

Predicting Academic Scholar Progress Using Supervised Machine Learning

Mohd Rafi Ahmed Taha ¹ Subramanian K.M ² Samdani Begum ³

¹ Research Scholar, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086.

² Professor, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086.

³ Assistant Professor, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086.

Abstract - The primary reason why machine learning has gained so much prominence nowadays is that it enables accurate and reliable decision making by extracting hidden relationships between various features present in the data. For this purpose the technique such as supervised methodologies and unsupervised methodologies are used. For this reason, machine learning can be used in almost any area of work to help in proper decision making and predictions. In our current project, we are trying to predict this student performance that utilizes supervised machine learning methodologies like support vector machines, logistic regression, random forests etc. We have also tried to publish this model to a web application so that it can be used by the academic community. The information extracted and the knowledge gained by extracting information from the educational data set would be helpful for predicting the student grades and their future performance. The main intention of the project used to predict student performance beforehand and help them get good grades in future. This would help in increasing the motivation levels of the students, improving their grades, decreasing their dropout ratio, and preparing better students for a better world.

Index Terms —Education Data mining, Learning Analytics, Prediction model, machine learning

I. INTRODUCTION

Data analytics in the area of learning and education has a significant role in refining the present education system so that the underlying problems with the present education system can be discovered from various perspectives like the administration perspective, the perspective of the faculty, and the perspective of the students[1][2]. Discoveries of data analytics help in proper planning, understanding real-time issues, and taking appropriate decisions that help in refining the system[3]. Due to the immense advantages that it offers, data analytics has paved its way into educational systems for activities like knowledge discovery[4] from existing databases, making predictions, text mining, artificial intelligence, etc. The primary objective of using learning analytics used to bring a shift from the existing learning practices to a technological paradigm.

A lot of student data has been gathered over the years and has been stored in the database but one is unable to bring out any useful information from the existing data and use it for analytical purposes[3],[5]. The data present in various source systems have to be integrated so that the power of analytics could be utilized in

the field of education [9]. These factors have increased the research conducted in the field of learning analytics as there are various avenues to explore [6].

Applying data analytics over educational data helps in understanding the weak areas of the students at an individual level, gathering information about the performance of the faculties, proper allocation of resources, and planning appropriately to mitigate the pain points and take necessary actions to improve the existing scenarios[10]. It is important to understand that the results of this analytics help in pinpointing the issues from various perspectives but things do not improve unless and otherwise all the people in the system like parents, students, teachers, and administrators come together to take appropriate decisions and implement them[7].

In this project, we have tried to analyze various features that are impacting the grades of the students in the educational data set to understand the underlying situations and help the teachers to take necessary actions I got that uplift the performance of the students[8]. We have also tried to publish this model to a web application that students and faculty can use in order to perform future predictions and improve the grades of the students[9].

II. SYSTEM ANALYSIS

Problem Statement:

Despite the fact that a large amount of student information has been obtained throughout the years and saved in the database, no valuable information can be extracted from the data and used for analysis. For the field of education to benefit from the power of analytics, the data from diverse source systems must be combined. These characteristics have boosted the amount of research done in the area of learning analytics because there are many different directions to go in.

Aim of the Project:

The main aim of the project is to forecast pupil performance by applying machine learning techniques in the field of education. We have used the Kalboard data set and fed it to the proposed model that is built using advanced machine-learning techniques.

Scope of the Project:

The scope of the project is limited to computing the accuracy of the proposed model and predicting pupil performance. The admin of the system trains the proposed model with training data and tests the model with test data. The users of the system can be students or faculty who can register themselves on the system and utilize the prediction model.

