ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

CLASSIFYING STUDENTS BASED ON THEIR ACADEMIC PERFORMANCE USING TRANSFER LEARNING

P N S Sowmya Bharadwaj, Dr. V Akila *Information Technology Gokaraju Rangaraju Institute of Engineering and Technology* Hyderabad **:** psreechami@gmail.com; akila_be@yahoo.co.in

***Abstract* —** A major problem in today's educational environment is the creation of technologies that aid students in learning, whether in a traditional classroom setting or an online one. It is important for a student to have a good academic performance along with other skills.

In existing system, the researchers have forecasted the students grade-level performance using a two-step process. First they formulated the problem as binary classification to overcome the limitation of student's poor performance and second is they gained insights to factors that lead to poor performance.

These features were generated using grades from a public university's first-year students from University of Minnesota based on which they identified the different students of interest and Gradient Boosting and Random Forest Classifiers performed best based on AUC and F1 score metrics.

In my project I am using a transfer learning technique which is used to train the pre-existing model in-order to increase the accuracy of the algorithms and will develop a user-login page for authentication and privacy reasons.

***Keywords*—** *Academic Performance, Transfer Learning.*

I. INTRODUCTION

A major problem in today's educational environment is the creation of technologies that aid students in learning, whether in a traditional classroom setting or an online one. It is important for a student to have a good academic performance along with other skills.

Higher educational institutions try to improve the retention and success of the students enrolled. According to US National Centre for Educational Statistics, 60% of students on four-year degree will not graduate at same institution where they started and 30% of students drop out after their first year of college. As a result, the college always look for ways to serve students more efficiently and effectively. Most of the existing approaches focus on identifying the students at risk who can be benefited further with better assistance for the successful completion of a course. The initial or the fundamental approach to predict student's performance is the grades that they achieve. While reasonable prediction accuracy has been achieved there is a significant weakness of the models proposed to identify the poor-performing students.

After identifying the latter group, additional resources and support can be provided to enhance the likelihood of their success. We will classify the students in different ways to define groups of students taking a course: failing students, students dropping the class, students performing worse than expected and students performing worse than expected and performing well, while taking into consideration the difficulty of a course.

Using these features, we present a comprehensive study to answer the following questions: which features are good indicators of a student's performance? which features are the most important? The findings are interesting as different features are the most important for different classification tasks.

II. RELATED STUDY

The development of technology that assist students in learning, whether in a regular classroom setting or an online one, is a significant issue in today's educational climate. In addition to other skills, a student needs to perform well in the classroom.

Agoritsa Polyzou, George Karypis, et al. have forecasted the student's academic performance using a two-step process. First they formulated the problem as a binary classification as pass or fail to overcome the limitations of student's performance. Second is they gained the insights to factors that lead to poor performance. They used four machine learning algorithms as decision tree, linear support vector machine, random forest and gradient boosting algorithms. Their highest accuracy score they concluded is 61%.

ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

Ansar Siddique, Asiya Jan, Fiaz Majeed, Adel Ibrahim Qahmash, Noorulhasan Naveed Quadri and Mohammad Osman Abdul Wahab, et al. have used have used three single classifiers including a Multilayer Perceptron (MLP), J48, and PART along with three well established ensemble algorithms encompassing Bagging (BAG), MultiBoost (MB), and Voting (VT) independently. To further enhance the performance of the above-mentioned classifiers, nine other models were developed by the fusion of single and ensemble-based classifiers. The evaluation results showed that MultiBoost with MLP outperformed the others by achieving 98.7% accuracy,98.6% precision, recall, and F-score.

Rasheed Mansoor Ali Sa, S Perumal, et al. proposed the CNN-Multiclass LDA method for predicting student performance and academic behavior. They implemented the CNN-based feature extraction which extracts valuable features from the student log then the best features are selected by the Minimum Redundancy Maximum Relevance(mRMR) feature selection algorithm thereby eliminating the least features. The features are fetched to the Multi-class Linear Discriminant Analysis (LDA) for classification, which classifies the result into low, medium, and high in order to assist tutors in predicting the low-ranking students. Based on the prediction, the tutors can easily find the low ranks of students who need a high preference for improving their academic performance. The researchers experiment showed that the proposed model achieved greater accuracy (96.5%), precision (094), recall (092), F-score (095), and requires less computation time than existing methods.

