Academic Performance Prediction

Burle Sravanthi, email: burlesravanthi@gmail.com
Chandolu Sravya, email: chandolusravya123@gmail.com
Suzanna William, email: suzannawilliam19@gmail.com
M. Vineela, Associate professor, email: vineela_m_99@yahoo.com

B.Tech 4th Year, Department of Computer Science and Engineering,
Bhoj Reddy Engineering College for Women, Hyderabad-59, Telangana, India.

Abstract - Digital data trails from disparate sources covering different aspects of student life are stored daily in most modern university campuses. However, it remains challenging to (i) combine these data to obtain a holistic view of a student, (ii) use these data to accurately predict academic performance, and (iii) use such predictions to promote positive student engagement with the university. To initially alleviate this problem, in this paper, a model named Augmented Education (AugmentED) is proposed. In our study, (1) first, an experiment is conducted based on a real-world campus dataset of college students (*N* = 156) that aggregates multisource behavioral data covering not only online and offline learning but also behaviors inside and outside of the classroom. Specifically, to gain in-depth insight into the features leading to excellent or poor performance, metrics measuring the linear and nonlinear behavioral changes (e.g., regularity and stability) of campus lifestyles are estimated; furthermore, features representing dynamic changes in temporal lifestyle patterns are extracted by the means of long short-term memory (LSTM). (2) Second, machine learning-based classification algorithms are developed to predict academic performance. (3) Finally, visualized feedback enabling students (especially at-risk students) to potentially optimize their interactions with the university and achieve a study-life balance is designed. The experiments show that the AugmentED model can predict students' academic performance with high accuracy.

Keywords – **A**cademic performance prediction, behavioral pattern, digital campus, Machine Learning (ML),Long Short-Term Memory (LSTM).

1. INTRODUCTION

Traditional programming differs significantly from machine learning. In traditional programming, a programmer codes all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

Machine learning is supposed to overcome this issue. The machine learns how the input and output data are correlated and it writes a rule. The programmers do not need to write new rules each time there is new data. The algorithms adapt in response to new data and experiences to improve efficiency over time.

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if it feeds a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the **learning** and **inference**.First of all, the machine learns through the discovery of patterns.One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a **feature vector.**We can think of a feature vector as a subset of data that is used to tackle a problem.

The machine uses few algorithms to simplify the reality and transform this discovery into a **model**. Therefore, the learning stage is used to describe the data and summarize it into a model.

2. RELATED WORK

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Real world campusdataset istaken from kaggle

Link: https://www.kaggle.com/c/1056lab-student-performance-prediction/data

Dataset:

There are 1044 number of records

In this data set we take34 columns in the dataset, which are described below.

Id - student's id

School - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

Class - student's class (binary: 'mat' - Mathematics or 'pot' - Portuguese language)

Sex - student's sex (binary: 'F' - female or 'M' - male)

Age - student's age (numeric: from 15 to 22)

Address - student's home address type (binary: 'U' - urban or 'R' - rural)

Famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than **3)**

Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

guardian - student's guardian (nominal: 'mother', 'father' or 'other')

traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

failures - number of past class failures (numeric: n if 1<=n<3, else 4)

schoolsup - extra educational support (binary: true or false)

famsup - family educational support (binary: true or false)

paid - extra paid classes within the course subject (Math or Portuguese) (binary: true or false)

activities - extra-curricular activities (binary: true or false)

nursery - attended nursery school (binary: true or false)

higher - wants to take higher education (binary: true or false)

internet - Internet access at home (binary: true or false)

romantic - with a romantic relationship (binary: true or false)

famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

freetime - free time after school (numeric: from 1 - very low to 5 - very high)

goout - going out with friends (numeric: from 1 - very low to 5 - very high)

Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

health - current health status (numeric: from 1 - very bad to 5 - very good)

absences - number of school absences (numeric: from 0 to 93)

G3 - the final grade (numeric: from 0 to 20, output target)

Data Preparation:

Wetransform the data by getting rid of the missing data and removing some columns. First, we create a list of column names that we want to keep or retain.

Next we drop or remove all columns except for the columns that we want to retain.

Finally we drop or remove the rows that have missing values from the data set.

Model Selection:

While creating a machine learning model, we need two datasets, one for training and other for testing. But now we have only one. So let's split this in two with a ratio of 80:20. We will also divide the data set into feature column and label column.