The objective of the Project:

Machine learning has been so popular in recent years mostly because it makes decisions more accurate and trustworthy by uncovering hidden connections between different aspects of the data. Techniques like supervised methodologies and unsupervised methodologies are employed for this goal. Because of this, machine learning may be used in practically any field of employment to aid in accurate prediction and decision-making. With the help of supervised machine learning techniques like support vector machines, logistic regression, random forests, etc., we are now working on a project where we are attempting to predict the performance of the students. We have also made an effort to offer this model as a web-based application so that the academic community may utilize it. The data that was retrieved from the educational data set and the knowledge that was obtained from it might be used to forecast student grades and performance in the future. The project's primary goal was to forecast student performance in advance and support their future academic success. This would aid in boosting students' motivation levels, enhancing their academic performance, lowering their dropout rate, and better preparing them for a better future.

Proposed System:

We propose to develop and deploy a web application that can predict the academic performance of a student based on various parameters. The machine learning model would be deployed to the web where the student can upload and view his academic parameters and receive predictions. We aim to build an automated prediction system that can accurately classify the academic grade of the student and make this information available to students, parents, and teachers.

Advantages:

- Automated system
- Can predict student grades before the beginning of the course
- Utilizes the large amount of data that is generated through educational systems.

III. PROPOSED MODULAR IMPLEMENTATION**The Algorithm/ Technique used:**

Below is the technical approach to address the problem:

1. Identification of dataset
2. Explorative Data Analysis
3. Cleaning the dataset and applying NLP techniques
4. Feeding the dataset to multiple algorithms and finding the best algorithm that suits the scenario
5. Training the final classifier and creating a model for the final classifier

Testing the final classifier and saving the results.

Below is the proposed modular implementation of the project. It consists of two modules:

1. Admin
2. Student

Admin Module:

The admin of the system is responsible for the activities like:

1. Uploading the dataset
2. Data Analysis of the dataset
3. Splitting the dataset for training and testing
4. Training the model for logistic regression and SVM
5. Review the performance of the algorithms on the given dataset
6. View student details and predictions

Student Module:

The system's user may take advantage of the following available machine learning services:

1. Logging into the system
2. Save one's academic parameters to the database
3. Receive predictions about academic performance.

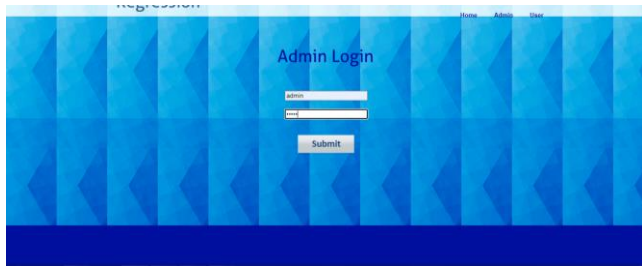
IV. PROJECT EXECUTION**Home page:**

This is the starting page of the application when the application is executed on Pycharm, the application is hosted on a web server and URL is generated to access the application once the user clicks on the URL the below page is opened on the browser.

**Admin Login:**

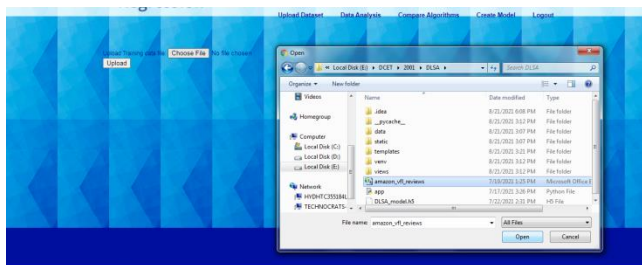
This is the login page for the admin module. The admin need to login into the system with his credentials in order to perform

operations like uploading the dataset, Training the dataset, Exploratory data Analysis of the dataset, Feeding the dataset to different Machine learning Algorithms to find the Algorithm that can meet the best accuracy and Create a model that can be hosted on the Flask Application to be used by the users.



Upload Dataset:

On this page, the administrator of the system can upload datasets that are used for training the machine learning models. The admin has to select the file by clicking on the Choose file button and click on the upload button to upload the file to the server. Once the upload is complete, a success message would be displayed that the file is successfully uploaded. For this project we are using Student performance reviews as a dataset.

