*Deepali Joshi et al Int. Journal of Engineering Research and Applications* www.ijera.com
*ISSN : 2248-9622, Vol. 4, Issue 5( Version 7), May 2014, pp.37-40*

RESEARCH ARTICLE OPEN ACCESS

# Prediction Of Study Track Using Decision Tree

Deepali Joshi*, Priyanka Desai**
*(Department of Computer Engineering, Thakur College of Engineering& Technology, Mumbai)
** (Department of Computer Engineering, Thakur College of Engineering& Technology, Mumbai)

**ABSTRACT**
One of the most important issues to succeed in academic life is to assign students to the right track when they arrive at the end of basic education stage. The education system is graded from 1st to 10th standard, where after finishing the 10th grade the student's are distributed into different academic tracks or fields such as Science, Commerce, Arts depending on the marks that they have scored. In order to succeed in academic life the student should select the correct academic field. Many students fail to select the appropriate field. At one instant of time they prefer a certain type of career and at the next instant they consider another option. To improve the quality of education data mining techniques can be utilized instead of the traditional process. The proposed system has many benefits as compared to traditional system as the accuracy of results is better. The problems can be solved via the proposed system. The proposed system will predict the streams through the decision tree method. With each and every input the proposed system evolves with better accuracy.
***Keywords***-Data Mining, Decision Tree, C4.5, NBTree, Decision Tree Algorithm, ssc marks, Accuracy

## I. INTRODUCTION

In today's competitive world everyone wants to be successful in every phase of life. In order to achieve success it depends on whether the correct field is selected or not. If the correct field or track is selected by students then definitely they will be successful in their careers. Sometimes it happens that proper field is not chosen in that situation it becomes very difficult for students. The education available to students can be categorized in different phases. The term basic education specifies the education from 1st to 10th can be specified as basic education[1]. After 10th number of options are available to students and the option selected by student will specify the career that they will opt for. After 10th various streams are available like Science, Commerce, Arts and after 10+2 the students will go for graduation courses. With all these courses and options in hand it becomes very difficult to choose or focus on one career option or to select suitable option. A method is required through which the student can find out which stream is more suitable to them. Once the stream is known they will get a direction of career.

Some solutions are there in order to achieve this but they do not provide appropriate results. One method which is used to specify the stream. Students have to give the basic details. Apart from this percentage obtained will also be used to predict the result. The mechanisms which are used currently do not provide the accurate result and as a result of this the student may not be able to select the proper stream and as a result may not be successful in the academic life.

The mechanism through which students can get the accurate track is by use of Decision tree. Basically a trained data set will be available and trough this trained data set the decision tree will predict the results. In this scenario the result will be the appropriate field or track that is suitable to the student. With each and every result that is predicted the system or the method is going to evolve. As the system evolves it provides the results which are highly accurate.

## II. BUILDING THE MODEL

Data Mining means extracting the meaningful knowledge from large set of data. Decision Tree is a classification technique[9]. Decision Tree is a tree structure which consists of various nodes, leaf node, non-leaf node, root node. There are various algorithms for decision tree, the algorithm used is C4.5, NBTree. The C4.5 algorithm uses the divide and conquer technique. The C4.5 is used due to various advantages, 1. It can handle the continuous and discrete values, 2. It can handle missing values.

The WEKA API is used to generate the result that is the suitable stream for the student. There are three files which will be used, 1. Train, 2. Test, 3. New. The training data is provided to the WEKA API so that it can learn from the trained data set and can generate the results for the given input. The generated result will again be stored in the training data set so it can be used for the next input.

## III. DATA COLLECTION

In order to implement the proposed system data is required as it is to be provided to the algorithm. The data is gathered in the form of ssc mark-sheets[1]. The SSC mark-sheets were collected

from the students and this data is used as training data set so that the result can be generated.

## IV. DECISION TREE ALGORITHM

A decision tree is similar to flow chart and it contains nodes such as root, non-leaf and leaf nodes. Decision tree can handle high dimensional data and they are simple and fast. There are various steps involved to create a decision tree. A data is needed so that the generation of decision tree can be started and such type of data is known as training data or training data set. The decision tree will learn from the training data set. The examination of the tree is to be carried out and this is done through the test data set. Once this is done the completely new data will be provided also known as new data set.

