Student Result Prediction in Covid-19 Lockdown using Machine Learning Techniques

Dr. D. Santhi Jeslet¹, Dr. Dinesh Komarasamy² J. Jenita Hermina³
¹Professor and Head
Department of Computer Science, M.G.R.College, Hosur, TN, India
E-mail: santhijeslet@gmail.com
²Assistant Professor (SRG)
Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, TN, India
E-mail: dinesh.nova@gmail.com
³PG Student
Department of Computer Science and Engineering, Er. Perumal Manimekalai College of Engineering, Hosur, TN, India
E-mail: jeniihermina@gmail.com

Abstract

Covid-19 virus has moved the world to a static state. This has confused the education field in conducting the end examinations. The school education has declared the results of a few classes as pass without conducting the end examinations. But, this may not be possible for all the education system, especially for final year college student results. For this reason, the prediction of result based on their previous performance is essential. Prediction can be effectively achieved by using machine learning algorithms. Machine learning automatically learns and improves from example and experience. Supervised machine learning algorithms Logistic regression and SVM are used for this work. The data set of 1460 students result of a college is considered for the study. Finally, the trained machine predicts accurately whether the student is eligible to acquire the degree or not and the same is viewed in the college portal.

Keywords: Machine Learning, Dataset, Logistic Regression, SVM

1. Introduction

The student result prediction based on the previous performance is carried out using machine learning techniques. Machine learning is a trending concept that outperforms data mining. The concept behind machine learning is that a model or a machine that automatically learns from past experiences to predict the unseen data [1]. The algorithms of machine learning analyze a large amount of data to predict new information. The application of machine learning has a huge impact in the field of education, medicine, automobile, etc., The best example is the search engine which identifies the phrases using the keywords with the help of machine learning techniques [2]. It is also used in image
recognition to identify an object in an image [3]. Besides, it can also predict the purchase behavior of the customer based on their previous purchase.

Based on data fed to the model, algorithms in machine learning are categorized into three categories namely supervised learning, unsupervised learning and reinforcement learning. The data fed to the model arrives with an established class structure in supervised learning [4] [1]. This input data acts as a teacher and its role is to train the model or the machine.

The model in supervised learning generally predicts the properties based on the other related features. Moreover, the model is used to predict the unseen data that has the same features. Whereas in unsupervised learning, there is no defined class structure for input data and the algorithm’s job is to find a structure in the dataset by constructing clusters [5] [1]. But, adding labels is not possible in clustering. Reinforcement learning is close to unsupervised learning in which the input data are unlabeled. It finds the best outcome through the ability of the agent which interacts with the environment. It follows the concept called the hit and trial method. The agent is praised or fined for a right or wrong answer and the model trains themselves based on the praised points. Once trained, it gets ready to predict the new data given to it.

The upcoming section 2 describes the literature review. In section 3 the proposed work is discussed and section 4 describes the experimental result of the work.

2. Literature Review

Dropout prediction was implemented at the Eindhoven University of Technology by Dekker et al., for evaluating the effectiveness of machine learning. Machine learning techniques like CART, Bayes and J48 classifier were used to compare the effectiveness of machine learning techniques. It is concluded that J48 classifier gives more accuracy than other techniques [6].

Researchers from various universities in India analyzed the dataset of university students using numerous algorithms and compared the precision and recall values. It was found that ADT decision tree model shows high accuracy [7].

Another study was carried out at the University of Minho, Portugal by Cortez and Silva for performance prediction. The data set contained information about the marks of Math and Portuguese language for predicting whether the student would pass or fail. Machine learning algorithms like decision tree, random forest, neural networks, and support vector machines were used and compared for accuracy. On the other hand, the examination dataset of a student in different colleges was compared. Inclusion of the past grades resulted in an improved performance [8].

A study by Jalota et al., depicts that an upcoming research area that uses data mining techniques called Educational Data Mining. It involves machine learning and statistical techniques to interpret the learning methodologies of the student, to predict their academic performance and suggests improvement if required [9]. The prediction result of the students is calculated using the confusion matrix [10]. From the confusion matrix accuracy is calculated to know the performance of the system [11].

