

Career Prediction System using Big Data and Data Mining Techniques

Deepak Mokashi¹, Darshan Rahude², Diksha Fegade³, Vaishnavi Dhembre⁴, Mahesh Shinde⁵, Balaji Bodkhe⁶

1-6Department of Computer Engineering, M.E.S. College of Engineering (Wadia), Pune, Maharashtra, India

Submitted: 05-06-2021

Revised: 18-06-2021

Accepted: 20-06-2021

ABSTRACT: Career selection is one of the important and difficult tasks as we all know that there are many different career opportunities and paths available for students. So, the situation of students is quite confusing during the selection of a career. If the student selects the right field of his choice then he can study in that field more precisely. Wrong career selection which he or she does not like may affect the low productivity of a student in that field and also may affect his professional life. This system is a web application that would use different data mining algorithms to help students to select for their appropriate career. The framework will assist students in selecting career opportunities based on their interest and desire to take up the course.

KEYWORDS: Career prediction System; Data mining Algorithm; Decision Tree; C5.0 Algorithm; Questionnaires set.

I. INTRODUCTION:

As various domains are exploring and also research is increasing, a number of career opportunities are available in different domains. This creates more confusion in students to choose the right career option. This may be due to unawareness about the fields that are available, not knowing his personal strength, less exposure, market situation, equal interest in multiple fields etc. There is a risk of choosing the wrong career path due to these confusions, and the effects of this may be anxiety, depression and poor work performance, and it could affect professional life. So, there is a requirement of a system which will help students to make decisions while choosing their career option by taking all the necessary aspects into consideration.

II. LITERATURE SURVEY:

[1] describes Best Career Options for Modern Society students using different analyses. They provide subjective as well as objective evaluation for opting individual career options and

career-based sessions for future growth. Quantitative Research involves the gathering of knowledge on logical and problem solving and the use of numerical methods of analysis to draw evaluation conclusions. It includes tests, structural observations and examinations. Qualitative measures refer to situational tests and behaving recording etc. They measure career acceptance of candidates by combining text algorithms and various item techniques. They have used text classification to provide more assured results and for increasing the efficiency of the system.

[2] discusses the system where they are going to help Final year CS students to select appropriate domains or fields after completing their engineering course. In this work, they are performing different data mining algorithms on dynamic data sets to check student vision based on factors like technical, interpersonal and academics. The motive behind this study is to predict the best career option for ongoing CS students by analysing college Alumni information by taking different parameters into consideration. These parameters mostly underline professional skill, interpersonal skill and academic results for prediction study. Then prediction is done by analysing data using classification algorithms. Fig.1 shows the steps they used to get information from alumni data.

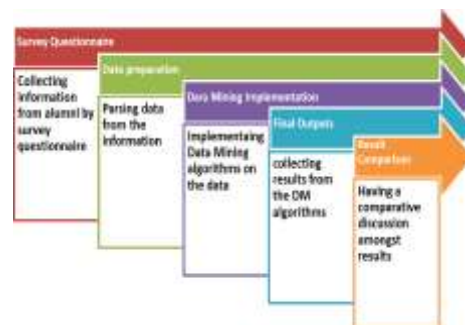


Fig -1: Steps to get information from alumni data.

Five different classification algorithms (ID3, CART, Random Forest, Support Vector

machine, Neural network) used. The most accuracy attained by the CART and Random Forest algorithm (95.04%).

[3] owing to the increasing career paths with diversified opportunities, choosing a suitable career has become the most important and crucial phase in life. So, they proposed that analytics would definitely help the students to select an institution and program or course in accordance with their field of interest, personality traits, and mental ability. Their system is useful for colleges, management and universities to review. They analyse the various aspects that play a keen role in the education sector like counselling performs a major function. This system appends factors like 1. Student population 2. Seat filling by category 3. performance 4. Placements 5. NAAC 6. NBA. They differentiate two or more colleges with respect to performance and student-staff ratio, affiliations and student rating.

[4] The aim of this study is to improve student satisfaction with their course by taking parameters such as academic workload, readiness for academic study, improvement in core technical skills and employability after graduation. It proposes a model called Map My Career. Using text mining and data analytics, they introduced one software application to support students in topic selection. The study suggested what students are studying in universities and how it would be useful for their future. The application provides students with knowledge about their course and curriculum and raises awareness among students about the skills needed and the skills students should have for any specific work.