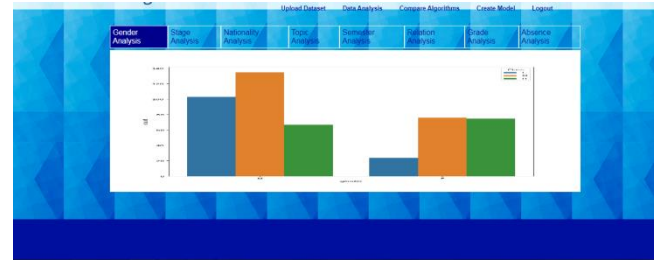


Data Analysis:

Exploratory Data Analysis is performed on the dataset in order to clean the dataset for any missing data, identify patterns, identify the relationships of various parameters of the outputs with the help of graphs, statistics etc.

Gender Analysis:

The below graph shows the Gender Analysis.



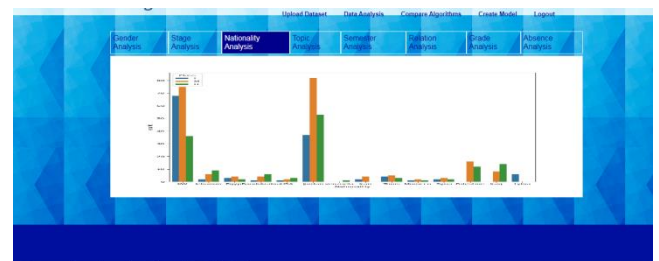
Stage Analysis:

The below graph shows the Stage Analysis.



Nationality Analysis:

The below graph shows the Nationality Analysis.



Topic Analysis:

The below graph shows the Topic Analysis.



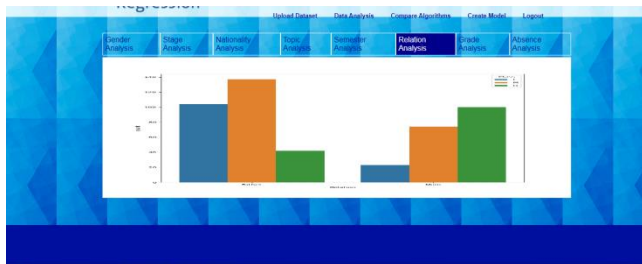
Semester Analysis:

The below graph shows the Semester Analysis.



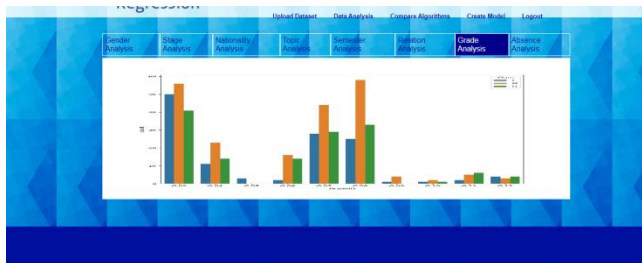
Relation Analysis:

The below graph shows the Relation Analysis.



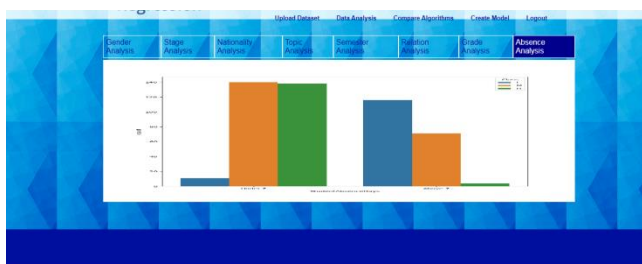
Grade Analysis:

The below graph shows the Grade Analysis.



Absence Analysis:

The below graph shows the Absence Analysis.

