Student’s performance using Classification Algorithms

Plabon Banik
Department of Computer Science and Engineering
East West University
Dhaka, Bangladesh
2019-1-60-167@std.ewubd.edu

Abstract

Student performance means a prediction about the performance of a student. This is a unique idea mostly invented by the youths. This survival depends on a lot of parameters and values. That's why we have come up with an idea to predict student performance. We have used Classification algorithms of machine learning to find a good way of success in a startup. We have taken the dataset from Kaggle to use our algorithms. We have implemented 7 different algorithms on this dataset to get which algorithm suits best for this dataset. The algorithms are Decision tree, Random forest, K-Nearest neighbors, MLP, Naive Bayes SVM, and Logistic regression. From different algorithms, we got different scores of efficiency. From them, we found the Random forest best fit for our dataset. This algorithm will help the newbies to know the condition.

Introduction

Recently, online systems in education have increased, and student digital data has come to big data size. This makes it possible to draw rules and predictions about the students by processing educational data with data mining techniques. All kinds of information about the student’s ethnicity, parental level of education, lunch, test preparation course, math score, reading score, and writing score can be uto for predict abou situation of a student. In this study, the successes of the students at the end of the semester are estimated by using the student data from our dataset. The aim of this study is to predict the students’ final grades to support the educators to take precautions for the children at risk. A number of data preprocessing processes were applied to increase the accuracy rate of the prediction model. A wrapper method for feature subset selection was applied to find the optimal subset of features. After that, some popular data mining algorithms (Decision tree, Random forest, K-Nearest neighbors, MLP, Naive bayes, SVM, Logistic regression) were used and compared in terms of classification accuracy rate.

Literature review

A critical increase in predicting student performance requirements is a result of the increasing interest by universities to raise the level of student performance and education to keep pace with the development taking place in various aspects of current life [12]. Various techniques were applied to predict student performance; each of these techniques belongs to different areas such as artificial neural network, machine learning, and collaborative filtering. Effectiveness of transfer from deep neural network learning was investigated in order to predict the performance of a higher-education student by [13]. Experiment was organized based on five compulsory courses dedicated for two undergraduate systems. Empirical results show that the prognosis of students at risk of failure can be achieved with good accuracy in the majority of cases. The system of student performance prediction based on MyMediaLight which is an open source recommendation system was proposed by [14]. This system is applied for GPA database of the student that was collected previously from university. A technique of biased matrix factorization (BMF) was proposed by authors to predict student performance to assess them and to choose appropriate courses. Techniques’ combination was suggested in [15] to develop a new prediction for a student performance model. The gray model and the Taylor approximation method were combined together to achieve the best results by computing approximations several times to improve the predictive accuracy of two gray models. The results obtained from this research can help both education administrators and educators to choose better solutions to raise the performance of a student who is experiencing instability in the learning process, as well as the matrix factorization, restricted Boltzman machine, and collaborative filtering techniques was used by [16] to analyze systematically the data that has been collected from the academic management system. The obtained results show that the better technique among the previously mentioned techniques is the restricted Boltzmann machines. Due to its effectiveness and simplicity, the collaborative filtering algorithm is used in the recommendation system. However, effectiveness of such techniques is limited due to the data sparsity, and it restricts the further improvement of prediction results. Thus, there is more interesting on the model that consist of combination of deep learning and historical data prediction algorithm. To achieve more accurate latent features, a model based on the quadratic polynomial regression model was proposed in [17]; in this model, traditional algorithm of matrix factorization was improved; then, the input data into deep neural network is the latent features. Implementation of the proposed model on three different datasets show significant improvement in efficiency prediction compared to the traditional models. The new model proposed in [18] consists of combination between deep learning and collaborative filtering model. The feedback in the neural network during the prediction process works in the form of simulations of the interaction process between the student and the educational institution. The preprocessed features will used as the input of neural network. The proposed model is implemented on ten million samples of MovieLens dataset and dataset of one million samples of Movielens to verify the performance of the proposed model which obtained very good results. Other approaches were suggested to improve prediction of student performance which is found in [19, 20]. Numerous research studies were proposed previously; these research studies take into account the issue of prediction of student performance by using the theory of machine learning; there is still room for improvement which is the student performance prediction factor analyzing based on data the transformation technique and the explanation model. The main aim of this research proposed a new approach by taking into account deep learning technique represented by long short-term memory (LSTM) with the use of time-based features.

