Feature Selection for Performance Prediction using Decision Tree

Anirudhd Soni
Oriental Institute of Science and Technology
Bhopal, MP, India

Anansha Gupta
Oriental Institute of Science and Technology
Bhopal, MP, India

ABSTRACT
With the proliferation of digitalization from past decades, there has been exponential growth in data. The data is considered as the new oil. Useful insights can be extracted from the data and can be used for the growth of industry or organization, the branch of computer science that deals with the discovery of novel information and insightful pattern from the raw data is called Data Mining, it is used almost in every field from banking, healthcare to entertainment and surveillance. Here, specifically, the paper discusses about the data mining used in the field of education called Educational Data Mining (EDM), it is an inchoate data mining research area that aims to improve the students' performance, provide quality education, helps students to determine career choices, etc. This research paper aims to describe feature selection criterion and how combination of features like internal marks, mid-semester marks, etc. helps in determining students' performance using decision tree classification method. The results obtain can be used by instructors and teachers to plan structured approaches for the students performing low in their academics, need attention and counseling from their tutor guardians. Thus, with early prediction action can be taken within time to improve the result of students and the overall performance of an institution.

General Terms
Data Mining, Decision tree

Keywords
Educational Data Mining, Classification and Regression Trees, Pearson Correlation, Feature selection

1. INTRODUCTION
Ever since the concept of data mining was introduced, humans are always looking for various ways of applying the concept in different fields, this results in a better understanding of the field which can help in achieving desired results most effectively. As of today, these techniques are used in many fields such as to assess the potential market risks, understanding human behaviour, diagnosing medical conditions, etc. helping us to better understand the science behind it.

This paper will discuss about the Educational Data Mining and how the appropriate features can be selected to use in a decision tree to predict the result of the students, using Pearson Correlation. Feature selection is a crucial step while training a decision tree, using unnecessary extra features or very few features can lead to a decision tree having an inadequate accuracy range. Using Pearson's correlation can help one decide what features are related to the target value and what features are redundant and unnecessary, this can help to create a more accurate decision tree using just the right number of features.

The predicted result of the students can help in identifying the ones who need to work more on their academics to improve their results in the future. Teachers can also use this data to monitor the expected performance of the students and accordingly they can arrange for extra classes for the students who need improvement. This will improve the result of the students, improve the rate of learning and ultimately help the educational institutes to create highly skilled graduates.

2. EDUCATIONAL DATA MINING
Data mining is a process used to process raw data and find meaningful patterns and information in large data sets. This information is then can be used in many ways such as predicting spam, fraud detection, creating business models, predicting the presence of cancer cells, etc. A new branch of data mining, Education Data Mining (EDM) is gaining more recognition these days[9]. In EDM, data mining techniques are used on student-related data and it is studied to make out useful information, this information and patterns then can further be used to predict students' results, their ability to get jobs, future career choices, User Behaviour Modelling, Trend Analysis, etc. This can be very helpful in understanding the way learning works and what are the factors that play a major role in students' life which results in different outcomes for different students.

This paper focuses on the part of predicting the students' result with the help of students' data which include information such as previous marks, attendance, previous educational background, gender, amount of time devoted to study every week, etc. This could help to identify the students in a class who need more attention and where they need improvement for which proper actions can be taken to improve the results.

3. DECISION TREE INDUCTION
3.1 Decision Trees
A decision tree is a supervised learning algorithm for classification and regression[7]. Most commonly it is used for establishing classification system which aims to learn simple decision procedure from the data set and create a model that predicts the class of the target value. It uses branch-like segments and nodes to construct a tree structure that can efficiently deal with large and complex datasets. At each node, one of the features of data is evaluated to separate the observations or to make a tractable path for the decision-making. The tree contains three nodes; root node, internal nodes and leaf nodes. Initially, the root node starts the tree structure, it evaluates the feature that best split the data, next are the internal nodes which are the decision making nodes based on which the tree splits into further branches. The leaf nodes are the final nodes where the final categorization of the target value is done.
Since the Decision tree is categorized under supervised

learning methodology, the construction of the tree requires a training dataset. The dataset contains data tuples describing the value of features and an associated class label. Based on the feature value the tree path is constructed from root to leaf node and followed. Large and complex training datasets are divided further into validation sets to check the accuracy of the model on data tuple whose class label is unknown. The class of the data tuple is ultimately predicted by the leaf node of the Decision Tree.

3.2 Decision Tree Algorithms

ID3 or Iterative Dichotomiser 3 was invented by Ross Quinlan, it is called so because it iteratively divides the features into two or more groups until it gets the node to represent the outcome(leaf node). To construct a tree it uses top to bottom greedy approach i.e. for each decision node it selects the feature that yields the largest information gain for the categorical targets.

