ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 5, Issue 4, April 2017

Career Counselling Using Data Mining

[*]Nikita Gorad[1], Ishani Zalte[2], Aishwarya Nandi[3], Deepali Nayak[4]

[1, 2, 3] Final year Students, [4]Assistant Professor Vidyalankar Institute of Technology, Mumbai, India.

ABSTRACT: Selecting an appropriate career is one of the most important decisions and with the increase in the number of career paths and opportunities, making this decision have become quite difficult for the students. According to the survey conducted by the Council of Scientific and Industrial Research's (CSIR), about 40% of students are confused about their career options. This may lead to wrong career selection and then working in a field which was not meant for them, thus reducing the productivity of human resource. Therefore, it is quite important to take a right decision regarding the career at an appropriate age to prevent the consequences that results due to wrong career selection. This system is a web application that would help students studying in high schools to select a course for their career. The system would recommend the student, a career option based on their personality trait, interest and their capacity to take up the course

KEYWORDS: Career prediction; Data mining; Personality traits; C5.0; Adaptive boosting.

I. INTRODUCTION

With the increase in research and exploration in various domains, there are many new career opportunities in every field. This creates more confusion to the students studying in tenth or twelfth grade to select one career option. The reasons for this confusion could be unawareness of self-talent and self-personality trait, unawareness of the various options available, equal interests in multiple fields, less exposure, market boom, assumed social life, peer-pressure etc. Due to theses confusions, the student may select a wrong career option and the consequences of this wrong decision could be work dissatisfaction, poor performance, anxiety and stress, social disregard etc.
Thus, there should be proper counseling of the student's psychology, interest and their capacity to work in a particular field.

II. RELATED WORK

There are various websites and web apps over the internet which helps students to know their suitable career path. But most of those systems only used personality traits as the only factor to predict the career, which might result in an inconsistent answer. Similarly, there are few sites that suggest career based on only the interests of the students. The systems did not use the capacity of the students to know whether they would be able to survive in that field or not.
The paper by [1] Beth Dietz-Uhler & Janet E. Hurn suggest the importance of learning analytics in predicting and improving the student's performance which enlightens the importance of student's interest, ability, strengths etc. in their performance. According to the paper by [2] Lokesh Katore, Bhakti Ratnaparkhi, Jayant Umale, the career prediction accuracy was determined using 12 attributes of students and different classifiers with c4.5 having the highest accuracy of 86%. [3] Another paper by Roshani Ade, P.R.Deshmukh suggested incremental ensemble of classifiers in which the hypothesis from number of classifiers were experimented and by using 'Majority voting rule', the final results was determined. The proposed ensemble algorithm gave an accuracy of 90.8% [4].The paper by Mustafa Agaoglu suggested the importance of different attributes in evaluating the performance of faculty. It also showed the comparison of different classifiers proved that the most accurate classifier was c5.0 which has the maximum attribute usage compares to other classifiers like CART, ANN-Q2H, SVM etc. Also, the suggestions provided by the system are very much generalized and not specific to a university or country/state. The suggestion for course is also generalized. For example, the results of few systems were a group of courses like data analyst, accountant, law etc. Thus, if a student gets such a recommendation then he/she might again get confused as the above specified course belongs to different streams.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 5, Issue 4, April 2017

III. OVERVIEW

The project was to develop a web application that can be used by any student who needs help in selecting the career path. The system displays questionnaire to the students which the student will have to answer. The three set of questionnaires provided to the students based on personality traits, interests and capacity [5].

3.1 Personality traits: A personality traits are characteristics that are distinct to an individual and are based on the psychology of a person. There have been many approaches to psychological traits theory but the one used in this project is the Myers-Briggs theory according to which there are four prominent pairs of personality characteristics which are:

- Introvert(I) vs. Extrovert(E)
- Sensing(S) vs. Intuitive(N)
- Thinking(T) vs. Feeling(F)
- Judging(J) vs. Perceiving(P)

Based on these four pairs, a total of 16 types of personality types can be obtained by combining four traits (selecting one trait from each pair). Example: ISTJ, ESTP, INFJ etc.

