Predictive Analysis Tool for Predicting Student Performance and Placement Performance using ML algorithms

Dr. B. Muthusenthil[1], Venkat Mugesh S[2], Thansh D[3], Subaash R[4]

[1] *Assistant Professor, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*
[2] *UG Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*
[3] *UG Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*
[4] *UG Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*

ABSTRACT

Recently, Predictive Analysis using machine learning and data science has had greater growth in almost every field. Predictive analysis for predicting student performance has always been difficult than other predictions such as House price prediction, stock market prediction, sales prediction, etc., Due to the difficulty in finding and collecting appropriate datasets, especially in college-level education. Since the grading system differs from universities to universities, it is even more difficult to collect proper datasets that can be used as a global dataset, created based on the system that is common to almost every university. In this paper, we have taken this difficult task and created a dataset, which contains student details of over 185 students. The dataset was collected from SRM Valliammai engineering college, it includes the data of 2018 and 2019 passed out students. Other than academic data, the dataset also covers other attributes of the students. The dataset contains around 20 attributes which would improve the accuracy score. Using this dataset our aim is to develop a training model that predicts the final CGPA of a student and also the placement result of the particular student. We have successfully predicted the student's final CGPA and their placement result. We have predicted the results using algorithms like Linear Regression (LR1), Decision tree (DT) algorithm, K-nearest neighbor algorithm (KNN), Logistical regression (LR2) and Lasso regression (LR3) algorithm. And we have obtained excellent results with accuracy over 94%. The predicted results are obtained from comparing the accuracy score of the above-given algorithms. The one with the highest accuracy was displayed as the final output.

Keyword: - *Machine Learning, Data Science, Predictive Analysis, Student Performance.*

1. INTRODUCTION

Machine learning is a subset of artificial intelligence (AI) that helps computers or teaching machines learn and make intelligent decisions from all previous data. The architecture for machine learning involves collecting and storing a rich set of information and turning it into a standardized knowledge base for various uses in different fields. Predicting student performance is a daunting issue faced every year by educational institutions such as colleges, schools, and training centers. As a result, forecasting the performance of students at an earlier stage would enable the educational institutions to find solutions to avoid negative student performance. Lectures should predict their student's success and find appropriate learning strategies to improve the performance of the students. It can also strengthen the institution's admission policies and help students resolve their grades.

Machine learning (ML) methods have been used successfully in a number of fields, such as healthcare [5], environmental studies [6], industrial [7] and education systems [3]. To date, the idea of machine learning still attracts researcher in the educational sector [8], [9]. In addition, the idea of e-learning and big data in education presents researchers with extremely large data that should be thoroughly analyzed to help educators and decision-makers develop the education systems.

In this paper, we aim to predict student's final CGPA and Placement result using five different ML algorithms. This can be done by creating a training model, which would be trained with the training dataset. The training data will be analyzed through the machine learning algorithms and an algorithm with the highest accuracy will be found. Using that algorithm, final prediction would be performed.
The remainder of this paper was structured as follows: Section 2 explores the related works about machine learning in the education systems. Section 3 sets out the approach proposed. The experimental datasets used in this paper are discussed in Section 4. The experimental results and analysis of the suggested solution are shown in Section 5. Section 6 draws conclusions and works to come

2. RELATED WORKS

Machine learning for education systems was extensively investigated [1], where five separate fields were defined: prediction, model discovery, data extraction for human judgment, clustering, and mining relationships. Most previous educational systems work, apply to universities or virtual learning [10]. The data collected in all previous works, either from surveys or from e-learning programs. Kapur et al. [11] used two different methods of machine learning (i.e. J48 Decision Tree, and Random Forest) to predict marks on students in the field of education. The data collected consists of 480 entries which relate to the enrollment of the student. Veracano et al. [12] used various methods of machine learning to estimate students dropping out for unbalanced data set. The authors collect 419 samples from a single high school in Mexico. Saif and. Al [13] looks at a number of courses and discusses whether successful or bad performance can be expected. Saarela et al. [14] introduced a method for predicting the level of difficulty of various mathematical problems and for predicting whether or not the students could answer those questions.

