International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST218448

Student's Performance Analysis with EDA and Machine Learning Models

Dr. A. Senthil Kumar[1], K. Joshna[2]

[1]B.E., M.E., PGDVLSI., DISM., Ph.D. (IITR), PDF. (TUT.SA), Senior PDF. (VSD. TUV, EUROPE)
Department of EEE, Principal, Sanskrithi School of engineering, Puttparthi, Andhra Pradesh, India
[2]Department of ECE, Sanskrithi School of Engineering, Puttparthi, Andhra Pradesh, India

ABSTRACT

Article Info
Volume 8, Issue 4
Page Number : 297-302

Publication Issue
July-August-2021

Article History
Accepted : 08 July 2021
Published : 15 July 2021

Educational data analytics is used to study the data which is available in the educational institutions and bring out the insights from it. Analytics is a process of discovering, analyzing, and interpreting meaningful patterns from large amounts of data. Predictive analytics can help in improving the quality of education by providing right information for decision makers to take better decisions. This paper focuses on the need for implementing the data analytics in educational system, suggests some strategies to use these needs. While implementing any system, the understanding of different components and their functions is necessary. The educational data analytics has potential to discover, analyze and predict meaningful knowledge from educational data which will help to education management system for flexible planning, execution and prediction for future.

Keywords : Data Analytics, Visualization Tools, Pandas, Matplotlib, Predictive Modeling, Random Forest.

I. INTRODUCTION

Information regarding the students plays a vital role in predicting their future enhancements. An information system should not only aim to collect, store data and process information but also help in the formulation of education policies, their management and their evaluation. There has been greater interest in highly specific indicators concerned such as attendance rates, students' achievement level and discipline problems.

Currently the huge amount of data is stored in educational databases. These databases contain the useful information to predict students' performance. Educational data analytics is used to study the data available in the educational field and bring out the hidden knowledge from it. Analytics is a process of discovering, analyzing, and interpreting meaningful patterns from large amounts of data.

Data analytics is necessary with great potential to help institutions to focus on the most important information in their data warehouses, Software tools and methodologies have been developed in the field of statistics and data analytics to process data and to let these data more informative to users who need them for decision making.

Data analytics relies on the techniques such as, classification, association, correlation, categorization, prediction, estimation, visualization. Higher education institutions can use classification, for analysis of student performance, or use estimation to predict the probability of a variety of outcomes, such as result, persistence and course success. Data analytics enables educational institutions to better allocate resources and staff, proactively manage student outcomes and improve the effectiveness of education management system.

The Data analytics in educational management system comprises of four main steps:

Data Collection

Data need to be collected regarding courses offered, students enrolled, and their results etc. Data need to be preprocessed and need to be converted in proper format to store.

Data wrangling

Programatically transforming data into a format that makes it easier to work with. It might mean modifying all of the values in a given column in a certain way, or merging multiple columns together. Data that is entered manually by humans is typically fraught with errors.

Data visualization

It uses the matplotlib module for the basic graph representation and also seaborn for the fascinating themes, visualizations etc.,

Model Building

The model is build using the ML techniques such as Random forest , KNN and Decision tree. These are compared with the predictions they get and the one with better prediction is considered as the best approach.

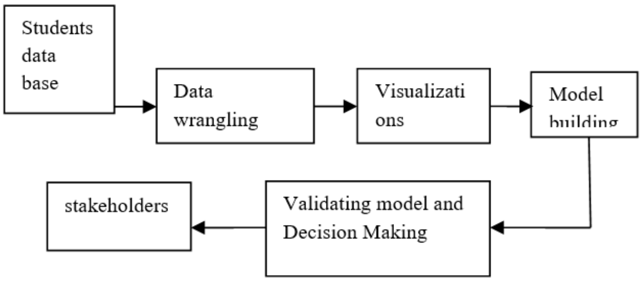


Figure 1. Steps for data analytics in Educational Management Analysis

Analytics enabled decision making

People at top level management as well as middle level management will be able to take better decision by analyzing the data. The model will be helpful for them in decision making.

