Student Career Prediction Using Decision Tree and Random Forest Machine Learning Classifiers

N. VidyaShreeram[1], Dr. A. Muthukumaravel[2]
{vidushreeram123@gmail.com[1], dean.arts@bharathuniv.ac.in[2]}

Research Scholar, Department of Computer Applications, Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India.[1],
Dean, Faculty of Arts and Science, Bharath Institute of Higher Education and Research, Chennai, Tamilnadu, India.[2]

Abstract. Education is so important for the youngsters; some of them don't have the interest towards schoolings, so they drop their study after certain time. As in this fastest world student are going through their academics and with their interested courses. So this is an important in the entire student to choose his future career. Machine learning approaches are applied in various domains. This proposed work deals with the career predication of the students as weather they will be going for their next level of higher education from their present graduation level using machine learning concepts like DT (Decision Tree) and RF (Random Forest). Applying the concept of DT it yields a result of about 91% of accuracy and applying RF it gives 93% of accuracy level. The result of the proposed system helps the recruiters to select the only needed and proper candidates.

Keywords: Random Forest, Machine Learning, Decision Tree, Classifier, Accuracy, Career Prediction

1 Introduction

The literacy in the world is the key for the social economic progress, as per the calculation the India has 81.1% of education people. The topmost country in the education is Kerala. The percent of the educated one includes from both the rural and urban areas. But the rest of the 18.9 percent of illiterates is noticeable. The reason for illiteracy in India is many and more. The most common reason for the illiteracy is the poverty condition of the family, inadequate of school and also because of the inadequate number of properly trained educators. There is more number of children in street or orphans who don't even get a chance for education. In this society competition is multiplying day by day. Mainly it is a very tremendous to beat the competition in this difficult world. Though it is a tough one student should compete it and reach the goal. So there is a need to evaluate the student with their performance in a constant manner and the interest towards their goal should be evaluated. And it is very important that they are directed towards a right path. There are various types of sectors for the candidates to choose their right path.

Machine learning is an application which makes the system to learn and improve from the experience without being explicitly programmed. It is a process of data analysis which automates analytical model building; machine learning is one of the branches of the AI (Artificial Intelligence). The systems that were built on the machine learning algorithms have

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308621

the ability to learn something from the past data. Machine learning is basically of three types they are supervised learning, unsupervised learning and reinforcement learning. Here in this proposed system machine learning concepts are applied to detect weather the particular candidate will be preceding his higher level education. Machine learning has many concepts and branches. Such as decision tree, naviebayes, random forest etc. By applying the concept of RF and DT on the student dataset and accuracy level of RF is better comparing with DT.

The section II conveys different reviews of the authors at machine learning concepts. Section three elaborates the concept in this proposed system. Section four is about the result and discussion of the proposed method. Finally in the last section concludes the result of the proposed system.

II LITEERATURE REVIEW

Students doing their favorite stream and doing their academic activities based on their own decision. Evaluate students capability is very important to find their career path. This evaluation is helpful to the students to increase the performance level and encourage their interested area. The job providing companies are also evaluating the students in various aspects. Career prediction systems are used to decide the suitable role of the students or they are interested to their higher studies. K. Sripath Roy et al., provides more importance of computer field student's career prediction with the help of three machine learning concepts. Here the data can be trained by using various three machine learning algorithms like SVM, XGBOOST and decision tree. From the above mentioned algorithms SVM produces better result in term of accuracy level [1].

Taking decision about their career in people life is very difficult. Normally the student's career is predicted by using questionnaires method. But in this method does not reflect the state of the students. Min Nie, et al., proposed a new model based on students behaviors in college campus. Student's behaviors also used to predict their career [3].

Due to the advancement of computing concepts deep learning approaches has considered for various domains. It is used in the education domain also. Bendangnuksung et al., proposes a new model called DNN (Deep Neural Network) to find the category of the students. This model is used to provide the solution for failed people. The same data set can be applied on existing machine learning concepts. But this proposed model produces 84.3% of accuracy and performed well compared with existing machine learning approaches [4].

