A Machine Learning Practice on NAS Dataset: Influence of Socioeconomic Factors on Student Performance

Kotrike Swetha, M. Imtiaz Ur Rahaman

***Abstract:** India's population is enormous and diverse due to which its education system is very complex. Furthermore, due to several reasons that they have grown up in different environmental situations. Over the years, several changes have been suggested and implemented by various stakeholders to improve the quality of education in schools. This paper presents a novel method to predict the performance of a new student by the analysis of historical student data records, and furthermore, we explore the NAS dataset using cutting edge Machine Learning Algorithms to predict the grades of a new student and take proactive measures to help them succeed. Similarly, NAS Dataset can also be worthwhile to the employee dataset and can predict the performance of the employee. Some of the Supervised Machine Learning Algorithms for Classification which have been successfully applied to the NAS dataset. Support Vector Machines and K-Nearest Neighbours algorithms did not crop results in coherent time for the given dataset; Gradient Boosting Classifier outperformed than all other algorithms reliably.*

***Keywords:** Indian Education System, National Achievement Survey, Machine Learning Algorithms, Supervised Learning, Gradient Boosting Classifier*

I. INTRODUCTION

Machine Learning is a field which is raised out of "Artificial Intelligence (AI)." Applying AI, We want to build better and smarter equipment. It is an idea that is to learn from many existing models and experiences without being programmed. Instead of typing the code, you have data about the general algorithm, and the logic is based on the given data. For example, we have a method to classify our data in a consistent classification, where we want to give each class a label and the number of specialized courses. In recent years, the growing interest in India's complex education sector has been associated with significant differences in social, economic, and geographical conditions. Every student in structured years grows in another spiritual environment [1]. Due to this, it becomes too ambiguous for students to find the best appropriate methodology for learning. The main aim of this paper is to predict the performance of the new student based on the old student record data. In this paper, the supervised machine learning algorithms have been applied on the NAS dataset among all Gradient boosting classifier has performed well and used in the dataset.

In this paper, we study the NAS dataset about VIII class students and their socioeconomic factors to build a system that can guide future students and parents. Performance of prospective students is predicted using the most popular and effective Supervised Machine Learning Algorithms. This system can further be easily extended to any education dataset for providing career guidance, etc. Gradient boosting is a particular type of technique[1] for reducing the errors and for building the predictive models. Gradient boost work involves three elements like Loss Function, Weak Learner, and Additive Model. The great success in the application was the Adaptive Boosting or AdaBoost, which is the First Boosting Algorithm. Gradient boosting involves four enhancements for fundamental of gradient algorithm like Tree constraints, Shrinkage, Random Sampling, Penalized learning.

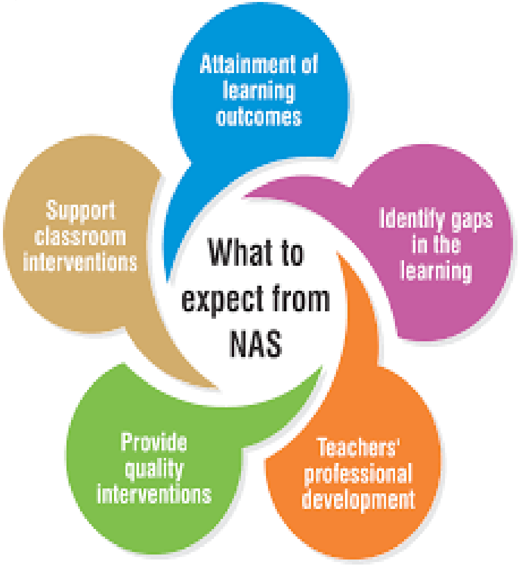

Attainment of learning outcomes
Support classroom interventions
What to expect from NAS
Identify gaps in the learning
Provide quality interventions
Teachers' professional development

Fig 1: Expectations of NAS

The above figure summarizes the purpose and processes of the National Achievement Survey. It has five different stages that begin with identifying the gaps in learning.

The remainder of the paper is organized as follows into seven sections supported by references Section II describes/outlines the literature review, Section III discusses/analyses the limit of the study, Section IV describes the functional architecture, Section V presents the Algorithm used and its performance, Section VI addresses/analyses (Experimental results) , Section VII concludes the paper.