3.2 Interest: Interest in this context implies that how much a student likes a subject and is keen to learn about it. Here the student will be first asked about the basic 3 streams i.e. Arts, Science and Commerce. The system then checks in which stream the student is more interested and then the further questions are comparison based questions which compare the subject of the selected stream and at the end determines one particular subject that the student is interested in.

3.3 Capacity: Capacity implies that how efficiently a student can learn their interested subject and survive in that particular career path. For this purpose, the student will get questions that they had in their school curriculum and each question will have 4 options and a timer associated with it. Here the system asses not only the correctness of the answer but also the speed of the student to answer the question. This helps in knowing the memory, ability to solve and grasping capacity of the student. From the answers obtained, the system predicts a particular course for the student [6]. The prediction is done using one of the Decision tree algorithms which is C5.0 on the personality traits of the student. To further enhance the accuracy of the algorithm Adaptive boosting was applied on the C5.0 algorithm and then C5.0 was applied on the dataset which included interests and capacity of the student also.

The algorithm was implemented using the C5.0 package in R. From the rule-set and decision tree obtained, the personality combinations for various courses was made. Then the interest and capacity results were also integrated. Based on these combinations the web application was developed.

IV. ALGORITHM USED

Data mining is all about extracting patterns from an organization's stored or warehoused data. These patterns can be used to gain insight into aspects of the organization's operations, and to predict outcomes for future situations as an aid to decision-making. C5.0 is a sophisticated data mining tool for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions.

C5 algorithm follows the rules of algorithm of C4.5. C5 algorithm has many features like: The large decision tree can be viewing as a set of rules which is easy to understand. C5 algorithm gives the acknowledge on noise and missing data. Problem of over fitting and error pruning is solved by the C5 algorithm. In classification technique, the C5 classifier can anticipate which attributes are relevant and which are not relevant in classification.

4.1 Boosting of Classifier

C5.0 supports boosting of the classifier to improve the accuracy. Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones.

[7] AdaBoost is one of the boosting methods. It is called adaptive because it uses multiple iterations to generate a single composite strong learner. AdaBoost creates the strong learner (a classifier that is well-correlated to the true classifier)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 5, Issue 4, April 2017

by it relatively adding weak learners (a classifier that is only slightly correlated to the true classifier). During each round of training, a new weak learner is added to the ensemble and a weighting vector is adjusted to focus on examples that were misclassified in previous rounds. The result is a classifier that has higher accuracy than the weak learner's classifiers.

V. IMPLEMENTATION

Step 1. The first step of implementation was to collect data from students studying in different courses.
For this purpose, an online survey was conducted using Google forms. The questions asked in the survey are based on personality traits. Based on the answers the personality type of the student will be determined which will be one among the 16 types. Then it was found that which courses were selected by which personality trait (Figure 1).

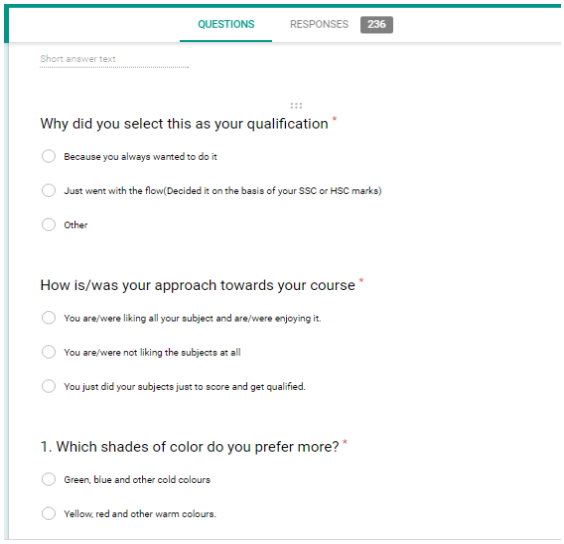

Figure 2: Google form sample.

