ANALYSIS OF EFFICIENCY OF CLASSIFICATION AND PREDICTION ALGORITHMS (DECISION TREE) ON SUITABLE JOBS PREDICTION FOR COMPUTER SCIENCE ENGINEERING STUDENTS DATASET

Awinash Kumar

Dept. Computer Applications PES University Bangalore India

Lekha Achuth

Associate Professor Dept. Computer Applications PES University Bangalore India

I.   ABSTRACT

There is a huge spike in number of computer science engineering graduated students who are unemployed or un- deremployed. This is the prime reason why mental health of young aged student (18-24) is also going down. It is also true that world is facing economic crisis. From student point of view, it is very important for them to be well planned and organized from initial stage of their academic. World of information technology is quite unpredictable and dynamic in nature. In present time, there are lot of specific jobs in this field which requires different level and combination of skills. In this research we are trying to analyze all academic details and skills of students at a time and then predict best matching jobs for students. The idea is to make them aware before final semester so that they can plan their study according to it and be hopeful, stress-free and happy in their life even if they are poor in some subjects. For this research data has been collected from reputed engineering colleges of Chennai. Data mining techniques are used to provide the analysis for the classification and prediction algorithms. The algorithms like Decision Tree, Random Forest has been used here which generate most accurate model in which after feeding level of academic details and skills of students is suggesting best matching jobs for them. The core idea behind this research is to visualize the trends of previously placed students and match it with skills of students who are in current academic session and then determine most suitable jobs according to their skills and help them in living stress free life.

Index Terms—Decision Tree. Random Forest. Data mining

II.   INTRODUCTION

Data mining is explained as the practice of deriving the information from huge sets of data. The data must be organized to extract useful information. It is an analytical process used to explore large amounts of data in search of some consistent patterns or common relationships [1]. World of Information technology is quite dynamic in nature. New things keep coming, older ones keep getting replaced. Competition in this field has increased significantly. It has got even worsened due to pandemic(covid) [2]. Reputed engineering colleges are also seeing lower placement in recent years. Moreover, there has been decline in admission of almost 10 percent in engineering courses [3]. Student Placement is also very important for reputation of college. Some research had already shown that the mental health of young aged students is going very down. Most of them are suffering from stress, anxiety, depression [4]. When

questions come how can student lead stress free life. It is found when they are happy and ensured that they have bright future, their performance in academic and ethical standards increases significantly [5]. But current field of information technology has changed a lot. Lot of Counselling research has been done to keep students stress- free life and helps in finding most suitable jobs but no one has yet done research which can help engineering students to find what is most suitable jobs according to their skills in current scenario. There are lot of different jobs like network engineer, software developer, network manager, technology analyst, Data Analyst, Full stack developer which requires specific level of knowledge. For example, if student apply for role of software developer, there is a need of excellent level of Data structure and algorithm, LINUX, Networking, there may be intermediate level of Python, SQL, Machine learning be OK for recruiters but however if you apply for role of Data analyst or Data scientist, then there is a need of excellent level of Machine learning, Artificial intelligence. Python, SQL. There it may happen that intermediate level of Data structure and algorithm would be OK for recruiters. So, for current placement scenario we need to analyze as many skills at a time. In this research we aimed to build an app which

can analyze all academic details and skills of student at a time and then predict best matching jobs for students. For doing this research data has been collected from reputed engineering college of Chennai and with the help of classification algorithms like Decision Tree and Random Forest, we are generating most accurate model which when given level of skills as input predict best matching jobs for students. We are using the concepts of Educational Data Mining for this research. We hypothesized that if students would know their best matching jobs at earlier stage, they would be happy and would be working continuously for achieving that job, and if they would be engaged and happy, they would lead a stress-free life.

A.   CLASSIFICATION ALGORITHM

Classification is one of the most frequently used technique used in data mining. Classification is the process of identifying a set of functions which explains and separates the data classes or concepts. The derived functions are based on the analysis of a set of training data in which the class value is known.[6]. The   main aim of classification is to precisely predict the target class for each case in the data. For example, a classification model can be used to identify loan applicants as low, medium, high credit risks.

