

# Mapping User Skills to Predict Suitable Job Domains Using Data Mining Techniques

Atharva Dhanwate<sup>1</sup>, Hariharan Iyer<sup>2</sup>, Kunal Chaudhary<sup>3</sup>, Ditty Varghese<sup>4</sup>

Department of Computer Engineering, D. B.I.T, Mumbai, India.atharvadhanvate@gmail.com<sup>1</sup>, hariharanr1000@gmail.com<sup>2</sup>, chaudharykunal612@gmail.com<sup>3</sup>, ditty.dbit@dbclmumbai.org<sup>4</sup>

**Abstract**—To develop an ingenious web application with the aim of solving challenges that concern predicting the most optimum and suitable job profile for a user. Job choosing is one of the most important decisions that every student has to take once they graduate. This project concentrates on making this decision making process more efficient and accurate for all the IT graduates so that they can opt for the best suitable domains available in the present IT market.

This project provides an optimum solution by considering various skill sets of the user viz. resume, aptitude, soft skills, and other technical skills. With the help of data mining techniques, all these skills are taken into consideration and a final result consisting of the top three most suitable domains is displayed to the user.

Also, the project helps the user to develop skills related to the suggested job domains by showcasing different skills and their importance in the present IT world.

**Keywords**—deep learning, Keras, skill computation, database, complexity computing, analysis and prediction.

## I. INTRODUCTION

A student studies several different courses in a span of 4 years and it is no surprise that this leaves the student confused in which field he should go forward to make his carrier. To solve this dilemma and to provide students with the most probable field that they would perform well in is our goal. One of the major reason behind unemployment of engineering graduates is due to the lack of knowledge on various job categories as well as an unclear understanding of skill requirements associated with those job categories, in addition, the unemployment issue is highly related to the graduates' performance at job.[5] Hence, the need of a prediction based web application to determine and evaluate the user's skill set and knowledge, by considering various factors such as academic grades, technical skills, soft skills along with some other attributes and their mapping into job domains using various machine learning and data mining techniques to predict the most suitable job domains for the user is seen.

This project concentrates on understanding the skill sets that the user has and helps the user to identify the best suitable job domain with the help of algorithms and other computations done by conducting various assessments necessary to test the user intelligence.

With professional knowledge of subjects and skills the system can be used by other organizations in terms of their work and ethics. This project will also open up a proper work environment where in students can give the test and get a proper knowledge of all the skills he/she possesses

and the job he can suitably plan for to target during placement. The system can work for large number of users with limited data and accurately predict the results pertaining to the best optimum job domains suitable for the user with respect to the present IT market.

This can even help the students to develop on the areas in which he is weak. The technical and other related assessments questions will be kept at par with the industry standards and hence will provide an efficient overview of the skills of the user and hence help in predicting the preferred job domain for the user.

## II. LITERATURE REVIEW

P1. Predicting Employability Skills among Information Technology Graduates of Philippine State University in their On-the-Job Training using J48 Algorithm: This paper talks about the employability skills which are build-up by the graduates of Leyte Normal University in the field of Information Technology, Tacloban City. In this paper data mining has been used to predict employability skills of the graduates especially on the J48 algorithm a C4.5 decision tree model. The result of this study can be a basis for policy measures for effective OJT program thereby focusing skills and competencies necessary for employability. Through the paper we can understand that the administration has been inspired to design a software that will help in decision making and strategic management of the university.

P2. Students' Employability Prediction Model through

Data Mining To predict the employability of Master of Computer Applications (MCA) students. Various algorithms used were Decision tree J48, Bayesian classifiers Naïve Bayes, Support Vector Machines algorithm SMO and Ensemble Methods Random Tree and Random Forest and Multi-layer Perceptrons were used on the data set using 10-fold cross validation. It considers only MCA students and aspects such as the salary package and the company of a person isn't taken into account.

P3. Performance Analysis and Prediction in Educational Data Mining, A Research Travelogue is a Comparative analysis on various papers which can be broadly classified as: Predicting Academic Performance with Pre/Post Enrollment Factors. Comparison of Data Mining Techniques in predicting academic performance Correlation among Pre/Post Enrollment Factors and Employability, Other areas of Education. They Identified common attributes used across various studies and Mentioned the lack of psychometric tests in most of the surveys and Highlighted the importance of the same. The methodologies used were Statistical and Clustering Processes, Naive Bayes Algorithm, Decision Tree, Bayesian Network, Cluster Analysis, Genetic Algo, K-means, Fuzzy, Classifiers, Random Tree, Random Forest.

