IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STUDENT MARKS PREDICTION USING MACHINE LEARNING TECHNIQUES

[1] A Harika, [2] Akshatha K A, [3] Anirudh A, [4] Eesha B S, [5] Prof. Sandhya A Kulkarni

[1] Student, [2] Student, [3] Student, [4] Student, [5] Assistant Professor

[1] Computer Science and Engineering,

[1] K S School of Engineering and Management, Bangalore, India

Abstract: To understand the student's rate of progress, it is crucial to forecast their performance. "Prevention is better than cure," goes the saying. The success of students can be significantly increased with early identification of at-risk students and preventive actions. The recommended task is utilized to assess a student's performance right now and forecast their outcomes in the future. Every year, many kids fall behind because of inadequate supervision and assistance. Based on the results, teachers can concentrate on the students who are more likely to receive lower grades in the final semester and can also help the student by identifying needs for the final exams. This project's major goal is to show how likely it is to train and model the dataset and how feasible it is to create a predictive model for student performance with a dependable accuracy rate.

Index Terms **- Marks Prediction, Random Forest, User interface, Model.**

I. INTRODUCTION

Machine learning is a part of man-made reasoning (AI) and software engineering which centers around the utilization of information and Algorithm to mimic the way that human learn, bit by bit working on its precision. ML algorithms fabricate a model in light of test information, known as preparing information, to make forecasts without being expressly modified to do as such. ML algorithms are utilized in a wide assortment of utilizations, like medication, discourse acknowledgment and PC vision. ML approaches are customarily isolated into three general classifications, contingent upon the idea of the sign or input accessible to the learning framework. Administered learning, solo learning, and Reinforcement learning are the three classifications of machine learning. ML is a part of artificial intelligence (AI) and software engineering which centers around the utilization of information and Algorithm to mimic the way that human learn, bit by bit working on its precision. ML algorithms fabricate a model in light of test information, known as preparing information, to make forecasts without being expressly modified to do as such. AI calculations are utilized in a wide assortment of utilizations, like medication, discourse acknowledgment and PC vision. ML algorithms are customarily isolated into three general classifications, contingent upon the idea of the sign or input accessible to the learning framework. Administered learning, solo learning, and Reinforcement learning are the three classifications of AI.

Principal objective for this task is to help instructors to examinations understudy execution effectively and if necessary, they can help her/him to work on their understudy's exhibition by making a few moves like expanding their understanding hours, giving a few tasks. This research depends on the classification procedure, order by and large alludes to the planning of information things into predefined gathering and classes. The preparation information is investigated by classification algorithm and during arrangement stage the test information are utilized to gauge the exactness of the grouping rules. In this paper, different Machine learning algorithms on the notable consequences of a course being shown in single men in software engineering frameworks program to figure out the expectation precision.

II. LITERATURE SURVEY

“Implementation of Student SGPA Prediction System (SSPS) Using Optimal Selection of Classification Algorithm” [1] In today’s world, there is competition in education institution every student plays a major role in the growth of the institution. An algorithm such as Logistic Model Tree, Random tree, and REP tree is used, the data set collected from the university may contain errors and noises which make the model less effective so data cleaning is done and the data set will reduce to 236 instances from 260 records. The REP tree algorithm has given more accuracy with 61.70%.

“Machine Learning Algorithm for Student’s Performance Prediction” [2]. The performance can be improved by predicting their marks by using the previous year’s marks and can groom the students to improve themselves. By using machine learning techniques, we can improve the performance of every student the dataset of 1170 data was collected from three subjects. Algorithm such as K-Nearest Neighbors, SVC, Decision Tree Classifier, and Linear Discriminant Analysis. The decision tree classifier model has given the highest accuracy of 94.44%.

"Prediction of Student's Performance by Modelling Small Dataset Size" [3] An educational institution's major objective is to give its students a high-quality education. Early performance forecasting for students can help them earn better grades and get into prestigious schools. The machine learning classification algorithm such as Naïve Bayes, Support vector machines, K-nearest neighbor, and Linear discriminant analysis. The Linear discriminant analysis has given accuracy of 79%.

"Prediction of Student Academic Performance Using Neural Network, Linear Regression, and Support Vector Regression: A Case Study" [4] Institutions have a significant impact on academic and pupil success. In the final year, pupils' academic standing has a big impact on their future jobs. The algorithm used is Neural Network (NN), Support Vector Regression (SVR), and Linear Regression (LR). The dataset of 134 data was collected, the linear regression has shown more accuracy compared to other algorithms.