EDA approach:

Exploratory data analysis or EDA is the critical and first step in analyzing the data. The main reasons to do the EDA are:

1. Detection of mistakes
2. Checking of assumptions
3. Preliminary selection of appropriate models
4. Determining relationships among the explanatory variables
5. Assessing the direction and rough size of relationships between explanatory and outcome variables.

The approach of Exploratory Data Analysis (EDA) process is used in this project in order to perform the initial investigations on data in order to discover patterns, to spot anomalies, to test hypothesis, find interesting relations among the variables and to check assumptions with the help of summary statistics and graphical representations. The EDA approach is precisely that an approach not a set of techniques.

The data obtained is collected into a rectangular array(e.g., spread sheet or database), most commonly with one row per experimental subject and one column for each subject identifier, outcome variable, and explanatory variable. Each column contains the numeric values for a particular quantitative variable or the levels for a categorical variable.

It is not possible for a human to look at a column of numbers or a whole spreadsheet and then determining important characteristics of data. They find looking at numbers to be tedious, boring, and/or overwhelming. Exploratory data analysis techniques have been devised as an aid in this situation. Most of these techniques work in part by hiding certain aspects of the data while making other aspects more clear.

Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate(usually just bivariate)

Most of the EDA techniques used are graphical in nature with a few quantitative techniques. The heavy reliance on graphics is that by its very nature the main role of EDA here is to open-mindedly explore, and graphics give us the unparalleled power to do so, enable the data to reveal its structural secrets, and being always ready to gain some new, often suspected insight into the data.

Specific statistical functions and techniques that are performed with EDA tools for this project include:

1. Univariate visualization of each field in the raw dataset, with summary statistics. It includes the plotting techniques like boxplots and histograms.

2. Multivariate visualizations and summary statistics used to assess the relationship between each variable in the dataset and the target variable. Analysis is done using the scatter plots and Bar charts.

Univariate analysis is the simplest form of data analysis, where the data being analyzed consists of only one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

Box plots

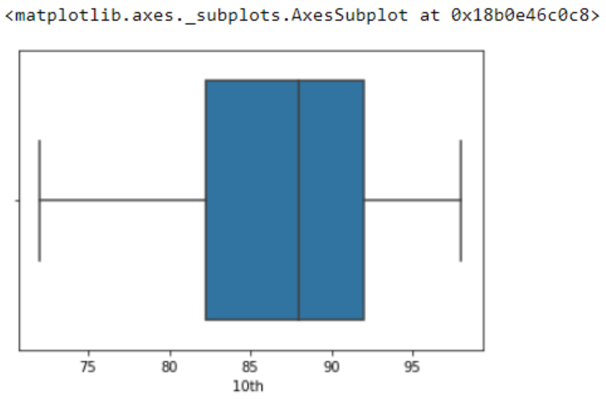
The boxplot is an important EDA tools for determining if a factor has a significant effect on the response with respect to either location or variation. It is an effective tool for summarizing large quantities of information.

Box plots are formed by

Vertical axis: Responsible variable

Horizantal axis: The factor of interest

A box and whisker plot-also called a boxplot-displays the five number summary set of data. The five number summary is the minimum, first quartile, median, third quartile, and maximum.



```
<matplotlib.axes._subplots.AxesSubplot at 0x18b0e46c0c8>
```

Histogram:

A histogram is a plot that discover and show the underlying frequency distribution (shape) of a set of continuous data.

The purpose of a histogram used in this context is to graphically summarize the distribution of a univariate dataset. This allows the inspection of the data for it's underlying distribution (e.g., normal distribution), outliers, skewness.