B.   PREDICTION

Prediction analysis is used in data mining to derive the information from the data and is used to predict trends and behavior patterns. It helps in identifying a missing value or unknown value. It may inform what might happen in future with reliability which is acceptable.[7]

Section III is the literature survey where the description of domain (Educational Data Mining), Algorithm used in this paper are specified. Section IV is experimental setup which has description of dataset, the attributes that are present in dataset

, algorithm used and technology used is specified. Section V is about methodology which describes the process and analysis of algorithm. Section VI is conclusion which describes end result or what we have observed by using algorithm.

III.   LITERATURE SURVEY

If we carefully analysis our paper we are using the concepts of Educational Data Mining. It refers to tools, techniques and research designed for extracting meaning from large repositories of data generated by or related to student's learning activities in educational settings.

Hence, it describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educa- tional universities and tutorials.[8] For example, several stu- dent learning management system track information such as when each student accessed each learning object, how many times they accessed it and for how much time they use the material. The precision of such data is that even fairly short session with computer-based-learning environment may produce a large amount of data for analysis. These analysis provide new information that would have been difficult to discern by looking at raw data which is very crucial for student performance analysis point of view. By doing such kind of analysis, we can easily make out differences between students. For instance, even if they score same marks, we can characterize them according various attributes like how much time they take to solve each question, how much negative answers of questions they have marked. Furthermore, analysis can be done on basis of what kinds of questions they are comfort with whether one of self-learning, numerical ability or memory-based concepts. By doing such kind of analysis we are able to analyze student comforts and difficulties with each kind of subjects. Earlier the concept of Educational Data Mining was only cornered around Mathematics and Statistics tools but with the introduction of Machine learning has completely revolutionized this field. Machine learning is a subfield of artificial intelligence, involving self-learning algorithms that derive knowledge from datasets in order to make prediction. Instead of requiring humans to manually derive rules and build models from analyzing large amount of data, machine learning offers a more efficient alternatives for capturing the knowledge in order to gradually improve the performance of predictive models and make data driven decisions.

In this paper we aimed at finding best matching jobs for students by analyzing levels of knowledge in all academic fields at a time according to present scenario. If we ana- lyze our problem, Problem is like if student is good in these subjects, intermediate in these subjects, very good in these subjects, poor in these subjects then what kind of jobs suits them most. Such kind of problems can be best solved by classification algorithms like Decision Tree and Random Forest.

In this paper, we are also focused with keeping students stress free. Lot of cases of suicide among students due to stress, depression had been reported
. We assume that if student would know their best matching jobs at earlier stage, even they have certain downfalls in life, they would be having patience and hope in their life. Having hope and patience is one of the best ways to keep yourself stress free.[5]
The reason for going classification algorithm like Random Forest, Decision tree is that we have many different values in class label of dataset. Previous such counselling works has been done using logistic regression, SVM but the problem was that using SVM, logistic regression individual accuracy rate was poor, this could have been done better by using ensemble learning techniques.[9] But the problem is that for increasing better performance of model we have to increase learning features to capture more information as the features used in dataset are very less correlated. The care we have to take while increasing learning features it that model should not be complex. For situation like this where learning features has to be increased and at same time over-fitting of model has to be avoided, Random Forest is best for such situation. But the problem with random forest is that it combines only decision tree model. That is why the use of Decision tree model, Random Forest is justified.

A.   DATASET

IV.   EXPERIMENTAL SETUP

investigate techniques to identify and store evidence from computer science from poor to average to excellent or even professionals.

Engineering Student Career dataset is in CSV (Comma Separated Value) format. This dataset was available at Kaggle in which data has been collected from successful placed student from reputed engineering colleges of Chennai. There are seventeen attributes for this dataset including the class values. The values of attributes are not interested, good, average, intermediate, poor, excellent, professional. The value of class values are various jobs in information technology field like   Database   manager, Database   engineer, Database administrator, Network engineer, Software developer, Data analyst, Data scientist, Business analyst, Graphic developer, Technical writer, Hardware engineer, Full stack developer. There is total 20000 samples   taken for study.

