez-Segura, Fuensanta Medina-Dominguez, and Laura Carvajal. "Balancing software engineering education and industrial needs." *Journal of systems and software* 85, no. 7 (2012): 1607-1620.

[8]Akman, Ibrahim, and Cigdem Turhan. "Investigation of employers' performance expectations for new IT graduates in individual and team work settings for software development." *Information Technology & People* (2018).

[9] Hiranrat, Chamikorn, and Atichart Harncharnchai. "Using text mining to discover skills demanded in software development jobs in thailand." In *Proceedings of the 2nd International Conference on Education and Multimedia Technology*, pp. 112-116. 2018.

[10] Aken, Andrew, Chuck Litecky, Altaf Ahmad, and Jim Nelson. "Mining for computing jobs." *IEEE software* 27, no. 1 (2009): 78-85.

[11] Bangladesh IT and ITES industry overview by BASIS, available at <<https://basis.org.bd/public/files/publication/5e123f2d5a6c6ba96136a3b168568073f9800e5b0f5b9.pdf>> , last accessed on 20-03-2021 at 07:00 PM.

[12] Giray, Görkem. "An Analysis of Desired Skills for Software Development Jobs in Turkey Using Text Mining."

[13]Matturro, Gerardo. "Soft skills in software engineering: A study of its demand by software companies in Uruguay." In *2013 6th international workshop on cooperative and human aspects of software engineering (CHASE)*, pp. 133-136. IEEE, 2013.

[14] Ahmed, Faheem, Luiz Fernando Capretz, Salah Bouktif, and Piers Campbell. "Soft skills requirements in software development jobs: A cross-cultural empirical study." *Journal of systems and information technology* (2012).

[15] Bailey, Janet L., and Greg Stefaniak. "Industry perceptions of the knowledge, skills, and abilities needed by computer programmers." In *Proceedings of the 2001 ACM SIGCPR conference on Computer personnel research*, pp. 93-99. 2001.

[16] Papoutsoglou, Maria, Nikolaos Mittas, and Lefteris Angelis. "Mining people analytics from stackoverflow job advertisements." In *2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp. 108-115. IEEE, 2017.

[17]Katore, Lokesh S., Bhakti S. Ratnaparkhi, and Jayant S. Umale. "Novel professional career prediction and recommendation method for individual through analytics on personal traits using C4. 5 algorithm." In 2015 Global Conference on Communication Technologies (GCCT), pp. 503-506. IEEE, 2015.

[18]Lou, Yu, Ran Ren, and Yiyang Zhao. "A machine learning approach for future career planning." Standford University (2010).

[19] One Hot Encoding of datasets in Python , available at << https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python//>>, last accessed on 22-04-2021 at 08:00 PM

[20] Save and Load Machine Learning Models in Python with scikit-learn , available at <<https://machinelearningmastery.com/save-load-machine-learning-models-python-scikit-learn/>>, last accessed on 24-04-2021 at 09:00 PM

.

Skill_Evaluatin_and_Carrer_Mapping_Spring_2021.pdf

ORIGINALITY REPORT

10%	9%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Daffodil International University Student Paper	3%
3	openaccess.iyte.edu.tr Internet Source	1%
4	mdp-toolkit.github.io Internet Source	<1%
5	repository.nwu.ac.za Internet Source	<1%
6	Submitted to East Delta university Student Paper	<1%
7	Submitted to Liverpool John Moores University Student Paper	<1%
8	Gerardo Matturro. "Soft skills in software engineering: A study of its demand by software companies in Uruguay", 2013 6th International Workshop on Cooperative and	<1%

	Human Aspects of Software Engineering (CHASE), 2013 Publication	
9	www.ijcai.org Internet Source	<1 %
10	"Proceedings of International Conference on Trends in Computational and Cognitive Engineering", Springer Science and Business Media LLC, 2021 Publication	<1 %
11	Limin Shen, Maosheng Pan, Linlin Liu, Dianlong You, Feng Li, Zhen Chen. "Contexts Enhance Accuracy: On Modeling Context Aware Deep Factorization Machine for Web API QoS Prediction", IEEE Access, 2020 Publication	<1 %
12	www.coursehero.com Internet Source	<1 %
13	Md. Yeasin Arafath, Mohd. Saifuzzaman, Sumaiya Ahmed, Syed Akhter Hossain. "Predicting Career Using Data Mining", 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018 Publication	<1 %
14	cmuir.cmu.ac.th Internet Source	<1 %

15	Hamza Salem, Manuel Mazzara. "ML-based Telegram bot for real estate price prediction", Journal of Physics: Conference Series, 2020 Publication	<1 %
16	theses.gla.ac.uk Internet Source	<1 %
17	pt.scribd.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		