The histogram used in the project shows the following observations:

1. Center of the data (location)
2. Spread of the data(the scale)
3. Skewness of the data

4. Presence of outliers
5. Presence of multiple modes in the data

The above features give the strong indications of the proper distributions model for the data

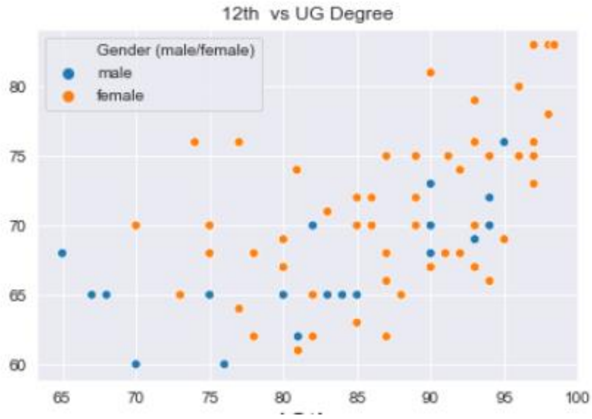
Multivariate Analysis

Multivariate data analysis refers to any statistical technique used to analyze data that arises from more than one variable. It models more realistic applications, where each situation, product, or decision involves more than a single variable.

Scatter plot

A scatter plot is a two-dimensional data visualization or chart type that is normally used here in order to observe and visually display the relationship between variables are represented by dots. The positioning of the dots on the vertical and horizontal axis will inform the value of the respective data point. These plots make use of Cartesian coordinates to display the values of the variables in a data set.

The most common use of the the scatter plot used in this project is to display the relationship between two variables and observe the nature of such a relationship. The relationships observed can either be positive or negative, non-linear, and/or, strong or weak.

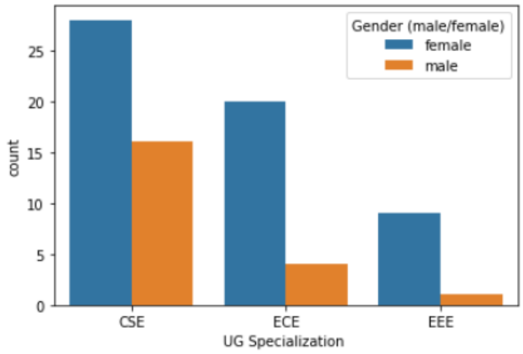


Bar chart

A bar chart shown in the below figure represents the numerical data, with rectangular bars having lengths proportional to the values that it represent.

Bar charts are used to distinguish objects between distinct collections in order to track variations over time. These bar charts are used in this case, because bar charts are very convenient when the changes are large.

```
<matplotlib.axes._subplots.AxesSubplot at 0x18b0d6b9848>
```



EDA helps stakeholders by confirming they are asking the right questions about standard deviations, categorical variables, and confidence intervals. EDA is a crucial step to take before diving into machine learning or statistic modeling because it provides the context needed to develop an appropriate model for the problem at hand and to correctly interpret its results.

EDA is valuable to make certain results they produce are valid, correctly interpreted, and applicable to the context of this project. Once the EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling including machine learning.

II. MACHINE LEARNING MODEL APPROACHES

The paper focuses on the use of machine learning models for predicting the better results after the EDA is done. The ML techniques used in this context helps us to get the still more better insights after EDA. Some of the machine learning models adopted in this project are Random forest, KNN and Decision tree classifier. The techniques are implemented with the data that is obtained after exploratory data analysis is done.

The data that is considered for implementing the model is basically divided in to the training and testing dataset. The dataset is trained with 80% of the data and tested with 20% of the data. After splitting the dataset into train and test, the different models that are mentioned can be used for fitting the model with train dataset. The models like KNN, Random forest and decision tree are considered one by one for the model fitting. After the model is trained with train dataset, it is tested with testing dataset. When the model is run with test dataset, predictions for the particular model mentioned are   are obtained.

The predictions are made individually for each of the model with the dataset considered. The predicted values obtained are always compared with actual values to know the better prediction and best model that is used.

We have the different performance metrics for knowing the better machine learning model. These performance metric helps us understand the difference between the actual values and the predicted and also provides the information about how good the model is for the particular data that is considered. The performance metrics that are considered for this projects are Mean absolute error(MAE), Mean square error(MSE) and Root mean square error(RMSE).

