International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 08, August-2018

ANALYSIS OF STUDENT PERFORMANCE FORECAST USING DATA MINING TECHNIQUES

P.T.Jamuna Devi[#1] , Dr. G.Kesavaraj [#2]

[#1]*M.Phil Scholar,*[#2]*Associate Professor,*
PG and Research Department of Computer Science,
Vivekanandha College of Arts and Sciences for Women, (Autonomous)
Tiruchengode, Namakkal-DT, TamilNadu, INDIA

Abstract:- Data mining is a study of classification, association, prediction, clustering of data for the various field. Classification deals with static data in other terms it is supervised learning. Clustering is a non-supervised techniques used to take decision in a particular problem. Data mining is based on complex algorithms that allow for the segmentation of data to identify patterns and trends, detect anomalies, and predict the probability of various situational outcomes. Predictive analytics is an area of statistics. It deals with extracting information from data and using it to predict trends and behavior patterns.

Education plays vital role to develop our nation. In this, there is a lot of research carried out in the field of education but no one is predict students pass rates. Here our research deals with student pass rates prediction using optimized Support Vector Machine (SVM) algorithm and Decision Tree (DT) algorithm.

We propose a new concept: features dependency algorithm, CART algorithm and machine learning algorithm to analysis relationship between the set of student features. And we also collect online and offline data of students from other schools and by using this algorithm we predict the student pass rates in blended learning. The purpose is to providing assistance to students who have greater difficulties in their studies and students who are at risk of graduating through data mining techniques. The result shows that to identify the week performance of the student and improve their performance by using this algorithm.

1. INTRODUCTION

Data mining is the procedure of discovering patterns in massive information sets regarding strategies at the intersection of machine learning statistics and information systems. It's a necessary scheme wherever intelligent strategies square assess applied to extract information patterns. It's associate knowledge domain subfield of engineering .The overall goal of the mining method is to extract information from a knowledge set associated convert it into an cheap structure for any use. Data mining is that the analysis step of the "knowledge discovery in databases" process, or KDD. The term may be a name, as a result of the goal is that the extraction of patterns and data from giant amounts of information, not the extraction (mining) of information itself. The goal is the mining of patterns and knowledge from large amounts of data. It is regularly applied to any form of large-scale data or information processing. It is also applied to purpose of computer decision support system, including artificial intelligence, machine learning, and business intelligence.

The actual data mining assignment is the semi-automatic or automatic psychiatry of large quantities of data to mine formerly unknown patterns. The patterns include groups of data records (cluster analysis), extraordinary records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. In machine learning and predictive analytics, these patterns can be summarized as input data and analysised. The data mining steps are data collection, data preparation, result interpretation and reporting. They are fit to KDD process.

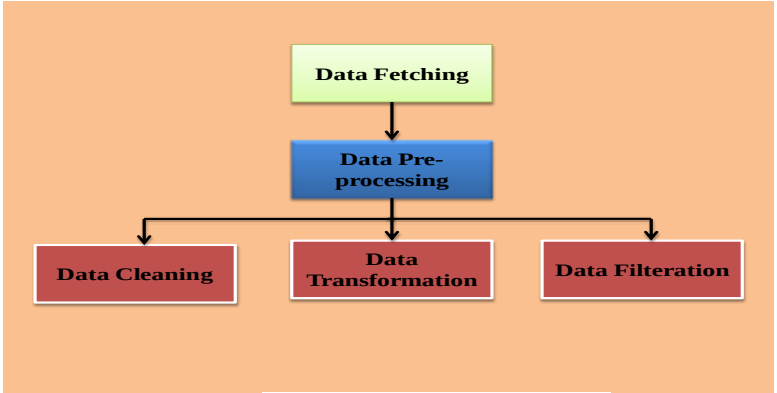

Fig 1.1 Data mining techniques

Data mining involves six common classes of tasks:

ANOMALY DETECTION

The unusual records can be identified.

ASSOCIATION RULE LEARNING

Relationships between two variables are searched. For example, a supermarket might gather data on consumer purchasing habits. For marketing purpose, association rule learning is used to verify which products are repeatedly bought by customer in super market. This is referred to as market basket analysis.

CLUSTERING

It is the task of identifiying groups and similar structures in the data without using known structures in the data.

CLASSIFICATION

It is the task of applying new structure to the data. For example, an e-mail program might try to classify an e-mail as "legitimate" or as "spam".

REGRESSION

For estimating the relationships among data or datasets, It attempts to find a function which models the data with the least error

SUMMARIZATION

It provides a more dense representation of the data set, including visualization and report generation.

DECISION TREE

A decision tree is a decision support tool.It uses a tree-like graph or model of decisions. It includes outcomes, resource costs, and utility. It is one way to demonstrate an algorithm that only contains conditional control statements.

In decision analysis, Decision trees are commonly used in operations research, to identify a approach to achieve a goal. It is one of the popular tool in machine learning.