Zafar Iqbal, Junaid Qadir, Adnan Noor Mian, and Faisal Kamiran, et al. have worked on Collaborative Filtering (CF), Matrix Factorization (MF), and Restricted Boltzmann Machines (RBM) techniques to systematically analyze a real-world data collected from Information Technology University (ITU), Lahore, Pakistan. They used a feedback model approach, where they measured the students' knowledge in a particular course domain, which provides appropriate counseling to them about different courses in a particular domain by estimating the performance of other students in that course which can be used as a component of an early warning system that will lead to students' motivation and provides them early warnings if they need to improve their knowledge in the courses. It also helps the course instructor to determine weak students in the class and to provide necessary interventions to improve their performance.

Sotiris Kotsiantis, Christos Pierrakeas, P. E. Pintelas, et al. have compared some of the state-of-the-art learning algorithms. Two experiments have been conducted with six algorithms, which were trained using data sets provided by the Hellenic Open University. Among other significant conclusions it was found that the Naive Bayes algorithm is the most appropriate to be used for the construction of a software support tool, has more than satisfactory accuracy, its overall sensitivity is extremely satisfactory and is the easiest algorithm to implement.

III. ALGORITHMS USED

Our work is implemented using built-in python libraries namely numpy used for mathematical operations, pandas used for Data Representation, seaborn and matplotlib used for data visualizations, sklearn used for modelling the data and pickle used as optimal solution for the model. The trained model can be zipped into a single pickle file and can be used as optimal solution which is capable of predicting the result.

Various supervised machine learning algorithms of classifications are used as trial-and-error basis for acquiring highest accuracies are Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes and Grid Search CV.

Decision Tree Classifier : Based on the tree structure with the conditions or rules, the Decision Tree Classifier generates the result as the optimum result. Decision Nodes, Design Links, and Decision Leaves are the three main parts of the decision tree algorithm. It works with splitting, which is used to divide data into subsets, pruning, which is used to reduce the length of the decision tree's branches, and tree selection, which is used to choose the shortest tree that best fits the data.

ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

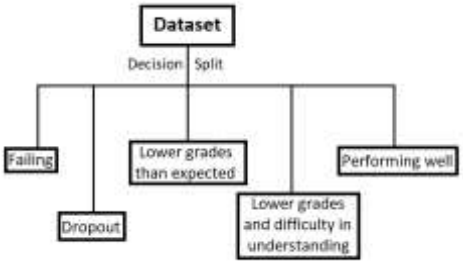

Fig 1. Decision Tree Classifier

Random Forest Classifier : For a more precise prediction, the Random Forest Classifier expands and merges various decision trees into a "forest". This technique is based on the idea that several different decision trees can work considerably more effectively together than they do separately. The bagging method (ensemble machine learning methodology known as Bootstrap Aggregation), which is used to lower the variance (error) of high variance algorithms, is used to train these trees.

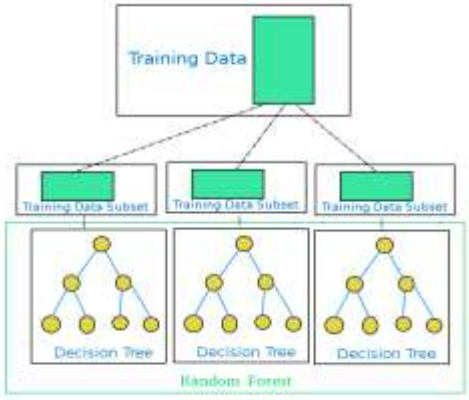

Fig 2. Random Forest Classifier

K Neighbors Classifier : K Neighbors Classifier (also known as KNN Classifier) places the new case into the category that is most similar to the available categories by assuming similarity between the new case/data and existing cases. Therefore, testing data is categorised into the category that is most similar to it after the model has been trained using a KNN classifier.