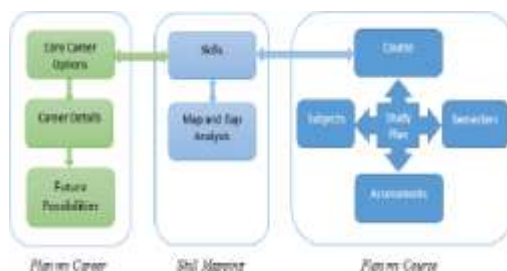


Fig -2: Architecture Diagram of proposed System.

[5] It has been proposed that for proper counselling and for selection of courses in engineering institutions data mining is an effective tool. All further careers of a student only depend on the branch selection of the student while taking admission. If a student chooses any branch because of their parents' pressure or due to their friends'

suggestions, then it is of no use. Based on the student's interest, career goals and skills of students one must have to select a career. This application helps students to select branches as per their skills and interests. Decision trees are used for this classification and regression problems. This includes :1]data collection from professional surveys. 2] This data from the survey is then used for making decision trees for choosing various courses. 3] Then the REP tree classifier algorithm was applied for the classification of students according to their skills and interests. The following diagram shows the information about students, choice of program and student's placement details.

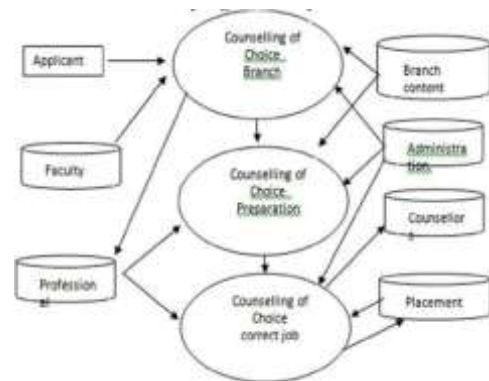


Fig -3: Counselling system.

[6] Where data storage is not feasible and more information is generated in that situation, the incremental learning algorithm is found to be a useful technique to provide students with the best career choices. The incremental learning algorithm is used in the event that the batch classifiers are of no use at that time. For future work, this algorithm uses data of various types such as web log, time series, multimedia purpose, and spatial since this data is like stream data. Data from students in the educational system is growing day by day, so incremental learning properties are essential for machine learning research. Various polling methods are used to merge voting with a weighted majority vote. In addition to these various important laws, there are to find the timeliness between the various mixing laws. The individual classifier and the collection of data used for that.

[7] To inspect course identification questions and address pieces, data mining is used. Classification is one of the most common application domains. The key task in classification is to assign a class mark from a set of possible class values to an unseen instance composed of a set of variables. Which is the most critical variable,

differentiating "satisfactory" and "not satisfactory" tutor outputs based on the appreciation of students. These can be assistive detectors to enhance their efficiency. Questions and responses to course evaluations in higher education institutions are assessed by the efficacy of the detector and, of course, by various proportions. Such results can be used to enhance measuring instruments. There are four different classification techniques: (a) decision tree algorithms, (b) support vector machines, (c) artificial neural networks, (d) discriminant analysis. Data mining correctly distinguishes "satisfactory" and "not satisfactory" detector outputs. And there are seven distinct categorizations used. Via this prediction, the usefulness of using data mining techniques in course assessment shows. In higher education mining, the efficacy and disclosure of data mining techniques, specifically decision tree algorithms, boosting, SVM, ANN, and DAA are provided over a dataset from daily life. For the classifiers, the understanding of the variable significant analysis is used, it is expressed that there are several potential areas of improvement in the nature of the measurement instruments. And a performance appraisal of the teacher was included in it.

[8] Pattern mode and Batch mode are used for fetching the data. In Data mining a well developed technique is used i.e. pattern mode. Instead of batch learning mode, pattern mode learning is preferred. The psychological condition of the students analysed by this system and suggest their career. In that for a career suggestion system and snapping of the node is based on information pick up the C4.5 algorithm is used. The system shows accuracy is up to 86% for the career data. The precision may vary relying on the data provided.



Fig -4: By using the C4.5 algorithm (Career Prediction System).