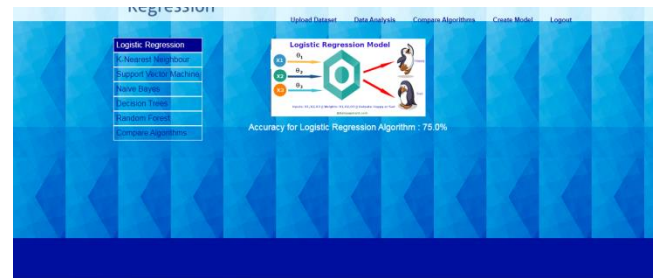


Compare Algorithms:

On this page, the admin can feed the dataset to various Algorithms to train them and get the test accuracy for each algorithm.

Logistic Regression:

When the dataset is feed to Logistic regression algorithm we observe that the test accuracy is 75.0%.



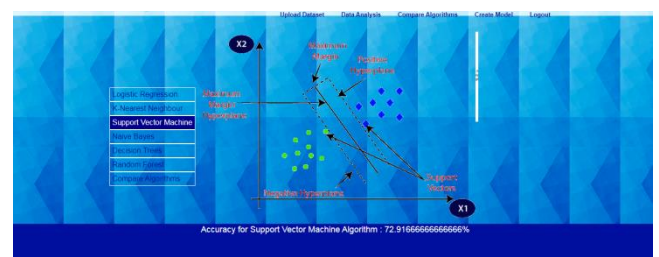
K-Nearest Neighbour:

When the dataset is feed to K-Nearest Neighbour algorithm we observe that the test accuracy is 72.22222222222221%.



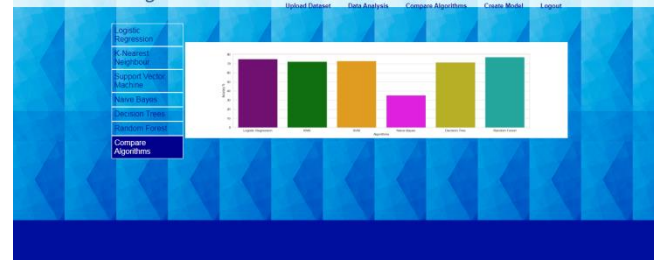
Support Vector Machine:

When the dataset is feed to Support Vector Machine algorithm we observe that the test accuracy is 72.16666666666666%.



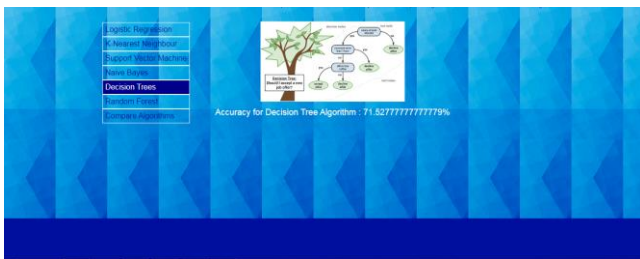
Naive Bayes:

When the dataset is feed to Naive Bayes algorithm we observe that the test accuracy is 35.41666666666667%.



Decision Trees:

When the dataset is feed to Decision Trees algorithm we observe that the test accuracy is 72.16666666666666%.



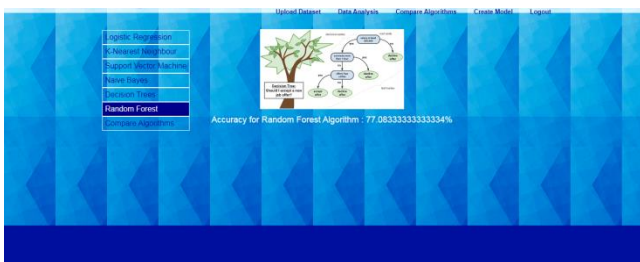
Create Model:

This screen shows the Accuracy of the Model as 0.8541666666666666%.



Random Forest:

When the dataset is feed to Random Forest algorithm we observe that the test accuracy is 77.08333333333333%.