P4. Student Performance Prediction Model based on Supervised Machine Learning Algorithms: This paper discussed the effectiveness of supervised machine learning algorithms in students' success prediction and academic performance in higher education. Different supervised machine learning algorithms were applied, and the performance criteria was evaluated. Logistic Regression emerged as the winner among various classification algorithms. One question this paper might leave in the mind of the reader is how to measure the grades of a student in the present lockdown situation and what other factors can be considered to evaluate the students' performance.

P5. Assessing Employ-ability of Students using Data Mining Techniques: this paper analyses student performance based on marks of general assessment test and suggest companies accordingly. The method used here involves filtering students based on minimum criteria set by company, Predict Employ-ability using Data Mining Algorithm, Analyse student strengths and weaknesses and suggest list of eligible companies, but in this solution the gap is that there are High model building time for certain algorithms, difficulty in company data gathering. Here, the Major focus is on the psychometric part with very little focus on the academic performance of the student. Does not mention domain specific requirements.

P6. Student Performance Analysis And Prediction Of Employable Domains Using Machine Learning addresses the problem of Lack of existing systems to map university curriculum subjects with job specific requirements. The proposed solution involves a web Application that computes

skill sets of user based on a series of questions and maps them into predefined clusters of subjects to suggest job domain. It involves Use of Data mining techniques to create clusters of subjects similar and mapping of user's skill sets to suggest best possible job roles to them. But this paper Emphasizes only on test performance with no consideration of academic performance, psychometric and soft skill attributes.

P7. Employ-ability prediction: a survey of current approaches, research challenges and applications is a Comparative analysis on various papers based on: Predicting students' employ ability, Skills mapping, Adjusting curriculum, Foreseeing long term market demand. It Identified common attributes used across various studies and Mentioned the lack of psychometric tests in most of the surveys and Highlighted the importance of the same. It used prediction algorithms such as decision tree, random forest, support vector machine, naive Bayes and nearest neighbor and ML models based on precision, recall, ROC area, root mean square error, root relative square error, and error rate. But does not mention company specific requirements.

P8. Job Prediction: From Deep Neural Network Models to Applications is a paper which concentrates on Determining a suitable job for a person based on their job descriptions as well as choosing the candidates that match the job the employers require. The main objective is to study Job prediction through different deep neural network models. It involves Implementation of four deep neural network models i.e. a single model Text CNN, two combination models and a custom proposed ensemble model, in addition to implementation of two pre-trained word embedding into these models. The fall backs of this paper are The data set contains very limited data with only 10,000 annotated job descriptions and only a few network models were used to study the data.

P9. Student Performance Analysis System (SPAS): The students which are most likely to score bad in specific subjects are categorized. The proposed system offers student performance prediction through the rules generated via data mining technique. It involves Use of data mining technique which classifies the students based on students' grade. It uses prediction technique using the decision tree generated from WEKA is not updated dynamically within the system's source code and the prediction can be offered to other courses as well. It doesn't consider any other factor other than marks for evaluating the students' performance.

P10. A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages explains that the construction process of DOM tree in the current extraction processes increases the time cost depending on the data structure of the DOM Tree. This study proposes a novel approach, namely UzunExt, which extracts content quickly using the string methods and

additional information without creating a DOM Tree. UzunExt, is twofold method which consists of extracting content from a web page and predicting suitable values of the stored additional data. This study only deals with the text of a web page, but this text can be changed with Ajax requests.