Tanya V Yadalam et al. [9] in this system, they used flask which is written in python and the forecast career of a student depends on the data and executes similarities between two vectors on that. To manage data in the format essential for similarities, use Pandas and NumPy files and use Pandas which is fast and for manipulation of data effective data frames is used. This system realizes the user area of interest and capabilities that makes relevant career ideas for users.

The structure provides extra capabilities to UG students. Users can lengthen their briefcase. Users can allow their estimation and observation to rely on their experience on various guidelines. For Engineering students this system is made, it can involve complementary streams like business and arts. By assigning codification, the user's profile can be managed in an extra assured way. By using Collaborative approach this system can be developed.

Min Nie et al. [10] It is said that to decide the career of a student is the pivotal role in anyone's life. For solving the post-graduation problem of students, the traditional machine learning method is limited. Based on the student's behaviour and skills, we can decide the career choices of students. For improving the model several insights are offered. For getting a priori information from college a prototypical cluster centre generation approach is used. Examples in the same cluster have the same label as the motivation of cluster assumption. We have introduced a novel regularization item to bridge the gap between the real-world examples and prototypical cluster centres. The results of multiple experiments demonstrate that our approach is superior to other approaches to career choice prediction. In future studies, three directions can be followed with interest. First Cluster Centres can be discovered in a more precise method. Our model can be extended from using only behavioural data to using multimodal data, such as adding school achievement and questionnaire data. It is meaningful to improve our model to not only predict career choices but also advise on career planning, such as advising on the courses required.

K Sripath Roy et al. [11] proposed that we know that students are going through their academics and opting for their interested courses, but it is very important to check ability, problem solving skills etc so they will know in what skillset they are putting themselves. This system will also help the recruiters to evaluate on the basis of different factors and which role is suitable for a particular candidate. And the system will help the candidates to analyse their strengths therefore they will get assurity about their domains. The system implements different types of algorithms like Decision trees for prediction.

Kalpesh Adhatrao et al. [12] proposed a system which can predict the performance of first year students based on the result obtained from students who are currently studying in second year of engineering. They are using classification techniques in data mining for processing. In classification, they are collecting information from

second year students and use that information as a training data set in the first phase. In the second phase they are collecting information like name, merit number, score obtained in board examination as X and XII, Score obtained in entrance examination etc. from first year students. They are using ID3 and C4.5 algorithms to obtain a decision tree on a training data set. They achieved accuracy 75.145% for both ID3 and C4.5 algorithms.

III. ALGORITHMS USED:

Data mining is the process of extracting patterns from data that has been saved or warehoused by an organisation. These designs can range from simple to complex. As a tool, it is used to obtain insight into parts of the organization's operations and to predict future results. We are using the C4.5 Decision tree algorithm for prediction of the best career option.

Decision Tree Algorithm:

Decision tree algorithms fall under the category of supervised learning. They can be used to solve both regression and classification problems. Decision trees use the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. In Decision Tree the major challenge is to identify the attribute for the root node in each level.

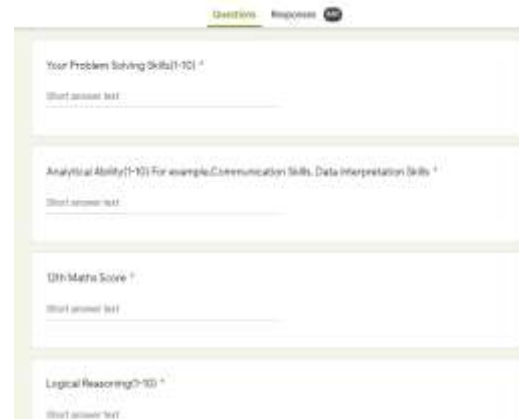
This process is known as attribute selection. We have two popular attribute selection measures:

1. Information Gain
2. Gini-Index

IV. IMPLEMENTATION:

Step 1:

In this step we collect datasets from students who are currently pursuing their career in different fields. For this we made one google form for questionnaires and did an online survey for data collection.



Step 2:

In this step we did some preprocessing on data which we got in an online survey.