DATASET DESCRIPTION

Startup data is a dataset of Kaggle on student performance prediction. It provides information about the ethnicity,parental level of education, lunch, test preparation course, math score, reading score, writing score. There are 9 columns which are the features of this dataset. There are 3 integers, 6 strings. The data types included in this dataset are categorical and numerical there and no missing values also. Data Sets have three general properties. Dimensionality, which the number of attributes that the objects in a data set have. This dataset has a high number of attributes that's why it has high dimensionality. Secondly, Sparsity happens when a significant portion of the variable's cells do not contain any actual data. This type of empty, or Null value and it includes missing values also. Lastly Resolution, the level of resolution affects the patterns in the data. A pattern may not be seen or take as noise data if the resolution is too fine; if the resolution is too coarse, the pattern may disappear. The machine learning algorithms which have been used in this paperwork with this student perfomance dataset. These ML models use this dataset for training purposes. A previously processed dataset is frequently split into different portions, which is required to verify how successfully the model's training went. A testing dataset is normally isolated from the data for this reason. Next, a validation dataset can help to avoid training the algorithms on the same sort of data and making a biased prediction on the student perfomance dataset whether it will good/average/score**.**

Dataset Cleaning

The data has different types of values; some are categorical, some are numerical and some or them are in strings. We have cleaned the dataset in such a way that we find a close & appropriate result. At first we selected such data sets which were unnecessary or irrelevant with our desired result or deciding factor. We have removed those columns of data using drop function which are Gender and Ethinic. Later on, we haven’t any columns with null or missing values. We don’t need filled those values. There are some string values in our dataset; we have replaced into integer with replace function. Finally doing so, our data cleaning or mining process will be done. In the next step we will use algorithms to find the accuracy rate.

Description of used algorithms

We have used 7 different types of algorithm in this dataset to find different results generated from this dataset. They are given bellow with a short description.

Decision Tree

Decision tree algorithms belong to a family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithms can also be used to solve regression and classification problems. The purpose of using the decision tree is to create a training model that can be used to predict the class or value of a target variable by learning a simple decision rule derived from previous (training) data.The decision tree starts at the root of the tree to predict the class label of the record. Compare the value of the root attribute with the attribute of the record. Based on the comparison, follow the branch that matches that value and jump to the next node.
Types of Decision Trees:

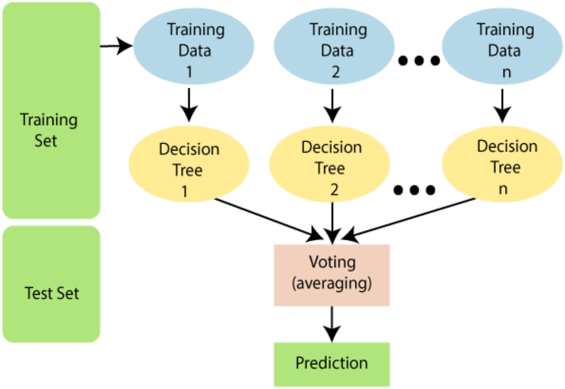
1. **Categorical Variable Decision Tree**
2. **Continuous Variable Decision Tree**

Random Forest

Random Forest is a well-known machine learning calculation that features a place to directed learning Strategy. It are often utilized for Classification and Regression issues also in Machine Learning. it's supported the concept of gathering learning, which might be a handle of mixing numerous classifiers to unravel a fancy issue and to progress the execution of the demonstrate.

