Available online @ https://jjem.jnnce.ac.in
https:www.doi.org/10.37314/JJEM.2021.050104
Indexed in International Scientific Indiexing (ISI)
Impact factor: 1.025 for 2018-19
Published on: 30 September 2021

The Road to Reopening Educational Institutes through Utilizing Machine Learning

Shashank G S[1], Yashas Vinay[2]

[1, 2]Department of Electronics and Telecommunication Engineering,
JNN College of Engineering, Shimoga, Karnataka, India

shashankgs1902@gmail.com, yashasv9@gmail.com

Abstract

The coronavirus transcended from epidemic to a pandemic in March 2020. Due to this many countries across the world went into lockdown, affecting daily activities and thereby severely hampering educational activities, specifically in rural areas and in particular in under-developed countries. This paper describes a supervised machine learning algorithm in detail, which provides a comfort rating by accounting the student demographics, available mode of transport, etc., across a given region, to help educational institutes across the world to come up with a suitable plan to restart their activities risk free amidst the pandemic.

Keywords: *Machine Learning, Covid-19, Comfort rating, demography, R programming, Data pre-processing, quantile function*

1. Introduction

Even amidst the pandemic, institutes are keenly looking forward to safely reopen and return to the more productive in-person classes. So, it became necessary to devise a pandemic class routine which had to account for all the constraints of the current situation.

This paper proposes a supervised ML method - multiple linear regression, trained using processed data obtained from a survey taken from students accounting for three major parameters-student demographics, regional provisions, available modes of transport-to-analyze the risks involved and determine a rating or a safety index which would help institute store-schedule their activities to host in-person classes risk free. Also, such statistical analysis of the risk equips students to plan their routine, especially that of transport.

2. Block Diagram of the working model

As shown in the Figure 1, first phase in building any machine learning model is training the model. The data is obtained through different methods. One of the most common ways is by conducting the surveys. The data collected via the survey is very crucial and is the building block for the model. The larger the amount of data obtained, the accuracy of the output model increase and gives output with greater precision.

The Linear Regression model is built using the data which is collected. Before building the model, it is important to clean the data and make sure it is free of outliers and skewness. This process is called as “data pre-processing” or “data cleaning”.