B.   ATTRIBUTES DESCRIPTION

1   Database Fundamentals - Students may vary in level of knowledge in Database, SQL from poor to average to excellent or even to professional level.

2   Computer Architecture - Students may vary in knowledge of hardware, operations as well as several components(organization) of computer system.

3   Cloud Computing - Students may vary in knowledge of distributed computing, cloud architecture from poor to average to excellent or even professional level.

4   Cybersecurity - Students may vary in knowledge of Cybersecurity, ethical hacking from poor to average to excellent or even professional level.

5   Computer Networking - Students may vary in knowledge of Computer network fundamentals, network programming, IoT from poor to average to excellent or even professional level.

6   Software Development - Students may vary in level of knowledge of Data structure and algorithm from poor to average to excellent or even professional.

7   Programming Skills - Each student may show more interest in particular programming language. It may happen that one may be strong in that language, but weak in other programming language.

8   Project Management - Students may vary in level of knowledge regarding concept of project evaluation and hence in project management skills from poor to average to excellent or even professional

9   Computer Forensic Fundamentals - Students may vary in level of knowledge of computer forensics that uses

10   Technical Communication skills - Some students may be particular good in technical communication; hence the post of technical writer can also be designated to them while someone will not be interested.

11   Artificial intelligence - Some students if they show interest in this subject may grab very good placement while many students may be poor or average in this subject but however this is very important subject according to present scene.

12   Software Engineering - Some students think this as complete theoretical subjects and hence show no interest; however, concepts of this subjects may be very important while monitoring software development process.

13   Data Science - This is other vast and very important subject which include concepts of Excel, SQL, Data analysis, Machine learning. If student is showing good interest, he has chance of huge success.

14   Communication skills - Although students may not be bothered about these skills, but student having good knowledge but poor communication skills may not form good image in eye of recruiters.

15   Graphic designing - Well if someone is interested in game development domain, he should show good interest in that subject and this field has attractive market while for someone this may not be interesting at all.

16   Aptitude and reasoning - This is first level of exam in almost entrance. The idea behind this paper is to check student logical ability and problem-solving capacity. Since world of information technology is quite dynamic, there is a high chance that students have to keep learning new tools. If someone having very good technical knowledge but with poor aptitude may not move to second round.

17   Web technologies - This is also new and important subject, the demand of full stack developer or certain web technologies is very high, however it is also found that some students may not show interest in this subject at all.

C.   DECISION TREE ALGORITHM

Decision Tree is one of the most important supervised classification algorithms. It is attractive because of its inter- pretability. Based on features in training set the decision model learn series of questions to infer the class label of samples. It is if-else like statement in which if statement is executed then else block will not be executed. It means we need not to test all variables if we got class label. It makes tree like representation in which attribute represent node, edge represent value in that attribute, leaf represent class label. This increases performance of model and hence consume less time. It works very best for classification problems.

Step1 First we should clearly define learning attribute and class label.

Step2 Among all attributes the best among them is chosen for split. It is chosen by finding one having more information gain by measuring entropy or Gini gain. Lower entropy or lower Gini -gain means more purity, hence high information gain.

Step3 We should check that all instance of these attributes belongs to same class if yes if we should stop constructing tree. This is stopping criterion.

Step4 While stopping criterion is not satisfied, again select best attribute based on gini index and start constructing tree.

D.   RANDOM FOREST ALGORITHM

Random Forest is one of ensemble learning technique which combines good features of one and many model and produces best accurate model. However Random Forest combines only decision tree and produces a meta model. It uses bagging technique and bootstrapping technique. Bagging technique is used to produce meta model from same kind of model. The advantage with random forest is that it allows for full development of tree, it means more

information because there is no pruning concept and at time it avoids overfitting of model. It can be used for both classification and regression problem. Random forest constructs de-correlated trees it means each of tree is independent of each other.