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author

Kotrike Swetha*, (MTech), Department of CSE, G. Pulla Reddy Engineering College, Autonomous, Kurnool.

M. Imtiaz Ur Rahaman, Associate Professor, Department of CSE, G. Pulla Reddy Engineering College, Autonomous, Kurnool.

II. LITERATURE REVIEW

The literature on the subjected area shows a variety of approaches, the investigator high lights briefly the significance of research in Secondary education and summarizes the relevant studies that have been conducted in this area.

Socio-economic status (SES)

Many researchers (Krasno, Jonesson & Elder 2004) [9] found that their writing school was associated with academic achievement with their parents' social and economic status (SES). These socio-economic variables include the following study: Among others, gender, race, family background, neighbors, parents' academies, work, and family income. Major Bancers (1996) [10] has a direct link between family experience and student achievement.

Children's living conditions at home or home affect social efforts and performances. This statement was added to a study carried out by Ziona (2002) [11], which also influenced the academic performance of their children in the parenting of SES and their responsibility in the child's school and their involvement in cultural activities.

III. DELIMITATION OF THE STUDY

For this study, we analyzed the NAS dataset collected from VIII class students and their socioeconomic factors to build a system that can guide future students and parents. Performance of prospective students is predicted using the most popular and effective Supervised Machine Learning Algorithms.

Dataset declaration and its attributes: The National Achievement Survey 2014 dataset contains about 185000 student records with 64 columns or parameters per student. Of these, 60 columns are input parameters, and 4 columns are output parameters. The 60 input parameters are STUID, State, District, Gender, Age, Category, Same language, Siblings, Handicap, Father education, Mother education, Father occupation, Mother occupation, Below poverty, Use calculator, Use computer, Use Internet, Use dictionary, Read other books, Books, Distance, Computer, Library, Like school, Subjects, Give Lang HW, Give Math HW, Give Science HW, Give SoSc HW, Correct Lang HW, Correct Math HW, Correct Scie HW, Correct SocS HW, Help in Study, Private tuition, English is difficult, Read English, Dictionary to learn, Answer English WB, Answer English aloud, Maths is difficult, Solve Maths, Solve Maths in groups, Draw geometry, Explain answers, SocSci is difficult, Historical excursions, Participate in SocSci, Small groups in SocSci, Express SocSci views, Science is complicated, Observe experiments, Conduct experiments, Solve science problems, Express science views, Watch TV, Read magazine, Read a book, Play games, Help in household. The 4 output parameters are Maths %, Reading %, Science %, and Social %**.**

Data wrangling: Data wrangling is an essential step in any project involving data analytics. In this step, among several things, we remove columns that may not be relevant, change null or missing values (in columns) to appropriate defaults, convert non-integer fields to integers, etc., before applying Machine Learning Algorithms[3].

The following columns were dropped from NAS dataset for our research: STUID, District, and Subjects.