Companies spent lot of time to appoint the students from various educational institutions. The educational intuitions also not able to provide 100% job facility within the qualified students. Students from the various intuitions are also unable to get proper training from their educational institutions. They are not providing the proper training based on organization requirements. The recruiting organizations also have no knowledge about the students. To solve the above problem KachiAnvesh et al., developed a new tool. It will be executed automatically. The main task of this tool is to find the student's eligibility based upon various qualifications such as CGPA, extra courses completed, projects completed, internship etc. The main goals of this project are 1.to make the decision to go the students to the interview or refuse them. 2. Educational institution provides training to students those who are refused by small reasons like programming skill, communication skill etc. This task is executed based on minimum eligibility and preferred eligibility. These two types of eligibilities are important for job providing companies [5].

Success is one of the important factor is everyone life. The public needs to be doing well should provide concentration to their younger people, because the youngest peoples are the future of the country. In this research article Vahide Nida Uzel et al., studied about the performance of various level students to be measure with the help of various classifiers like MLP (Multilayer Perceptron), RF (Random Forest) ,NBC(Naïve Bayes Classifier), DT (Decision Tree) and classifier based on voting. This study also used to find the academic related characteristics of the students. Various features affect the achievement of the people. The important features are student absent, satisfaction of parents about school, activity on the class and the responsibility of the parents. Here Apriori concept is used to identify the association between the various features. According this research classifier based on voting concept provides better result than existing ANN(Artificial Neural Networks) technique. ANN and classifier based on voting method is also use the same dataset [6].

Vladimir L. Uskov et al., presented the output of their research project. The aim of their research work is to evaluate eight types of machine learning approaches. This research work mainly focuses on student's performance in academic. The authors also conducted a survey on computer science graduates to identify their attitude about using of machine learning based analysis in education field and the student's responses also added in the research paper [7].

III PROPOSED SYSTEM

In earlier days student's career is forecasted by using questionnaire method. But it is a time consuming process and it is very difficult to find the status of student's opinions. Various computing techniques are used to predict the career of the student's. In this research work DT and RF machine learning concepts are used to forecast the student's career. Compare to existing traditional methods new computing concepts like machine learning approaches produces better result. The following Fig 1 shows the flow diagram of proposed system.

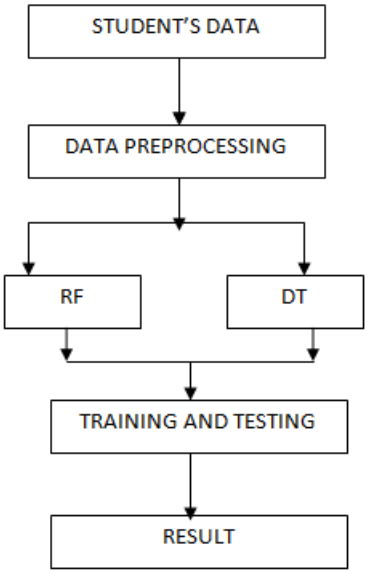

Fig1 Proposed System Flow Diagram

Student's data is given to the input of the proposed system. Data collected from colleges and do the preprocessing task. Preprocessing task is important for data analysis task because in this process remove the unwanted data from the original data and fill the mission values. DT and DT algorithms are applied on the preprocessed data. In the next level train and test the data. This research work uses 16 various attributes like age, father and mother education status, health condition, family economy level etc. are used to predict their career. The following correlation heat map fig 2 shows the various attributed in this research.

