Student Career Prediction Using Machine Learning Approaches

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. India is blessed with the number of good schools and colleges. But most of the students are dropping their next level of education because of various reasons. The reason is many and more, some of the students have some economic problem with their family, some of the students don't have interest towards their next level of education, some matters about the gender and some rural areas don't have good schools and educators. So this proposed method deals weather the students will be going to the next level of higher education. This can be evaluated with the concepts of machine learning which the subset of artificial intelligence. Machine learning is made up with the Mathematics and Science concepts. This paper deals with the students' career prediction by using various machine learning concepts like Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Adaboost. RF classifier yields better accuracy of 93% compared with other machine learning classifier. Machine learning classifiers are implemented by using Python programming language.

Keywords: Machine Learning Classifier, Support Vector Machine, Random Forest, Decision Tree, Adaboost

1 Introduction

Traditionally student's career can be predicted by using questionnaire. But this method takes lot of time. Now, computing technologies play a vital role in various fields. Machine learning is one of latest computing technique. In this digital world Machine learning is used in various fields and industries such as clinical analysis, image processing, classification, regression then more and more. It has the capability to develop and study automation without being explicitly. Machine learning is of three types supervised machine learning, unsupervised machine learning and reinforcement machine learning algorithms. Machine learning in simple words it is the science of obtaining to learned and behave like humans. It is so important to analyze the ability of the students and they should be directed in the right path way.

In this research work the concepts of machine learning are applied to detect the next level of education of the particular candidate. This prediction is important for all type of educational institutions and recruiters. Based upon the result of this prediction accuracy the educational institutions find the low level people and provide the proper training to them to improve their performance. Job providing companies also spent lot of amount of selecting a proper candidate. The output of the prediction model is also used to find the status of the students, if

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308642

they are interested to go to the job or they interested to do their higher studies. This research work mainly focus on the career prediction of undergraduate level students. Machine learning algorithms such as DT, RF, SVM and Adaboost classifiers are used to construct the model. From the above classifiers RF produces better result. These classifiers are implemented with the help of python programming, because most of the real time problems are easily implemented by this language.

Section two deals with the views and approaches are used by various authors in career prediction research domain. Section three explains the proposed system with appropriate algorithms. In section four elaborates the proposed system result part. Finally section five concludes the proposed work.

2Literature Review

Career selection is playing a major role in the life student’s life. Traditionally specialized profession people using questionnaires to identify the important factors affecting career paths. But it is very difficult to predict the career path because of complication of every student aim and dreams. A recent fact provides the suggestion using student’s data based on their behavioral aspects to forecast the career path. Min Nie et al., proposed a novel model known as ACCBOX (Approach Cluster Centers Based On XGBOOST) to forecast student’s choices in their career. The final result clearly states that the current method is better than other predicting methods. In this model uses 13 behavioral data collected from 4000 students [1].

Mining education data is also one of the important tasks in education field. In the beginning days data mining methods are used in education field by using less number of arguments, because low record maintenance in concern institutions. Recently the large volume of data can be stored on student. In India 0.3 % people only move from their PG level to research level. This prediction task concentrates performance of the student’s by using various arguments and the students are classified as low, high and medium type. To execute the above process the authors K. B. Eashwar et al., combines SVM and K-means method. A SVM concept is used for classification purpose and K-mean technique is mainly used for cluster the student’s data [2].

In education domain predicting the student performance level is one of the important tasks. Data mining concepts are used to predict the student’s performance by using various types of tests. AnkitaKadambande et al., uses semantic rules and SVM concepts to do predictions. Semantic rules are used to improves the educational content quality and convey education action to every student. Here the authors helps the students by providing better suggestion and issues recommendations for improving student’s performance level in forthcoming exams. This system will provide the helps to low level and high level students and also to increase the student’s interest about their education. The main purpose of this research work is to increase the quality of learning measures and support the students by forecasting their academic level and help the students [3].