Similarly, Raj kumar et al., proposed a system that predicts the result of the college students based on the total marks and percentage. The system used two machine learning algorithms SVM and KNN. The two algorithms were compared upon accuracy. The comparison shows that SVM outperforms KNN [12].
Many of these researches are similar that uses different algorithms to build prediction models and prediction are made. Then the models are compared using accuracy and precision. But none of these studies have considered the marks of all the semesters for predicting whether the students pass or fail in the final semester to acquire the degree. Moreover, a logistic regression algorithm is not used in most of the predictions. So this work develops the prediction model using both linear regression and SVM on the data set which has the information of the students such as their register number, number of arrears, total mark of all the previous semester etc.

3. Proposed Work

The flow diagram of the proposed work is depicted in Fig.1 First, the dataset needed for the work has to be selected. Here the students result of the college is selected as the dataset. Next, the selected dataset is divided into a training dataset and test dataset.

After splitting the dataset, prediction models are trained by two algorithms namely Logistic Regression and Support Vector Machine (SVM). Finally, based on the output of the algorithms the result of the students who are eligible for acquiring the degree is predicted. And the results are uploaded to the portal.

a) Dataset Collection

In general, the result of the students in different colleges has a different format. In this work, the data set is selected from a college which comprises 15 different attributes like college code, Date of Birth, register number, name of the student, etc.,. Out of these 15 attributes only three main attributes are selected for prediction. Those attributes are student roll number, the total number of arrears up to V semester, total marks up to V semester and the result (Pass/Fail). Here the value of the result is either 0 (Fail) or 1(Pass). The result attribute is the class label. The dataset which is in Excel format is converted to CSV format. The collected dataset is preprocessed for further steps of the work.
b) Data preprocess

Preprocessing is a process of making the data suitable for the model. And by preprocessing, the accuracy can be increased. Data preprocessing involves two steps like data split and feature scaling.

i) Data Split.

The data set of the college selected for the work contains more than 5350 records. Since this work concentrates only on the final year student’s result, only 1460 records that belong to the final year students are considered. Among 1460 students record 25% of the data set is used for testing and the remaining 75% is used for training.

ii) Feature Scaling.

The next step of the data preprocessing is feature scaling. It is a technique for standardizing the dataset’s independent variables in a given range. Here the independent variables are the number of arrears and the total mark of the students. The data has to be standardized because computing any two values from the number of arrears and total marks, total mark values will dominate the number of arrears values, and it will produce an incorrect result. So the data are preprocessed using the standardization method. Equation (1) shows the standardization formula.

\[
X_i = \frac{x_i - \text{mean}(x_i)}{a}
\]  

(1)

Where \(X_i\) - new value  
\(x_i\) - each value of independent variable in dataset  
\(a\) - Standard deviation  
\[\text{Standard deviation}(a) = \frac{\sum(x_i - \text{mean}(x_i)^2)}{N}\]  

(2)

Where \(N\) is the total size of dataset.

From the testing data, one of the independent variables Number of Arrears (\(x_i\)) is taken as an example for calculating the feature scaling. The size of the testing dataset (\(N\)) is 365. By using the formula, the mean value is calculated as 2.909836. From (2) Standard Deviation is calculated as 4.111. With the values, new standardized values of the number of arrears are formed as -0.4645, 0.2651…,-0.7078.

iii) Managing missing data.

Handling the missing values in the dataset plays a major role. All the missing data is replaced by -1 in this case.

c) Training

After data preprocessing, the preprocessed data are used for training. The training data is used to train the machine as per the machine learning algorithm selected and the test data is used to find out how perfectly the machine can predict new answers based on its training. There are various types of algorithms under the three categories of machine learning. The categories are supervised learning, unsupervised learning and reinforcement learning. Among these, supervised learning is used because training is based on the class label. In supervised learning, there are two main categories of algorithms namely classification and regression. The work has to predict and classify whether the student passes and acquire the degree or fail, classification algorithms are the best choice. Out of many algorithms under classification, Logistic Regression and SVM algorithms are used for prediction in this work.
i) Logistic regression.