3. PROPOSED METHOD

Prediction results for Educational Systems are very dynamic. They are not limited to very few attributes. To get the accuracy in prediction all the dynamic attributes to the system has to be included in the dataset. Our proposed system consists of several ML algorithms such as KNN (k –Nearest Neighbor) algorithm, Linear Regression (LR1), Logistical Regression (LR2), Lasso Regression (LR3) and Decision Tree (DT) algorithms. The control flow of our method is illustrated in Fig. 1.

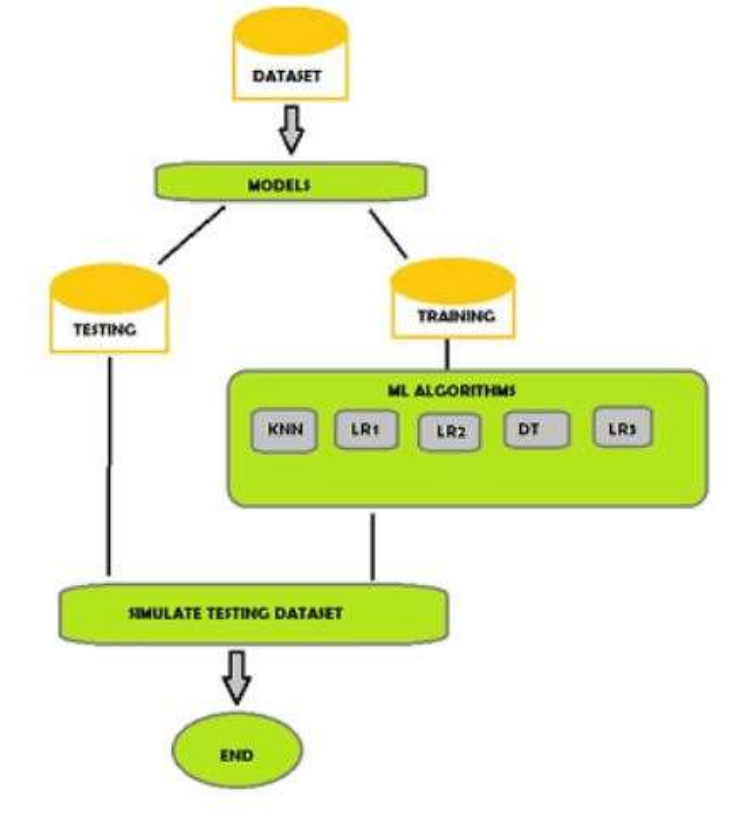

Fig -1: A pictorial representation of the proposed methodology.

3.1 ML Algorithms

One of the most accurate algorithms, that was suited for predicting Student's final CGPA is Linear Regression (LR1). Linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm. The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient. For example, in a simple regression problem (a single x and a single y), the form of the model would be: y = B0 + B1*x

Using the above representation, the prediction of input can be done in a very simple and effective manure. Linear Regression is one of the attractive models because of its simple representation.The next important algorithm which was used in this tool is Decision Tree (DT) algorithm. This algorithm showed a dynamic nature in predicting student's placement results. A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Cost functions are an important part in decision trees. Cost functions are used for classification and regression.

$$\text{Regression: } \text{sum}(y - \text{prediction})^2$$

Here, we are predicting the company in which the student will be placed. Now the decision tree will start splitting by considering each feature in training data. The mean of responses of the training data inputs of a particular group is considered as prediction for that group. The above function is applied to all data points and the cost is calculated for all candidate splits. Again, the split with the lowest cost is chosen.

Classification: G = sum(pk * (1 — pk))

This is how the decision tree will classify the trained data and predict the result based on the given input data.

3.2 Dataset Distribution

Our project starts from collecting the dataset. This would be 50% of the project since the dataset is an important part of predictive analysis. We have collected student's information from SRM Valliammai Engineering College. The dataset includes information of 185 students of 2018 and 2019 passed out batch. It consists of over 20 attributes. Which covers some of the dynamic areas of the student, that affect their final Cgpa and their placement results. Fig. 2 shows the contents and how the dataset is distributed.