The authors S.A. Oloruntoba et al., identifies the association among initial academic profile level and the last academic profile level. For this research work Federal Polytechnic student database can be used. The initial academic profile is represented as O level. The performance of the students in academic is defined by using GPA (Grade Point Average). Current research work mainly focused on develops a new model for predicting performance of the student with the help of data mining approaches. Preprocessing task is used to delete the unwanted data. Here student's performance can be predicted by using SVM concept. The result of this classifier is compared with other machine learning concepts like linear regression, KNN and decision tree. The accuracy level of SVM is better than other machine learning concepts [4].

Student's performance prediction is important in higher educational institutions. The prediction result is used to spot and increase the performance level of the students. Various factors are influenced to improve the performance level. AhmedSharafElDen et al., uses classification concept to increase the quality of advanced learning system. Here the authors uses Adaboost technique with genetic technique is known as Ada-GA to increase the performance of the classifiers. Ada-Ga technique is useful to identify the student's risk level in earlier manner with large amount of data. This output is used by the tutor to issue the proper advice to the concern students [5].

Student's data are increased day by day. Among the various prediction methods machine learning concept is one of the outstanding model. Meimei Han et al., proposed a Adaboost model to predict the students level. The result of the Adaboost classifier is compared with other machine learning concepts like neural network, decision tree, SVM and random forest. Initially association policy and correlation study are used to find the characteristics of the model. In next level various prediction models are used to predict the data. Finally compare the accuracy level of the prediction models. In term of accuracy level AdaBoost is higher but the time and cost is high compared with other models. Association rule mining is used to assist the students locate their issues from the origin of the issue to assist them to resolve the issues [6].

Level of student's performance is one of the major significant values of the all type of educational organizations. To improve the value of the institutions, need to forecast the performance of the student's. Special type of treatment is needed for low level performer. FaridJauhari, et al., proposes three various boosting approaches to construct the classifier for forecasting the level of student's. In this research 1UCI dataset is used for developing model [8].

3Proposed System

In this proposed work the four concepts of machine learning are applied that are decision tree, random forest SVM and Adaboost.

DT is one of the predictive modeling methods that are helpful in the statistics data mining and machine learning. It is very popular and easiest concept to develop a model for real time issues. Decision tree is of two types they are categorical variable decision tree and continuous variable decision tree. DT is basically a graph which is the branching method that is used to explain all the possible results of the decision. Decision tree can be drawn or created by graphics program or with some specific programs the various outcomes from the drawn tree that can be used as a decision making tool for research analysis or for planning strategy. RT is also called as random decision forest is an ensemble learning process for the classification, regression and some other tasks. This concept of machine learning gives a good accuracy level for the result.

RF classifies will handle the missing values and maintain the accuracy for a large proportion of data. RF is a branch of machine learning techniques. The main advantage of RF classifier is it performs classification task and regression task. RF classifier having the capability of managing large amount of dataset and it avoids the issue like overfitting. The steps are shows the working procedure of the RF classifier.

1. Choose “K” number of features from the total “m” number of features.
2. From K features, compute the value d by using split point
3. Divide the nodes into sub nodes using the data best split
4. Repeat the steps 1-3 until “l” nodes reached
5. Construct the forest by repeat the 1-4 steps for n times to construct number of trees ‘n’

SVM is also a branch of machine learning SVM is also called as support vector machine or support techniques that is associated with the learning algorithms which analyze the dataset used for classification and regression analysis. It is one of the suitable methods of linear and nonlinear data type. Each student has various identifiers, and each of them is represented as a multidimensional items. Hyperplane separates the data from one class to another class. SVM find the hyperplane and using vectors and edges. Many data analysts says that SVM takes more time for training; but the accuracy level is high compared with other techniques. SVM approach is best method when the training samples are classified by using large amount of arguments. Using nonlinear approach compare selected argument of one student with others to forecast the performance of the students. The sample data falls on HP(hyperplane) are called as support vectors. Normally HP value is closer to MMH (maximum marginal hyperplane). SVM approach is best suitable for fewer amounts of data with less than training set 2000. MMH has been defining by using Lagrangian formulation is represented as follows.

$$d(X^{T}) = \sum_{i=1}^{l} y_{i} \alpha_{i} X_{i} X^{T} + b_{0} \qquad \text{----------(1)}$$

Here $y_i$isrepresented as label of call vector X_i .

