International Journal of Advanced Development in Science and Technology
Volume : x Issue : xx xx /xx /xxxx www.ijadst.com e-ISSN: 2582-1059

# STUDENT PERFORMANCE ANALYSIS USING ENSEMBLE ALGORITHM

[1]Anupriya. V, [2] Supraja.V,[3]Mrs R.Gayathri
[1,2,3]Department of IT, Meenakshi Sundararajan Engineering College.
[1]anupriyavasu2000@gmail.com.

**ABSTRACT**

*Performance analysis of outcome based on learning is a system which will strive for excellence at different levels and diverse dimensions in the field of student's interests. This paper proposes a complete EDM framework in a form of a rule-based recommender system that is not developed to analyse and predict the student's performance only, but also to exhibit the reasons behind it. The proposed framework analyses the students' demographic data, study related and psychological characteristics to extract all possible knowledge from students, teachers and parents. Seeking the highest possible accuracy in academic performance prediction using a set of powerful data mining techniques. The framework succeeds to highlight the student's weak points and provide appropriate recommendations. The realistic case study that has been conducted on 200 students proves the outstanding performance of the proposed framework in comparison with the existing ones.*

***Keywords: dataset, machine learning algorithm Ensemble methods, python, predicting the accuracy of result, student result analysis, EDM frame work.***

## 1. INTRODUCTION

An education system is one of the most important parts for the development of any country. So it should be taken very seriously from its start. Most of the developed countries have their own education system and evaluation criteria. Now a day's education is not limited to only the classroom teaching but it goes beyond that like Online Education System, MOOC course, Intelligent tutorial system, Web-based education system, Project based learning, Seminar, workshops etc. But all these systems are not successful if they are not evaluated with accuracy. So, for making any education system to success, a well-defined evaluation system is maintained. Every educational institution generates lots of data related to the registered student and if that data is not analysis properly then all afford is going to be wasted and no future use of data happen.This institutional data is related to the student admission, student family data, student result etc. Every educational institution applies some assessment criteria to evaluate their students. For decades, the students are analysed through various numbers of processes. Student's performance evaluation is as important as that of the study process of the student.The regular and traditional procedure is to take examinations or class assessments etc. This process takes a lot of manual effort to complete the evaluation and it is very time consuming as well. This paper proposed the usage of machine learning in the evaluation of the performance of the students. It does not only evaluate the performance but also helps in improving various aspects of the students.A lot of evaluation process may put the student into a lot of stress. Keeping the student in stress is not going to help the student perform well. Performance evaluation of students is essential to check the feasibility of improvement. Regular evaluation not only improves the performance of the student but also it helps in understanding where the student is lacking. It takes a lot of manual effort to complete the evaluation process as even one college may contain thousands of students. This project proposed an automated solution for the performance

evaluation of the students using machine learning. Machine Learning is a system that can learn from example through self-improvement and without being explicitly coded by programmer.The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results.

Machine learning combines data with statistical tools to predict an output. This output is then used by corporate to makes actionable insights. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input, use an algorithm to formulate answers.

A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of task like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

## 2. PROPOSED SYSTEM

Performance evaluation of students is essential to check the feasibility of improvement. Regular evaluation not only improves the performance of the student but also it helps in understanding where the student is lacking. It takes a lot of manual effort to complete the evaluation process as even one college may contain thousands of students. This project proposed an automated solution for the performance evaluation of the students using machine learning.

Performance analysis of outcome based on learning is a system which will strive for excellence at different levels and diverse dimensions in the field of student's interests. This paper proposes a complete EDM framework in a form of a rule based recommender system that is not developed to analyze and predict the student's performance only, but also to exhibit the reasons behind it. The proposed framework analyzes the students' demographic data, study related and psychological characteristics to extract all possible knowledge from students, teachers and parents Seeking the highest possible accuracy in academic performance prediction using a set of powerful data mining techniques. The framework succeeds to highlight the student's weak points and provide appropriate recommendations. The realistic case study that has been conducted on 200 students proves the outstanding performance of the proposed framework in comparison with the existing ones.