Here, we importe train_test_split function of sklearn. Then use it to split the dataset. Also, *test_size = 0.2*, it makes the split with 80% as train dataset and 20% as test dataset.

The *random_state* parameter seeds random number generator that helps to split the dataset.

The function returns four datasets. Labelthem as *train_x, train_y, test_x, test_y.* If we see shape of this datasets we can see the split of dataset.

Weuse Random Forest Classifier, which fits multiple decision tree to the data. Finally,we train the model by passing *train_x, train_y* to the *fit* method.

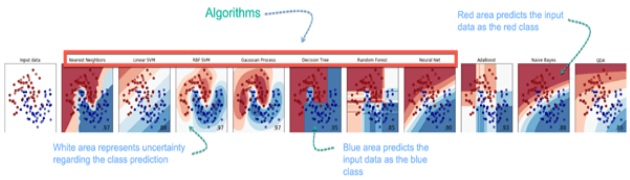
Once the model is trained, we need to Test the model. For that we pass *test_x* to the predict method.

Random Forest is one of the most powerful methods that is used in machine learning for classification problems. The random forest comes in the category of the supervised classification algorithm. This algorithm is carried out in two different stages, the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the classification.

B.PREDICTION ALGORITHMS

There are plenty of machine learning algorithms. The choice of the algorithm is based on the objective.

In the Machine learning example below, the task is to predict the type of flower among the three varieties. The predictions are based on the length and the width of the petal. The picture depicts the results of ten different algorithms. The picture on the top left is the dataset. The data is classified into three categories: red, light blue and dark blue. There are some groupings. For instance, from the second image, everything in the upper left belongs to the red category, in the middle part, there is a mixture of uncertainty and light blue while the bottom corresponds to the dark category. The other images show different algorithms and how they try to classified the data.



Challenges and Limitations of Machine Learning

The primary challenge of machine learning is the lack of data or the diversity in the dataset. A machine cannot learn if there is no data available. Besides, a dataset with a lack of diversity gives the machine a hard

time. A machine needs to have heterogeneity to learn meaningful insight. It is rare that an algorithm can extract information when there are no or few variations. It is recommended to have at least 20 observations per group to help the machine learn. This constraint leads to poor evaluation and prediction.

Application of Machine Learning

Augmentation:

Machine learning, which assists humans with their day-to-day tasks, personally or commercially without having complete control of the output. Such machine learning is used in different ways such as Virtual Assistant, Data analysis, software solutions. The primary user is to reduce errors due to human bias.

Automation:

Machine learning, which works entirely autonomously in any field without the need for any human intervention. For example, robots performing the essential process steps in manufacturing plants.

Finance Industry

Machine learning is growing in popularity in the finance industry. Banks are mainly using ML to find patterns inside the data but also to prevent fraud.

3.PROPOSED STRATEGY

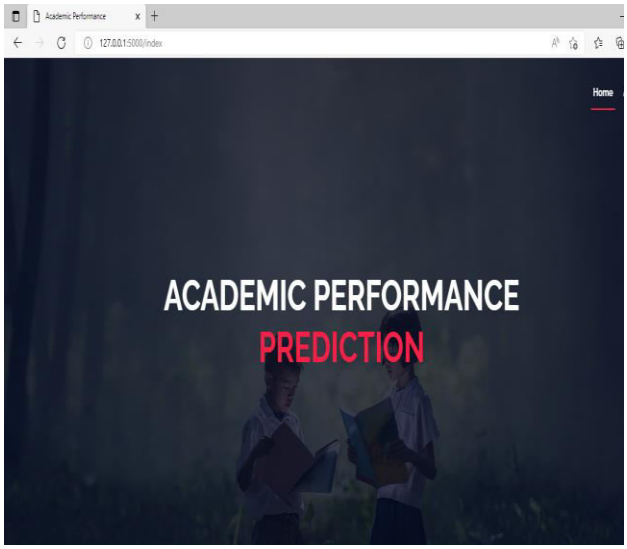
- In our study, academic performance prediction is considered as a classification problem.
- In a digital campus dataset, the main task is to first extract features from the raw multisource data; then select the features that are strongly correlated with academic performance and use these features to train the classification algorithm; and finally provide visualized feedback based on the prediction results.
- The main task of this system is to select features and use the features to train the prediction algorithm.
- Subsequently, the selected features are used to train the ML based classification algorithm for the academic performance prediction.