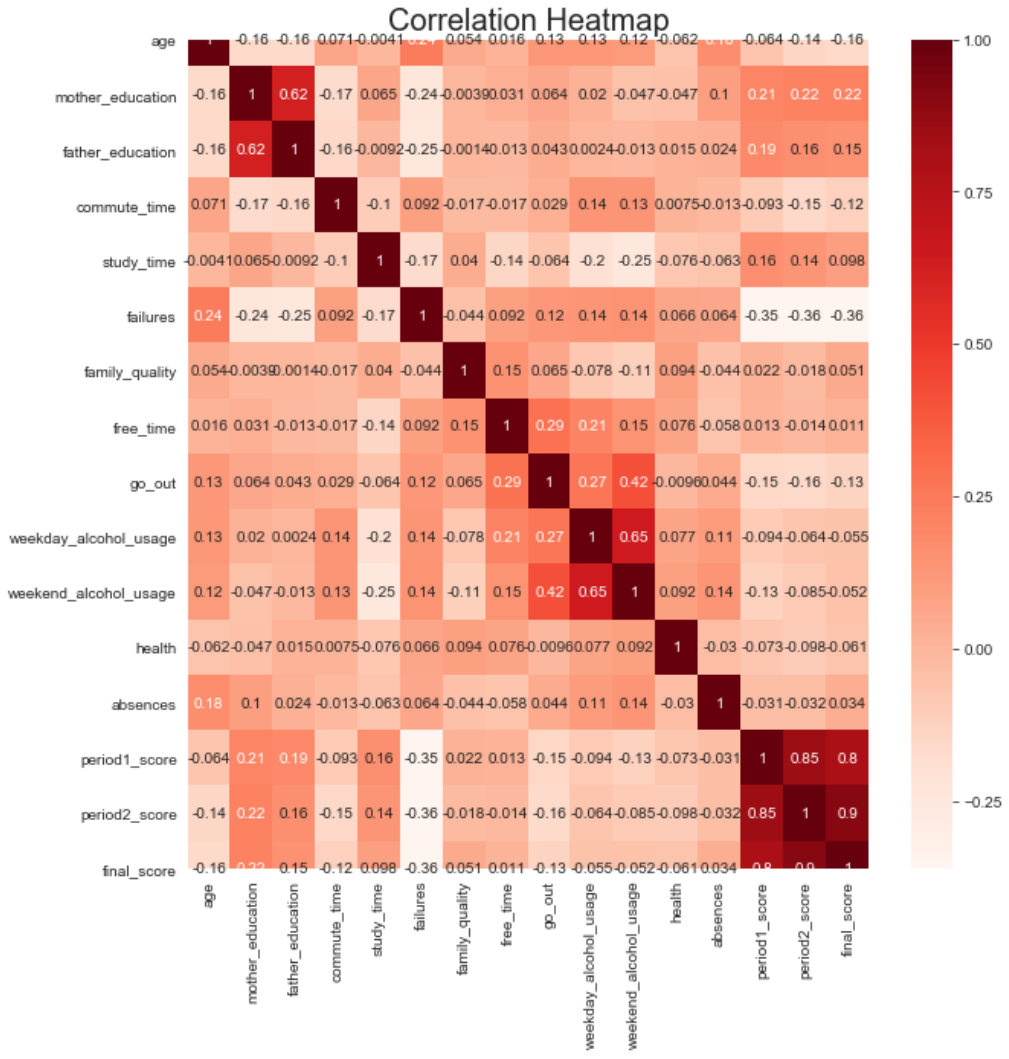

Fig 2 Correlation Heat map

Decision Tree (DT)

Decision tree are most common in research operation, particularly in decision analysis to identify a strategy which is most likely to reach a goal and it tree and turns out that it has influenced a wide area of machine learning, covering both classification and regression. Using the basic concepts of decision tree various advance concepts also implemented like RF,

bagging and gradient boosting. XG Boost technique is also one of the advance concepts of DT. Normally CART, C5, ID3 and C4.5 type of decision trees are used to classify the data. in DT node represented as the input identifier(X) and divide the variable, assumption of identifier is numerical value. The leaf of the decision tree is called as terminal node and represented by using the identifier(Y) which is important for forecasting. The intital step of the DT is selecting the root value. In the next stage compute the IG (Information Gain) and Entropy of every node before division. Then choose the nodes which contain more IG or low entropy value. Again divide the node and repeat the same process until no option to divide or the value of entropy is low. Entropy is one of the important metric to assess the chance of information. IG is used to evaluate how much the entropy value is decreased prior to past division.

Construct DT, to compute two kinds of entropies with the help of frequency data table. Entropy value is calculated by using only one attribute in frequency table is as the following equation.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad \text{--------(1)}$$

Calculate the Entropy value of frequency table by using two attributes as the following equation.

$$E(T, X) = \sum_{c \in X} P(c) E(c) \qquad \text{-------(2)}$$

The IG value can be calculated by using the following equation.

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \qquad \text{------(3)}$$

RANDOM FOREST (RF)

RF is on the best prediction instrument in the field of education. It yields better prediction result in terms of accuracy compared with other machine learning concepts and it is working like as bagging and boosting techniques. RF approach creates many decision trees by using sample predictor identifiers. It makes the assumptions of nominal type of identifiers (regressions) and continuous identifiers (classifications). Mean value of every tree are used in regression type of problems and weighted value of decision trees are used for classification type of problems. Aggregate values of all decision trees produce the better prediction result. RF concept is unique for handling large amount of data and it is easy to handle missing data. The RF approach can be explained as the following steps.