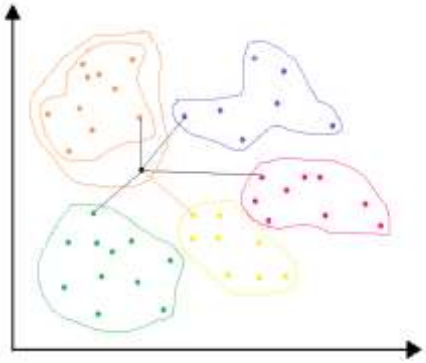

Fig 3. K Neighbors Classifier

Support Vector Machine : To make it simple to place the new data point in the appropriate category, SVM constructs a decision boundary that can divide n-dimensional space into classes. Support vectors are data points that are closer to the hyperplane and have an impact on the hyperplane's position and orientation. By utilising these support vectors, we increase the classifier's margin.

ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

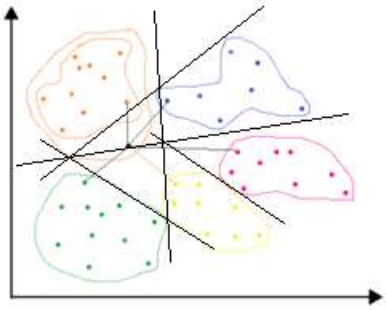

Fig 4. Support Vector Machine

Logistic Regression : Based on a collection of independent factors, the chance of an event occurring is estimated using logistic regression. As the dependent variable has five alternative outcomes—students failing, dropping out, receiving grades below expectations, receiving grades below expectations, having difficulties understanding and doing well—we are using multinominal logistic regression.

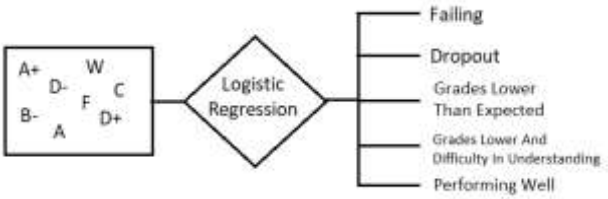

Fig 5. Logistic Regression

Gaussian Naive Bayes : The Gaussian Naive Bayes model is an extension of the naive Bayes framework for strong independence assumptions. By replacing the parameters with the new input value for the variable, one can utilise the Gaussian probability density function to generate predictions. As a result, the Gaussian function will provide an estimate for the likelihood of the new input value.

IV. IMPLEMENTATION

This project is implemented into two parts first part includes code, visualization and predicting performance using pickle file and the second part includes user login page for predicting performance.

In first part, a dataset with labelled data is given as input. Training and testing are applied on each algorithm and checks for the accuracy. Output is visualization of pie chart after applying Grid Search CV on each algorithm and bar chart of accuracies before and after applying Grid Search CV algorithm, also the prediction of performance using pickle file in which algorithm with highest accuracy is dumped and it predicts the performance with a grading scale as 0 for failing or 1 for dropout or 2 for grades lower than expected or 3 for grades lower and difficulty in understanding or 4 for performing well.

In Coding part, we first import libraries, Load the dataset, select the features for input and output, define train and test values, apply various algorithms as Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes get the accuracy and again on each algorithm we apply Grid Search CV algorithm and find out the increased accuracies.

In each algorithm, we import the algorithm, fit the train and test values, predict using the test value for existing data and get the accuracy score. Once we get the accuracy score we apply Grid Search CV algorithm on the same algorithm then again we fit the train and test values, predict using the test value for existing data and get the accuracy score. This accuracy will be more than the previous accuracy.

In second part, we have to sign up by entering username, name, email, mobile number and password for the first time and can login by using the username and password.

Once we sign up we are directed to a login page to enter username and password, Academic Performance Page is opened.

ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

Here we enter the no of students who received the grades as A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F and W and click on submit. We get the result as students failing or dropout or grades lower than expected or grades lower and difficulty in understanding and performing well in Performance Prediction Page.

V. RESULTS

From anaconda prompt firstly we change the directory to the project folder path then we enter a command to open python using python app.py code then we get an IP Address which is to be opened in browser where we can run our front-end code.

Fig 6. Academic Performance Prediction Page

The above fig shows the first page to signup, to sign in or signup we click on signup, the following page is displayed.