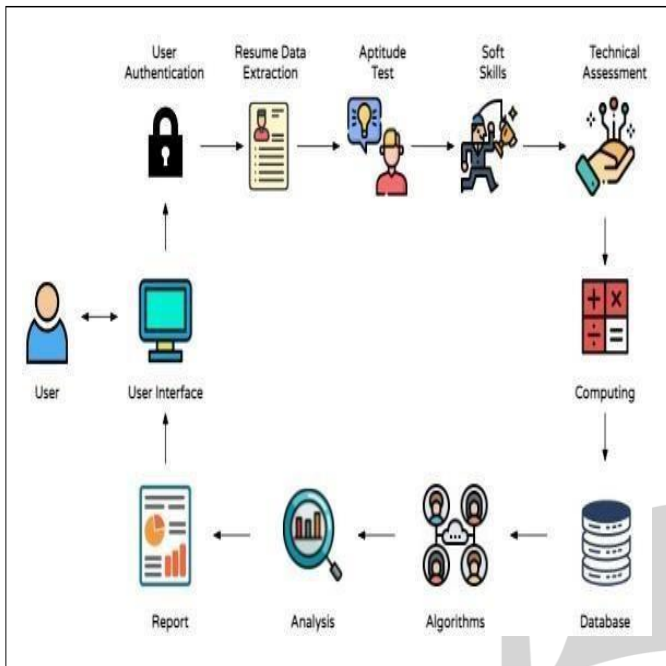


Fig. 1. Block diagram with steps

### III. PROPOSED SYSTEM

Psychology has also been used in this hunt of finding the skills which make a student employable. The below is what we have decided to do:

#### A. System Analysis

Application of data mining techniques to educational data has been essential in providing help to the institutions in various activities.[2] In order to make a web application with the target of solving challenges which occur with predicting the most optimum and suitable job profile for a user.

The proposed system has been designed in such a way that it can be divided into 6 major modules. The working of all these modules will give us the desired/intended output.

Going into the details of these components we first have one of the most important module that is the UI of the system and along with that we also have to keep a user authenticated so these two both constitute of the first module. The second module is getting the skills of the user from his/her resume with the help of scraping. We later use these details in our system. All the assessments can also be included inside one module. Then we move towards the computation involving the database and mapping of the data. The fifth module is the one where the various algorithms is used to give an efficient result for the

user. And the last, one of the most important is the report generation.

Fig.1. shows the various modules and the flow of the system

Let us go in depth of each module to see what it exactly is for:

**M1. User:** The user is a human who will be using the project to give various inputs, wanting to know the predictions for suitable job domains. This user is basically an IT Graduate or any other person who is willing to opt for a job in the near future.

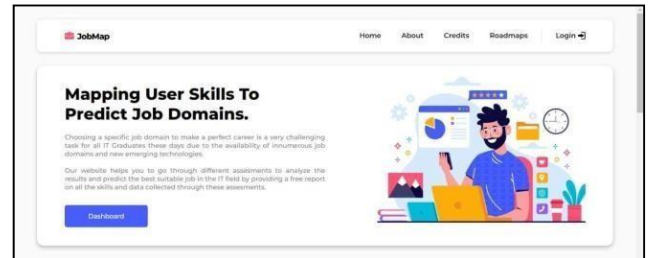


Fig. 2. Home page User Interface

**M2. User Interface:** Its the module that will act as an interface between the user and the system. All the inputs from the user and the computed data from the system will be displayed on the UI. This is a very important module as it is active throughout the process of job prediction. It includes various web pages as follows:

1. Home page
2. Dash board
3. Log in page
4. Sign up page
5. Dashboard page
6. Quiz Page
7. Result Page

**M3. User Authentication:** This module involves all the user authentication processes right from account creation to logging into the website and creating a database of all the users. The highlight of this part is the use of google firebase to create an option for the user to use google sign in if required. Also, we have used MySQL for the DBMS part.

**M4. Resume Data Extraction:** This part of the project involves taking the resume of the user in the form of pdf. This will further be stored in the database and used to scrape out the important skills the user has mentioned in the resume for mapping in the future. There is an option on the Dashboard page through which the user can upload his resume.

**M5. Aptitude Test:** This module involves an aptitude test consisting of various questions which are obtained through web scraping of various questions from indiabix.com. The user can opt this test from the dashboard. It is a non timed test which is in the form of MCQs. The result of the test will be considered while mapping the skills of the users with the data set. Also the results of the test will be shown at the end of the complete assessment process on the results report.

**M6. Soft Skills:** This is another quiz module similar to the aptitude test which is a self assessment test for the user

where he will rate himself on the range of 5 for various softskills like leadership, team work, etc. The user can opt this test from the dashboard. The result of the test will be considered while mapping the skills of the users with the data set.

**M7. Technical Assessment:** The technical assessment involves evaluating the technical skills of the user through a test on various subjects like DBMS, OOPM, DS, etc. The questions for these subjects will be obtained through web scraping different websites particular to that subject. The user can opt this test from the dashboard. It is a non timed test which is in the form of MCQs. The result of the test will be considered while mapping the skills of the users with the data set. Also the results of the test will be shown at the end of the complete assessment process on the results report.