X^T represented as a test record

α_i and b_0 are used to represent the numeric arguments that are find in automatic manner by SVM classifier.

l identifier represented a s the total number of support vectors.

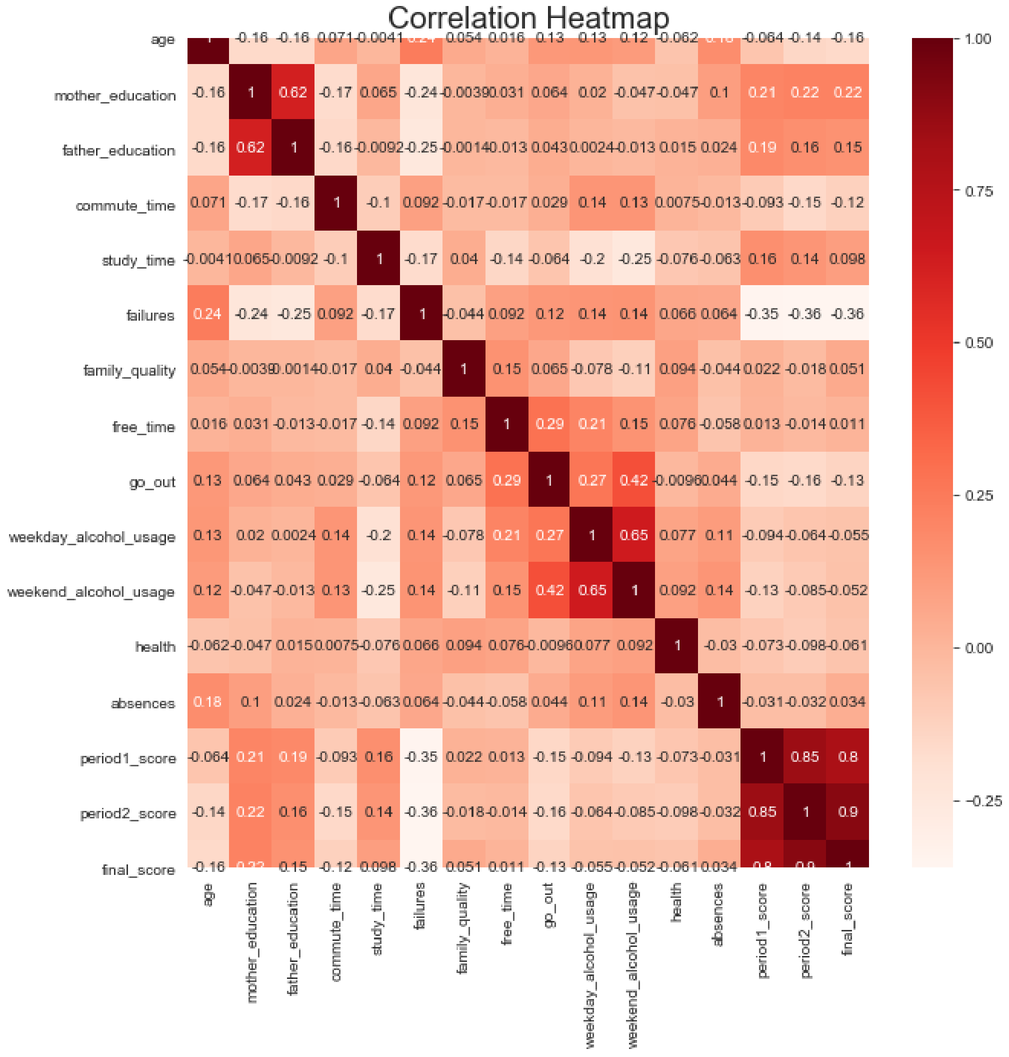
ADABOOST:

Adaboost is a machine learning mean concept algorithms. This Adaboost works with the principle to generate different weak learners and combine their prediction to forms one strong rule. Boosting is use to create a group of predicators. The following steps are used to understand the concept of Adaboost classifier.

