International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 4, Issue 1, January 2017, PP 46-51
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0401006
www.arcjournals.org

Predicting Academic Achievement in Programming for Problem Solving using Supervised Machine Learning Techniques

N Rajasekhar

Professor, Department of Computer Science and Engineering Institute of Aeronautical Engineering, Hyderabad, India

Abstract: *This project is an examination of factors that impact programming aptitudes. It examines the advancement of machine learning models to anticipate approaching students' performance. Our variables anticipate whether students will be strong or weak developers with 60 to 70% precision. Students find computer programming problematic and fight to pro the focus thoughts. Recognizing students who face difficulties in programming, is inconvenient and habitually educators don't have the thought about how well students get along until after the main examination. This examination may not occur until a couple of months after the module has started and whether or not the assessment is expressive of likely overall execution on the module, it may be passed the final turning point for students to pull back from the course or for educators to intervene to continue students from missing the mark. The factors investigated often as possible depending upon the students being involved with the module material and consequently, it is difficult to tell how farsighted comparative components would be at whatever point assessed before on the module. A model that could anticipate likely programming execution in the underlying stages and help with diminishing this issue isrecommendable.*

Keywords: *Machine learning, Prediction, Programming, Supervised techniques.*

1. INTRODUCTION

Programming is the process of execution of instructionsto perform a set of activities. It can be performed in many languages depending upon the type of application, platform availability, and time and space constraints. Itinvolves under- standing the problem, building algorithms, verification and validation of requirements, and implementation of algorithms in the target language. Programming helps to carry out the required activities in the desired wayand isthe onlywayto communicate with the computer. Hence a Programming language is difficult to be understood by a normal person. With the rapid evolvement of technology, there has been a vast increase in users of computers and smartphones. Most of the companies purely rely on computers to deal with huge datasets. It is necessary to learn programming as it provides solutions to maintain data, retrieving of information, make analysis, etc. There is a huge demand for programmers as the world is becoming tech-centric every day. Programming cannot be learned in a single day and hence needs a lot of practice and exposure to the conceptsinvolved.

Our thesis is to develop such a model which predicts the programming skills of students using supervised machine learning techniques. This prediction analysis can be made after a few weeks

of programming classes so that studentscan make their decision of either continuing or withdrawing from the course as early as possible. This analysis also helps them inimproving their weaker areas. Machine learning is the study of algorithms, patterns, and statistical analysis which provides the thinking ability to the computer. It develops computer programs that can access data and studies the patterns to use on their own. It automates the computer to do tasks by the knowledge of previously solved problems. Machine learning makes better predictions when compared to other techniques as it devises complex algorithms with historical data and outputs such that these algorithms analyze the patterns of the provided data and make predictions on new datasets. Machine learning is of four categories namely supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In our project pre- dictions using machine learning are done by supervised techniques. In supervised machine learning techniques, both input and output variables are provided, and a mapping function is determined that can be used for a new set of input and output variables. This helps in optimizing performance and is highly accurate when compared to other machine learningtechniques.

Predicting students' academic achievement in programming within the first few weeks of the course using supervised machine learning techniques. This prediction helps students develop the domains in which they scored low and also helps in the decision making of either continuing or withdrawing from course. The results of the prediction save time for students, helps the faculty identify struggling students, therefore students can be given enough attention to cope up with the programming courses.

Thorough analysis of supervised algorithms of machine learning and proposing algorithm(s) based on the accuracy depicted by them; therefore they can be used for future prediction purposes. The investigation led right now upon built up look into by applying the utilization of an alot bigger and various arrangement of highlights than are nor- mally considered in investigations of this nature. Its primary commitments are developing existing exploration by fusing various information mining methods into a solitary pipeline, includinghighlightascription,includedeterminationutilizing hereditary calculations, and irregular backwoods with hyper- parametertuning.