User Registration:

This is the Registration page for the user module. The user need to get registered into the system in order to utilize the services of the system.

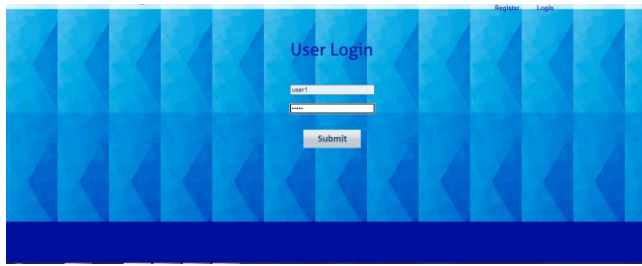


Compare Algorithms:

This screen shows the comparison of various test accuracies of the Algorithms.

User Login:

This is the login page for the user module. The admin need to login into the system with his credentials in order to receive predictions based on health parameters and save those values into the database.



CONCLUSION

The key factor in a student's job success is academic proficiency. The majority of nations have converted their traditional educational institutions to online learning platforms. As a result, it is highly challenging to evaluate students' behavior during distance learning with regard to their academics, participation in tests or discussion groups, and engagement with teachers and other students in the classroom. Academic institutions may profit greatly by studying and retrieving the hidden information that is acquired through these training management systems. And by making use of this knowledge to make improvements to students' academic brilliance by projecting their performance in the future and making the appropriate adjustments. In the study that follows, we looked at a student model, and logistic regression and support vector machines are the two classifiers that assist assess the effectiveness of the model. Based on the characteristics chosen using the feature selection approach, these classifiers are chosen using two separate technologies. The amount of student involvement with the system, timeliness in the classroom, and the three main categories are the elements that determine the student's academic grades. Through the use of support vector machines and gain ratio feature selection, the findings for the three or so factors are achieved. If we demonstrate the support vector machine technique utilizing sequential minimum optimization, the results for the forecast of students' future performance will be more accurate. In order to reduce dropout rates and assist students in their weak areas, we also suggest implementing the model into a web-

based application and making it accessible to the students, parents, and instructors.

REFERENCES

- [1] N.Padhy , D.Mishra ,R. Panigrahi , "The survey of data mining applications and feature scope," arXiv preprint arXiv:1211.5723, November, 2012.
- [2] R.S.Baker, S.Gowda, and A.Corbett, "Automatically Detecting a Student's Preparation for Future Learning: Help Use Is Key," In Proceedings of the 4th International Conference on Educational Data Mining, pp. 179–188,2011.
- [3] R.S.Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," Journal of Educational Data Mining, vol.1(1) , pp. 3–17,2009.
- [4] A.Merceron,K. Yacef, C. Romero, S.Ventura and M. Pechenizkiy, "Measuring correlation of strong symmetric association rules in educational data," Handbook of educational data mining, pp. 245-256, 2010.
- [5] A.El-Halees, "Mining Students Data to Analyze Learning Behavior: A Case Study," The 2008 international Arab Conference of Information Technology (ACIT2008) Conference Proceedings, University of Sfax, Tunisia, December , 2008.
- [6] C.R.Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol.40 (6), pp.601–618, 2010.
- [7] A.Wolff, Z.Zdrahal, D.Herrmannova, J. Kuzilek and M. Hlosta, "Developing predictive models for early detection of at-risk students on distance learning modules," 2014.
- [8] E.A.Amrieh, H.Thair, A.Ibrahim, "Preprocessing and analyzing educational data set using X-API for improving student's performance," 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), December, 2015.
- [9] H.Agrawal and H.Mavani, " In Student Performance Prediction using

Machine Learning,” International Journal of Engineering
Research
and Technology,2015.

[10] E.AAmrieh, H.Thair, A.Ibrahim, “Mining Educational
Data to
predict student's academic performance using ensemble
methods,”
International Journal of Database Theory And Application ,
Vol. 9 ,
2016.