Attribute	**Description**
REG_NO	University register number of the student (numeric)
NAME	Name of the student
GPA_1 to GPA_8	Individual Gpa obtained by the student for 8 semesters (numeric: 1 to 10)
CGPA	Final CGPA obtained by the student (numeric: 1 to 10)
MATHS	Marks obtained in mathematics in high school (numeric: 0 to 200)
PHYSICS	Marks obtained in physics in high school (numeric: 0 to 200)
CHEMISTRY	Marks obtained in chemistry in high school (numeric: 0 to 200)
TOATL	Total marks obtained in high school (numeric: 0 to 1200)
PLACED	Placement status of student (binary: 1 placed and 0 not placed)
PLACED_COMPANY	If placed name of the company else Nil
AREA	Native area of the student (binary: 0 rural and 1 urban)
BOARD	Board of education till high school (binary:0 state board and 1 central board)
QUOTA	Quota through which the student joined the college (binary: 0 govt and 1 management)
SPORTS	Time spent in playing sports (numeric: 1 very low to 5 very high)
JOURNALS	Journals published by the student (numeric: 1 very low to 5 very high)
COURSES	External courses done by the student (numeric: 1 very low to 5 very high)

Fig -2: Dataset Distribution.

4. EXPERIMENTAL DATA

The experiment begins from collecting the dataset. Once the collection of the dataset is completed, the pre-processing of that experimental data begins. The pre-processing includes the cleaning of the dataset, analysis of the dataset and training the model with dataset.

4.1 Training the Model

In our project, we have created the training model using Jupyter notebook. It uses python as its programming language. The first step in training the model is loading the dataset into a data frame through this notebook. Once the dataset is loaded, the attributes must be analyzed as the training data.

4.2 Data Cleaning

Data cleaning is a process of creating a data frame that contains only the essential data needed for the prediction. The cleaning includes the removal of unwanted rows and columns from the trained dataset. It also includes creating dummies of certain attributes. This would be done for any binary columns or any columns which have string data in them. Since the prediction cannot be done for character values or the string values, they have to be converted into dummy columns (i.e) every string values in a column would be converted into a new column with binary values in it. The Fig. 3 shows the cleaned dataset and dummies created for the columns BOARD, AREA, and QUOTA.

COURSES	JOURNALS	SPORTS	PREDICT_VAR	BOARD_0	BOARD_1	AREA_0	AREA_1	QUOTA_0	QUOTA_1
3	1	1	0	1	0	1	0	1	0
3	1	4	10	1	0	1	0	1	0
5	5	1	10	1	0	1	0	1	0
3	1	4	0	1	0	1	0	1	0
4	1	1	10	1	0	1	0	1	0

Fig -3: A cleaned dataset with dummies.

5. EXPERIMENTAL RESULT AND ANALYSIS

Once the training and cleaning of the dataset is completed, the analysis and prediction part of the project begins. The dataset after getting cleaned will be converted into .csv file (comma separated values) and gets loaded into the training model.

5.1 Data Analysis

One of the important parts of this project is data analysis. Here the loaded data will be analyzed graphically to understand the dataset. With the findings from the analyzed graphs, the prediction process will be performed. The attributes in the dataset would be plotted against one another and its findings would give a better understanding of the dataset. Out of 24 attributes present in our dataset, only 7 dataset makes an impact on the CGPA. Those findings are represented in Fig. 4, Fig. 5 and Fig. 6.