### A. C4.5 ALGORITHM

The C4.5 algorithm is a successor to ID3. C4.5 handles both categorical and continuous attributes to build a decision tree[6]. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values.

### B. NBTREE ALGORITHM

NB Tree is a lies between Naive Bayes classifier and classification[6]. The NB Tree model can be described as a decision tree with nodes and branches. Given the A set of instances of the algorithm evaluation node "Practice" for each division of the property. If the greatest value the property is significantly better than the practices Instance, based on the current node, will be divided into Property. The division is to be done properly, providing an important to better the effectiveness of a naive Bayesian classifier to create the current node.

## V. IMPLEMENTATION

To predict the suitable stream, the student have to give input such as name, gender, phone number, email id, hobbies, railway line, nearby station. Once this is complete various options are provided to the student. The student should enter $10^{th}$ marks where marks have to be entered subject wise. The various subjects are English, Marathi, Sanskrit/Hindi, Maths, Social Science, Science and Technology, ssc percentage. As soon as the marks are entered by the student the option will be visible, Result based on marks. If the student wants to find out the suitable stream with respect to marks, the option should be selected and the result will be displayed. The result is generated by the WEKA API. The input is taken from the student that is the ssc marks and these marks will be provided as input and along with this the training data set will be supplied and the result will be generated.

The training data set is the ssc marks of students consisting of subject marks as well as percentage. With the help of the training data set the results will be predicted that is the suitable stream. The training data set is stored in database and the connection is made between the proposed system and the database. As the result is generated it will be stored in the database and now will be used as training data.

The proposed system generates accurate results for the marks but in order to increase the accuracy the result based on combination of subjects marks is also considered. If the student has good marks in languages that is English, Hindi/Sanskrit, Marathi then it can be specified Arts is more suitable for the student.

## VI. ACCURACY

The algorithms used C4.5 and NBTree, the two algorithms are very efficient and generate the appropriate result. Accuracy is very essential factor for any algorithm. **Accuracy** is the proportion of the time that the predicted class equals the actual class, usually expressed as a percentage.WEKA API provides a method through which the accuracy of various algorithms can be checked. There are various parameters like correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error, relative absolute error, root relative squared error, total number of instances. The basic definition for these parameters is given below.

**1.Root mean-squared error:**

TheRoot mean-squared error value is computed by taking the average of the squared differences between each computed value and the correct value.

**2.Mean absolute error:**

Mean absolute error is the average of the difference between predicted and actual value in all test cases.

**3.Root relative squared error:**

Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value.

**4.Relative absolute error:**

Relative absolute error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

## VII. SOLUTION

The proposed system provides essential information to the students. The students are not

aware of the various courses which are available, so the proposed system will provide the detailed data about the courses that they can take up after their SSC. Not only the courses but the colleges offering the courses along with their informationwill be provided to the students. In order to gather this information the student have to spend large amount of time either by visiting the places or by using the internet.

The proposed system takes input from the students like their detailed information and marks obtained in SSC. The input taken is the name, contact number, email-id, address and subject wise marks. Depending upon the marks the appropriate stream will be specified to the student. In order to generate the results a trained dataset is utilized. The trained dataset is the authentic data which was collected from the students. The WEKA API will learn from the trained data and accordingly it will specify the stream for the student. Now-a-days the college and its location is also taken into consideration by the student as well as their parents. So the location that is entered by the student according to that the colleges which are nearby will be displayed to the student which are offering the specified stream.

In order to predict the stream the combination of subjects will be considered as well. If the student has good marks in languages then the student can select the ARTS field. Similarly if the student has good score in maths then Commerce field is more suitable for the student.

## VIII. RESULTS

The proposed system is more efficient as compared to the existing system and generates appropriate results. The input is taken from the student and depending upon the marks obtained by the student the appropriate stream will be predicted. Information is gathered from the student such as their details and the ssc marks. Depending upon the combination of subjects in whichthe student has scored good marks the stream can be specified to the student.

In order to find the stream students need information about the courses and the colleges offering the courses. This complete information is provided to the students. According the predicted stream the colleges which are near to the location of student will be provided to the student. The students will get all the information that is needed.