1. From the top training data set create the weaker classifier
2. Construct the decision stump of every identifiers
3. Assign high weight value to wrongly classified samples
4. Repeat the process from step 2 until all data points properly classified.

4Result And Discussion

The main concept of this proposed system is to find whether a student is interested towards the next level of higher education. This proposed works mainly is based on the career aspect prediction with the help of ML concepts. Here machine learning concepts are applied by using Python programming language. Compare with other programming languages Python programming mainly used for implement real time problems. For this research work the data collected from various educational institutions. Preprocessing concepts are applied on the collected data. During the preprocessing task the missing values are filled on the collected student's data. The unwanted data also removed from the original data during preprocessing process. Then the important features are extracted by using feature extraction techniques. In this research work 16 features are used to develop the prediction model. The important features are age, health condition, parent's status, study time etc. The following Correlation Heatmap figure 1 shows that what are the features are used to construct a prediction model.



	age	mother_education	father_education	commute_time	study_time	failures	family_quality	free_time	go_out	weekday_alcohol_usage	weekend_alcohol_usage	health	absences	period1_score	period2_score	final_score
age	1	-0.16	-0.16	0.071	-0.0041	0.24	0.054	0.016	0.13	0.13	0.12	-0.062	0.18	-0.064	-0.14	-0.16
mother_education	-0.16	1	0.62	-0.17	0.065	-0.24	-0.0039	0.031	0.064	0.02	-0.047	-0.047	0.1	0.21	0.22	0.22
father_education	-0.16	0.62	1	-0.16	-0.0092	-0.25	-0.0014	-0.013	0.043	0.0024	-0.013	0.015	0.024	0.19	0.16	0.15
commute_time	0.071	-0.17	-0.16	1	-0.1	0.092	-0.017	-0.017	0.029	0.14	0.13	0.0075	-0.013	-0.093	-0.15	-0.12
study_time	-0.0041	0.065	-0.0092	-0.1	1	-0.17	0.04	-0.14	-0.064	-0.2	-0.25	-0.076	-0.063	0.16	0.14	0.098
failures	0.24	-0.24	-0.25	0.092	-0.17	1	-0.044	0.092	0.12	0.14	0.14	0.066	0.064	-0.35	-0.36	-0.36
family_quality	0.054	-0.0039	-0.0014	-0.017	0.04	-0.044	1	0.15	0.065	-0.078	-0.11	0.094	-0.044	0.022	-0.018	0.051
free_time	0.016	0.031	-0.013	-0.017	-0.14	0.092	0.15	1	0.29	0.21	0.15	0.076	-0.058	0.013	-0.014	0.011
go_out	0.13	0.064	0.043	0.029	-0.064	0.12	0.065	0.29	1	0.27	0.42	-0.0096	0.044	-0.15	-0.16	-0.13
weekday_alcohol_usage	0.13	0.02	0.0024	0.14	-0.2	0.14	-0.078	0.21	0.27	1	0.65	0.077	0.11	-0.094	-0.064	-0.055
weekend_alcohol_usage	0.12	-0.047	-0.013	0.13	-0.25	0.14	-0.11	0.15	0.42	0.65	1	0.092	0.14	-0.13	-0.085	-0.052
health	-0.062	-0.047	0.015	0.0075	-0.076	0.066	0.094	0.076	-0.0096	0.077	0.092	1	-0.03	-0.073	-0.098	-0.061
absences	0.18	0.1	0.024	-0.013	-0.063	0.064	-0.044	-0.058	0.044	0.11	0.14	-0.03	1	-0.031	-0.032	0.034
period1_score	-0.064	0.21	0.19	-0.093	0.16	-0.35	0.022	0.013	-0.15	-0.094	-0.13	-0.073	-0.031	1	0.85	0.8
period2_score	-0.14	0.22	0.16	-0.15	0.14	-0.36	-0.018	-0.014	-0.16	-0.064	-0.085	-0.098	-0.032	0.85	1	0.9
final_score	-0.16	0.22	0.15	-0.12	0.098	-0.36	0.051	0.011	-0.13	-0.055	-0.052	-0.061	0.034	0.8	0.9	1

Fig 1: Correlation Heatmap

The prediction accuracy value changed depends upon the number of features. The following figure 2 shows the association between number of features and accuracy value.