The proposal aims to analyze student's demographic data, study related details and psychological characteristics in terms of final state to figure whether the student is on the right track or struggling or even failing. In addition to extensive comparison of our proposed model with the other previous related models. These recommendations are based on experimented studies for enhancing the student's academic performance. In addition to the mentioned above functionalities, the System will also alert all parties with the possible upcoming mental illnesses that the student might suffer from. Bayesian networks, k-nearest neighbor classifier, the goal of this study is to provide a comprehensive review of different classification techniques in data mining.

## 3. ALGORITHM

### 3.1. ENSEMBLE

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias(boosting), or improve predictions(stacking)

Ensemble methods can be divided into two groups:

- Sequential ensemble methods where the base learners are generated sequentially (e.g. AdaBoost). The basic motivation of sequential methods is to exploit the dependence between the base learners. The overall performance can be boosted by weighing previously mislabeled examples with higher weight.

- Parallel ensemble methods where the base learners are generated in parallel (e.g. Random Forest)

The basic motivation of parallel methods is to exploit independence between the base learners since the error can be reduced dramatically by averaging. Most ensemble methods use a single base learning algorithm to produce homogeneous base learners, i.e. learners of the same type, leading to homogeneous ensembles. There are also some methods that use heterogeneous learners, i.e. learners of different types, leading to heterogeneous ensembles. In order for ensemble methods to be more accurate than any of its individual members, the base learners have to be as accurate as possible and as diverse as possible.

### 3.2. BAGGING

Bagging stands for bootstrap aggregation. One way to reduce the variance of an estimate is to average together multiple estimates. For example, we can train M different trees on different subsets of the data (chosen randomly with replacement) and compute the ensemble: Bagging uses bootstrap sampling to obtain the data subsets for training the base learners. For aggregating the outputs of base learners, bagging uses voting for classification and averaging for regression.

### 3.3. BOOSTING

Boosting refers to a family of algorithms that are able to convert weak learners to strong learners. The main principle of boosting is to fit a sequence of weak learners− models that are only slightly better than random guessing, such as small decision trees− to weighted versions of the data. More weight is given to examples that were misclassified by earlier rounds. The predictions are then combined through a weighted majority vote (classification) or a weighted sum (regression) to produce the final prediction. The principal difference between boosting and the committee methods, such as bagging, is that base learners are trained in sequence on a weighted version of the data. The algorithm below describes the most widely used form of boosting algorithm called AdaBoost, which stands for adaptive boosting.

### 3.4. STACKING

Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features. The base level often consists of different learning algorithms and therefore stacking ensembles are often heterogeneous.

### 3.5. RANDOM FOREST

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

### 3.6. DECISION TREE

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is

a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

### 3.7. SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane

### 3.8. TENSOR FLOW

The most famous deep learning library in the world is Google's Tensor Flow. Google product uses machine learning in all of its products to improve the search engine, translation, image captioning or recommendations. To give a concrete example, Google users can experience a faster and more refined the search with AI. If the user types a keyword in the search bar, Google provides a recommendation about what could be the next word. Google wants to use machine learning to take advantage of their massive datasets to give users the best experience. Three different groups use machine learning:

- Researchers
- Data scientists
- Programmers

They can all use the same toolset to collaborate with each other and improve their efficiency. Google does not just have any data; they have the world's most massive computer, so TensorFlow was built to scale. TensorFlow is a library developed by the Google Brain Team to accelerate machine learning and deep neural network research. It was built to run on multiple CPUs or GPUs and even mobile operating systems, and it has several wrappers in several languages like Python, C++ or Java

### 3.9. ARCHIETECTURE OF TENSOR FLOW

TensorFlow architecture works in three parts:

- Preprocessing the data
- Build the model
- Train and estimate the model

It is called TensorFlow because it takes input as a multi-dimensional array, also known as tensors. You can construct a sort of flowchart of operations (called a Graph) that you want to perform on that input. The input goes in at one end, and then it flows through this system of multiple operations and comes out the other end as output. This is why it is called TensorFlow because the tensor goes in it flows through a list of operations, and then it comes out the other side

## 4. MODULE

### 4.1. DATA COLLECTION

ML depends heavily on data. It’s the most crucial aspect that makes algorithm training possible and explains why machine learning became so popular in recent years. But regardless of your actual terabytes of information and data science expertise, if you can’t make sense of data records, a machine will be nearly useless or perhaps even harmful. Data collection is the process of gathering and measuring information from countless different sources. In order to use the data, we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand. In a nutshell, data preparation is a set of procedures that helps make your dataset more suitable for machine learning. In broader terms, the data prep also includes establishing the right data collection mechanism. And these procedures consume most of the time spent on machine learning.