Properies of random forest are given bellow:

1. It takes less preparing time as compared to other calculations.
2. It predicts yield with tall precision, indeed for the expansive dataset it runs effectively.
3. It can to keep up exactness when an expansive extent of information is lost.



K-Nearest Neighbour

A k-nearest-neighbor algorithm, often abbreviated k-nn, is an approach to data classification that estimates how likely a knowledge point is to be a member of 1 group or the opposite counting on what group the info points nearest thereto are in. The k-nearest-neighbor is an example of a "lazy learner" algorithm, meaning that it doesn't build a model using the training set until a question of the info set is performed.The k-nearest neighbors (KNN) calculation could be a basic, easy-to-implement administered machine learning calculation which can be utilized to illuminate both classification and relapse issues. The k-nearest neighbors (KNN) calculation could be a basic, easy-to-implement administered machine learning calculation which can be utilized to illuminate both classification and relapse issues. "Lazy learner" algorithms, just like the k-nearest-neighbor, don't create a model of the data set before using it. Calculations are only performed if the data point's neighbors are asked to be polled. This makes k-nn a breeze to use in processing applications.

Support vector machine

A support vector machine (SVM) may be a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for every category, they're ready to categorize new text. Compared to newer algorithms like neural networks, they need two main advantages: higher speed and better performance with a limited number of samples (in the thousands). This makes the algorithm very suitable for text classification problems, where it's common to possess access to a dataset of at the most a few of thousands of tagged samples.

Naive Bayes

A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of knowledge points. Popular uses of naive Bayes classifiers include spam filters, text analysis and diagnosis . These classifiers are widely used for machine learning because they're simple to implement.Naive Bayes is additionally referred to as simple Bayes or independence Bayes.

Logistic regression

Logistic regression is that the acceptable statistical method to conduct when the variable is dichotomous (binary). Like all regression analyses, the logistic regression could also be a predictive analysis. Logistic regression is used to elucidate data and to explain the connection between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.It's important to understand the following assumptions regarding logistic regression before going into its implementation:

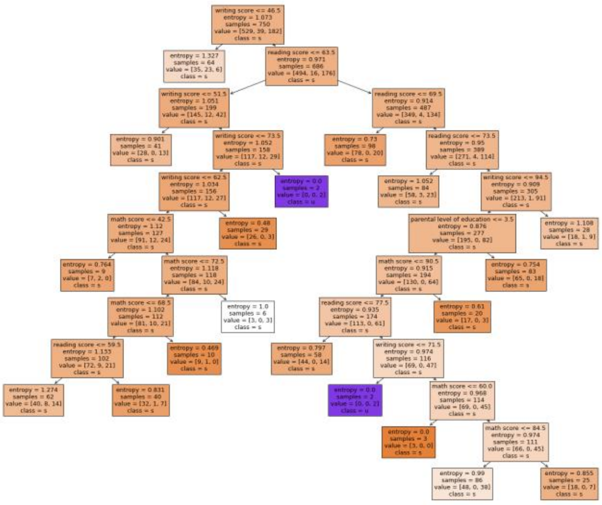
1. It's important to note that when using binary logistic regression the target variables must always be binary. The intended outcome is indicated by factor level 1.
2. As a result, the independent variables in the model must be independent of each other.
3. Meaningful variables must be included as part of the analysis process.
4. For logistic regression, we need use a big sample size.

Multi-layer perceptron

A multilayer perceptron (MLP) may be a feedforward artificial neural network that generates a group of outputs from a group of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network. MLP may be a deep learning method.