2. LITERATURE SURVEY

Many types of research have been carried out to estimate students' performance for the past few years. Different pa- pers were published describing the need for predictions and the procedure for carrying out the predictions. Few kinds of research focused on the process of predictions using machine learning and the other few pointed out the advantages of performing predictions using machine learning techniques.

Hussein Altabrawee, Osama Abdul Jaleel Ali and Samir Qaisar Ajmi [1], to tackle the issue of recognizing the students who have a poor scholastic exhibition in the soft- ware engineering subject offered by Al-Muthanna University, College Of Humanities, and four order models have been worked to foresee the presentation of the students. Four AI systems, completely associated feed-forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, have been utilized. The models have been contrasted with each other utilizing the ROC file execution measure and the grouping precision. ANN model has the most elevated ROC file that rises to 0.807 and exactness of77.04. Likewise, the choice tree model demonstrated that not all the qualities include in the arrangement procedure. PC Grades-Course1, Accommodation, Interest in examining PC, Educational Environment Satisfaction, and the Residency are the property utilized by the choice tree model. Four arrangement models have been made and tried utilizing four AI methods, completely associated feed-forward Artificial Neural Network, Naïve Bayes, Logistic Regression, and De- cision Tree. Table 1 shows the exactness and the presentation measures for each model just as thelattices.

Table1. *Performance metrics of different models.*

Model	TP	FP	TN	FN	Prec.	Recall	F-M.	Acc.	Error.	ROC
ANN	67	18	57	19	79.17	77.62	78.47	77.04	22.96	0.807
DT	67	19	56	19	77.96	77.83	77.88	76.93	23.61	0.762
LR	62	17	58	24	79.23	71.91	74.87	74.57	25.47	0.767
NB	55	23	52	31	70.51	64.27	67.21	66.52	33.48	0.697

Mushtaq Hussain, WenhaoZhu, Wu Zhang, Syed Muham- mad Raza Abidi, Sadaqat Ali [2], This examination explored thecapacitytoforeseeanstudent'sdifficultyforaconsequent coursework meeting utilizing a TEL framework and the MATLAB programming language. They separated students' info highlights and yield factors (e.g., the mean evaluations of students' in a meeting) from the TEL framework. To start with, they prepared the models (LR, ANN, SVM, NBC, and DT) utilizing preparing information (from meetings 1–4) that depended on completely input includes and tried the models on testinginformation.

Geurts, Irrthum, and Wehenkel [3], contend that learning trees are among the most mainstream calculations of Ma- chine Learning because of three principle qualities: inter- pretability, adaptability, and usability. Interpretabilityimplies that the model built to delineate element space into theyield space is straightforward since it is a guide of on the off chance that rules. Attention to that the tree models are simpler to disclose to individuals than straight relapse sinceit reflects more human dynamic than other prescient models. Moreover, it is reasonable to the effect of extra factors to the model, being particularly applicable to the investigation of steady legitimacy. It likewise evaluates which variableor blends of them, better predicts a given result, just as computes which cut off esteems are maximally prescient of it.

S.B. Kotsiantis, G.E. Tsekouras, and P.E. Pintelas [4], the technique that utilizes various subsets of preparing information with a solitary learning strategy is the boosting algorithm. It allows loads to the preparation occurrences, and this weight esteems are changed relying on how well the related preparing occasion is found out by the classifier; the loads for misclassified cases are expanded. After a few cycles, the forecast is performed by taking a weighted vote of the expectations of every classifier, with the loads being relative to every classifier's exactness on its preparation set. Boosting has three primary fundamental tuningparameters:

- The size of the set, which is equivalent the number of trees togrow,
- The shrinkage parameter, which is the pace of gaining starting with one tree then onto the next, and
- The unpredictability of the tree, which is the number of conceivable terminal hubs is normally set to 0.01 or to 0.001, and that the littler the estimation of, the most elevated should be the number of trees, so as to accomplish greatforecasts.