III. METHODOLOGY

Figure 1 represents the block diagram of this project. Initially the client/student visits the web application where the home page of the website is displayed which contains navigation buttons to various other pages that includes 'about us', 'contact us', 'student login' and 'admin login'. If the client is a teacher, then they can navigate to the admin login page and enter the admin credentials.

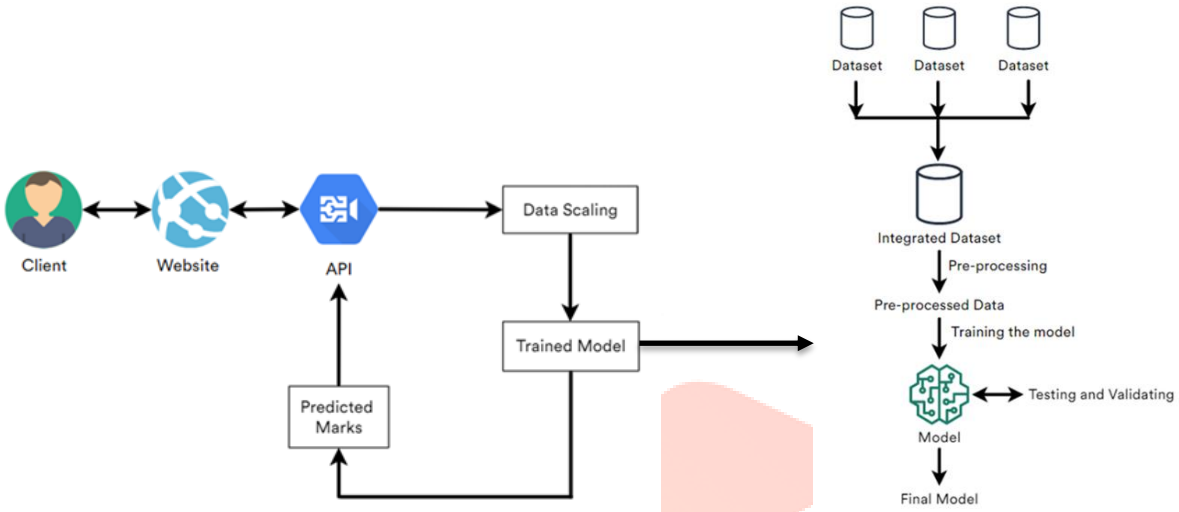

Fig 1: Block Diagram

Once the login button is clicked it takes the client to the admin dashboard wherein the client i.e., the teacher can enter student details into the database and can also edit the details of the students whose data is already registered in the database. If the client who visited the website is a student, then they can navigate to the student login page and enter the credentials given to that particular student. After the student is login with the credentials, a page containing the student's details (USN of the student, percentage scored by the student in all the seven semesters) is displayed. When the student clicks on the predict button present in the page, the particular student's form details are sent to the trained machine learning model trough an API. Trained machine learning model is created by first collecting the data from K S School of engineering and management. All the small data is combined into a large dataset called the integrated dataset.

This dataset is pre-processed by removing the errors and noises which can make the model less effective. Also, the missing data is handled by computing mean method and the outliers are handled by using inter quartile method based on the distribution of the columns. This pre-processed dataset is divided in a ratio of 80:20 where 80 percentage of the dataset is separated for training and 20 percent of the dataset is assigned for testing the model. The model is implemented support vector machine, random forest, decision tree, and K-Nearest Neighbor (KNN) algorithms. The model is trained using the training dataset. The model that is trained on various algorithms is tested using the testing dataset. The accuracy obtained from different machine learning algorithms is noted and the best algorithm amongst them which gives high accuracy and less error rate is selected as a final model. The final machine learning model takes the details of the student from the form as an input and predicts the grade that particular student is likely to score in the eighth semester. The grade predicted by the machine learning model is sent back to API and in return these predicted grades is sent to the website. Finally, these predicted grades are displayed on the website for the student.

IV. IMPLEMENTATION

4.1 Data Collection

The data used in this research is collected from K S School of Engineering and Management. Real historic data of the students who studied undergraduate course in computer science and engineering is collected across 2010, 2011 ,2012 ,2013 ,2014 ,2015 ,2016 ,2017 batches. The dataset constitutes a total of 662 instances. The information of each student include percentage procured in all the 7 semesters (Sem_1=60, Sem_5=80) as attributes and grade, this particular student is likely to secure in the final semester, as target.

4.2 Data Cleaning

Irrelevant perceptions are any sort of information that is of no usage to us and can be eliminated instantly. Structural errors that emerge during estimation, transfer of data, or other comparable circumstances are removed.