Acquire K samples information data from the training information

For every sample construct RF tree

*choose m_{try} identifier

*choose the best identifier / division point with in m_{try} identifier

*divide the node into 2 sub nodes

c) Do again the step b until n_{min} value reached

d) Forecast the new information from model tree

Here K number of trees can be described as follows:

$$\{T_k\}_1^K$$

The point x , predicted can be developed as follows:

$$Regression : \hat{f}_{rf}^{K}(x) = \frac{1}{K}\sum_{k=1}^{K} T_k(x)$$

$$Classification : \text{Given that } \hat{C}_k(x) \text{is the class prediction of the } k^{th} \text{ tree}, \hat{C}_{rf}^{K}(x)$$
$$\text{is the majority vote of the random forest trees}(\{\hat{C}_k(x)\}_1^K)$$

In this research work RF and DT concepts are used to predict the student's career. This concept can be implemented with the help of python programming.

IV RESULT AND DISCUSSION

Machine learning approaches are used in various fields. In education domain also machine learning concepts are used. This research mainly focuses on student's career forecasting. Student's historical data can be used as the input of this work. From the entire data the various features are extracted by using feature extraction methods. The prediction accuracy value changed depends on the number of features extracted from the original data. The following figure 3 shows the relation between the number of features and the prediction accuracy level.

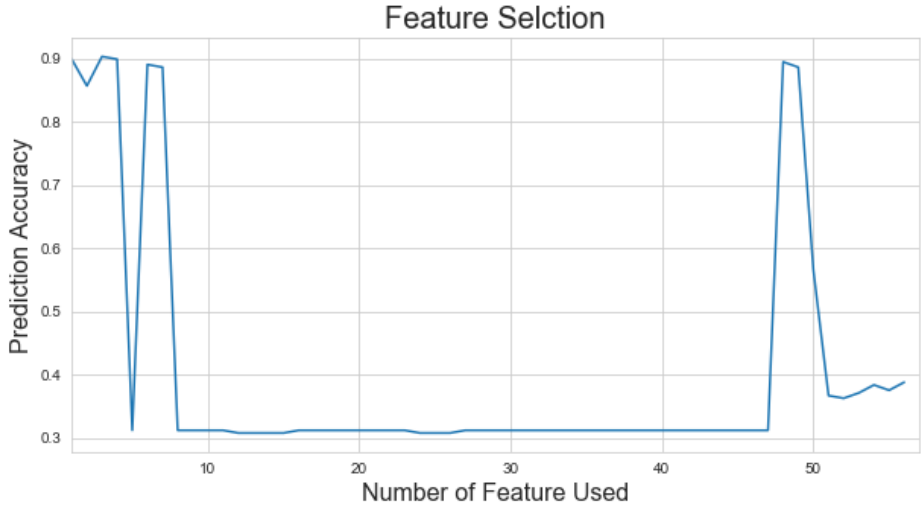

Figure 3 Relation between number of features and Accuracy level

Initially, preprocessed data can be classified by using decision tree and the next level DF classifier also applied on the preprocessed data. The following figure 4 a) and b) shows the accuracy level of Decision tree and random forest classifier.