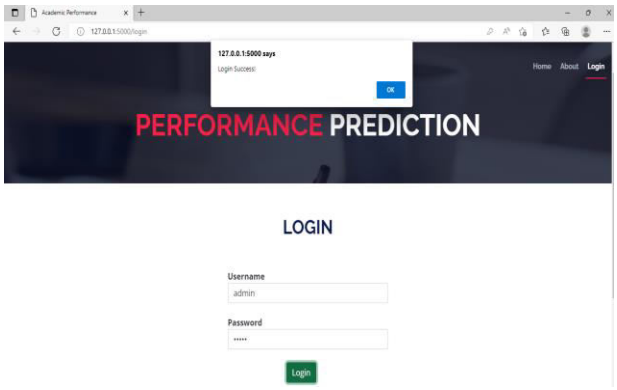
4.EXPERIMENTAL RESULTS

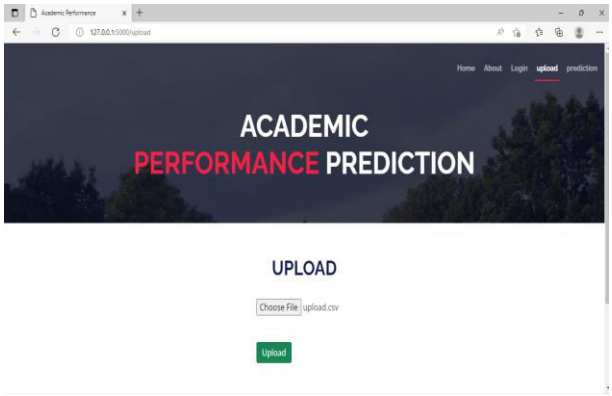
- In Existing system, HMM(Hidden Markov Model) is being used; whichuses entropy as one of its parameters. HMM does not consider the independence of the features.

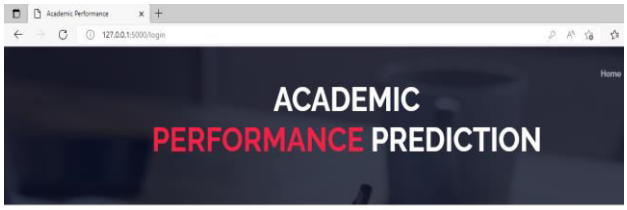
- Lyapunov Exponent (LyE) is used to recognize student's activities and discover their nonlinear behavioral patterns.