Boosting is the general technique used to lessen the mistake of learning calculations. Freund and Schapire [5] presented two variations of boosting calculation, adaBoost.M1, and ad- aBoost.M2 which can join with different calculations as the frail student. The consequence of shows that AdaBoost beat C4.5, and AdaBoost can improve the presentation of C4.5 when C4.5 was utilized as its powerless student. There are numerous enquiries about characterization and expectationsin different fields that utilizeAdaBoost.

Thomas colloney, Wilson Adaljo [6], Bagging is a free gathering based strategy. The point of this technique is to build the exactness of shaky classifiers by making a composite classifier, at that point consolidate the yields of the scholarly classifiers into a solitary forecast. The Bagging calculation begins with resampling the first information into various preparing informational collections which called bootstraps, and each bootstrap test size is equivalent to the size of the first preparing set. All bootstrap tests will be prepared to utilize various classifiers. Singular classifiers results are then consolidated through larger part voteprocess, the class picked was by mostclassifiers is the troupe choice. Hastie et al., Flach, James et al [7], Bagging is the shortform for bootstrap Aggregation, and is a general strategy for diminishing the fluctuation of arrangement trees. The strategy produces various bootstraps from thepreparation

set, growing a tree that allocates a class to the districts of the element space for each. Finally, the class of areas of each tree is recorded and the greater part vote is taken. The greater part vote is essentially the most regularly happening class overall trees. The out-of-pack blunder can be registered as a valid gauge of the testing mistake for the stowed model since the reaction for every perception is anticipated utilizing just the trees that were not fit utilizing that observation. Bagged trees have two primary fundamental tuningparameters:

- The quantity of highlights utilized in the forecast isset as the absolute number of indicators in the element space.
- The size of the bootstrap set, which is equivalent the number of trees todevelop.

Noah H. Gilbert [8], Research on hereditary calculations and decision trees has additionally been investigated in detail. Specialists at Zhejiang Gongshang University characterized cell phone clients into various utilization levels utilizing hereditary calculations to advance the bitwise portrayals of the list of capabilities and trait loads. In crafted by Balaet. al the attention was on general example arrangement, and not explicit to student information. Essentially, crafted by Hansen et. al. concentrated on the grouping of Peptides utilizing arbitrary backwoods and hereditary calculations to lead includedetermination.

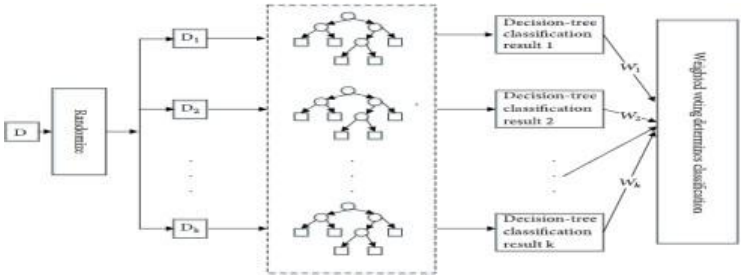

Fig1. *Methodology of weighted Random Forest.*

The calculation proposed shown in figure 1, has likewise been applied in the forecast of worker turnover in enterprises, for example, instruction, clinical, fund, and different fields. Hudson f. Golino, and Cristiano mauro a. Gomes [9], Random forest takes an irregular subsample of the first informational collection with substitution to developing the trees, just as chooses a subsample of the component space at every hub, so the quantity of the chose highlights (factors) is littler than the number of complete components of the element space: As focuses, the estimation of is held steady during the whole system for developing the forest and normally is set to an arbitrarily subsampling thefirst example and the indicators, [10]Random Forest improves the packed away tree technique by de-correlating thetrees.

3. PROPOSED METHODOLOGY

The prediction process contains various steps such as Data collection, Data Preprocessing, Implementation of Learning trees, Bagging, Boosting, Random forest, Prediction out- comes and graphical representation of results. These elements thoroughly discussed in the following sections.