A decision tree is a flowchart-like arrangement. A "test" on an attribute is represented by internal node.(e.g. whether a coin flip comes up heads or tails). The outcome of the test is represented by branches. Each leaf node represents a class label (resolution taken after computing all attributes). The structure symbolizes classification rules.

In decision analysis the expected values of competing alternatives are calculated. A decision tree and the closely related influence diagram are used as a visual and logical decision support tool.

The three types of nodes are

1. Decision nodes – represented by squares
2. Chance nodes – represented by circles
3. End nodes –represented by triangles

DECISION RULES

The decision tree can be summarized into decision rules. The content of the leaf node is the outcome. In if clause, the conditions along the path form a conjunction.

The rules are in the form:

if condition1 and condition2 and condition3 then outcome.

DECISION TREE TYPES

The two types of decision tree are

Classification tree

Regression tree

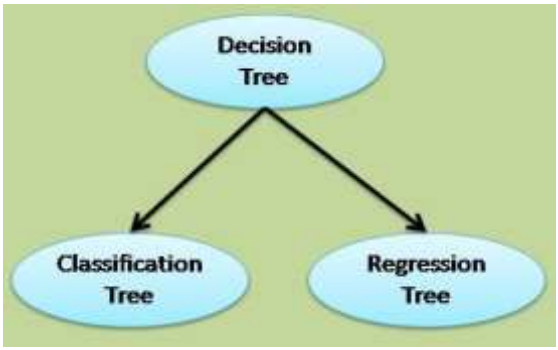

Fig 1.2 Decision tree Diagram

C4.5 (successor of ID3)

The C4.5 decision tree algorithm is an algorithm developed by Ross Quinlan, which was the successor of the ID3 algorithm. The C4.5 algorithm uses pruning in the generation of a decision tree, where a node could be removed from the tree if it adds little to no value to the final predictive model. The C4.5 algorithm was able to forecast the class of 95 substance out of 270, which gives it an Accuracy value of 35.19%.

CHAID (CHi-squared Automatic Interaction Detector)

A CHAID categorization tree is a category of decision tree. Chi-squared Automatic Interaction Detection (CHAID) is an additional decision tree algorithm which uses chi-squared based splitting condition instead of the usual splitting criterions used in additional decision tree algorithms. Chi-square tests are also used to combine categories in single nodes. If the counts for categories are not considerably dissimilar, they are combined into a node. The CHAID algorithm was able to guess the class of 92 items out of 270, which gives it an Accuracy value of 34.07%.

ID3 Decision Tree

The ID3(Iterative Dichotomiser 3) decision tree algorithm is urbanized by Ross Quinlan. The algorithm generates an unpruned full decision tree from a dataset. The ID3 algorithm begins with the unique set S as the root node. The ID3 algorithm was able to forecast the class of 90 substance out of 270, which gives it an Accuracy value of 33.33%.

2. STUDENT PERFORMANCE ANALYSIS

Students are the main asset for various institutions.Students play an important role in producing graduates of high qualities with its academic performance achievement. Academic performance achievement is the level of achievement of the students' educational goal that can be measured and tested through examination, assessments and other form of measurements. However, the academic performance achievement varies as different kind of students may have different level of performance achievement. The details of the student are collected from individual students. The details include both curricular and extra curricular activities. These are maintained as student data set. The student details are used to analyse the details of the student and identify the week performance of the student. The performance of the students are increasing by taking special attention.