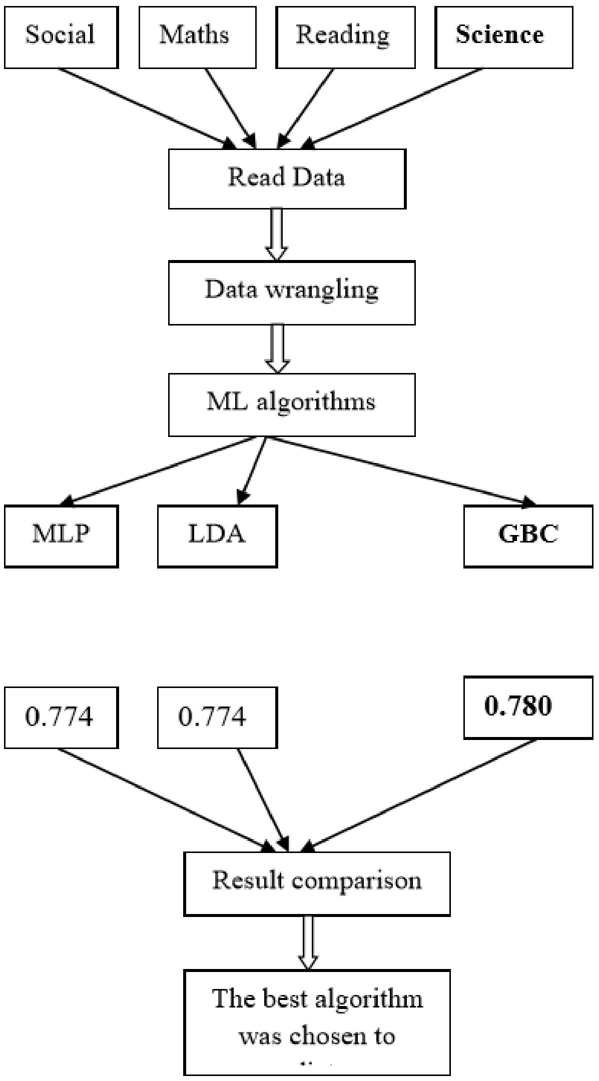
Further, the output columns: Maths %, Reading %, Science %, and Social % were removed after introducing a corresponding pass (1) or fail (0) output column. These output columns are named Maths Grade, Reading Grade, Science Grade, and Social Grade.

A student record is marked as fail (0) if the score is below 35. It is marked as a pass (1) if the score is above or equal to 35.

The missing or null values were filled with 0 before running Machine Learning Algorithms.

In this paper, we demonstrate the application of ML algorithms in the education sector, specifically, the NAS 2014 dataset about VIII class students' performance.

IV. FUNCTIONAL ATTRIBUTE



V. ALGORITHM USED AND ITS PERFORMANCE

The NAS [National Achievement Survey] dataset is performed on different algorithms of Supervised Machine Learning classifier such as

Logistic Regression[5], Multi-Layer Perceptron, Random Forest, Linear Discriminant Analysis, Gaussian Naive Bayes, Decision Trees, and Gradient Boosting has been applied on the dataset. Different test results are yield from different algorithms. Among all the algorithms Gradient boosting classifier has performed better results in all the four various fields. This NAS dataset is not having any existing algorithm [6].

Algorithms and parameters:

The results that are obtained from

```
pipeline = Pipeline([('normalizer', StandardScaler()), ('clf', LogisticRegression())])
pipeline.steps
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.0, random_state = 10)
fc=X_train.shape[1]
clfs = []
clfs.append(MLPClassifier(hidden_layer_sizes=(fc, fc, 1), max_iter=100, alpha=1e-4, solver='adam', verbose=10, tol=1e-6, random_state=1, learning_rate_init=.1))
clfs.append(RandomForestClassifier())
clfs.append(GradientBoostingClassifier())
clfs.append(LogisticRegression(multi_class='multinomial', solver='lbfgs', max_iter=1))
clfs.append(LinearDiscriminantAnalysis())
clfs.append(GaussianNB())
clfs.append(DecisionTreeClassifier())
clfs.append(SVC())
clfs.append(KNeighborsClassifier(n_neighbors=10))

for classifier in clfs:
   pipeline.set_params(clf = classifier)
   scores = cross_validate(pipeline, X_train, y_train)
   print('---------------------------------')
   print(str(classifier))
   print('-----------------------------------')
   for key, values in scores.items():
          print(key,' mean ', values.mean())
          print(key,' std ', values.std())
```

VI. RESULTS

Results of Supervised Machine Learning Algorithms for Classification like Logistic Regression, Multi-Layer Perceptron, Random Forest, Linear Discriminant Analysis, Gaussian Naive Bayes, Decision Trees, and Gradient Boosting are presented below. While Support Vector Machines and K-Nearest Neighbours algorithms did not yield results in a reasonable time for the given dataset, it has been observed that Gradient Boosting Classifier outperformed all other algorithms consistently.