Step1 Assume number of cases in training set is N. Then samples of these N cases are taken at random but with replacement. This sample will be training set for growing the tree. The technique used is bootstrap technique. Splitting of dataset in both horizontal or vertical way is possible and then generate decision tree from these datasets and then feed it to random forest

Step2 If there are M input variables, a number m<M is specified such that each node, m variables are selected at random out of M. The best split on these m is used to split the node. The value of m is held constantly while we grow forest.

step3 Aggregating the result of generated from these models. Well known method such as voting or averaging is used. Since we have classification problem voting is used.

E.   PROPOSED SOLUTION
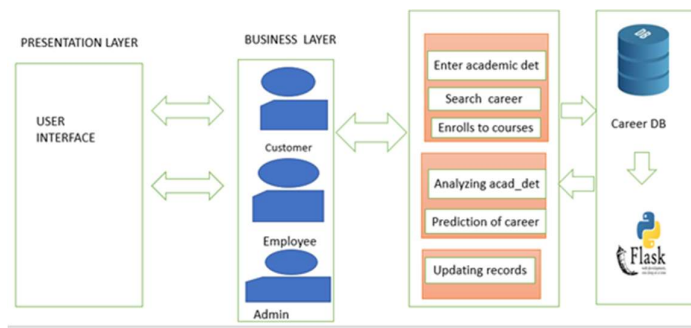
Students, Employee, Admin are the one around whom this app will be revolving. All of them will be able to access app through web browser with active internet connection. For student to receive their best matching job, they need to register themselves, At the time of registration they have to enter the level of knowledge of academic field. Their academic details will get stored in database and will be identified through their unique email-id and password. Now employee after giving their successful login-id can see the academic details of students and then will visualize it and grade it accordingly and using in built model he feed to model and then predict best matching jobs for students. After this employee has responsibility of sending student best matching jobs to admin by filling a form. This will also be stored in database and student best matching jobs will be identified with help of their unique email id. After that admin has responsibility of sending these job details to students on their e-mail id. In building this application strong concept of database is required. At certain point some validation criterion has also been considered. For example, at the time of registration students should fill all the question, no question should be left. With valid e-mail id, only one student is allowed for registration

V.   METHODOLOGY
A.   INPUT

The data used for this analysis is taken from Kaggle where data has been collected from successful placed student of reputed engineering college of Chennai. Data collected shows the details of level of students in each subject whether student is good, poor, intermediate, excellent

in many subjects as possible. If someone has not chosen that subject because of choice of optional, it has been assumed that student is not interested in this paper. However, in this dataset, data from personal traits like number of hours of study, number of hours you watch TV, number of hours you play games, whether you drink alcohol, travel a lot, how much time you offer pray

[8] were excluded assuming students to behave according to ethical standards. Total such 20000 samples were recorded in dataset. The algorithm accept file only in CSV format. The user need is to load the file for execution.

B.   PROCESS
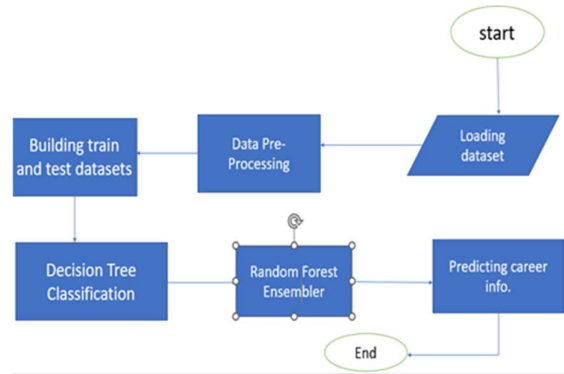
First CSV file has been loaded with the help of read_csv   . For this first pandas library has been imported.
 Pandas core datatype was Data Frame.
For initial inspection we have mostly used three methods
df. head () - this method is used to see first five rows
df. tail ()- this method is used to see last five rows.
df.info () - this method is used to list column name with datatype

Moving on Data pre-processing steps, it includes cleaning data and making it into perfect shape for feeding into model. For this first it checked that whether data is clean or not by is null () method. Since dataset has no null or missing values. Then we move into second step of data pre-processing where categorical data is converted into numerical one by label encoder. Since in this dataset there is all categorical values like good, intermediate, poor, excellent. Label encoder has been used to convert all these categorical data into numerical ones. Then we move into final step of pre-processing where data is shaped into common scale by using MinmaxScaler. In this data gets scaled to a fixed range usually 0 to 1.