The essential step in any machine learning model is to evaluate the accuracy of the model. The MSE, MAE, RMSE are used to evaluate the performance of the model in regression analysis.

1. The Mean absolute error represents the average of the difference between the actual and predicted values in the data set. It measures the average of the residuals in the dataset.

2. Mean squared error is the average of the squared difference between the original and predicted values in the dataset. It measures the variance of the residuals.

3. Root mean squared error is the square root of mean squared error. It measures the standard deviation of residuals.

COMPARISION OF EVALUATION METRICS OF ALL THE MODELS CONSIDERED:

The evaluation metrics of all the models obtained are placed in one table and the table shows the values of MAE, MSE and RMSE for different models.

grade	algorithm	MAE	MSE	RMSE
10th	Random forest	0.45	3.1	0.55
	KNN	0.757	7.0	0.84
	Decision tree	0.58	5.2	0.729
12th	Random forest	0.602	6.5	0.82
	KNN	0.78	10.09	0.1004
	Decision tree	0.819	11.3	0.106
UG	Random forest	0.445	3.6	0.59
	KNN	0.552	6.0	0.80
	Decision tree	0.638	5.6	0.75

By the above table, the comparision of evaluation metrics of different models is presented. The metrics values obtained for the Random forest model are near to zero and low values, which indicates that after evaluating the metrics for different models, the table shows the low values for all the metrics in Random forest model compared to KNN AND Decision tree classifier.

As the Random forest model has got the low evaluation metrics, it directly indicates that the error between the actual values and predicted values is very low. Using the Random forest model the predicted values are very close to the actual values.

So, from the above mentioned table, Comparing with the KNN AND decision tree, Random forest model can be considered as the best suited Machine learning model for the data that is considered for this project and it is the best predictive model in this context.

III. TOOLS OF DATA COLLECTION & ANALYSIS

In order to build an efficient decision support system, there must be combined several techniques and methods that can improve the performance and the accuracy of the analysis from two major perspectives: knowledge base data and future predictions. Various tools are needed for that project some for analyzing data, some for designing, implementation and some developing software tools such as Anaconda Platform, Jupyter Notebook, Advanced Excel,Python Modules(Pandas, Matplotlib, Seaborn) And Random Forest Regression.

IV. CONCLUSION

In the current education system, the institutions are still using the conventional way of analyzing the student's academic performance, their attendance rates and retention rates. All these things are analyzed either by the teachers manually or the college heads, there is no involvement of using the latest technologies. So, this paper focuses on using the data analytics in improving the student's performance as well as helping the institution in taking the better decision with the EDA and the machine learning techniques.

V. REFERENCES

[1]. Jinan Fiadhi, Lakehead University, The Next Step for Learning Analytics, Published by the IEEE Computer Society 1520-9202/14/$31.00 © 2014 IEEE

[2]. "Learning Analytics Pilot," University of Wisconsin-Maddison, Spring 2014; www.cio.wisc.edu/learninganalytics-pilot.aspx.

[3]. Jindal Rajni and Dutta Borah, Predictive Analytics in a higher education context, August 2015

[4]. Sheila MacNeill, Lorna M. Campbell, Martin Hawksey, Analytics for Education, 2014

[5]. Michelle Davis, "Data Tools Aim to Predict Student Performance," Education Week Digital Directions, February 8, 2012

[6]. B.K. Bharadwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp.

[7]. S. K. Yadav, B.K. Bharadwaj and S. Pal, "Data Mining Applications: A comparative study for predecting students' performance", International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol 1, No. 12, ISSN: 2045-8711, 2011.

[8]. Author name, paper name, title ,conference name, volume, pages.

Cite this article as :

Dr. A. Senthil Kumar, K. Joshna, "Student's Performance Analysis with EDA and Machine Learning Models", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 4, pp. 297-302, July-August 2021. Available at doi : https://doi.org/10.32628/IJSRST218448
Journal URL : https://ijsrst.com/IJSRST218448