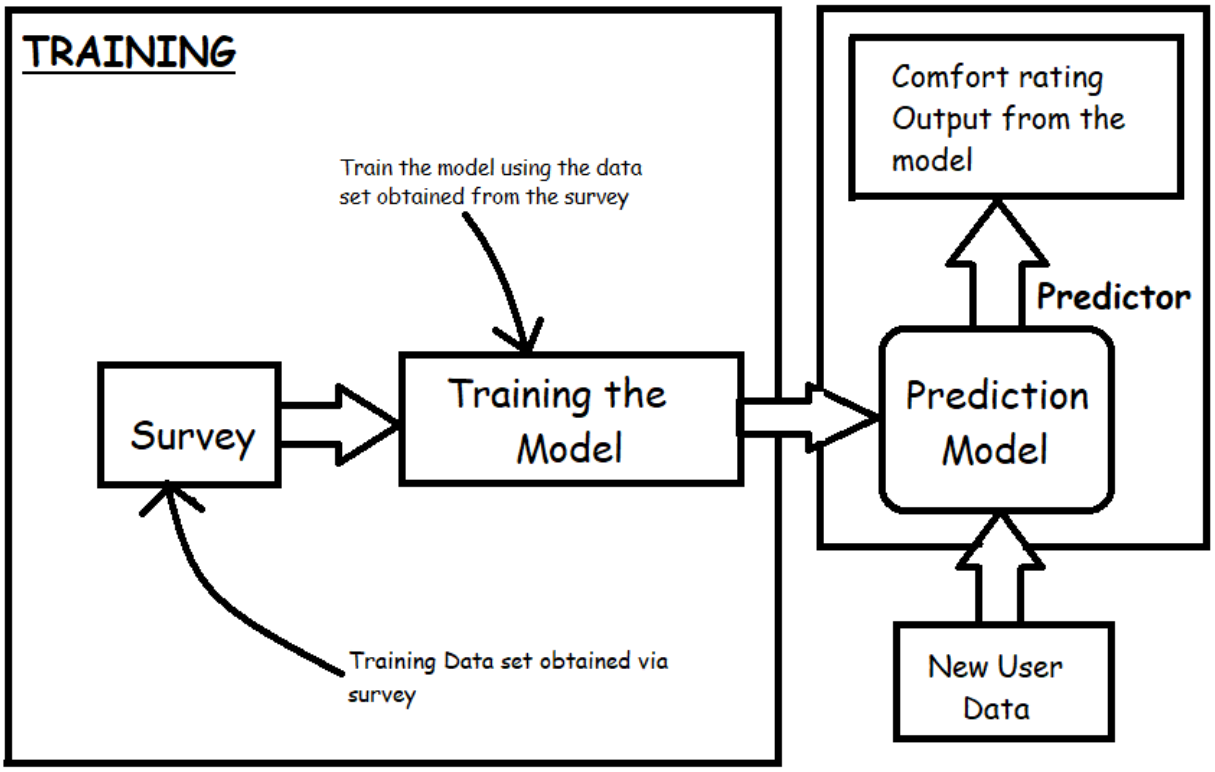

Figure 1: Block diagram of proposed methodology

If the values of the feature obtained by the survey vary over a large scale, we reduce the scale for the purpose of convenience while building the model. Although the scale is reduced the significance of the values will not change. Different processes which could be used to do this procedure are “Normalization” or “Standardization”.

After all the pre-processing and data-cleaning the resultant dataset is fit to be used for the purpose of building the model.
In general, the steps involved in building a simple linear regression model are as follows: -

1. Collecting the Data
2. Data pre-processing
3. Generating the model

Each of the above mentioned steps is explained below in detail.

2.1 Collecting the data

A survey was conducted which was taken up by 400+ students making up a diverse demographic, who responded to all the questions asked based on their status and situation.

The questions regarding their stream of study, current residence, post-reopen residence, approximate distance from institute, regional provision – containment zones, preferred modes of transport, likelihood of attending after re-opening, etc., were asked.

Collecting individual data, over synthetic, constrained generation of data seemed sane as it would account for the practical ground realities.

Answers to the questions such as,

- Whether students' residence comes under the containment zone/sealed zone/red zone.

- The distance of students' residence from the college.

- The type of residence of students:
  - Home
  - Hostel
  - PG

- Mode of transport to college:
  - Public
  - College
  - Personal

were identified to be used to train the correct fit algorithm.

It must be noted that all data collected will contain outliers, may be skewed, and can disrupt the model which eventually leads to over-fitting or under-fitting. This necessitates "Data pre-processing".

2.2 Data Pre-processing

In data pre-processing, the raw data collected is cleaned or organized and prepared to make it suitable for building the Machine Learning model. This is a very crucial step while we are in the process of building the model.

A real-world dataset generally contains noise, missing values, and may exist in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required for cleaning the dataset and making it suitable for a machine learning model and as a secondary effect increases the accuracy and efficiency of said machine learning model.

These unusable data points or data set are called as outliers or skewness. It is important to account for the outliers by removing them to help in streamlining the machine learning and calculation processes.

In this case quantile function is used since there were some values which went above the range of other data points.

These data-points were normalized with the remaining data points using quantile function.

Using the features, the model predicts the comfort rating – prediction variable, of a student which in turn is used to determine the student safety index in terms of percentage.

If the feature contains values which are in binary - yes/no format, "dummy variables" were used further dividing the features into sub-features using R's data pre-processing the library called "dummy". Figure 2 shows the features before adding dummy variables.

```
> summary(df)
 red_zone   dist_frm_clg      residence
 No :237   Min.   :  0.20   Home  :239
 Yes:122   1st Qu.:  6.00   Hostel: 91
           Median : 20.00   PG    : 28
           Mean   : 44.74   Room  :  1
           3rd Qu.: 60.00
           Max.   :225.00
   transport   comfort_rating
College :233   Min.   :1.00
Personal: 69   1st Qu.:1.00
Public  : 57   Median :2.00
               Mean   :2.05
               3rd Qu.:3.00
               Max.   :3.00
```

Figure 2: Features before adding dummy variables

The following features are subdivided into sub-features as presented in Table 1.

Table 1: Features and their sub-features

Features	**Dummy Variables**
RedZone	RedZoneYes
	RedZoneNo
Transport	TransportPublic
	TransportPrivate
	TransportCollege
Residence	ResideHome
	ResidePG
	ResideHostel

After normalizing and assigning the dummy variable the linear regression model is derived using the inbuilt functions in R programming.