4.3 Data Preprocessing

This phase in this study deals with the missing instances and other problems associated with the dataset. Initially the missing values are filled by calculating the mean with respect to the particular row. All the rows containing less than six non-NA values are discarded. If a particular row has missing values of more than two columns, then this row is deleted.

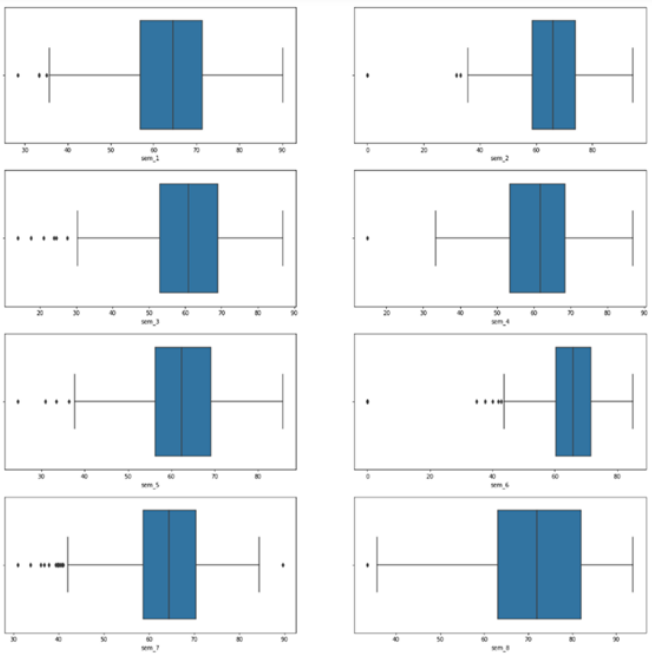

Fig 2: Box Plot

Outliers are the values which vary significantly from all the other values. These outliers are caused by measurement or execution error. Most of the data mining strategies eliminate these outliers, however, there are also some outliers which are found to be good type and these can be ignored. This study uses correlation and scatter plot to detect these outliers. The outliers in the dataset used for this study is displayed in fig.2 IQR is utilized to measure variability by partitioning a data index into quartiles.

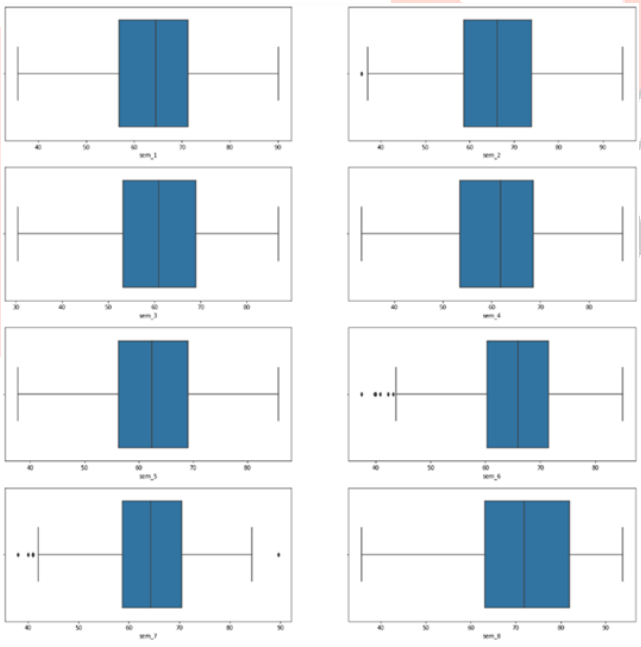

Fig 3: Box Plot without Outliers

The above fig.3 represents the box plot after removing the outliers using IQR method.
The fig.4 depicts the heat map that is created for this study. It shows that wherever there is a greater correlation between the points a lighter color is observed that is, 1.0 to 0.8. the darker color described that the points differ at a high level from each other. As observed in the above heat map eighth semester has a darker color because this column has a very less correlation with all the other columns in the dataset.