### 4.2. PRE PROCESSING

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are:

International Journal of Advanced Development in Science and Technology
Volume : 3 Issue : 06 July 2021 www.ijadst.com e-ISSN: 2582-1059

- Formatting
- Cleaning
- Sampling

**Formatting:** The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flatfile, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

**Cleaning:** Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonym zed or removed from the data entirely.

**Sampling:** There may be far more selected data available than you need to work with. More data can result in muchlonger running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

### 4.3. FEATURE EXTRACTION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python.

### 4.4. EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation to avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

## 5. RESULT ANALYSIS

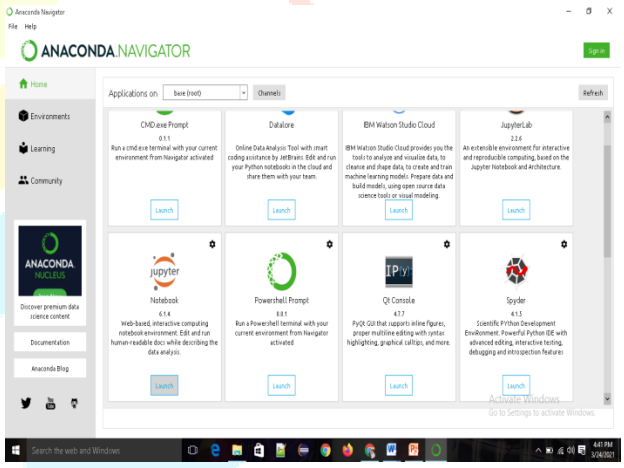



***Fig 1: Opening Jupyter Notebook using Anaconda Navigator***

(649, 33)

| | school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | ... | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | ... | 4 | 3 | 4 | 1 | 1 | 3 | 4 | 0 | 11 | 11 |
| 1 | GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | ... | 5 | 3 | 3 | 1 | 1 | 3 | 2 | 9 | 11 | 11 |
| 2 | GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | ... | 4 | 3 | 2 | 2 | 3 | 3 | 6 | 12 | 13 | 12 |
| 3 | GP | F | 15 | U | GT3 | T | 4 | 2 | health | services | ... | 3 | 2 | 2 | 1 | 1 | 5 | 0 | 14 | 14 | 14 |
| 4 | GP | F | 16 | U | GT3 | T | 3 | 3 | other | other | ... | 4 | 3 | 2 | 1 | 2 | 5 | 0 | 11 | 13 | 13 |

(395, 33)

| | school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | ... | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | ... | 4 | 3 | 4 | 1 | 1 | 3 | 6 | 5 | 6 | 6 |
| 1 | GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | ... | 5 | 3 | 3 | 1 | 1 | 3 | 4 | 5 | 5 | 6 |
| 2 | GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | ... | 4 | 3 | 2 | 2 | 3 | 3 | 10 | 7 | 8 | 10 |
| 3 | GP | F | 15 | U | GT3 | T | 4 | 2 | health | services | ... | 3 | 2 | 2 | 1 | 1 | 5 | 2 | 15 | 14 | 15 |
| 4 | GP | F | 16 | U | GT3 | T | 3 | 3 | other | other | ... | 4 | 3 | 2 | 1 | 2 | 5 | 4 | 6 | 10 | 10 |

(1044, 33)

| | school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | ... | famrel | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | ... | 4 | 3 | 4 | 1 | 1 | 3 | 6 | 5 | 6 | 6 |
| 1 | GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | ... | 5 | 3 | 3 | 1 | 1 | 3 | 4 | 5 | 5 | 6 |
| 2 | GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | ... | 4 | 3 | 2 | 2 | 3 | 3 | 10 | 7 | 8 | 10 |
| 3 | GP | F | 15 | U | GT3 | T | 4 | 2 | health | services | ... | 3 | 2 | 2 | 1 | 1 | 5 | 2 | 15 | 14 | 15 |
| 4 | GP | F | 16 | U | GT3 | T | 3 | 3 | other | other | ... | 4 | 3 | 2 | 1 | 2 | 5 | 4 | 6 | 10 | 10 |