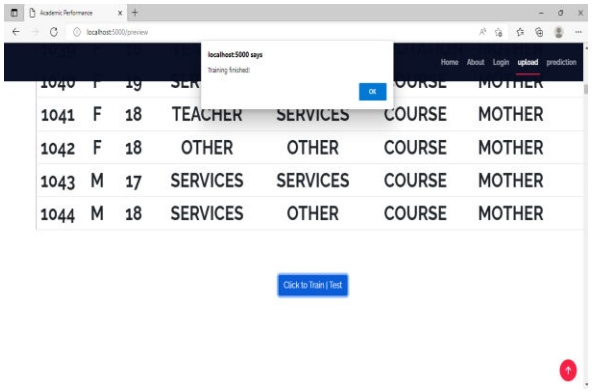
5. RESULTS

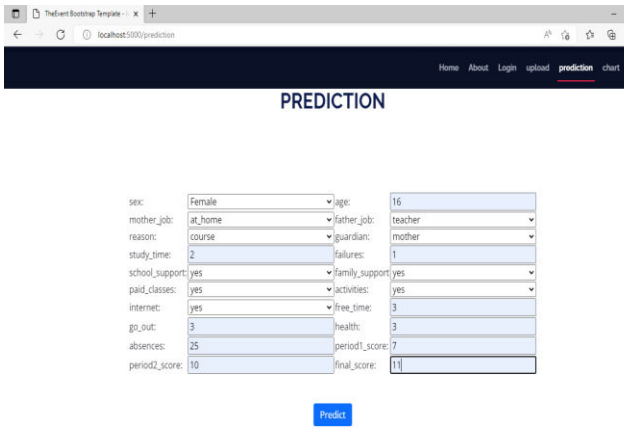


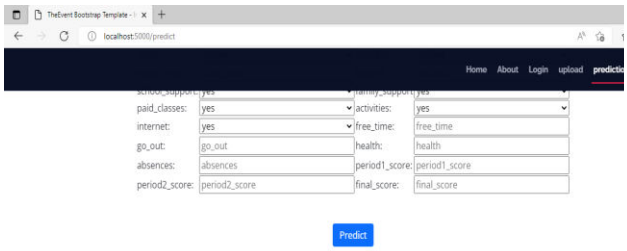




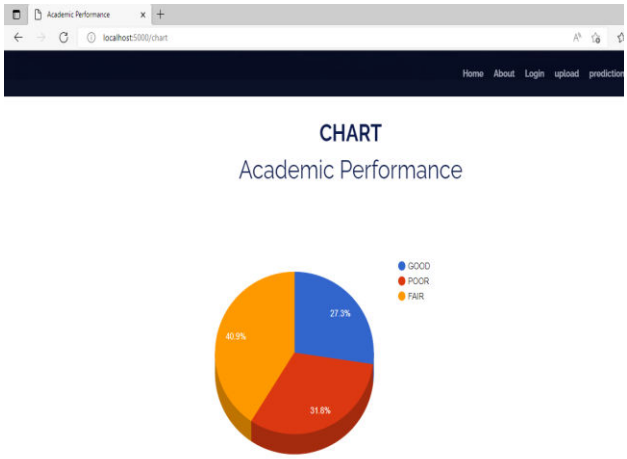








Prediction is:fair



6. CONCLUSION

In conclusion, our study is based on a complete passive daily data capture system that exists in most modern universities. This system can potentially lead to continual investigations on a larger scale. The knowledge obtained in this study can also potentially contribute to related research among students.

7. FUTURE SCOPE

We can extend the Academic performance Prediction project in which the students will receive their feedbackalong with the necessary links and pdf's through SMS or email.

8. REFERENCES

[1] A. Furnham, and J. Monsen, "Personality traits and intelligence predict academic school grades," Learning and Individual Differences,
vol. 19, no. 1, pp. 0-33, 2009.

[2] M. A. Conard, "Aptitude is not enough: How personality and behavior predict academic performance," Journal of Research in Personality,
vol. 40, no. 3, pp. 339-346, 2006.

[3] T. Chamorropremuzic, and A. Furnham, "Personality predicts academic performance: Evidence from two longitudinal university

samples," Journal of Research in Personality, vol. 37, no. 4, pp. 319-
338, 2003.

[4] R. Langford, C. P. Bonell, H. E. Jones, T. Pouliou, S. M. Murphy, and E. Waters, "The WHO health promoting school framework for improving the health and well-being of students and their academic achievement," Cochrane Database of Systematic Reviews, vol. 4, no. 4, pp. CD008958, 2014.

[5] A. Jones, and K. Issroff, "Learning technologies: Affective and social issues in computer-supported collaborative learning," Computers &
Education, vol. 44, no. 4, pp. 395-408, 2005.

[6] D. N. A. G. Van, E. Hartman, J. Smith, and C. Visscher, "Modeling relationships between physical fitness, executive functioning, and
academic achievement in primary school children," Psychology of Sport & Exercise, vol. 15, no. 4, pp. 319-325, 2014.

[7] R. Wang, F. Chen, Z. Chen, T. Li, and A. T. Campbell, "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," In Proc. of the ACM International Joint Conference on Pervasive & Ubiquitous Computing,
Seattle, WA, USA, 2014.

[8] R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell, "SmartGPA: How smartphones can assess and predict academic performance of college students," In Proc. of the ACM International Joint Conference on Pervasive & Ubiquitous Computing, Osaka, Japan, 2015.

[9] M. T. Trockel, M. D. Barnes, and D. L. Egget, "Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors," Journal of American College Health, vol. 49, no. 3, pp. 125-131, 2000.

[10] D. M. Hansen, S. D. Herrmann, K. Lambourne, J. Lee, and J. E.
Donnelly, "Linear/nonlinear relations of activity and fitness with children's academic achievement," Med Sci Sports Exerc. vol. 46, no. 12, pp. 2279-2285, 2014.