Figure 3 shows the features after adding dummy variables.

```
> summary(df)
   red_zoneNo      red_zoneYes       residencePG
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
 Median :1.0000   Median :0.0000   Median :0.00000
 Mean   :0.6602   Mean   :0.3398   Mean   :0.07799
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
 transportPersonal transportPublic transportCollege
 Min.   :0.0000    Min.   :0.0000  Min.   :0.000
 1st Qu.:0.0000    1st Qu.:0.0000  1st Qu.:0.000
 Median :0.0000    Median :0.0000  Median :1.000
 Mean   :0.1922    Mean   :0.1588  Mean   :0.649
 3rd Qu.:0.0000    3rd Qu.:0.0000  3rd Qu.:1.000
 Max.   :1.0000    Max.   :1.0000  Max.   :1.000
  dist_frm_clg    residenceHome    residenceHostel
 Min.   :  0.20   Min.   :0.0000   Min.   :0.0000
 1st Qu.:  6.00   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 20.00   Median :1.0000   Median :0.0000
 Mean   : 44.74   Mean   :0.6657   Mean   :0.2535
 3rd Qu.: 60.00   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :225.00   Max.   :1.0000   Max.   :1.0000
  comfort_rating
 Min.   :1.00
 1st Qu.:1.00
 Median :2.00
 Mean   :2.05
 3rd Qu.:3.00
 Max.   :3.00
```

Figure 3: Features after adding dummy variables

2.3 Generating the model

The pre-processed data which has been derived as explained in the previous sections is used as a data-frame to generate a multiple linear regression model. Equation (1) is the general equation for multiple linear regression models:

$$Y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_p x_{ip}+\epsilon \qquad (1)$$

where, for $i=n$ observations
y_i=dependent variable
x_i=explanatory variables
β_0=y-intercept (constant)
β_p=slope coefficients for each explanatory variable
ϵ=the model's error term (residuals)

The model generated as per the data preprocessed and collected is as follows (Equation 2).

Comfort_rating = 2.1075283
+ 0.2695046[RedZone]
– 0.0011319[dist]
– 0.2308670[resideHome]
– 0.1995267[resideHostel]
– 0.0201616[residePG]
+0.251986[transpoCollege]
+0.013585[transpoPersonal] (2)

where,
RedZone = 1 if NO and 0 if YES
dist = any real number distance (in km)
resideHome, resideHostel, residePG = 1 if YES and 0 if NO
transpoCollege, transpoPersonal = 1 if YES and 0 if NO
comfort_rating- it is a number between 1 and 5

Figure 4 shows the coefficient and statistics of the model.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.35409 -0.97044  0.00962  0.80187  1.48378