***Fig 2: Training Data***

International Journal of Advanced Development in Science and Technology
Volume : 3 Issue : 06 July 2021 www.ijadst.com e-ISSN: 2582-1059

| | age | Medu | Fedu | traveltime | studytime | failures | famrel | freetime | goout | Dalc | Walc | health | absences | G1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 | 1044.000 |
| mean | 16.726 | 2.603 | 2.388 | 1.523 | 1.970 | 0.264 | 3.936 | 3.201 | 3.156 | 1.494 | 2.284 | 3.543 | 4.435 | 11.214 |
| std | 1.240 | 1.125 | 1.100 | 0.732 | 0.834 | 0.656 | 0.933 | 1.032 | 1.153 | 0.912 | 1.285 | 1.425 | 6.210 | 2.983 |
| min | 15.000 | 0.000 | 0.000 | 1.000 | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 |
| 25% | 16.000 | 2.000 | 1.000 | 1.000 | 1.000 | 0.000 | 4.000 | 3.000 | 2.000 | 1.000 | 1.000 | 3.000 | 0.000 | 9.000 |
| 50% | 17.000 | 3.000 | 2.000 | 1.000 | 2.000 | 0.000 | 4.000 | 3.000 | 3.000 | 1.000 | 2.000 | 4.000 | 2.000 | 11.000 |
| 75% | 18.000 | 4.000 | 3.000 | 2.000 | 2.000 | 0.000 | 5.000 | 4.000 | 4.000 | 2.000 | 3.000 | 5.000 | 6.000 | 13.000 |
| max | 22.000 | 4.000 | 4.000 | 4.000 | 4.000 | 3.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 75.000 | 19.000 |

*Fig 3: Final outcome*

| | school | sex | age | address | famsize | Pstatus | Medu | Fedu | Mjob | Fjob | ... | freetime | goout | Dalc | Walc | health | absences | G1 | G2 | G3 | final_grade |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | GP | F | 18 | U | GT3 | A | 4 | 4 | at_home | teacher | ... | 3 | 4 | 1 | 1 | 3 | 6 | 5 | 6 | 6 | poor |
| 1 | GP | F | 17 | U | GT3 | T | 1 | 1 | at_home | other | ... | 3 | 3 | 1 | 1 | 3 | 4 | 5 | 5 | 6 | poor |
| 2 | GP | F | 15 | U | LE3 | T | 1 | 1 | at_home | other | ... | 3 | 2 | 2 | 3 | 3 | 10 | 7 | 8 | 10 | fair |
| 3 | GP | F | 15 | U | GT3 | T | 4 | 2 | health | services | ... | 2 | 2 | 1 | 1 | 5 | 2 | 15 | 14 | 15 | good |
| 4 | GP | F | 16 | U | GT3 | T | 3 | 3 | other | other | ... | 3 | 2 | 1 | 2 | 5 | 4 | 6 | 10 | 10 | fair |