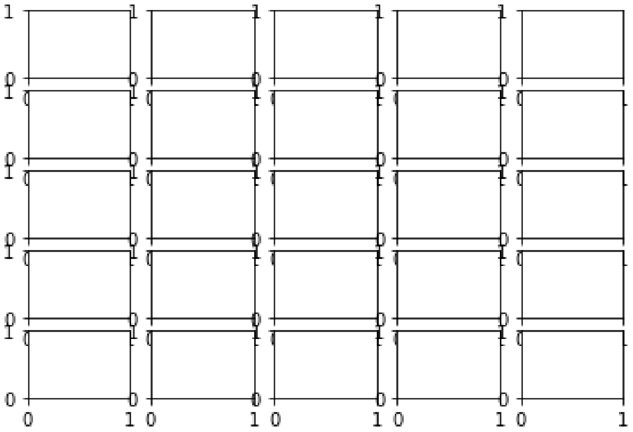
Result & Analysis

We have implemented 7 different types of algorithm on this data set and analyzed the percentage of score for different types of algorithms. The deciding factor in this data set is "student performance" column which had three values, good, average and bad. We have only keeped those columns which adds a real value this dataset. Those values like gender or droped is irrelevant that's why they were droped. We have kept neccesarry columns in from the data set. Then we have split the dataset into two part. One is train data other is test data. From the data set we have taken 25% data as test data and other 75% data as train data.


writing score <= 46.5 entropy = 1.073 samples = 750 value = [529, 39, 182] class = s
entropy = 1.327 samples = 64 value = [35, 23, 6] class = s
reading score <= 63.5 entropy = 0.971 samples = 686 value = [494, 16, 176] class = s
writing score <= 51.5 entropy = 1.051 samples = 199 value = [145, 12, 42] class = s
reading score <= 69.5 entropy = 0.914 samples = 487 value = [349, 4, 134] class = s
entropy = 0.901 samples = 41 value = [28, 0, 13] class = s
writing score <= 73.5 entropy = 1.052 samples = 158 value = [117, 12, 29] class = s
entropy = 0.73 samples = 98 value = [78, 0, 20] class = s
reading score <= 73.5 entropy = 0.95 samples = 389 value = [271, 4, 114] class = s
writing score <= 62.5 entropy = 1.034 samples = 156 value = [117, 12, 27] class = s
entropy = 0.0 samples = 2 value = [0, 0, 2] class = u
entropy = 1.052 samples = 84 value = [58, 3, 23] class = s
writing score <= 94.5 entropy = 0.909 samples = 305 value = [213, 1, 91] class = s
math score <= 42.5 entropy = 1.12 samples = 127 value = [91, 12, 24] class = s
entropy = 0.48 samples = 29 value = [26, 0, 3] class = s
parental level of education <= 3.5 entropy = 0.876 samples = 277 value = [195, 0, 82] class = s
entropy = 1.108 samples = 28 value = [18, 1, 9] class = s
entropy = 0.764 samples = 9 value = [7, 2, 0] class = s
math score <= 72.5 entropy = 1.118 samples = 118 value = [84, 10, 24] class = s
math score <= 90.5 entropy = 0.915 samples = 194 value = [130, 0, 64] class = s
entropy = 0.754 samples = 83 value = [65, 0, 18] class = s
math score <= 68.5 entropy = 1.102 samples = 112 value = [81, 10, 21] class = s
entropy = 1.0 samples = 6 value = [3, 0, 3] class = s
reading score <= 77.5 entropy = 0.935 samples = 174 value = [113, 0, 61] class = s
entropy = 0.61 samples = 20 value = [17, 0, 3] class = s
reading score <= 59.5 entropy = 1.155 samples = 102 value = [72, 9, 21] class = s
entropy = 0.469 samples = 10 value = [9, 1, 0] class = s
entropy = 0.797 samples = 58 value = [44, 0, 14] class = s
writing score <= 71.5 entropy = 0.974 samples = 116 value = [69, 0, 47] class = s
entropy = 1.274 samples = 62 value = [40, 8, 14] class = s
entropy = 0.831 samples = 40 value = [32, 1, 7] class = s
entropy = 0.0 samples = 2 value = [0, 0, 2] class = u
math score <= 60.0 entropy = 0.968 samples = 114 value = [69, 0, 45] class = s
entropy = 0.0 samples = 3 value = [3, 0, 0] class = s
math score <= 84.5 entropy = 0.974 samples = 111 value = [66, 0, 45] class = s
entropy = 0.99 samples = 86 value = [48, 0, 38] class = s
entropy = 0.855 samples = 25 value = [18, 0, 7] class = s