In order to get the results input is provided by the student. The input consists of student name, gender, phone number, email id, hobbies, railway line, nearby station and the ssc subject-wise marks and ssc percentage. Once the required input is supplied the proposed system generates the output and displays the comparison of accuracy.

The table shows the comparison of C4.5 and the NBTree algorithm. The table contains various parameters and along with that the instances. Instances are basically the number of students for whom result have been generated.The result of comparison is:

| For C4.5 | For NBTree |
|---|---|
| Correctly Classified Instances 14 100 % | Correctly Classified Instances 14 100 % |
| Incorrectly Classified Instances 0 0 % | Incorrectly Classified Instances 0 0 % |
| Kappa statistic 1 | Kappa statistic 1 |
| Mean absolute error 0 | Mean absolute error 0.0833 |
| Root mean squared error 0 | Root mean squared error 0.0962 |
| Relative absolute error 0 % | Relative absolute error 100 % |
| Root relative squared error 0 % | Root relative squared error 100 % |
| Total Number of Instances 14 | Total Number of Instances 14 |

Table no 1. Comparison of C4.5 and NBTree

## IX. CONCLUSION

The proposed system takes input from the students and predicts the stream which is best for them in order to make the student successful in the academics as well as their life. The proposed system provides better and accurate results as compared with the traditional method. The marks of the student are considered and along with that marks of combination of subject are verified in order to give the results. Theinformation about the colleges as well as various courses will be provided to the student. So that the searching time can be minimized. Depending on the location that is entered by the student the colleges providing the specific courses will be displayed to the student. The students as well as their parents will have clarity about the course to be selected. Not only the stream but also the colleges which are nearest to the student's location are displayed. Apart from this information about the college like address and contact number is provided to the student.

## REFERENCES

[1] Ahmad I., Manarvi, I., Ashraf, N. "Predicting university performance in a subject based on high school majors", *978-1-4244-4136-5/09/ ©2009 IEEE*

[2] Zhiwu Liu, Xiuzhi Zhang. "Predictionand Analysis for Students' Marks Based onDecision Tree Algorithm", Intelligent Networks and Intelligent Systems (ICINIS), 2010 3rd International Conference on Digital Object Identifier:10.1109/ICINIS.2010.59 Publication Year: 2010 , Page(s): 338 – 341

[3] Anupama Kumar S, Vijayalakshmi M.N. "Mining of student academic evaluation records in higher education", Recent Advances in Computing and Software Systems (RACSS), 2012 International Conference on Digital Object Identifier: 10.1109/RACSS.2012.6212699 Publication Year: 2012 IEEE , Page(s): 67 – 70

[4] Bunkar, K, Singh U.K., Pandya B, Bunkar R. "Data mining: Prediction for performance improvement of graduate students using classification", Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on Digital Object Identifier: 10.1109/WOCN.2012.6335530 Publication Year: 2012 IEEE, Page(s): 1 – 5

[5] Garcia, E.P.I. ; Mora, P.M. "Model Prediction of Academic Performance for First Year Students", Artificial Intelligence (MICAI), 2011 10th Mexican International Conference on Digital Object Identifier: 10.1109/MICAI.2011.28 Publicat ion Year: 2011 IEEE , Page(s): 169 – 174

[6] Pumpuang, P., Srivihok, A., Praneetpolgrang, "Comparisons of classifier algorithms: Bayesian network, C4.5, decision forest and NBTree for Course Registration Planning model of undergraduatestudents", Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on Digital Object Identifier: 10.1109/ICSMC.2008.4811865 Publication Year: 2008 IEEE, Page(s): 3647 – 3651

[7] Qasem A. Al-Radaideh, Ahmad Al Ananbeh, and Emad M. Al-ShawakfaA "Classification Model For Predicting The Suitable Study Track For School Students". *IJRRAS 8 (2) August 2011*

[8] Sunita Beniwal, jitender Arora, "Classification and Feature Selection Techniques in Data Mining", *IJERT AUGUST 2012*

[9] Heena Sharma, Navdeep Kaur Kaler."Data Mining with Improved and Efficient Mechanism in Clustering Analysis and Decision Tree as a Hybrid Approach" *(IJITEE) ISSN: 2278-3075, Volume-2, Issue-5, April 2013*