**M8. Computing:** This module involves computing of various scores of the tests conducted in the earlier modules. All the scores are put up in a way and the user skills are then put into a database which will be used to map with the training data set to predict the job domain. This module involves creation of a dataset for training with the help of web scraping. The details of the tables are previously mentioned in the proposed solution. It also involves creation of a proper user database for mapping.

**M9. Algorithms:** This module involves the major work of classifying the job domains with the help of algorithms to predict the best suitable domains. This module involves the analysis of the job domains and further cleansing of the data and putting forward a proper result in a very neat format that can be used to display to the user. It also involves analysis of the jobs that are best suited for the domains that are predicted. These jobs will be further shown in the results module.

**M10. Reports:** The reports page of the website is accessible to the user after attempting all the various test modules. This displays the results of all the different tests taken by the user and also the final 3 best suited domains for the user in a neat format. This page also shows the trending skills for the respective job domains that the system has classified for the user based on the said criteria.

## B. System Design

The system design architecture can be explained as a compound mechanism which is interdependent on various modules for its effective functioning. The front end of the system basically consists of the GUI framework which facilitates the end user into using the system and accessing data. Furthermore, we have the questionnaire part wherein the incorporated test is accessed by the user and subsequent results are computed by the system algorithm. This computed skill level is then correlated and mapped with a database created which houses the various job domains taken into consideration. Hereafter we make use of machine learning and the chosen algorithms to ensure

that correct mapping is done on the data and eventually an analysis report is generated which predicts the job domain. The frontend user consists of uploading resume of the user in order to get the skills from his/her resume. The results obtained are used to suggest domains to work on. A person feedback is the system which is used to get the improvement factor.

## IV. EXPERIMENT RESULTS

Initial step is for the user to register on the website. The user can also login using his mail id, which then marks the start of the session for that particular user.

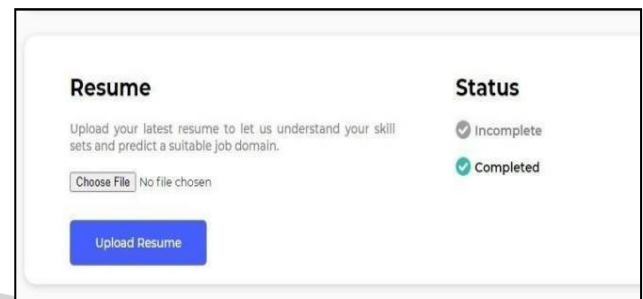


Fig. 3. Resume Upload

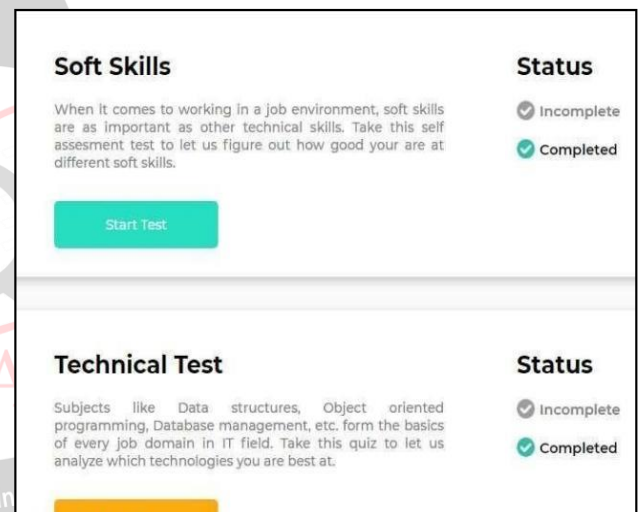


Fig. 4. Assessments

After a user registers and logs in next step is uploading of resume. The resume is stored in the database and a scraper is run on it which extracts the skills from the resume.

The next step is to take the assessments which include aptitude test, soft skills and the most important technical assessment. The scores of these tests are calculated and given as input to the computational model which helps in predicting the job domains with respect to the user's score.

Since the results we are giving to the user have more than one job domains so it is using a multi-classifier and hence one of the major factors for multi-classifiers is the AUC (Area Under Curve) which tells how good the classification for a particular object has been done and in our case the AUC is 95% and hence we can say that this gives quite precise results.