Decision Tree generated in python

Then we have implemented algorithms using sklearn. In this dataset we have used decision tree, Random forest, K-Nearest Neighbour, MLP, SVM, Naive byes and logistic regression in this data set. After using these 7 algorithms we get different types of success rate score for each algorithm. The best algorithm according to the success rate is Decision tree and Random forest as we have got 98.26% for decision tree and 97.83% for random forest success rate by using this algorithm. The rest are bellow them. For example the MLP table is given bellow.

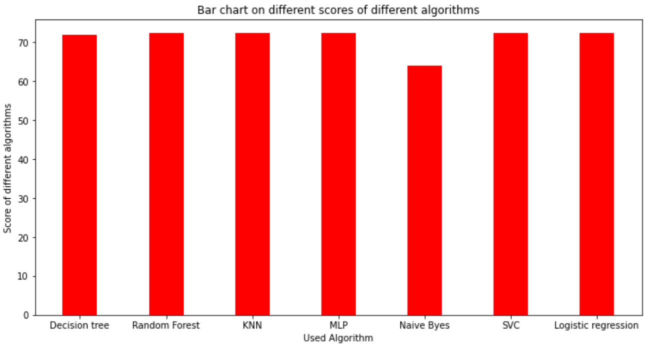


MLP graph

The other score for different algorithms are given bellow in the table.

Name	Accuracy Score(in Percentage)
Decision Tree	72.00
Random Forest	72.59
K-Nearest Neighbour	71.32
MLP	72.25
Naïve Bayes	64.00
Support Vector Mahine	71.66
Logistic Regression	75.55

From the table we can see Naïve Bayes and K-Nearest Neighbour gives the lowest and Logistic Regression gives the highest score. If we visualise them in a bar chart then they would look like this.



Bar chart on accuracy score of different Algorithms

The results satisfy our desired score. As the Logistic Regression best fits for this dataset, So we can predict a students performance.

Conclusion

While working on this dataset we have came acrossed different types of problems and challenges and we have overcome them by learning the solution. We have worked on this type of problem keeping in mind on the usage of different Machine learning algorithm and give benefit to others. This dataset is a real life dataset and we have come to an assumption that if anyone want best result from it, he should take the Logistic Regression in consideration as Logistic Regression has the highest accuracy rate for this types of dataset. If any student want to understand his/her quality about own studyby following the way generated from the Logistic Regression will be the best for him/her.

We hope this findings of ours will help others and add some value in machine learning industry.

REFERENCES

[1] Bruner, J.S. (2009) The Process of Education. Harvard University Press, Cambridge.

[2] Chinyoka, K. and Naidu, N. (2013) Uncaging the Caged: Exploring the Impact of Poverty on the Academic Performance of Form Three Learners in Zimbabwe. International Journal of Educational Sciences, 5, 271-281. https://doi.org/10.1080/09751122.2013.11890087

[3] Chindanya, A. (2012) Effects of Parental Involvement in the Education of Children. Unpublished D. Ed. Thesis, University of South Africa, Pretoria.

[4] Mustapha, H.S. and Benjamin, N. (2021) Ethics: An Insight into Psychological Research and Practice. Open Access Library Journal, 8, e7110. https://doi.org/10.4236/oalib.1107110

[5] Balali, G.I., Yar, D.D., Dela, V.G.A. and Adjei-Kusi, P. (2020) Microbial Contamination, an Increasing Threat to the Consumption of Fresh Fruits and Vegetables in Today's World. International Journal of Microbiology, 2020, Article ID: 3029295. https://doi.org/10.1155/2020/3029295