A. Data Collection

The dataset is gathered from the Archeology office and the Sociology branch of the school of Humanities at Al- Muthanna University during the 2015 and 2016scholarly

years. Two information sources have been utilized, overview gathered from thestudents and the students' evaluations of information records. The dataset contains 151 student records, 66 male and 85 females. The dataset contains twenty properties. The characteristics can be isolated into five classifications which are close to home and way of life, examining style, family related instructive condition fulfillment, and student'sevaluations.

B. Data Preprocessing

The training information is utilized to ensure the machine perceives designs in the information, the testing dataset information is utilized to guarantee better exactness and proficiency of the calculation used to prepare the machine, and the test information is utilized to perceive how well the machine can predict for new datasets dependent on its preparation.

C. Learning Trees

A learning tree segments the component space into a few unmistakable fundamentally unrelated areas (non-covering). Every area is fitted with a model that plays out the marking capacity, assigning one of the classes to that specific space. The class is relegated to the locale of the component space by recognizing the larger part class in that district. To show up in an answer that best isolates the whole element space into progressively unadulterated hubs (locales), recursive paired segments are utilized. A hub is viewed as unadulterated when 100% of the cases are of a similar class, for instance, low scholarly accomplishment. A hub with 90% of low accomplishment and 10% of high accomplishment understudies is progressively "unadulterated" at that point a hub with half of each. Recursive parallel parcels function asfollows. The component space is part into two districts utilizing a cutoff from the variable of the element space that prompts the most virtuesetup.

Within the sight of overfitting, the mistakes will introduce an enormous fluctuation from the preparation set to the test set utilized. Also, the characterization tree does not have a similar prescient precision as other old-style AI draws near. To forestall overfitting, the difference issue and further-more to expand the expectation precision of the arrangement trees, a procedure named outfit trees can be utilized.

D. Bagging

Bagging is short form of Bootstrap aggregation. It is an example of a dataset with substitution. This implies another dataset is made from an arbitrary example of a current dataset where a given line might be chosen and added more than once to the example. It is a helpful way to deal with use when evaluating qualities, for example, the mean for a more extensive dataset, when you just have a restricted dataset accessible. By making tests of your dataset and evaluatingthe mean from those examples, you can take the normal of those appraisals and improve thought of the genuine mean of thehiddenissue.Thisequivalentmethodologycan beutilized with AI calculations that have a high change, for example, decisiontrees.

E. Boosting

Boosting utilizes all in positions at every reiteration, except keep up a load for each example in the preparation set that mirrors its significance; altering the loads makes the student centreon various occurrences thus prompts various classifiers. With both, the numerous classifiers are then joined by Voting to frame a composite classifier. Boosting allows diverse democratic qualities to part classifiers based on their precision. The boosting technique centreson the examples for an informational index which includes preparing each new model occurrences from the blunder or mis- classification of the past one to produce prescient models. Although, it has demonstrated to have higher expectation precision contrast with sacking however endure significant confinement of overfitting. Boosting alludes to a general and provably effective technique for creating an extremely exact forecast rule by consolidating unpleasant and decently erroneous guidelines. The accompanying subsections depict bothtechniques.

F. Random Forest

Rather than specifying all qualities for input traits in search if the split with the most reduced cost[10], we can make an example of the info credits to consider. This example of info characteristics can be picked haphazardly and without substitution, implying that each information quality needs possibly considered once when searching for the split point with the most reduced expense. We can see observe a rundown of highlights is made by haphazardly choosing highlight files and adding them to a rundown (called highlights), this rundown of highlights is then identified and explicit qualities in the preparation dataset assessed as split focuses. The fundamental thought behind this is to join different choice trees in deciding the last yield instead of depending on singular choicetrees.