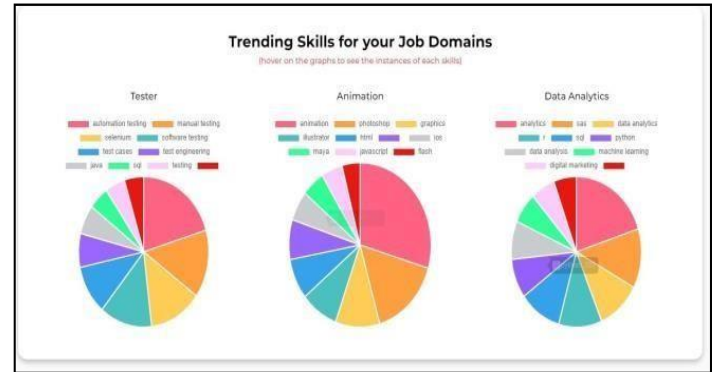
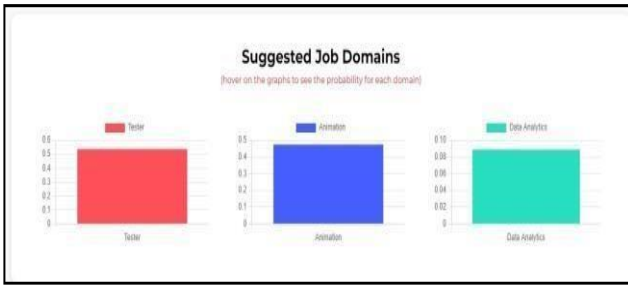


Fig. 7. Sample Output

17610	Data Scientist, Data Analyst, R, Python, Machine Learning	data scientist
17611	Machine Learning, Artificial Intelligence, Statistical Modeling, Big Data	data scientist
17612	Python, Machine Learning, Spark, R, Hadoop, SQL, MySQL, Data Structures	data scientist
17613	Data Scientist, Data Science, Google Analytics, machine learning algorithms	data scientist
17614	Machine Learning, SQL, Python, Spark, SCALA, R, Hadoop, Java, Data Mining	data scientist
17615	C, Opencv, Image Processing, C++, MATLAB, Pattern Recognition, Opengl	data scientist
17616	Machine Learning, Deep Learning, Data Analysis	data scientist
17617	Deep Learning, r, machine learning, data science, regression analysis, time	data scientist
17618	Machine Learning, Customer Segmentation, Big Data	data scientist
17619	C, Machine Learning, Python, Data Science, C++, SQL, Data Mining, Spark	data scientist
17620	Bi, DW, ETL, Big Data, Test Cases, Unit Testing, Data Processing, End User,	data scientist
17621	C, Machine Learning, Python, Data Science, C++, SQL, Data Mining, Spark	data scientist
17622	Machine Learning, Python, R, Spark, Algorithms, Data Analysis, Product Life	data scientist
17623	data science, machine learning, python, r, spark, algorithms, statistics	data scientist
17624	Machine learning, data scientist, data mining, NLP, Python	data scientist
17625	Deep Learning, Data Science, R, Machine Learning, Banking, Python, Big Data	data scientist
17626	spark, aws, ml, Machine Learning	data scientist
17627	data scientist, machine learning, python, analytics specialist, analytics	data scientist
17628	r, web analytics, cluster analysis, sql, google analytics, data analysis	data scientist
17629	Machine Learning, R, Python, Statistical Modeling, Opencv, Data Analytics	data scientist
17630	r, deep learning, linear regression, time series, statistical modeling, nlp	data scientist
17631	deep learning, linear regression, data science, r, machine learning, python	data scientist
17632	Deep Learning, NLP, Java, Text Mining, Python, Spark, Machine Learning	data scientist

Fig. 5. Screenshot of the data-set

A Classification model has been made on the basis of the attributes/skills required and its corresponding job-domain. This model is trained on around 70% of our dataset and is later used for predicting the job domains suitable for a person having an accuracy above 65%.

The testing/training of the model has been done using various algorithms which were : KNN Algorithm, Naive Bayes, Decision Tree. Out of these three the maximum efficiency was given by Naive Bayes Algorithm.