Maths Grade Output:

Table 1. Maths Grade Output

Algorithm	**Results**
Logistic Regression	0.844665
MLPClassifier	0.846580
RandomForestClassifier	0.844416
linear discriminant analysis	0.846580
Gaussian	0.700606
Decision Tree Classifier	0.749573
GradientBoostingClassifier	**0.846758**

Reading Grade Output:

Table 2. Reading Grade Output

Algorithm	**Results**
Logistic Regression	0.617675
MLPClassifier	0.680902
RandomForestClassifier	0.667819
linear discriminant analysis	0.681102
Gaussian	0.550283
Decision Tree Classifier	0.585628
GradientBoostingClassifier	**0.698140**

Science Grade Output:

Table 3. Science Grade Output

Algorithm	**Results**
Logistic Regression	0.731370
MLPClassifier	0.774548
RandomForestClassifier	0.767405
LinearDiscriminantAnalysis	0.774548
Gaussian	0.605342
Decision Tree Classifier	0.661150
GradientBoostingClassifier	**0.780348**

Social Grade Output:

Table 4. Social Grade Output:

Algorithm	**Results**
Logistic Regression	0.690420
MLPClassifier	0.745543
RandomForestClassifier	0.734100
LinearDiscriminantAnalysis	0.745543
Gaussian	0.583113
Decision Tree Classifier	0.633753
GradientBoostingClassifier	**0.751418**

VII. CONCLUSION AND FUTURE SCOPE

In this paper, we applied various Machine Learning Algorithms to the NAS 2014 dataset to predict the pass or fail outcome of future students in 4 different subjects. It has been observed that Gradient Boosting Classifier has consistently outperformed other classification algorithms. These results help parents and teachers take appropriate proactive measures to improve the academic outcome of future students. While our study was conducted on binary output (pass or fail), further experiments and research can be performed with multi-class output on below lines:

O	10.00	Outstanding
A+	9.00	Excellent
A	8.00	Very Good
B+	7.00	Good
B	6.00	Above Average
C	5.00	Average
P	4.00	Pass
F	0.00	Fail

These Machine Learning Algorithms can further be applied to employment datasets to help employers predict the performance of future employees. Likewise, these algorithms can also help aspirants choose appropriate career tracks in various domains based on the performance of existing and similar aspirants in those domains.

REFERENCES

1. How to Configure the Gradient Boosting Algorithm by Jason Brownlee on September 12, 2016 in XGBoost
2. R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn and N. Radi, "Machine learning approaches to predict learning outcomes in Massive open online courses," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 713-720.
3. F. Sciarrone, "Machine Learning and Learning Analytics: Integrating Data with Learning," 2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET), Olhao, 2018, pp. 1-5.
4. F. Zhang, B. Du and L. Zhang, "Scene Classification via a Gradient Boosting Random Convolutional Network Framework," in IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 3, pp. 1793-1802, March 2016
5. Z. Wen, B. He, R. Kotagiri, S. Lu and J. Shi, "Efficient Gradient Boosted Decision Tree Training on GPUs," 2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Vancouver, BC, 2018, pp. 234-243.
6. A. Riccardi, F. Fernández-Navarro and S. Carloni, "Cost-Sensitive AdaBoost Algorithm for Ordinal Regression Based on Extreme Learning Machine," in IEEE Transactions on Cybernetics, vol. 44, no. 10, pp. 1898-1909, Oct. 2014.
7. A. Gulati, A. Batra, R. Khurana and M. M. Tripathi, "Cognitive learning recommendation system in Indian context," 2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH), Hyderabad, 2017, pp. 1-6
8. Gifty Esi Barnes (Author)James Sunney Quaicoe (Author), 2012, The influences of selected socio-economic factors of parents and parenting attitudes on the academic achievements of their wards, Munich, GRIN Verlag.
9. Crosnoe, R., Johnson, M. K., & Elder, G. H. (2004). School size and the interpersonal side of education: An examination of race/ethnicity and organizational context. Social Science Quarterly, 85(5), 1259-1274.
10. Majoribanks, Kevin. 1996. Family Learning Environments and Students' Outcomes: A Review.Journal of Comparative Family Studies, 27(2), 373-394.
11. Jeynes, William H. 2002. Examining the effects of parental absence on the academic achievement of adolescents: the challenge of controlling for family income. Journal of family and economic issues, 23(2)