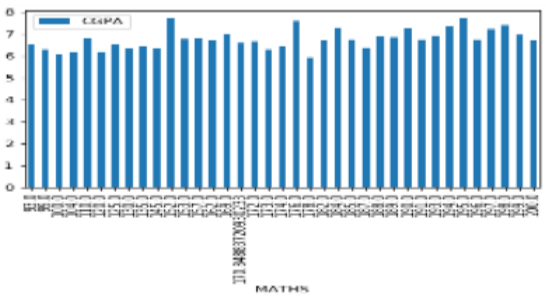

Fig -4: CGPA vs MATHS

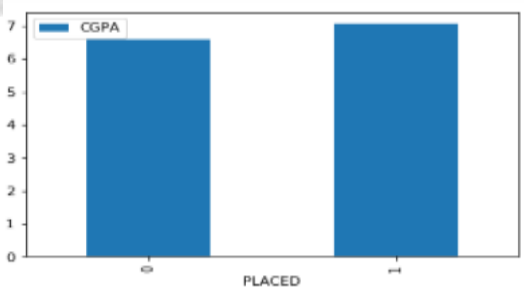

Fig -5: CGPA vs PLACED

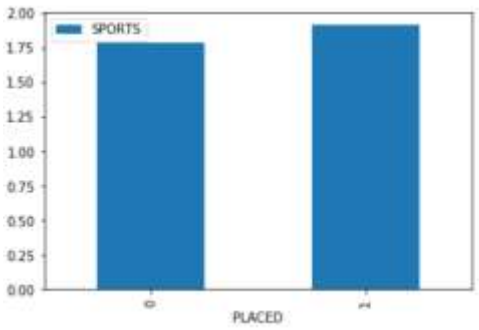

Fig -6: SPORTS vs PLACED

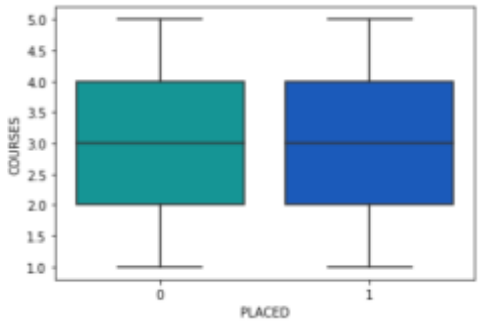

Fig -7: COURSES vs PLACED

5.2 Predicting CGPA

After the data was analyzed completely, it is found that attribute MATHS has a very big impact on the attribute CGPA as you can see in Fig. 4. Thus, for predicting CGPA the attributes GPA_1 to GPA_7 and MATHS are required as the test input. With this test data, the prediction has to be done by comparing the test_x and test_y in various ML algorithms. Where test_x is the input columns obtained from the user and test_y is the column that is to be predicted (here test_y = CGPA and test_x = GPA_1 to GPA_7 and MATHS). The input values will be tested on the algorithms and the one with the highest accuracy will get to predict the test_y and displays the final output. From our findings, the best algorithm for predicting CGPA is Linear Regression. Fig. 8 shows the findings.

	model	best_score	best_params
0	linear_regression	0.930854	{'normalize': False}
1	lasso	0.311272	{'alpha': 1, 'selection': 'random'}
2	decision_tree	0.830450	{'criterion': 'friedman_mse', 'splitter': 'best'}

Fig -7: A table with best scores of various algorithms

Thus, it was found that the highest accuracy can be reached from LR1 and this algorithm will be used for predicting CGPA. This model will be imported as a pickle file which will be accessed by a server python program. This server will call the pickle file and the result will be returned in JSON format. The server can be accessed through URL http://127.0.0.1:5000/predict_cgpa. The front end would make an HTTP request through JavaScript to this URL. The user input would be an excel file with all the test_x columns. That will be read by JavaScript and test_x values will be sent as a parameter to the given URL. Then the predicted CGPA will be returned in JSON format. Which will be displayed as the output in the UI screen, which is created using FLASK. It is shown in the Fig. 8

Fig -8: Front end showing predicted CGPA

5.3 Predicting Placement Result

Predicting the placement result is a bit tricky compared to CGPA. Because here we have to predict the name of the company which the student might get placed. Which is a string value, that cannot be predicted. So, we have given a unique number to all the 12 companies that the students got placed. Now the model will predict the unique number. Which will then be processed into the company's name through JavaScript. For this prediction, we have found the decision tree (DT) to be more effective than all the other algorithms. Fig. 9 shows our findings. Thus, we proceeded with DT and followed the process same as the CGPA and displayed the result. Fig. 10 shows the output screen for Placement Result.

	model	best_score	best_params
0	linear_regression	0.053334	{'normalize': True}
1	lasso	-0.014089	{'alpha': 1, 'selection': 'random'}
2	decision_tree	-0.686641	{'criterion': 'mse', 'splitter': 'random'}

Fig -8: Table showing the best scores of various algorithms

For placement prediction also we have used the same URL and the JSON would have both CGPA results and Placement result. The one which has to be displayed will be manipulated through JavaScript.