*Fig 4: Final grade distribution*

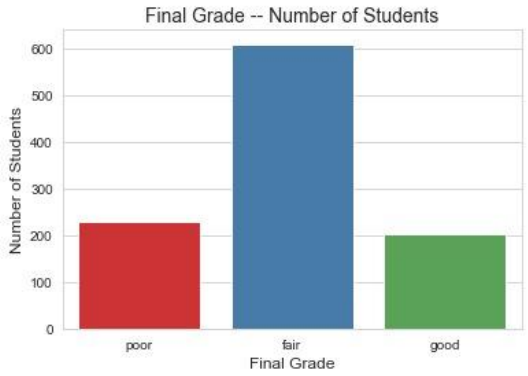
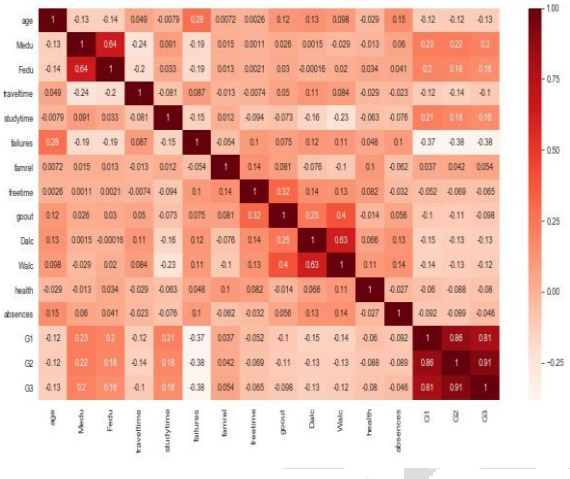
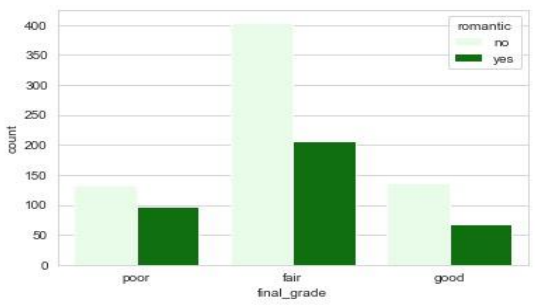
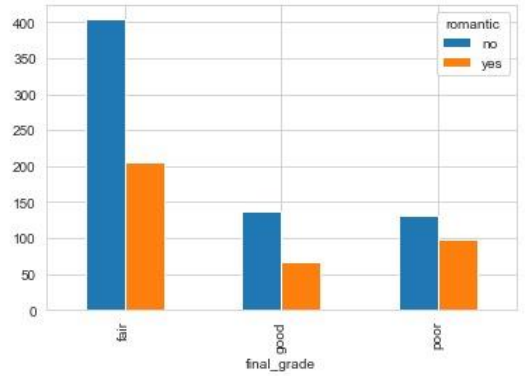
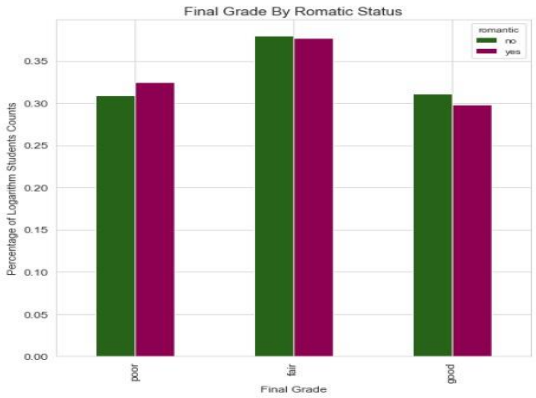