C4.5 is a popular algorithm to generate decision tree, it is the successor of the ID3 algorithm used to overcome its limitation, it handles both continuous attribute values and discrete sets of intervals. C4.5 uses the Gain ratio for the feature selection method.

C5.0 is Ross Quinlan's latest version release and an extension of C4.5. C5.0 builds a smaller rule set while without compromising the accuracy of the decision tree.

CART, Classification and Regression Trees, was introduced by Leo Breiman, which can be used for classification and regression predictive modeling problems. CART is similar to C4.5 but supports numerical target variables. CART constructs a binary tree using the feature that yields the largest information gain. In this research Python's Scikit Learn library is used to implement a decision tree that uses an optimized version of CART decision tree and does not support categorical variables for now.

4. METHODOLOGY

4.1 Data collection and preprocessing

Data collection is the first crucial step to train a machine learning model. Discrepancies in the data can lead to futile results. The dataset used in this study is from two Portuguese schools, which includes attributes such as student grades, demographic, social and school-related features. This data was collected by schools using school reports and questionnaires[5]. It consists of 649 instances and 33 attributes. The First 32 of the 33 instances are independent and the last one is dependent on all the other attributes. The data does not consist of any missing values hence it makes it a bit easy to pre-process the data.

Data pre-processing is required to remove redundant and garbage data from the dataset to maximise the accuracy of the trained model, because, the presence of garbage and redundant data can adversely affect the performance of the trained model. The decision tree algorithm which is used in this study requires integer values for every attribute therefore every string value is to be converted in an integer form. The below table contains all the attributes that are used in the data.

1.School - student's school is binary, Gabriel Pereira = 1 or Mousinho da Silveira = 0.

2.Sex - student's sex is binary, Female=0 or Male=1

3.age - student's age is numeric, from 15 to 22.

4.Address - student's home address type is binary, Rural=0, Urban=1.

5.Famsize – student's family size is binary, Greater_Than_3 = 0, Less_or_equal_3 =1.

6.Pstatus - parent's cohabitation status is binary, Apart=0, Together=1

7.Medu - mother's education is numeric, 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.

8.Fedu - father's education is numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.

9.Mjob – mother's job is numeric, at_home = 0, health=1, other = 2, services=3, teacher=4.

10. Pjob - father's job is numeric, at_home = 0, health=1, other = 2, services=3, teacher=4.

11.Reason - reason to choose this school is numeric: home =0, reputition=1, course=2, other=3.

12.Guardian- student's guardian is numeric, mother=0, father=1, other=2.

13.traveltime - home to school travel time is numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - 1 hour.

14.studytime - weekly study time is numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours.

15.failures - number of past class failures is numeric: n if 1<=n<3, else 4.

16.schoolsup - extra educational support is binary: yes=0 or no=1.

17.famsup - family educational support is binary: yes=0 or no=1.

18.paid - extra paid classes within the course subject is either Math or Portuguese which is binary, yes=0 or no=1.

19.activities – involvement in extra-curricular activities is binary, yes=0 or no=1.

20.nursery - attended nursery school is binary: yes=0 or no=1

21.higher - wants to take higher education is binary, yes=0 or no=1.

22.internet - Internet access at home isbinary, yes=0 or no=1.

23.romantic - with a romantic relationship is binary: yes=0 or no=1.

24.famrel - quality of family relationships isnumeric: from 1 - very bad to 5 – excellent.

25.freetime - free time after school is numeric, from 1 - very low to 5 - very high.

26.goout - going out with friends is numeric, from 1 - very low to 5 - very high.

27.Dalc - workday alcohol consumption is numeric: from 1 - very low to 5 - very high

28.Walc - weekend alcohol consumption is numeric: from 1 - very low to 5 - very high

29.health - current health status is numeric: from 1 - very bad to 5 - very good

30.absences - number of school absences is numeric: from 0 to 93.

31.G1 - first period grade is numeric: from 0 to 20

32.G2 - second period grade is numeric: from 0 to 20

33.G3 - final grade is numeric: from 0 to 20, (output target)

This pre-processed data is then used to select appropriate features that will be used to train the decision tree classifier. The G3 (33rd column) is the target value that is to be predicted by the model.