Algorithm	Accuracy	Percentage	
		Train	Test
KNN	0.53212	70	30
Naive Bayes	0.64631	70	30
Decision Tree	0.57565	70	30
KNN	0.54437	80	20
Naive Bayes	0.64922	80	20
Decision Tree	0.57784	80	20
KNN	0.65047	90	10
Naive Bayes	0.65047	90	10
Decision Tree	0.56241	90	10

Fig. 6. Various Algorithms and Accuracy

Fig. 7. shows the sample output for a user after attempting all the different tests and other procedures.

The output presented in the report tab represents the top three domains that are suggested on the basis of all the different tests given by the user. This is in the form of a bar graph which shows the probability of that particular job domain in the range of [0,1], which can be taken as a quantity to measure the specialization of the user in that particular field.

Apart from the the output shows 3 piecharts, each representing the proportions of skills required to have for the top 3 suggested job domains. This data is taken from naukari.com website through web scrapping and the instances of the skills mentioned in the requirements section in the job postings of those particular job domains are taken into consideration.

Also, the scores for technical and aptitude test scores are shown below so that the user can get a fair idea of his knowledge from the scores.

As a part of analysis, the website was shown to more than 30 students aged 20-22 years pursuing engineering and the algorithms proved to be correct most of the times. To improve the accuracy more research with respect to which job domain requires more of which skill is to be done.

## V. CONCLUSION

Job domain selection is a very important part of career decision, we have successfully designed a solution to overcome this problem through prediction technology with the help of data mining techniques.

We also completed studying various literature and papers from trusty resources to understand the different problems and solutions which will help us in the development of our project in the future. We have also put up a proper chart for the same with all the learning and the gaps in those papers.

Hence, we were successfully able to create a website with a rich UI which takes the various tests and required document of the user into consideration to provide the best domains for the respective user.

## REFERENCES

- [1] Rommel L. Verecio, Predicting Employability Skills among Information Technology Graduates of Philippine State University in their On-the- Job Training using J48 Algorithm, Indian Journal of Science and Technology.
- [2] Tripti Mishra, Dharminder Kumar, Sangeeta Gupta, Students' Employability Prediction Model through Data Mining, International Journal of Applied Engineering Research.
- [3] Ali Salah Hashim, Wid Akeel Awadh and Alaa Khalaf Hamoud, Student Performance Prediction Model based on Supervised Machine Learning Algorithms, IOP Conference Series: Materials Science and Engineering.
- [4] Yogesh Bharambe, Nikita More, Manisha Mulchandani, Dr. Radha Shankarmani, Sameer Ganesh Shinde, Assessing Employability of Students using Data Mining Techniques, 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).
- [5] Ankit Patel, Savio Mascarenhas, Akhil Thomas, Ditty Varghese, Student Performance Analysis And Prediction Of Employable Domains Using Machine Learning, International Conference on Recent Advances in Computational Techniques (IC-RACT) 2020.
- [6] Harsh Namdev Bhor, & Dr. Mukesh Kalla. (2022). Analysis of Performance Comparison of Intrusion Detection System Between SVM, Naïve Bayes Model, Random Forest, K-Nearest Neighbor Algorithm. 17(05), 128–144. <https://doi.org/10.5281/zenodo.6601187>, 2022.
- [7] Nesrine Mezhoudi, Rawan Alghamdi, Rim Aljunaid, Gomathi Krichna & Dilek Düşteğör. Employability prediction: a survey of current approaches, research challenges and applications, 2020 RIVF International Conference on Computing and Communication Technologies.
- [8] Tin Van Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, Anh Gia-Tuan Nguyen, Job Prediction: From Deep Neural Network Models to Applications, 2020 RIVF International Conference on Computing and Communication Technologies.
- [9] Chew Li Sa, Dayang Hanani bt. Abang Ibrahim, Emmy Dahliana Hossain, Mohammad bin Hossain, "Student Performance Analysis System (SPAS) The 5th International Conference on Information and Communication Technology for The Muslim World.
- [10] Bhor H N, Kalla M. TRUST-based features for detecting the intruders in the Internet of Things network using deep learning. Computational Intelligence. 2021;1–25. <https://doi.org/10.1111/coin.12473>
- [11] ERDİNÇ UZUN, A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages, IEEE Access ( Vol 8).
- [12] Pooja Thakar, India Anil Mehta ,Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue, IOP Conference Series: Materials Science and Engineering