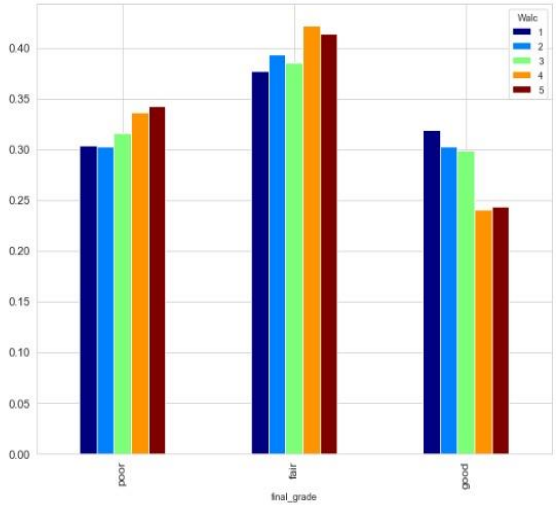

*Fig 5: Corelation heatmap*

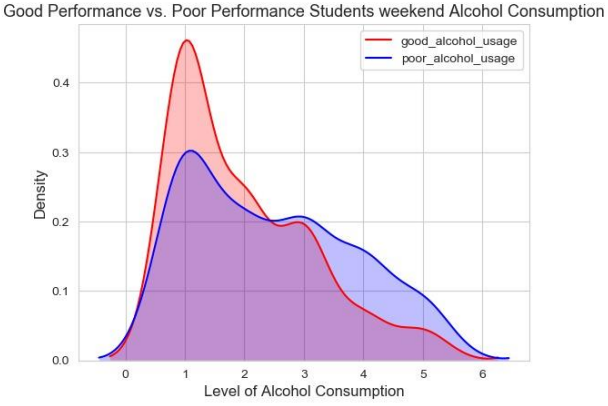




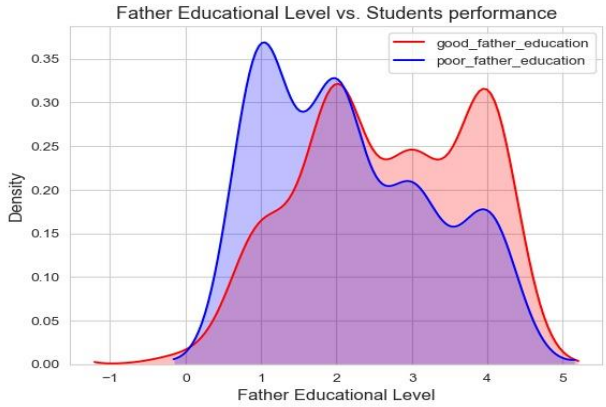



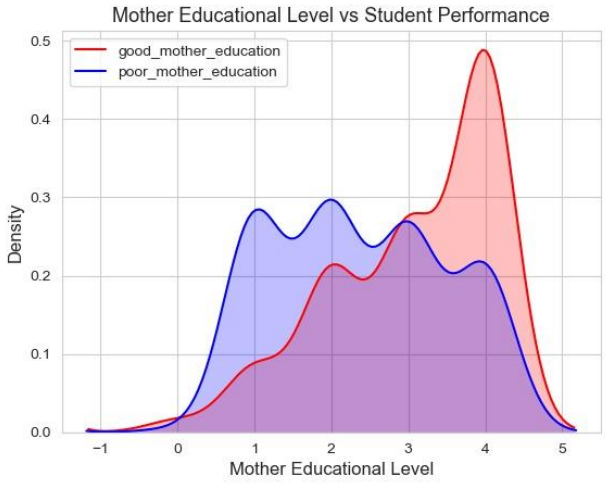



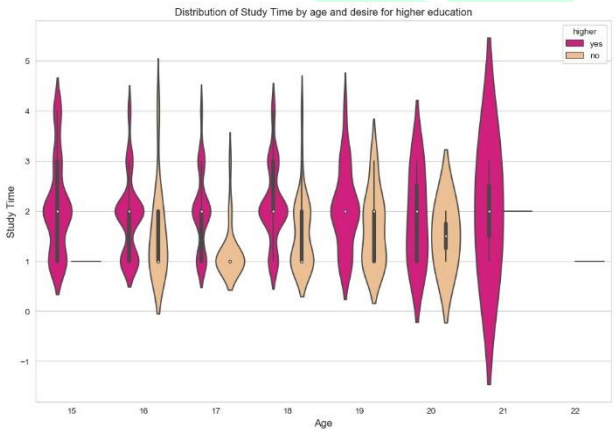



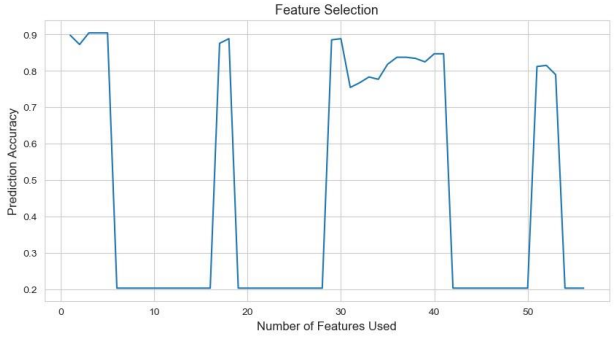



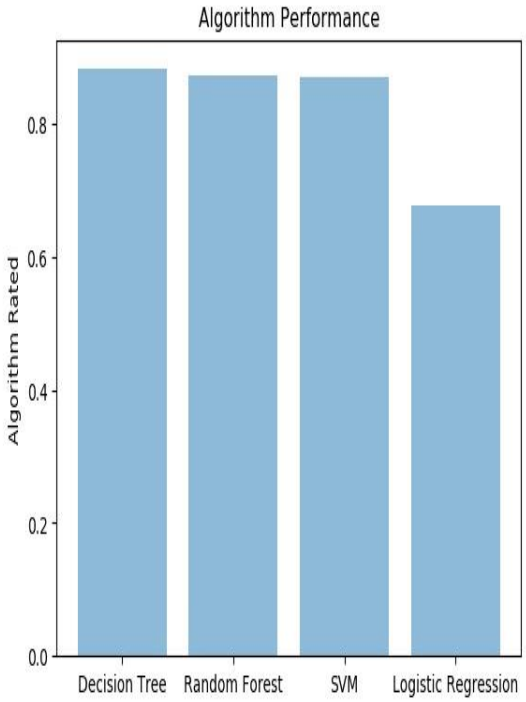
International Journal of Advanced Development in Science and Technology
Volume : 3 Issue : 06 July 2021 www.ijadst.com e-ISSN: 2582-1059