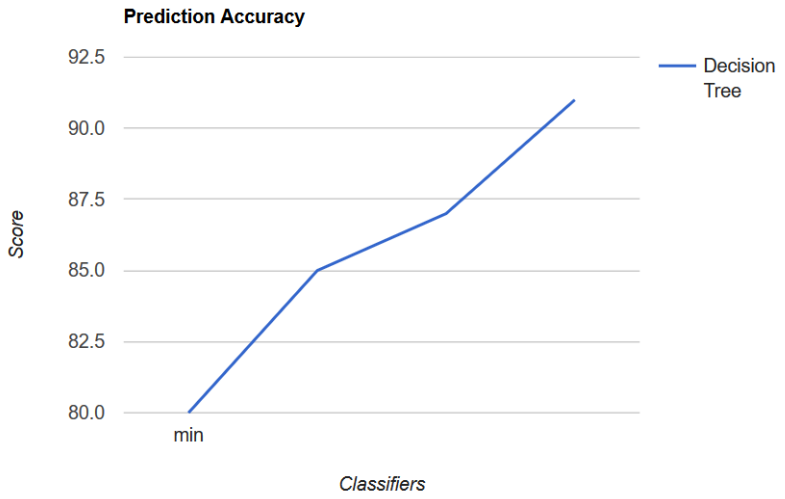

Figure 4 a) Accuracy Level of Decision Tree

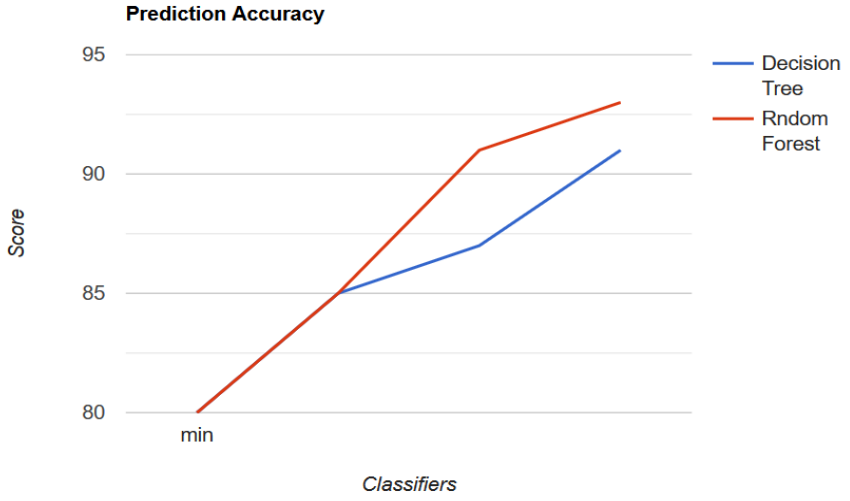

Figure 4 b) Accuracy Level of Decision Tree and Random Forest

Based upon the output of our proposed system accuracy level of Random Forest is better than Decision Tree.

V CONCLUSION

The thing that is very important for the student who chooses their higher level education is that they should know their capability and select their courses with their interest towards the particular course. The output of the career prediction is also mainly used by the recruiters, because the job providing organizations spent large amount of money for selecting the suitable candidates. This research work forecast the undergraduate students' career by using RF and DT machine learning classifier. The classifiers are implemented by using Python programming languages. Compared with other programming language it is easy to implement real world problems. Final output shows that Random Forest classifier produces better result than the Decision Tree classifier.

References

[1] K. SripathRoy ,K.Roopkanth , V.UdayTeja , V.Bhavana and J.Priyanka (2018), "Student Career Prediction Using Advanced Machine Learning Techniques", International Journal of Engineering & Technology, Vol. 7, pp. 26-29.

[2] Min Nie , ZhaohuiXiong , RuiyangZhong , Wei Deng and Guowu Yang (2020), " Career Choice Prediction Based on Campus Big Data—Mining the Potential Behavior of College Students" , Journal of Appl. Sci. pp. 1-14.

[3] Min Nie, Lei Yang, Jun Sun, Han Su, Hu Xia, DefuLian& Kai Yan (2018), "Advanced forecasting of career choices for college students based on campus big data, Springer Frontiers of Computer Science , Vol. 12, pp. 494–503.

[4] Bendangnuksung andPrabu P (2018)," Students' Performance Prediction Using Deep Neural Network", International Journal of Applied Engineering Research ISSN 0973-4562, Vol. 13, No. 2, pp. 1171-1176.

[5] KachiAnvesh, B. Satya Prasad, V. Venkata Sai Rama Laxman, B. Satya Narayana (2019), Automatic Student Analysis and Placement Prediction using Advanced Machine Learning Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Vol. 8, No, 12, pp. 4178-4183.

[6] Vahide Nida Uzel, Sultan SevgiTurgut and Selma Ayşe Özel (2018), "Prediction of Students' Academic Success Using Data Mining Methods", IEEE 2018 Innovations in Intelligent Systems and Applications Conference (ASYU).

[7] Vladimir L. Uskov , Jeffrey P. Bakken , Adam Byerly and Ashok Shah (2019), "Machine Learning-based Predictive Analytics of Student Academic Performance in STEM Education", 2019 IEEE Global Engineering Education Conference (EDUCON).

[8] S.Umadevi, S. Nirmala SugirthaRajini, A. Punitha and Viji Vinod (2020)," Performance Evaluation of Machine Learning Algorithms In Dimensionality Reduction", International Journal of Advanced Science and Technology, Vol. 29, No. 9s, (2020), pp. 3845-3853.