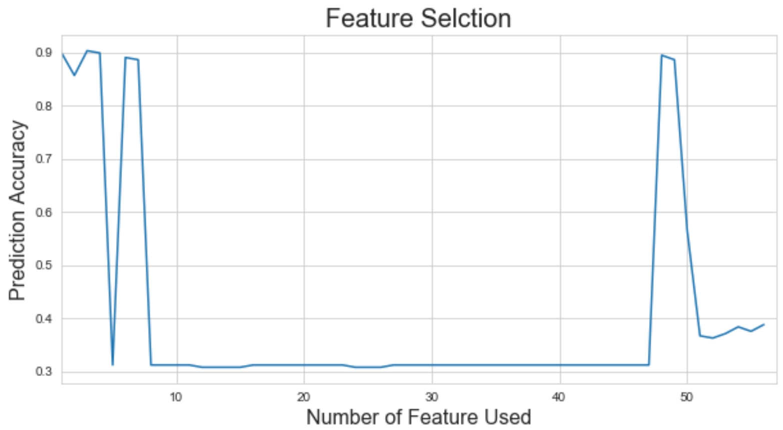


Figure 2 Relation between number of features and prediction accuracy

In this research work the authors predict the student's career by using four machine level classifiers. The following table 1 shows the accuracy level of the four different classifiers.

Table 1 Classifier name with Accuracy Value

Sl. No	Classifier Name	Accuracy Value
1.	Adaboost	87.00%
2.	SVM	88.50%
3	Decision Tree	91.00%
4.	Random Forest	93.00%

The following figure 3 a) and 3 b) demonstrates the different classifier accuracy value using line and bar chart.