Fig -8: Front end showing predicted Placement result

5. CONCLUSION

In this paper, we proposed a method with a set of machine learning algorithms to predict student performance. Five different machine learning algorithms are examined they are Linear Regression (LR1), Decision Tree (DT), Logistical Regression (LR2), Lasso Regression (LR3) and K-nearest neighbor (KNN). The obtained results show that Linear Regression is the best performing algorithm for predicting student performance. It has outperformed all the other methods and showed an accuracy of 94%.

In our future works, we will integrate many domains like stock market prediction, sales prediction, weather prediction, etc., into this tool and predict results on a large scale.

6. REFERENCES

[1] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," JEDM, vol. 1, no. 1, pp. 3–1

[2] C. Romero, S. Ventura, and E. Garc´ıa, "Data mining in course management systems: Moodle case study and tutorial," Computers & Education, vol. 51, no. 1, pp. 368 – 384, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360131507000590

[3] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. V. Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil," Journal of Business Research, vol. 94, pp. 335 – 343, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0148296318300870

[4] L. H. Son and H. Fujita, "Neural-fuzzy with representative sets for prediction of student performance," Applied Intelligence, vol. 49, no. 1, pp. 172–187, Jan 2019. [Online]. Available: https://doi.org/10.1007/s10489-018-1262-7

[5] C. M. Hatton, L. W. Paton, D. McMillan, J. Cussens, S. Gilbody, and P. A. Tiffin, "Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare," Journal of Affective Disorders, vol. 246, pp. 857 – 860, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165032718319931

[6] E. Fijani, R. Barzegar, R. Deo, E. Tziritis, and K. Skordas, "Design and implementation of a hybrid model based on two-layer decomposition method coupled with extreme learning machines to support real-time environmental monitoring of water quality parameters," Science of The Total Environment, vol. 648, pp. 839 – 853, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0048969718331851

[7] D. D. Clercq, D. Jalota, R. Shang, K. Ni, Z. Zhang, A. Khan, Z. Wen, L. Caicedo, and K. Yuan, "Machine learning powered software for accurate prediction of biogas production: A case study on industrial-scale chinese production data," Journal of Cleaner Production, vol. 218, pp. 390 – 399, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095965261930037X

[8] K. S. Rawat and I. V. Malhan, "A hybrid classification method based on machine learning classifiers to predict performance in educational data mining," in Proceedings of 2nd International Conference on Communication, Computing and Networking, C. R. Krishna, M. Dutta, and R. Kumar, Eds. Singapore: Springer Singapore, 2019, pp. 677–684.

[9] S.Carnell,B.Lok,M.T.James,andJ.K.Su,"Predictingstudentsuccess in communication skills learning scenarios with virtual humans," in Proceedings of the 9th International Conference on Learning Analytics & Knowledge, ser. LAK19. New York, NY, USA: ACM, 2019, pp. 436– 440. [Online]. Available: http://doi.acm.org/10.1145/3303772.3303828

[10] P. Ducange, R. Pecori, L. Sarti, and M. Vecchio, "Educational big data mining: How to enhance virtual learning environments," in International Joint Conference SOCO'16-CISIS'16-ICEUTE'16, M. Gra˜na, J. M. L´opez-Guede, O. Etxaniz, ´A. Herrero, H. Quinti´an, and E. Corchado, Eds. Cham: Springer International Publishing, 2017, pp. 681–690.

[11] B. Kapur, N. Ahluwalia, and R. Sathyaraj, "Comparative study on marks prediction using data mining and classification algorithms," Int. J. Adv. Res. Comput. Sci., vol. 8, pp. 632 – 636, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095965261930037X

[12] C. M´arquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: a case study with high school students," Expert Systems, vol. 33, no. 1, pp. 107–124, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12135

[13] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Computers & Education, vol. 113, pp. 177 – 194, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/ S0360131517301124

[14] M. Saarela and B. Yener, "Predicting math performance from raw largescale educational assessments data : A machine learning approach," 2016.