Fig 7. Sign Up Page

If we don't have login credentials we have to sign up. To sign up, we should enter a username, name, email, mobile number and password.

Then it is automatically directed to Sign in Page. We are asked enter the username and password.And if we have login credentials, click on Login here! Sign In. Sign in page is opened asking to enter the username and password.

ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

Fig 8. Login Page

After we enter username and password, Academic Performance Prediction page is opened. Here we enter the no of students who received the grades as A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F and W and click on submit. We get the result as students failing or dropout or grades lower than expected or grades lower and difficulty in understanding and performing well in Performance Prediction Page.

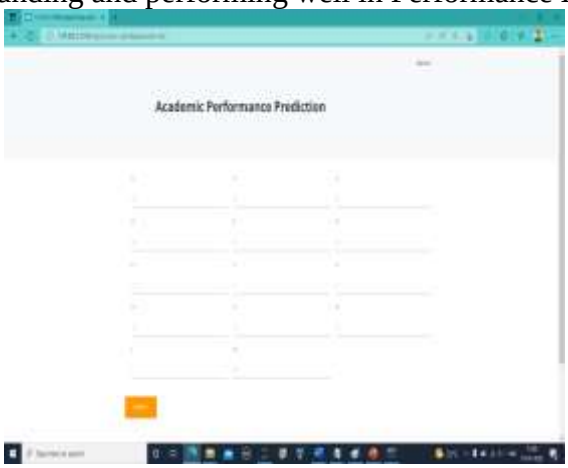

Fig 9. Academic Performance Prediction Page

For example, if we enter 1 student got A+ grade, 1 student got A, 2 students got A-, 5 students got B+, 60 students got B, 0 students got B-, 2 students got C+, 3 students got C, 35 students got C-, 54 students got D+, 5 students got D, 12 students got D-, 65 students got F and 3 students got or W. The Performance Prediction is Grades lower than expected.

Fig 10. Performance Prediction Page.

VI. COMPARISON OF ACCURARIES OF VARIOUS ALGORITHMS USED

Once we understand how the students are understanding the course it is easy to implement new techniques for them to understand the subject in more efficient manner.

We used various algorithms Decision Tree Classifier, Random Forest Classifier, K Neighbors Classifier, Support Vector Machine, Logistic Regression and Gaussian Naive Bayes.

To find out which algorithm gives more accuracy we visualized the data using a pie chart after applying Grid Search CV on each algorithm.

In the below pie chart, KN represents K Neighbours Classifier with 79% accuracy, RF represents Random Forest Classifier with 82% accuracy, DT represents Decision Tree Classifier with 74% accuracy, GNB represents Gaussian Bayes with 53% accuracy, LR represents Logistic Regression with 83% accuracy and an exploded part SVM represents Support Vector Machine with 85% accuracy which is highest when compared to other algorithms accuracies.

ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

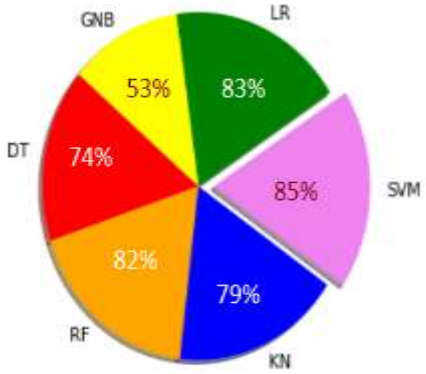

Fig 11. Algorithm Accuracy Comparison Using a Pie Chart

We also visualized each algorithm accuracies before and after applying Grid Search CV algorithm using Multiple Bar Charts

In below bar chart, DT represents Decision Tree Classifier with 68.7% before Grid Search CV and 73.8% after Grid Search CV, RF represents Random Forest Classifier with 81.8% before Grid Search CV and 82.2% after Grid Search CV, KN represents K Neighbours Classifier with 76.7% before Grid Search CV and 79.3% after Grid Search CV, SVM represents Support Vector Machine with 79.1% before Grid Search CV and 84.7% after Grid Search CV, LR represents Logistic Regression with 83.3% both before and after applying Grid Search CV and GNB represents Gaussian Bayes with 34.7% before Grid Search CV and 53.4% after Grid Search CV.