[6] Baskerville, D.J. (2020) Mattering; Changing the Narrative in Secondary Schools for Youth Who Truant. Journal of Youth Studies, 23, 1-16. https://doi.org/10.1080/13676261.2020.1772962

[7] Maynard, B.R., McCrea, K.T., Pigott, T.D. and Kelly, M.S. (2012) Indicated Truancy Interventions: Effects on School Attendance among Chronic Truant Students. Campbell Systematic Reviews, 8, 1-84. https://doi.org/10.4073/csr.2012.10

[8] Mchelu, A. (2015) The Effect of Long Commuting on Students' Academic Performance in Day Community Secondary Schools in Tanzania: A Case of Songea Municipal Council. The Open University of Tanzania, Dar es Salaam.

[9] Jackson, M., Jonsson, J.O. and Rudolphi, F.J. (2012) Ethnic Inequality in Choice-Driven Education Systems: A Longitudinal Study of Performance and Choice in England and Sweden. Sociology of Education, 85, 158-178. https://doi.org/10.1177%2F0038040711427311

[10] Mallett, C.A. (2016) Truancy: It's Not about Skipping School. Child and Adolescent Social Work Journal, 33, 337-347. https://doi.org/10.1007/s10560-015-0433-1

[11] Romero, D. and Molina, A. (2011) Collaborative Networked Organisations and Customer Communities: Value Co-Creation and Co-Innovation in the Networking Era. Production Planning & Control, 22, 447-472. https://doi.org/10.1080/09537287.2010.536619

[12] Keppens, G. and Spruyt, B. (2017) The Development of Persistent Truant Behaviour: An Exploratory Analysis of Adolescents' Perspectives. Educational Research, 59, 353-370. https://doi.org/10.1080/00131881.2017.1339286

[13] Onyele, C.V. (2018) Influence of Truancy on Academic Performance of Secondary School Students in Enugu East Local Government Area of Enugu Stste.

[14] Gottfried, M.A. (2019) Chronic Absenteeism in the Classroom Context: Effects on Achievement. Urban Education, 54, 3-34. https://doi.org/10.1177%2F0042085915618709

[15] Roman, M.D. (2014) Students' Failure in Academic Environment. Procedia-Social and Behavioral Sciences, 114, 170-177. https://doi.org/10.1016/j.sbspro.2013.12.679

[16] Walker, J.M., Shenker, S.S. and Hoover-Dempsey, K.V. (2010) Why Do Parents Become Involved in Their Children's Education? Implications for School Counselors. Professional School Counseling, 14, 2156759X1001400104. https://doi.org/10.1177%2F2156759X1001400104

[17] Shifrer, D. (2013) Stigma of a Label: Educational Expectations for High School Students Labeled with Learning Disabilities. Journal of Health and Social Behavior, 54, 462-480. https://doi.org/10.1177%2F0022146513503346

[18] Akben-Selcuk, E. and Altiok-Yilmaz, A. (2014) Financial Literacy among Turkish College Students: The Role of Formal Education, Learning Approaches, and Parental Teaching. Psychological Reports, 115, 351-371. https://doi.org/10.2466%2F31.11.PR0.115c18z3

[19] Thompson, L.J., Clark, G., Walker, M. and Duncan Whyatt, J. (2013) 'It's Just Like an Extra String to Your Bow': Exploring Higher Education Students' Perceptions and Experiences of Extracurricular Activity and Employability. Active Learning in Higher Education, 14, 135-147. https://doi.org/10.1177%2F1469787413481129

[20] Park, H., Byun, S.-Y., and Kim, K.-K. (2011) Parental Involvement and Students' Cognitive Outcomes in Korea: Focusing on Private Tutoring. Sociology of Education, 84, 3-22. https://doi.org/10.1177%2F0038040710392719

[21] Boi, B.L. (2020) The Influence of Home Environment on Learning Achievements among Students' in Public Day Secondary Schools in Mbulu Town Council-Tanzania. The University of Dodoma, Dodoma.