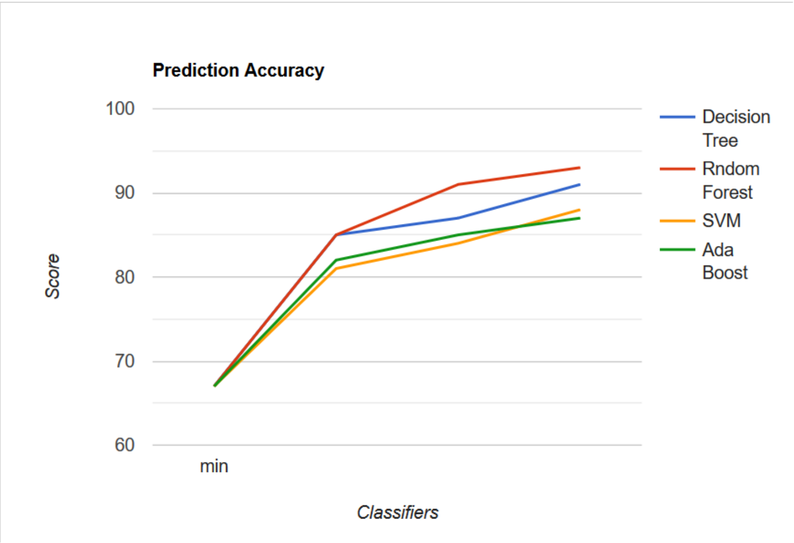


Figure 3 a) Line graph for comparing accuracy level of 4 classifiers

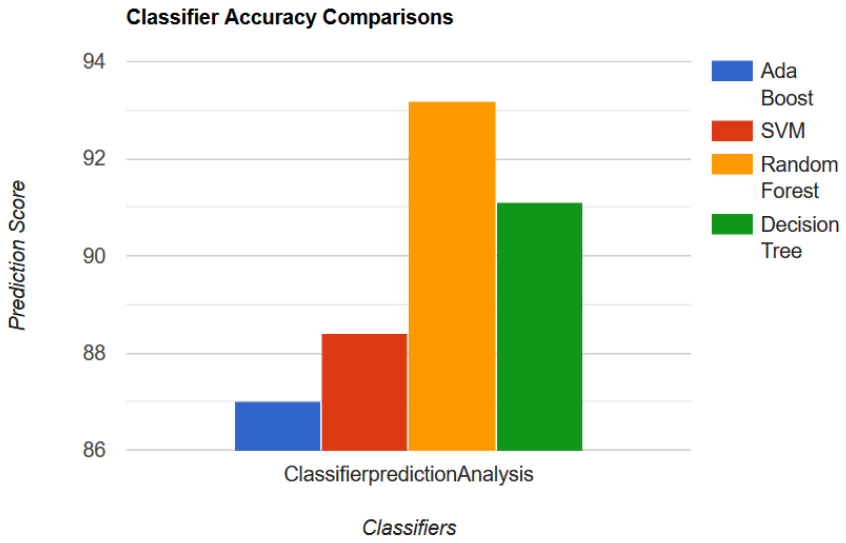


Figure 3 b) Bar graph for comparing accuracy level of 4 classifiers

Conclusion

Student’s career prediction is one of the important research areas in this current digital world. Traditionally various survey methods are used to predict the student’s career. But those methods take large amount of time to predict the result. This current digital world various computing techniques are used to predict the result in various domain. Student’s career data also predicted by using computing concepts like machine learning. Compared with traditional methods it takes less amount of time and it yields better result. In this research paper to predict the student’s career by using Adaboost, SVM, RF and DT approach. From the above mentioned concepts RF produces better result in term of accuracy value. The machine learning concepts are implemented by using the programming language Python. Using the output of this research work educational institutions take more care about the low level students and the recruiters select the suitable candidates for their companies.

References

[1] 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.

[2] K. B. Eashwar, R. Venkatesan and D. Ganesh (2017), “Student Performance Prediction Using SVM”, International Journal of Mechanical Engineering and Technology (IJMET) Vol. 8, No. 11, pp. 649–662.

[3] AnkitaKadambande, Snehal Thakur, AkshataMoholand A.M.Ingole (2017), “Predict Student Performance by Utilizing Data Mining Technique and Support Vector Machine”, International Research Journal of Engineering and Technology (IRJET), Vol. 04, No. 05, pp 2818-2821.

[4] S.A. Oloruntoba1 and J.L.Akinode (2017), “Student Academic Performance Prediction Using Support Vector Machine”, International Journal Of Engineering Sciences & Research Technology, pp. 588-598.

[5] AhmedSharafElDen ,Malaka A. Moustafa, Hany M. Harb and AbdelH.Emara (2013), “Adaboost Ensemble With Simple Genetic Algorithm For Student Prediction Model”, International Journal of Computer Science & Information Technology (IJCSIT) , Vol. 5, No 2, pp 73-85.

[6] MeimeiHan ,Mingwen Tong , Mengyuan Chen , Jiamin Liu and Chunmiao Liu (2017), “Application of Ensemble Algorithm in Students' Performance Prediction“, IEEE 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI).

[7] V. Prasannakumari B. Bavani, S. Nirmala SugirthaRajini, M.S. Josephine (2019), “Heart Disease Prediction System based on Decision Tree Classifier”, Jour of Adv Research in Dynamical & Control Systems, Vol. 11, 10-Special Issue, 2019, Volume 11, Issue 10.

[8] FaridJauhari and Ahmad AfifSupianto (2019), “Building student’s performance decision tree classifier using boosting algorithm”, Indonesian Journal of Electrical Engineering and Computer Science Vol. 14, No. 3, June 2019, pp. 1298 - 1304.