Step3:

In this step, for building a decision tree model we used learn, pandas, NumPy like inbuilt libraries from python and import all in Jupyter notebook. We cannot use categorical variables directly for prediction models. So, for categorical variables we used dummy variables to convert into binary variables. After that we built the model on which we got accuracy 76%.

```

from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import Decomposition
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score

# Create the pipeline
encoder = OneHotEncoder(sparse_output=False)
decoder = Decomposition()
model = Pipeline([
    ('encoder', encoder),
    ('decoder', decoder)
])

# Fit the pipeline
model.fit(X_train, y_train)

# Predict on new data
y_pred = model.predict(X_test)

# Print the report
print(classification_report(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))

```



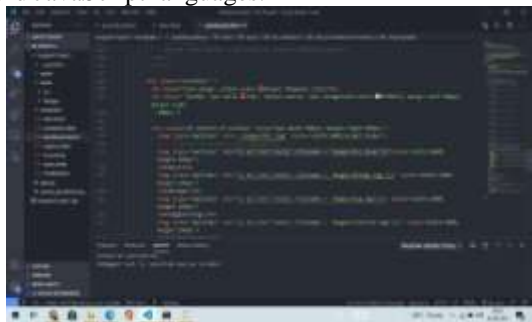
Step4:

We have used Python Flask server for connection of prediction model with frontend UI.



Step 5:

For frontend UI development we used Html, CSS and JavaScript languages.



Step 6: Result of Implementation

The accuracy of the C4.5 algorithm on the dataset was 76 percent. After giving answers to all questions, users can get an appropriate career option on screen.

Congratulations! Your career is predicted ,
Suggested Career Field
Medical



V. CONCLUSION:

The motive behind this study is to help students to select appropriate career options among all other available options. The study discovered that selecting the best career option not only depends on the personality trait of students but also the system will also concentrate on the student's interest. For this system to be used, students have to answer all the questions and based on the answers the system recommends a particular course along with the list of colleges providing those courses and other necessary information about that course. So, it also minimizes students' effort in

collecting information about colleges and other extra tasks.

REFERENCES:

- [1]. Lihui Zang, Xin Fang, Jung Wang on "Assessment of Career Adaptability: Combining Text Mining and Item Response Theory Method" Received July 30, 2019, accepted August 25, 2019, date of publication September 2, 2019, date of current version September 17, 2019.
- [2]. Sarath Tomy, Eric Pardede on "Map My Career: Career Planning Tool to Improve Student Satisfaction" IEEE Access Received July 23, 2019, accepted August 29, 2019.
- [3]. Jinka Thirunarayana department of cse Anantapur, Andhra Pradesh on "Counselling Guidance Using Big Data Analytics" Received 10 April 2018, Accepted 24 April 2018.
- [4]. Md. Yeasin Arafath Mohd. Saifuzzaman Sumaiya Ahmed on "Predicting Career Using Data Mining" 2018 International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018
- [5]. International Journal of Engineering Research Research in Computer Science and Engineering, Vol 5, Issue 4, April 2018 by Voore Subba Rao and Kalva Supriya Reddy.
- [6]. Roshani Ade and P.R.Deshmukh on "Efficient Knowledge Transformation System Using Pair of Classifiers for Prediction of Students Career Choice" , International Conference on Information and Communication Technologies (ICICT 2014).
- [7]. Mustafa Agaoglu on "Predicting Instructor Performance Using Data Mining Techniques in Higher Education" access received May 6, 2016 accepted May 10, 2016.
- [8]. Lokesh S. Katore, Bhakti S. Ratnaparkhi, Dr. Jayant S. Umale on "Novel Professional Career prediction and recommendation method for individuals through analytics on personal Traits using C4.5 Algorithm" Global Conference on Communication Technologies (GCCT 2015)
- [9]. Tanya V Yadalam, Vaishnavi M Gowda, Vanditha Shiva Kumar, Disha Girish, Namratha M on "Career Recommendation Systems using Content based Filtering" International Conference on Communication and electronic Systems (ICCES 2020) IEEE Conference

- [10]. Min Nie ,Zhaohui Xiong,Ruiyang Zhong ,Wei Deng and Guowu Yang on “Career Choice Prediction Based On Campus Big Data- Mining And Potential Behaviour Of College Students” , published on 20 April 2020.
- [11]. K sripath Roy, V Uday Teja, J Priyanka on “Student Career Prediction System” International Journal of Engineering and Technology (2018).
- [12]. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao on ““Predicting students’ performance using ID3 and C4.5 classification algorithms” in International Journal of Data Mining & Knowledge Management Process (IJDMP) Vol.3, No.5, September 2013.