4.2 Feature Selection

Feature selection or variable selection is the process of selecting the best set of relevant features used for model creation. It is the robust method for data reduction and an essential preprocessing phase in productive machine learning applications. [8]Feature selection aims to remove irrelevant and redundant features which can confuse the model when there are limited training examples. In a decision tree model predicting the class of target value with irrelevant features can construct a tree with high depth and a large number of nodes which will ultimately increase the complexity and training time of the model. Thus, to avoid the above reasons, feature selection is used to increases the classification accuracy and reduces the complexity of the model.

The correlation coefficient is used to find how strong a relationship is between two variables. In this paper, Pearson correlation is used for feature selection. It was developed by Karl Pearson[1] from a related concept introduced by Francis Galton[2] in the 1880s. Pearson correlation coefficient(r) is calculated to find the strength of linear correlation between two variables.

Statistically, the Pearson correlation coefficient(r) is calculated as the covariance of the two variables divided by the product of their standard deviations.

Pearson Correlation Formula-

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$r = Pearson\ Correlation\ coefficient$
$x_i = x\ variable\ samples$
$\bar{x} = mean\ of\ values\ in\ x\ variable$
$y_i = y\ variable\ samples$
$\bar{y} = mean\ of\ values\ in\ y\ variable$

The range of coefficient is between -1 to 1, where -1 means a strong negative relationship, 1 means a strong positive relationship and 0 means no relationship at all.

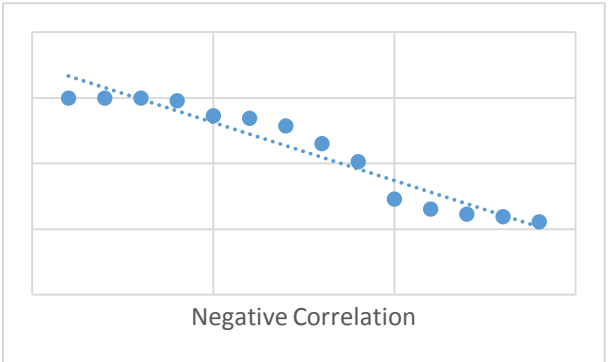

Figure 1: Negative Correlation

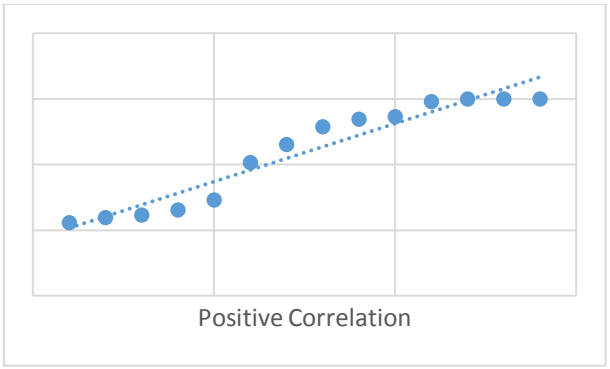

Figure 2: Positive Correlation

Table 1: Range of Correlation

Strength of Association	Positive	Negative
Small	0.1 to 0.3	-0.1 to -0.3
Medium	0.3 to 0.5	-0.3 to -0.5
Large	0.5 to 1.0	-0.5 to -1.0

[3] In this paper, the Pearson correlation coefficient(r) is calculated of all 32 features with the final marks to find out which feature is strongly related. Such features are then used to train the model to simplify classification accuracy and improve or maintain the accuracy of the decision tree.

4.3 Training model using CART

In this research the scikit-learn library in python is used for creating the decision tree. scikit-learn uses an optimized version of the CART algorithm; however, scikit-learn implementation does not support categorical variables for now [4]. The term CART was first introduced by Leo Breiman to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems. CART algorithm uses Gini's impurity index as a splitting criterion.[6] CART algorithm creates a binary tree build by splitting node into two child nodes repeatedly. The algorithm works repeatedly in the following three steps:

1. Find each feature's best split. For each feature with N different values, there exist N-1 possible splits. Find the split, which maximizes the splitting criterion. The resulting set of splits contains the best splits (one for each feature).

2. Find the node's best split. Among the best splits from Step 1 find the one, which maximizes the splitting criterion.

3. Spliting the node using the best node split from Step 2 and repeating from Step 1 until stopping criterion is satisfied.

As splitting criterion Gini's impurity index is used. The Gini impurity criterion is defined as:

$$\Delta i(s, t) = i(t) - pLi(tL) - pRi(tR)$$

where $\Delta i(s, t)$ is decrease of impurity at node t with split s, pL (pR) are probabilities of sending case to the left (right) child node tL (tR) and i(tL) (i(tR)) is Gini impurity measure for left (right) child node.

5. RESULTS

While calculating the correlation between all the features and the target value which is the final grade, it was found that the final grade is highly dependent upon the factors such as previous grades, number of failures in the past, and whether the student wants to take higher education or not. All the correlation values are listed below.