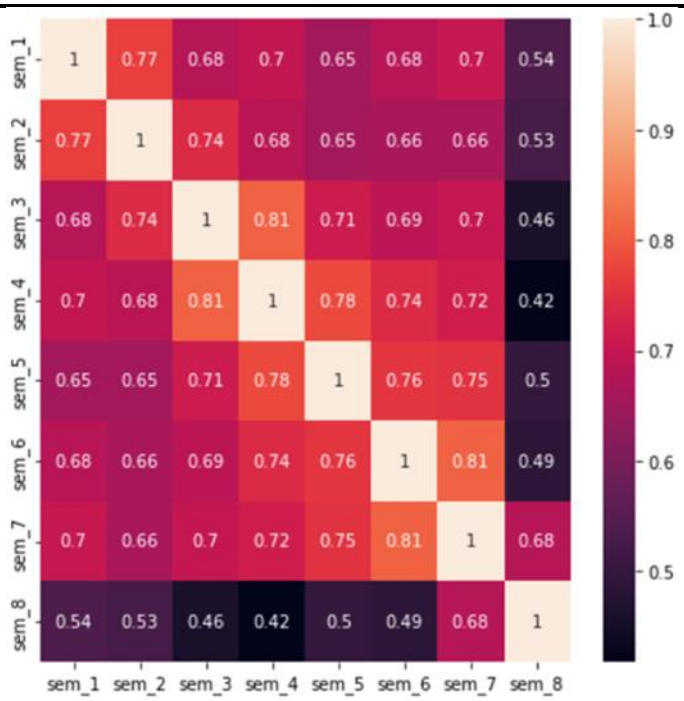

Fig 4: Heat Map

4.4 Data Scaling

There are numerous methods for doing this data scaling, this study uses MinMax scaler method. In this method, the lowest of feature is made equivalent to zero and the highest of feature is equivalent to one. MinMax Scaler compresses the data inside the given limit, typically of 0 to 1. It changes data by scaling features to a given limit. It scales the values to a particular value limit without changing the state of the primary distribution of the data.

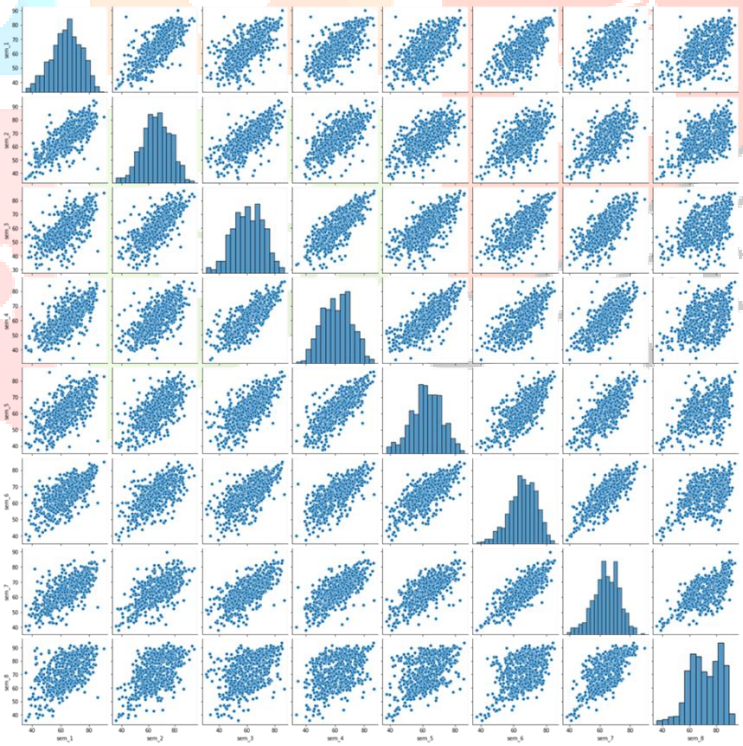

Fig 5: Pair Plot

In Fig 5 the correlation ranges between [50 - 80], positive co-relation is observed.

4.5 Cross Validation

In machine learning, we were unable to fit the model on the training data and can't say that the model will turn out precisely for the real data. For this, we should guarantee that our model got the right patterns from the data, and it isn't getting up a lot of noise. For this reason, the cross-validation procedure is used. Cross-validation is a method wherein the model is trained using the subset of the data collection and afterward assess using the correlative subset of the dataset. This procedure primarily saves some piece of sample dataset, utilizing the rest of the dataset to train the model. The model is then testing using the piece of the dataset. This study uses the K-fold cross-validation technique for this procedure.

V. COMPARISON WITH OTHER ALGORITHMS

This step comprises of model training, pattern identification, testing, assessment results. As referenced prior dataset was separated into testing and training sets. In the training set, the model is built from the classification techniques. Testing set is utilized to evaluate the model. After those results will be assessed. To really look at which of the algorithm will best suit in prediction, the following algorithms have been tested:

Naïve-Bayes algorithm
Logistic Regression
Random Forest
Support Vector Machine (SVM)
K Nearest Neighbour (KNN)

The drawn-out key pointers, which was extracted from the visualization analysis, were augmented in the five chosen classification algorithms. However, it's worth focusing on that since the picked classification algorithms have the capacity to train two distinct attributes' types, that is nominal and numeric. After training the model with these algorithms, the accurate results were assessed to see which variable type can work effectively with every classification algorithm in training the datasets of interest. As a result, Random Forest model reported the highest accuracy rate in predicting students' grades. This algorithm recording an accuracy rate of 71%.