Coefficients: (2 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.207662   0.910024   2.426   0.0158 *
red_zoneNo         0.259117   0.101156   2.562   0.0108 *
red_zoneYes              NA         NA      NA       NA
dist_frm_clg      -0.002186   0.001072  -2.040   0.0421 *
residenceHome     -0.289057   0.917407  -0.315   0.7529
residenceHostel   -0.199612   0.918298  -0.217   0.8280
residencePG       -0.027071   0.929032  -0.029   0.9768
transportCollege   0.078331   0.136919   0.572   0.5676
transportPersonal -0.189795   0.164577  -1.153   0.2496
transportPublic          NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9043 on 351 degrees of freedom
Multiple R-squared:  0.04668,   Adjusted R-squared:  0.02767
F-statistic: 2.455 on 7 and 351 DF,  p-value: 0.01806
```

Figure 4: coefficient and statistics of the model

Student Comfort Index (SCI), is derived which is basically the percentage notation of the comfort_rating output variable (Equation 3)

SCI= (comfort_rating/5) x 100 (3)

3. Features

The main features or parameters one would like to refer to are as follows:

1. Distance
2. Mode of travelling to college
3. Residence during the academic session
4. Regional provisions i.e. whether or not they were in a containment zone.

These were chosen as the main features as they gave sufficient insight without having to ask for more detail. The answers to these questions detail out how the student demographic of an institute is distributed. Plugging in this data an output value out of 5 to say as to whether students should or should not attend college, with 5 being- student shall attend college if possible and 1 being –student shall continue with online classes.

4. Linear Regression Diagnostics

Before using the regression model to make predictions it is necessary to ensure that the model is statistically significant.

The values which define the significance of the model are:

1. p-value/confidence interval
2. t-value
3. R-squared
4. Adjusted R-squared
5. Standard error
6. F-statistic

Common statistic ad their criterion that should be made sure for the best fit linear regression model as follows (Table 3).

Table 3: Standard Statistics

Statistic	**Criterion**
R-Squared	Higher the better
Adjusted R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05

5. Results

As shown in the previous sections a good fit model was obtained with a low p-value or confidence interval of about 1.8%. The low p-value is a good sign and hence indication of the efficient prediction model.

To check the working of the model generated, testing was done on a separate set of datasets whose comfort rating was already known. Here is the result of comfort rating obtained from the model in comparison with the comfort rating obtained from the students via survey.

Table 4 shows the test cases and Table 5 shows the result comparison.

Table 4: Test Cases

Student	Red Zone	Distance	Residence	Transport
Student 1	No	3	Home	College
Student 2	Yes	6	Home	College
Student 3	No	20	Home	College
Student 4	No	60	Hostel	College
Student 5	No	90	Hostel	College

Table 5: Result Comparison

Student	**Comfort rating (from new user)**	**Comfort rating (output from model)**
Student 1	2	2.249495
Student 2	2	1.98382
Student 3	3	2.2122
Student 4	3	2.2143
Student 5	2	2.32124

As the above table showcases the output from the model, it is evident that for some cases the model fits perfectly and for others there is a small deviation from the expected output. There is significant room for improvement as with a margin of error in some cases exceeding 20%. This is where the introduction of more features/variables and a larger dataset for training which can help further improve the overall accuracy of the model.

The accuracy of the model was calculated by taking the comfort rating, output generated by the model divided by the comfort rating from the survey corresponding to the student and then converted into percentage format (equation 4).

$$Accuracy = \frac{\left(Comfort\ Rating_{model_{output}}\right)}{\left(Comfort\ Rating_{user_{input}}\right)} \times 100 \qquad (4)$$

Average accuracy in given by equation 5.

$$Average\ Accuracy = \frac{\sum Accuracy}{Number\ of\ samples} \quad (5)$$

The accuracy of the model obtained is shown in Table 6.

Table 6: Accuracy

Student	Accuracy
Student 1	87.53%
Student 2	99.19%
Student 3	73.74%
Student 4	73.81%
Student 5	83.94%
Average Accuracy = 83.64%	

6. Further Possibilities

Although the features used in the current algorithm are producing results in the vicinity of the expected output, for more accurate results/output, more poignant and statistically significant features are required to fine tune the machine learning model. This improves the reach of the model and increases the significance of the output deciding which students can or should attend in-person classes if colleges were to open. These features are as follows:

1. Attaching a numerical value to the confidence of a student to write the end of semester exams.
2. Attaching significance to any online courses a student may have taken outside of college courses.
3. Numerical value attached to the student's willingness to attend college.
4. The assignments and projects that have been conducted through online classes, as a rating with regards to conceptual understanding.

These features complemented with the existing core features of distance, housing, zone status and mode of travelling will help determine the knowledge dissemination and assimilation by teachers and students, respectively alongside measuring the ability of students to apply what has been taught. This equips the machine learning model to give more appropriate and useful recommendations to the student as to the necessities of the situation based on their responses to the core variables and the variables dependent on education. This data when given to an institute can then be utilized to accommodate the students who are most in need of help as well as provide an opening to those students who want to utilize college facilities.

7. Conclusion

In conclusion, after running multiple simulations of the model and testing it over a variety of datasets, it is believed that the model is representative of the larger population given that the training dataset was filled in a systematic and logical manner by the individuals who participated in our survey. Assuming this is the case, the model was able to predict the safety rating of a student to an accuracy of 83.64% with a confidence interval 1.8%.

The other conclusion that can be drawn is that the model is able to provide an accuracy of 83.64% for a special case as our primary training dataset consisted of 351 discrete inputs, which is open to human error. If this is the case then to validate our model and to improve upon it, we suggest a larger input dataset to obtain a more representative machine learning model.

Finally, this model can be applied to, albeit with changes in the statistical regression models, other working environments to improve employee safety and increase productivity by utilizing features that are more suitable to the chosen environment.

References

1. Sina F. Ardabili , Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M. Atkinson, COVID-19 Outbreak Prediction with Machine Learning, Algorithms, Vol.13, Issue.10, 249, October 2020, pp.1-36. https://doi.org/10.3390/a13100249

2. Samuel Lalmuanawma, Jamal Hussain, Lalrinfela Chhakchhuak, Applications of Machine Learning and Artificial Intelligence for Covid 19(SARS-CoV-2) Pandemic: A Review, Chaos Solitons & Fractals, Vol. 139, 110059, June2020.
doi: 10.1016/j.chaos.2020.110059

3. Uma K, M.Hanumanthappa, Data Collection Methods and Data Preprocessing Techniques for Healthcare Data Using Data Mining, International Journal of Scientific & Engineering Research Vol.8, Issue 6, June-2017, pp.1131-1136.

4. Anjali Pant and R S Rajput, Linear Regression Analysis Using R for Research and Development In Book:Writing Qualitative Paper of International Standard, 2019. pp.180-195

5. Erhard Rahm, Hong Hai Do, Data Cleaning: Problems and Approaches, IEEE Data Engineering Bulletin, Vol.23,No.4, Dec 2000, pp.3-13.

6. Lokanath Mishra, Tushar Gupta, Abha Shree, Online teaching-learning in higher education during lockdown period of COVID-19 pandemic, International Journal of Educational Research Open, Vol.1, Jan 2020.pp.100012.
https://doi.org/10.1016/j.ijedro.2020.100012