## 6.CONCLUSION

Performance analysis for students is a major problem which has not been addressed by the universities and schools till now. It is important that they are countered and with the help of this analysis the schools and universities can give better care and attention to the students who need more than the others. The work reported in this thesis indicates the machine learning techniques with supervised learning algorithms to understand the performance of algorithms with respect to student records where we analysed the performance of students and categorized it into three classes as high , average, low with the accuracy of 79% .We use a ensemble model to find out the model with the most accurate analysis so that there is no error while the prediction takes place.Since we use a total of four algorithms here we are able to make sure that no feature is left unattended and thus it will help us gain more insight than the normal one algorithm method.

## 7. REFERENCES

[1] Y. Chen, Y. Wang, Kinshuk, and N.-S. Chen, "Is FLIP enough? Or should we use the FLIPPED model instead?," Computers & Education, vol. 79, pp. 16- 27, 2014.

[2] J. O'Flaherty and C. Phillips, "The use of flipped classrooms in higher education: A scoping review," The Internet and Higher Education, vol. 25, pp. 85-95, 2015.

[3] V. A. Romero Zaldívar, A. Pardo, D. Burgos, and C. Delgado Kloos, "Monitoring Student Progress Using Virtual Appliances: A Case Study," Computers & Education, vol. 58, no. 4, pp. 1058- 1067, 2012.

[4] A. Pardo et al., "Connecting data with student support actions in a course: a hands-on tutorial," presented at the Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, British Columbia, Canada, 2017

[5] J. McDonald et al., "Cross-institutional collaboration to support student engagement: SRES version 2," in 33rd International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education, 2016, pp. 397-405, Adelaide, SA, Australia: ASCILITE, 2016.

[6] J. Bergmann and A. Sams, Flip Your Classroom: Reach Every Student in Every Class Every Day. International Society for Technology in Education, 2012.Android-Shahrear Iqbal, et al 2018

[7] J. Yabro, K. M. Arfstrom, K. McKnight, and P. McKnight, "Extension of a review of flipped learning," Flipped Learning Network2014, Available: http://www.flippedlearning.org/domain/41, Accessed on: 11/March/2016.

[8] L. P. Galway, K. K. Corbett, T. K. Takaro, K. Tairyan, and E. Frank, "A novel integration of online and flipped classroom instructional models in public health higher education," BMC medical education, vol. 14, no. 1, p. 181, 2014.

[9] J. L. Bishop and M. A. Verleger, "The flipped classroom: A survey of the research," in ASEE National Conference Proceedings, Atlanta, GA, 2013, vol. 30, no. 9, pp. 1-18.

[10] S. Zappe, R. Leicht, J. Messner, T. Litzinger, and H. W. Lee, "Flipping" the classroom to explore active learning in a large undergraduate course," in American Society for Engineering Education Annual Conference and Exhibition, 2009.

[11] J. F. Strayer, "How learning in an inverted classroom influences cooperation, innovation and task orientation," Learning Environments Research, vol. 15, no. 2, pp. 171-193, 2012.

[12] A. S. Burke and B. Fedorek, "Does "flipping" promote engagement?: A comparison of a traditional, online, and flipped class," Active Learning in Higher Education, vol. 18, no. 1, pp. 11- 24, 2017.

[13] P. H. Winne, "How Software Technologies Can Improve Research on Learning and Bolster School Reform," Educational Psychologist, vol. 41, no. 1, pp. 5-17, 2006.

[14] P. H. Winne and A. F. Hadwin, "Studying as self-regulated learning," in Metacognition in educational theory and practice, D. J. Hacker, J. Dunlosky, and A. C. Graesser, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers, 1998, pp. 227-304.

[15] G. Lust, N. A. Juarez Collazo, J. Elen, and G. Clarebout, "Content Management Systems: Enriched learning opportunities for all?," Computers in Human Behavior, vol. 28, no. 3, pp. 795-808, 2012.