Table 1: Accuracy Results

Classification Algorithm	Accuracy	Macro Average	Weighted Average
Random Forest	0.71	0.72	0.71
K-NN	0.70	0.70	0.69
Support Vector	0.68	0.67	0.66
Logistic Regression	0.61	0.61	0.60
Naïve Bayes	0.60	0.58	0.57

The above Table.1 displays the accuracy recorded for every classification algorithm used in this study. Accuracy column depicts how accurate are the results produced by the particular algorithms.

VI. APPLICATION DEVELOPMENT

The implementation of machine learning tools in application development is a boundless practice today. A ton of arrangements that you utilize consistently depend on ML. For instance, in the event that you watched a film yesterday, Netflix would suggest comparable content for you the following day. Furthermore, assuming you began watching a film, deserted it, and gave a terrible review, machine learning algorithms would comprehend that you don't like such content and won't propose similar suggestions. Similarly in this study a user interface is developed wherein the students will be assigned with an individual login credentials.

Students can login to the web application where they can view their marks obtained from first to seven semesters and can predict their grade by clicking on the button 'predict'. Admin or teacher can login to the admin account where they can enter student data into the application manually or can also upload bulk student data in a .ods (open document spreadsheet) file format into the database directly. Admin is also able to view the uploaded data in the web application and has the access to edit or delete the student details.

VII. CONCLUSION AND FUTURE WORK

In this study, Classification techniques are used to predict student performance. Classification generally refers to the mapping of data items into predefined groups and classes. A classification algorithm analyses the training data during the learning phase, and the test data are used to evaluate the accuracy of the classification algorithms during the classification phase. Classification techniques such as Support vector machine (SVM), Random Forest (RF), Decision tree, and K-nearest neighbor (KNN) algorithms are used to predict the performance of the students.

The success of the student in competition is significantly influenced by their performance in the semester exams. this approach aids students who are at risk of having a weak performance by offering an early prediction that improves their performance in the upcoming semesters. The early prediction of marks helps the students to reach their goals and can perform better by studying hard. In this analysis, the Random Forest has given more accuracy of 74% compared to other algorithms. A user interface was developed so that after logging in, individuals can use their preview semester marks to predict their results.

In future work, we can include more factors, and using more datasets the performance of the model can be improved and students' performance can be improved.

REFERENCES

[1] Kaur, P. and Singh, W., 2016, August. Implementation of student SGPA Prediction System (SSPS) using optimal selection of classification algorithm. In 2016 International Conference on Inventive Computation Technologies (ICICT) (Vol. 2, pp. 1-8). IEEE.

[2] Hasan, H.R., Rabby, A.S.A., Islam, M.T. and Hossain, S.A., 2019, July. Machine learning algorithm for student's performance prediction. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[3] Zohair, A. and Mahmoud, L., 2019. Prediction of Student's performance by modelling small dataset size. International Journal of Educational Technology in Higher Education, 16(1), pp.1-18.

[4] Obsie, E.Y. and Adem, S.A., 2018. Prediction of student academic performance using neural network, linear regression and support vector regression: a case study. International Journal of Computer Applications, 180(40), pp.39-47.

[5] Xu, J., Moon, K.H. and Van Der Schaar, M., 2017. A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing, 11(5), pp.742-753.

[6] Bujang, S.D.A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H. and Ghani, N.A.M., 2021. Multiclass prediction model for student grade prediction using machine learning. IEEE Access, 9, pp.95608-95621.

[7] Gull, H., Saqib, M., Iqbal, S.Z. and Saeed, S., 2020, November. Improving learning experience of students by early prediction of student performance using machine learning. In 2020 IEEE International Conference for Innovation in Technology (INOCON) (pp. 1-4). IEEE.

[8] Shah, M.B., Kaistha, M. and Gupta, Y., 2019, November. Student performance assessment and prediction system using machine learning. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 386-390). IEEE.

[9] Turabieh, H., 2019, October. Hybrid machine learning classifiers to predict student performance. In 2019 2nd international conference on new trends in computing sciences (ICTCS) (pp. 1-6). IEEE.