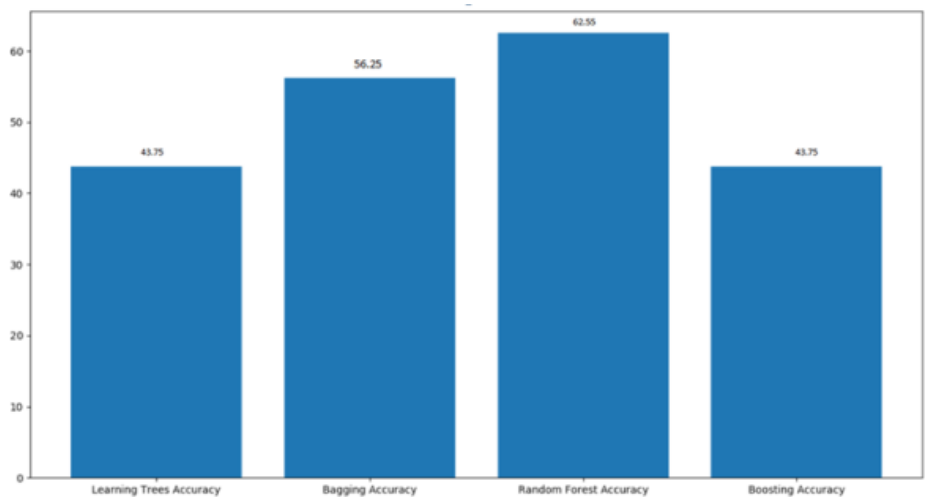

Figure 2 shows the Graphical presentation of performance of Algorithms.

Random selection feature of Random Forest algorithm made it achieve the highest accuracy among the fouralgorithms

4. CONCLUSION

Learning Tree's and Boosting's accuracy was least of all with 43.75% followed by Bagging's with 56.25%. Random Forest algorithm made the most stable predictions with 62.55%.

REFERENCES

[1] DorinaKababchieva. (2012). Student Performance Prediction using Data Mining Classification Algorithms. International Journal of Computer Science and Management Research,vol. 1.

[2] Edin Osmanbegovic and Mirza Suljic. (2012). Data mining approach for predicting studentperformance. Journal of Economics and Business, vol. X, Issue 1.

[3] Carlos Marques-Vera and Alberto Cano. (2013). Predicting student failure at school usinggenetic programming. ApplIntell, vol. 38, pp.315–330.

[4] Shaobo Huang and Ning Fang. (2013). predicting student academic performance in an engineering dynamic course: A comparison of four types of predictive mathematical models.Computers & Education, vol. 61, pp. 133–145. https://doi.org/10.1016/j. compedu.2012.08.15

[5] Ajay Kumar Pal and Saurabh Pal. (2013). Data Mining Techniques in EDM for Predicting the Performance of Students. International Journal of Computer and Information Technology, vol. 02, Issue 06.

[6] Ramanathan, Saksham Dhanda and Suresh Kumar D. (2013). Predicting Student Performance using Modified ID3 Algorithm. International Journal of Engineering and Technology, vol. 5 No 3.

[7] Alaa Khalaf Hamoud. (2016). Selection of Best Decision Tree algorithm for prediction andclassification of student Action. American International Journal of Research in Science,Technology, Engineering & Mathematics, vol 1, pp. 26-32.

[8] Dech Thammasiri, Dursun Delen, Phayung Meesad and NihatKasap. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications, 41, pp.321–330. https://doi.org/10.1016/ j.eswa.2013.07.046

[9] Ya-Han Hu, Chia-Ling L and Sheng-Pao Shih. (2014). Developing early warning systemsto predict students' online learning performance. Computers in Human Behavior, 36, pp.469–478. https://doi.org/10.1016/j.chb.2014.04.002

[10] SreckoNatek and Moti Zwilling. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert Systems with Applications, 41,pp.6400–6407. https://doi.org/10.1016/j.eswa.2014.04.024