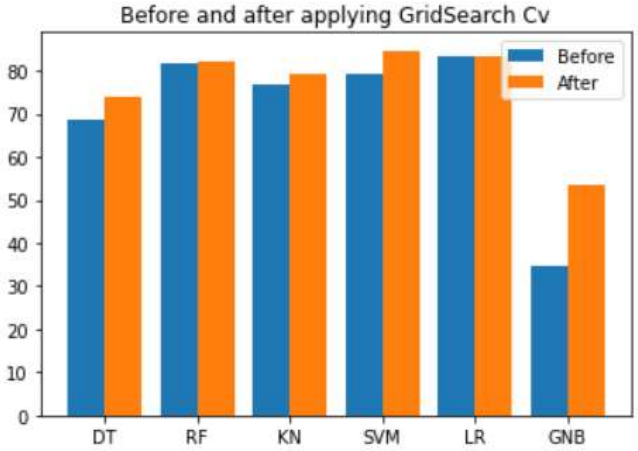

Fig 12. Algorithm accuracy Comparison Before and After Applying Grid Search CV Using a Bar Chart

VII. REFERENCES

[1] Agoritsa Polyzou, George Karypis, Feature extraction for classifying students based on their academic performance.

[2] Ansar Siddique, Asiya Jan, Fiaz Majeed, Adel Ibrahim Qahmash, Noorulhasan Naveed Quadri and Mohammad Osman Abdul Wahab, Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers.

[3] Rasheed Mansoor Ali Sa, S. Perumala, Multi-class LDA classifier and CNN feature extraction for student performance analysis during Covid-19 pandemic.

[4] Zafar Iqbal, Junaid Qadir, Adnan Noor Mian, and Faisal Kamiran, Machine Learning Based Student Grade Prediction: A Case Study.

[5] Sotiris Kotsiantis, Christos Pierrakeas, P. E. Pintelas, Predicting students' performance in distance learning using machine learning techniques.

[6] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch. Predicting student performance: an application of data mining methods with an educational web-based system.

ISSN: 2278-4632
Vol-12 Issue-09 No.01 September 2022

[7] J. McFarland, B. Hussar, C. de Brey, T. Snyder, X. Wang, S. Wilkinson-Flicker, S. Gebrekristos, J. Zhang, A. Rathbun, A. Barmer, et al. Undergraduate retention and graduation rates
[8] S. Morsy and G. Karypis. Cumulative knowledge-based regression models for next-term grade prediction.
[9] B. N. Sari, Implementasi Teknik Seleksi Fitur Information Gain pada Algoritma Klasifikasi Machine Learning untuk Prediksi Performa Akademik Siswa.
[10] E. A. H. T. A. I. Amrieh, Mining Educational Data to Predict Student's academic Performance using Ensemble Methods.
[11] D. Sartika and D. I. Sensuse, Perbandingan Algoritma Klasifikasi Naive Bayes Nearest Neighbour dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian.
[12] Kamran, S.; Nawaz, I.; Aslam, S.; Zaheer, S.; Shaukat, U. Student's performance in the context of data mining.
[13] Kaur, A.; Umesh, N.; Singh, B. Machine Learning Approach to Predict Student Academic Performance.
[14] Imran, M.; Latif, S.; Mehmood, D.; Shah, M.S. Student Academic Performance Prediction using Supervised Learning Techniques.
[15] Alturki, S.; Alturki, N. Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions.
[16] Aucejo, E.M.; French, J.; Ugalde Araya, M.P.; Zafar, B. The impact of COVID-19 on student experiences and expectations: Evidence from a survey.
[17] Sakri, S.; Alluhaidan, A.S. RHEM: A robust hybrid ensemble model for students' performance assessment on cloud computing course.
[18] Panigrahi, R.; Borah, S. Rank Allocation to J48 Group of Decision Tree Classifiers using Binary and Multiclass Intrusion Detection Datasets.
[19] Bashir Khan Yousafzai, Sher Afzal Khan, Student-Performulator: Student Academic Performance Using Hybrid Deep Neural Network.