Then we move into model building part where we first splitted our dataset into 80:20 ratio. It means our train dataset is 80 percent of original dataset and test size is 20 percent.

Training Dataset: The sample of data used to fit the model. Test Dataset: The sample of data used to provide an unbiased evolution of a final model fit on training dataset.

For this train test split method has been imported from sklearn. model selection, there test size=0.2, shuffle=True and random state equal to 1 is chosen to ensure that all class labels have been equally selected. All academic details of student have been kept as learning features and job roles as target features

.

Finally, we move into model deployment part where we have used Decision Tree and Random Forest

For building decision tree, decision tree classifier is im- ported from sklearn. Tree. Hyperparameter like criterion = Gini, min-samples-split, min-samples-leaf, min-impurity-split has been used. After building various decision tree, finally random forest has been used as meta model which com- bines best features of all decision tree and produce best accurate model. For this random forest classifier has been imported from sklearn. Ensemble. Some hyperparameters

Like n-estimators, with replacement equal to False has been used. Now once meta model is ready, we are predicting best matching jobs for students by predict method. For evaluation of model some parameter like accuracy, confusion matrix has been imported from sklearn. Metrics. For better analyzing of tree plot-tree method has been used.

C.   RESULT

First of all, after loading data set it is found that 20000 samples were successfully loaded. After checking its dimension, it is found that it is 20000 X 23 it means it has 20,000 rows and 23 columns. All column were of string datatype

Moving to result of data preprocessing steps, first after checking the null values it is found that there is no null values or missing values that means dataset is pure. After using label encoder all categorical values change into numerical ones. After using minmax scaler all values of each class get converted to common scale that between 0 to 1 and are precisely sex digit after decimal points.
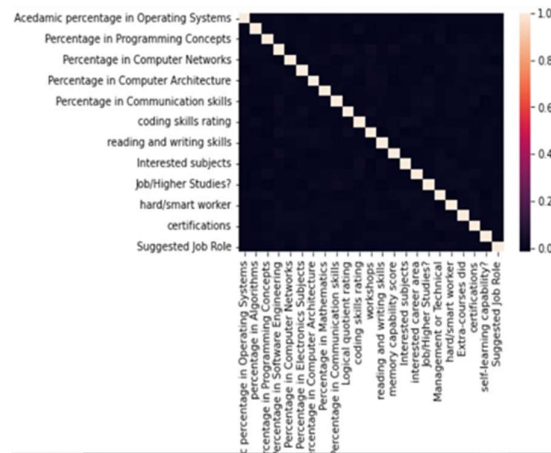
After drawing heatmap it is found that no features are strongly correlated with each other. Once during model build up test size is found to be 4000 which is 20 percent of dataset, train size is found to be 16000.

Once after model deployment using Random Forest when certain inputs were given, output was produced like

D.   CASE ANALYSIS

Case1:

Number of attributes taken as input: 5

 Success rate: 49 percent

Error rate: 51 percent Confusion matrix generated:

```
array([[2318,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0, 2333,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0,    1, 1696,    0,    0,    0,    0,    0,    0,    0],
       [   0,    1,    0, 1691,    0,    0,    0,    0,    0,    0],
       [   0,    0,    1,    0, 2261,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0, 2310,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0, 2240,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0, 1697,    0,    0],
       [   0,    0,    1,    0,    0,    0,    1,    0, 1684,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0, 1765]],
      dtype=int64)
```

Case2:

Number of attributes taken as input: 7 Success rate: 55 percent

Error rate: 45 percent Confusion matrix generated:

```
array([[2318,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0, 2333,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0, 1697,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0, 1692,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0, 2262,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0, 2310,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0, 2240,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0, 1697,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0, 1686,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0, 1765]],
      dtype=int64)
```