Table 2: Attribute no and its correlation

Attribute number	Correlation	Attribute number	Correlation
32	0.918	25	-0.122
31	0.826	3	-0.106
15	-0.393	29	-0.098
21	-0.332	30	-0.091
1	0.284	23	0.09
14	0.249	26	-0.087
7	0.24	16	0.066
8	0.211	24	0.063
27	-0.204	19	-0.059
28	-0.157	17	-0.059
4	0.167	18	0.054
11	-0.157	10	0.052
22	-0.15	5	0.045
9	0.148	12	-0.029
2	-0.129	20	-0.028
13	-0.127	6	-0.0007

Including the features, one by one in the order of decreasing absolute value of its correlation, as input, resulted in decision trees with a range of accuracies given below.

Table 3: No. of attributes taken and accuracies of the tree

No. of attributes taken	Training accuracy	Testing accuracy	Average accuracy
2	90.57	89.14	89.855
3	90.96	89.99	90.475
4	91.15	89.92	90.535
5	91.54	90.69	91.115
6	92.3	88.37	90.335
7	95.15	80.62	87.885
8	96.53	79.84	88.185
9	97.88	84.5	91.19
10	98.46	85.86	92.16
11	98.65	82.17	90.41
12	99.8	79.66	89.73
13	100	78.29	89.145
14	100	79.06	89.53
15	100	79.84	89.92

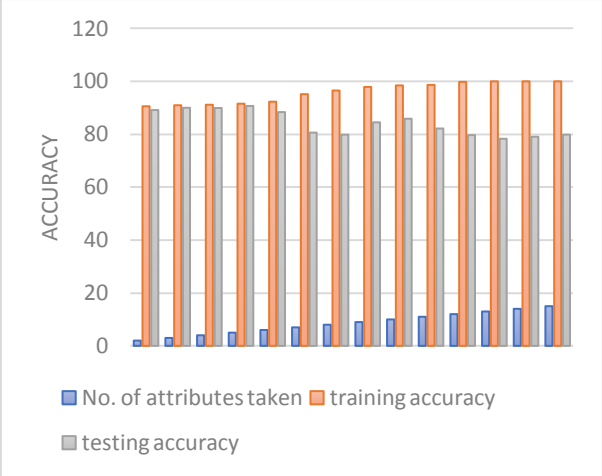

Figure 3: Accuracy percentage with no. of attributes taken

Initially, the testing accuracy and the training accuracy, both increases on increasing the features in the input dataset but after a point testing accuracy starts decreasing and the training accuracy starts reaching towards 100%. The most optimal and accurate decision tree was formed when 5 attributes with the highest absolute value of correlation were used resulting in a tree with a testing accuracy of 90.69% and training accuracy of 91.54%.

6. CONCLUSION

Educational Data Mining is fairly a new field with a lot of potentials if used effectively. In this paperdiscusses howfeatures can be selected fora Classification and Regression Tree using Pearson Correlation and how helpful the result prediction can be for the students, teachers and the universities. The use of Pearson's Correlation in selecting features helps in creating a more accurate decision tree effectively. This method is used on a single dataset of 649 students, the effectiveness of this method is yet to be checked on larger and different datasets if it is to be implemented on a larger scale.

6. REFERENCES

[1] Pearson, Karl (20 June 1895). "Notes on regression and inheritance in the case of two parents". Proceedings of the Royal Society of London. 58: 240–242.

[2] Galton, F. (1886). "Regression towards mediocrity in hereditary stature". Journal of the Anthropological Institute of Great Britain and Ireland.

[3] statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[5] Nidhi, Mukesh Kumar, NandiniNayar, Gaurav Mehta, “Student’s Academic Performance Prediction in Academic using Data Mining Techniques”. International Conference on Intelligent Communication and Computational Research.

[6] Breiman L (1984) Classification and regression trees. The Wadsworth and Brooks-Cole statistics-probability series. Chapman & Hall.

[7] AgungTriayudi, WahyuOktriWidyarto, Educational Data Mining Analysis Using Classification Techniques. Virtual Conference on Engineering, Science and

Technology (ViCEST) 2020

[8] M. Ramaswami and R. Bhaskaran , A Study on Feature Selection Techniques in Educational Data Mining. JOURNAL OF COMPUTING, VOLUME 1, ISSUE 1, DECEMBER 2009.

[9] Amjad Abu Saa, Educational Data Mining & Students' Performance Prediction. International Journal of Advanced Computer Science and Applications,Vol. 7, No. 5, 2016.