Step 2. Then data obtained from the survey had to pre-processed and consolidated into a common format as required by the system.

Introvert/Extrovert	Sensing/Intuiion	Thinking/Perceivir	Judging/Perceiving	Can Do Engineering
Introvert	Sensing	Thinking	Judging	Yes
Extrovert	Sensing	Thinking	Judging	Yes
Extrovert	Sensing	Thinking	Percieving	Yes
Extrovert	Sensing	Thinking	Percieving	Yes
Extrovert	Sensing	Thinking	Judging	Yes
Extrovert	Sensing	Thinking	Percieving	Yes
Extrovert	Sensing	Thinking	Judging	Yes

Figure 2: Pre-processed dataset.

Step 3. The dataset was then used to derive the decision tree for various courses. Using the C5 package in R, the algorithm was applied on the dataset for different courses. This figure 2 shows the people with thinking and perceiving personalities can take engineering .
This way, the algorithm was applied other courses and the personality types that can take up those courses were determined.

Step 4. To improve the accuracy of the prediction, adaptive boosting was applied on the algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 5, Issue 4, April 2017

Step 5. The next step was to improve the accuracy even more by incorporating the interests and capacity of the students along with personality trait. After obtaining the results, the combinations of personality, interest and capacity were generated (figure3).

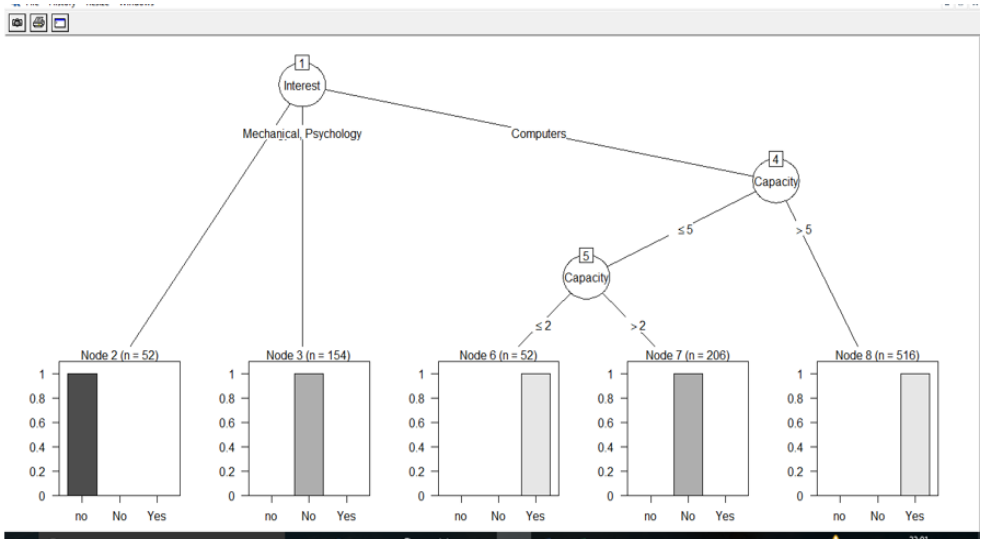

Figure 3: Decision tree for computer engineering.

Step 6. After completing the entire backend decision tree generations, the next step was to develop the web application. The application was developed using Microsoft's Visual studio. Asp.net is the server side technology used with C# as the programming language. The database was stored and processed using the Microsoft SQL Server.

VI. RESULTS OF IMPLEMENTATION

The C5.0 algorithm applied on the dataset had the accuracy of 66% which is quite low for a prediction system. C5.0 algorithm was then applied on the new dataset which has all the three parameters i.e. Personality trait, interest and capacity of the student.

```
Decision tree:

T.F = Feeling: No (151/42)
T.F = Thinking:
:...J.P = Judging: No (164/59)
    J.P = Percieving: Yes (85/38)


Evaluation on training data (400 cases):

          Decision Tree
        ----------------
        Size      Errors

           3  139(34.8%)   <<


         (a)   (b)    <-classified as
        ----  ----
         214    38    (a): class No
         101    47    (b): class Yes
```