Case 3:

Number of attributes taken as input: 11 Success rate: 66 percent

Error rate: 34 percent Confusion matrix generated:

```
array([[2318,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0, 2333,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0, 1697,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0, 1692,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0, 2262,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0, 2310,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0, 2240,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0, 1697,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0, 1686,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0, 1765]],
      dtype=int64)
```

Case 4:

Number of attributes taken as input: 15 Success rate: 88 percent

Error rate: 12 percent Confusion matrix generated:

```
array([[2318,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0, 2333,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0, 1697,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0, 1692,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0, 2262,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0, 2310,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0, 2240,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0, 1697,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0, 1686,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0, 1765]],
      dtype=int64)
```

Case 5:

Number of instances taken as input:17

Success rate: 97 percent

Error rate: 3 percent Confusion matrix generated:

```
array([[2318,    0,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0, 2333,    0,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0, 1697,    0,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0, 1692,    0,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0, 2262,    0,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0, 2310,    0,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0, 2240,    0,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0, 1697,    0,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0, 1686,    0],
       [   0,    0,    0,    0,    0,    0,    0,    0,    0, 1765]],
      dtype=int64)
```

VI.   CONCLUSION

This paper provides analysis of Classification and Predicting data mining techniques. These results are based on the Engineering Student Career Prediction dataset present on Kaggle. The dataset is focused more on academic details and skills of students, rather than personal trait of student to meet visualize current scenario of placement. Random forest shows best result when number of learning features has been increased. The advantage with Random Forest is that it has avoided over fitting of model and helps to capture more information from trees. However individual Decision tree is not producing much accurate result because if we increase our learning features it may happen decision tree model become complex, when learning features are less provided to tree, there also it has not been showing much accurate result because features of this dataset is very weak correlated. This is why finally we have to go for Random Forest. Finally, it has been predicting 97 percent accurate result.

A. FUTURE ENHANCEMENTS

This app is predicting career on basis of academic details entered by student. Furthermore, enhancements can be done if students are made to give test in every subject and fur- thermore concept of Educational data mining is used. This is because whatever students have entered academic details,

app is predicting career on basis of that. We are not sure how much students have understood their level of knowledge in particular subject. This will be better if employee himself will analyze the academic detail and level of skills of student. Further analysis can be done on basis of types of questions which students are entering, like questions based on memory based, self -learning capacity, numerical concepts, how much time is taken to solve each question. Such type of analysis is very helpful considering the vast competition of present time.

REFERENCES
[1]   Data   Mining   techniques, https://documents.software.dell.com/Statistics/Textbook/Data- Mining-Techniques
[2]   Engineering   Student   Placement   Statistics   , https://hindustantimes.com/Placement Statistics/Engineering College

[3]   Admission Survey, HTTPs://indianexpress.com/article/education/ministry/- survey-btech-only-ug-programme

[4]   Student Stress Statistics, https://research.com/education/student-stress- statistics

[5]   Stress Health, https://sensa.health/stress-anxiety

[6]   Classification algorithm, https://towardsdatascience.com/top-machine learning-algorithm-for-classification

[7]   Predictive Modelling Analytics and Machine learning, https://www.sas.com/en-gb/insights/articles/analytics/a-guide- to-predictive-analytics-and-machine-learning

[8]   Student Career and Personality Prediction System   by Prof.Pallavi Bharambe, Ms. Sonali, Nehre, Ms. Manasi Paste, Ms. Nandini Patil, International Journal of Advance Research in Science, Communication and Technology (IJARSCT), volume 2, Issue 1, April 2022

[9]   Student Career Prediction by Vidyapriya C, Vishhnuvardhan. RC, https://ijcrt.org

[10]   Student Career Prediction using Machine learning approaches by N. VidyaShreeram, Dr. A. muthukumaravel, https://eudl.com

[11]   Student Career Prediction Using Advanced Machine Learning Techniques by K. Sripath Roy, K. Roopkanth, V. Uday Teja, V. Bhavana, International Journal of Engineering and Techniques volume 3