Logistic regression is a technique that is generally used for binary classification problems like yes or no, pass or fail etc. Logistic Regression examines the relationship between the label to be predicted and other features using the logistic function to estimate the probabilities. To make a prediction, the probabilities from the logistic function must be converted into binary values. The function used is known as a sigmoid function which takes any real-valued number and visualizes it to a value ranging between 0 and 1, but never be 0 or 1 exactly.

![Logistic Regression Diagram]

To make the value 0 or 1 threshold classifier is used. The threshold value is fixed to 0.5, where all the values are equal to or greater than the threshold value are assigned to one class and all the other values are assigned to another.

ii) SVM.

SVM is a machine learning algorithm that produces a good accuracy rate with less computation power. Moreover, both regression and classification problems can be solved using SVM. However, the classification problem is focused on SVM. SVM locates the hyperplane in N-dimensional space (N — the number of features) that helps to classifies the data points. Data points on either side of the hyperplane can be treated as distinct groups and the scale of the hyperplane is based on the number of attributes.

d) Prediction

After training the models with the training set, the next step is to predict the data based on the test set. In general, the output of the prediction shows the overall predictions of test data. This work also predicts the result for the user given data. By assigning the No. of Arrears and Total marks of the student, the results either pass or fail are predicted. From the prediction, the result of the student for the final semester can be visualized.

e) View result in portal

Once after the data is predicted, the predicted results are gathered in the form of a CSV file. And the CSV file is uploaded to the college portal. From which the students can view their semester result.

4. Experiment and Result

The research aims to implement the two machine learning algorithms (logistic regression and SVM) to predict whether a student acquires the degree or not based on their previous performance up to V semester. The prediction models are created using the Python language. The sample data for both algorithms are shown in table 1.
Table 1. Sample dataset

The evaluation criteria are used to calculate the consistency of the model. There are several different types of evaluation criteria available for evaluating a model. They are confusion matrix, accuracy, precision and recall. With the help of confusion matrix, accuracy can be calculated. A confusion matrix is used to check the accuracy of the classification. Table 2 shows the general confusion matrix.

Table 2. General confusion matrix

The confusion matrix obtained using the test data and the output of prediction for both the models is shown in table 3.

Table 3. Confusion Matrix of Models

Table 3 shows that the correct output is 364 and the incorrect output is 1. The accuracy can be measured using confusion matrix. Equation (3) gives the accuracy of the model.

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]

Accuracy = (137+227)/(137+1+0+227) = 0.997

The Accuracy for both the model is 99.72%

Precision is a measure that quantifies the number of accurate results that are made correctly. Accordingly, precision estimates for the minority class. Equation (4) gives the Precision value.

\[
Precision = \frac{TP}{TP + FP}
\]

Precision = 137/137+0 = 1
The Precision for both the models is 100%.

A Recall is a metric that evaluates the number of correct positive predictions that could have been made from all the positive predictions. The equation (5) is used to calculate the recall value.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

Recall = $\frac{137}{137+1} = 0.9927$

The Recall for both the model is 99.27%

Using the training data collection both the model is well trained. Now, for new observations (Test set) result is visualized. For visualization, the rectangular grid is created. The resolution of the pixel points is 0.01. The X-axis is the number of arrears and the Y-axis is the total mark of the student.

The fig. 3 and 4 show the visualization of the test set. Both the classifiers have divided the student’s dataset into two regions (Pass/Fail). Fig.3 shows the logistic regression model in which the students who have no arrears with high total marks are indicated in the brown region with brown points. And the students who have many numbers of arrears and few total marks are in the blue region with blue scatter points. Fig.4 shows the SVM model in which the students who have no arrears with high total marks are indicated in the red region with red points. And the students who have many numbers of arrears and few total marks are in the green region with green scatter points.