This has been improved in the adaptive boosting algorithm to an accuracy rate of 94%.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 5, Issue 4, April 2017

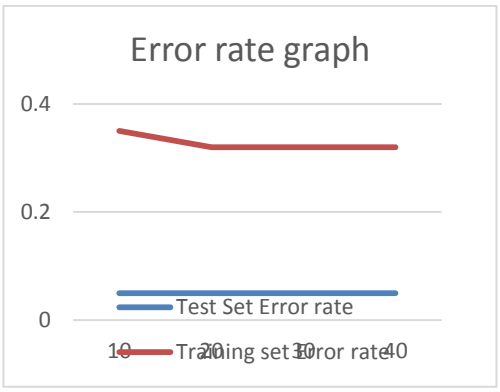

Figure 4: Error rate graph.

In the above figure 4, the red colored line indicates the error rate of training data which is quite high as compared to the error rate of test data which is shown by green line. Then the c5.0 algorithm was applied on the dataset with interest and capacity and the accuracy increased to 100%. The visualization of the decision tree is.

```
Decision tree:

Interest = Biology: no (52)
Interest in {Mechanical,Psychology}: No (154)
Interest = Computers:
:...Capacity > 5: Yes (516)
    Capacity <= 5:
    :...Capacity <= 2: Yes (52)
        Capacity > 2: No (206)

Evaluation on training data (980 cases):

          Decision Tree
        ----------------
        Size      Errors

           5    0( 0.0%)   <<


         (a)   (b)   (c)    <-classified as
        ----  ----  ----
          52                (a): class no
               360          (b): class No
                     568    (c): class Yes
```

From the following graph, we can get a clear idea of the comparison of the two methods (Figure 5).

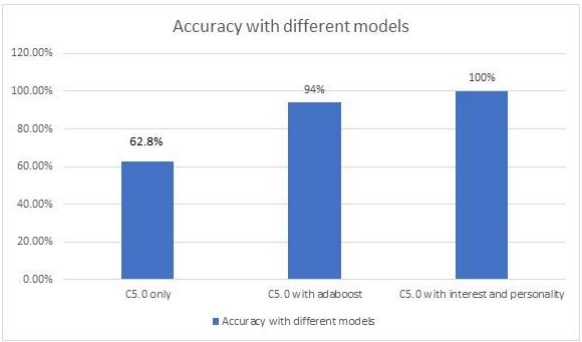

Figure 5: Accuracy with different models.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Vol. 5, Issue 4, April 2017

VII. CONCLUSION

The comparison of using C5.0 with adaptive boosting and C5.0 on dataset with personality, interest and capacity was shown before. This shows that for selecting a career not only the personality trait of a student is important, but also the interest of the student and the capacity of the student to take that courses is also important. Using this system, the student just needs to answer the question displayed by the system and based on the answers the system recommends a particular course along with the list of colleges providing those courses. Thus, the effort required to search the colleges will also reduce.

VIII. REFERENCES

[1] UD Beth, HE Janet, Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. Journal of Interactive Online Learning 2013; 12:17-26.
[2] KS Lokesh, RS Bhakti, et al. Novel Professional career prediction and recommendation method for individual through analytics on personal traits using C4.5 algorithm. IEEE Communication Technology (GCCT) on 3 December 2015.
[3] A Roshani, PR Deshmukh, An incremental ensemble of classifiers as a technique for prediction of student's career choice. IEEE Networks & Soft Computing (ICNSC) on 25 September 2015.
[4] A Mustafer, Predicting Instructor performance using data mining technique in higher education. IEEE 2016; 4:2379-2387.
[5] C Ling, R Dymitr, et al. Big Data: Opportunities for Big Data Analytics. IEEE Digital Signal Processing (DSP) on 10 September 2015.
[6] M Yannick, X Jie, et al. Predicting Grades. IEEE transactions on signal processing on 15 February 2016
[7] M Jiri, S Jan, AdaBoost. Centre for Machine Perception, Czech Technical University, Prague.