![Figure 3 Logistic Regression](image1.png)

![Figure 4 SVM](image2.png)

![Figure 5](image3.png)

![Figure 6](image4.png)

![Figure 7](image5.png)
Figure 5, 6, 7 shows the result of the models using the unseen data. Fig. 5 shows the prediction of the student who scored a total of 1900 with no. of arrears 10. Since the total is low and the no. of arrear is more, the output of the prediction shows that there are more possibilities for the student to be failed in the final semester. Fig. 6 shows the prediction of the student who scored total of 2500 with nil arrear. As the student has no arrears, the output of the prediction shows that there are more possibilities for the student to be passed in the final semester. Fig. 7 shows the prediction of the student who scored a total of 3000 along with the no. of arrear as 1. Even though the student has 1 arrear the total of the student is higher hence the output of the predictions shows that there are more possibilities for the student to be passed in the next semester shows that there are more possibilities for the student to be passed in the next semester.

Figure 8 shows the accuracy comparison between the existing model (KNN) [12] and the proposed work. From the above graph, it is clear that the KNN shows less accuracy when compared to the proposed work. Because the student dataset is larger, KNN model computes less accuracy. Therefore, for predicting student results the existing algorithm is not suitable. Hence, the proposed work outperforms the existing work.

5. Conclusion

Machine learning’s effectiveness in predicting student output depends on the good use of algorithms for data. To achieve the best results, it is important to choose the correct machine learning approach for the right data. In this work, two supervised machine learning algorithms namely Logistic regression and SVM are used for developing a prediction model. The data set of 1460 student’s results of a college is considered for the study. The training set has 1095 students record and the remaining 395 students record are test data. The model predicts the result based on the number of arrears and total marks of the student. Based on the confusion matrix of test data, the evaluation metrics like accuracy, precision and recall are calculated. It is found that for both the models, the accuracy is 99.72%, precision of 100% and recall value of 99.27%. Both the models show the same result. The trained machine can predict accurately whether the student is eligible to acquire the degree or not. During this covid 19 virus lockdown it is not possible to conduct the examination for all the college students. Instead, it is possible to predict the result of the final year students through this trained machine since it shows more accuracy. So it can be concluded that this prediction system using machine learning techniques is one of the needs of the hour in this pandemic period.
6. References

[1] Mitchell T M 1997 Does machine learning really work?. *AI magazine* 18 3 11-11

[2] Witten I H, Frank E, Hall M A and Pal C J 2016 Data Mining: Practical Machine Learning Tools and Techniques

[3] Alpaydin E 2014 *Introd. to Mach. Learn* Introduction to Machine Learning Ethem Alpaydin

[4] Mohri M, Talwalkar A and Rostamizadeh A 2012 *Foundations of machine learning (adaptive computation and machine learning series)* Mit Press Cambridge.

[5] Sugiyama M 2015 *Introduction to statistical machine learning* (Morgan Kaufmann)

[6] Dekker G W, Pechenizkiy M and Vleeshouwers J M 2009 Predicting Students Drop Out: A Case Study *International Working Group on Educational Data Mining*

[7] Yadav S K, Bharadwaj B and Pal S 2012 Mining Education data to predict student’s retention: a comparative study *arXiv preprint arXiv* 1203-2987

[8] Cortez P and Silva A M 2008 Using data mining to predict secondary school student performance

[9] Jalota C and Agrawal R 2019 *Analysis of educational data mining using classification* In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing COMITCon (IEEE) pp 243-247

[10] Fawcett T 2006 Introduction to receiver operator curves *Pattern Recognit. Lett* 27 pp 861-74

[11] Powers D M 2020 Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation *arXiv preprint arXiv:2010.16061*

[12] Kumar R and Sharma A 2016 Comparative Analysis of SVM & KNN For Academic Prediction of students *International Journal of IT & Knowledge Management* 10 pp 15-19