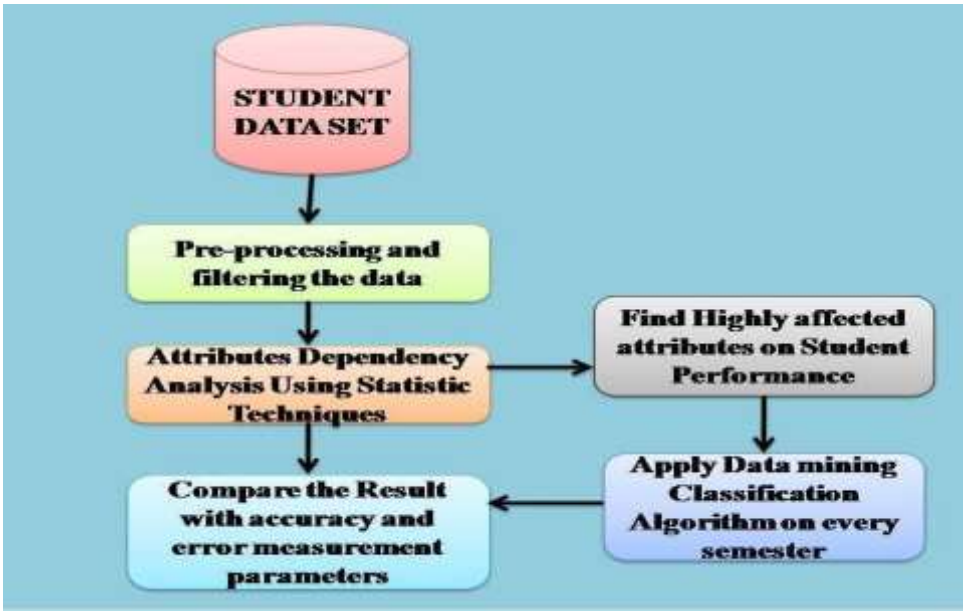

Fig 2.1 Student Performance Analysis

GOALS OF PROBLEM STATEMENT

The goals of the problem statement are

- Ideal
- Reality
- Consequences
- Proposal

EXISTING SYSTEM

All education system, it is critical to improve student pass rates and reduce dropout rates. Features dependencies and the grid search algorithm is used to optimize DT and SVM, in order to improve the accuracy of the algorithm. The

results show the experiment can achieve better results in this work. The purpose is to providing assistance to students who have greater difficulties in their studies, and students who are at risk of graduating through data mining techniques. The existing system uses ID3 algorithm to improve the student pass rate. The ID3 algorithm uses a greedy search. It selects a test using the in sequence gain measure, and then never explores the opportunity of exchange choices.

DISADVANTAGES

- The theoretical level of the set of student features using machine learning algorithm is not implemented.
- A small amount of data is used.
- Only offline data of the students can be used.
- Data might be over-fitted or over-classified, if a small model is tested.
- Only one characteristic at a occasion is tested for building a decision.
- Does not hold numeric attributes and missing values.

PROPOSED SYSTEM

A new concept features dependency algorithm, CART algorithm and machine learning algorithm to analysis relationship between the set of student features. And we also collect online and offline data of students from other schools and by using this algorithm we predict the student pass rates in blended learning. The purpose is to providing assistance to students who have greater difficulties in their studies and students who are at risk of graduating through data mining techniques. The result shows that to identify the week performance of the student and improve their performance by using this algorithm. CART can hold both numeric and categorical variables and it can simply hold outliers.

ADVANTAGES

- The theoretical level of the set of student features using machine learning algorithm is implemented.
- Using CART algorithm offline and online data of the students can be collected.
- The system calculates the score and provides outcome immediately.
- It removes individual errors that regularly occur through physical checking.
- The system provides a balanced outcome.
- The system excludes individual efforts and saves time and resources.
- Reduce the time and cost.
- Paper less examination.
- Answers are verified immediately.
- Accurate results.

MODULES

- Staff Module
- Student Module
- Parent Module
- Attendance Module
- Transport Module
- Admin Module

ATTENDANCE MODULE

The Attendance module is considered for teachers to be capable to take attendance. It is used students to be able to analysis their own attendance record. A teacher can spot the attendance category of a student. These category descriptions are configurable, and more can be added. The teacher adds Attendance as an movement of a course, and then sets up the sessions whose attendance is to be tracked. The Attendance module can create reports for both entire class and for individual students. Students may also see their own attendance record if the activity is not hidden. It allows students rapid access to a summary information for their own attendance. The attendance of the students in the class can be very easily monitored through this module. Teachers or administrative staff can update the attendance manually and specific search results can be obtained for the attendance of a particular students by putting in the name or roll number.

STUDENT MODULE

The student module is designed for teachers to update the personal information of the individual student. It includes Maintenance and reporting of student data, Student accounts and financial, Handling records of examinations, assessments, marks, grades and academic progression Maintaining discipline records. Student Module helps Class Teacher, Principal to have a students' development or perfection discussion with the parent without concerning too many people. Class wise consolidated Progress report can be generated from Student module. Number of Boys and Girls below each part and class for the complete School can be generated from Student Module.

TRANSPORT MODULE

Transport module directly deals with various issues like best exploitation of school transport and manages route and drivers details. This structure can be predefined. Transportation module is an skillful route development which makes the day to day task for transport very simple as its main focal point is on the safety of students. This module tracks all vehicle particulars along with driver particulars which get better safety and exact location can be easily tracked. This module also improves communication with parents as they will get alerts concerning delay of the bus through SMS.

STAFF MODULE

Staff module helps to manage all the staff information about photographs, qualification, and experience, any references, salary, address, contact information, marital status, maintain attendance details . This module helps to track and analyze the performance of staff. This module also prepares essential reports like staff details, academic performance report, staff attendance details, attendance register. This module also manages teacher's syllabus with date, subject, chapter, and status.

PARENT MODULE

Parent Module helps parent to view their child detail about performance, attendance, Marks of both internal and external and so on. This module help parent to find the respective subject staff details of their children. This module help parents to interact with staff about the regular activities of their child.

ADMIN MODULE

Admin Module enables supervision of all the administrative activities of the institutions by the authorized users. All the functions of the management system can be operated from the admin panel and access can be provided to different users. The authorized users can manage a wide range of activities of the system from anywhere and anytime by logging in their accounts. The privileges for each user can be assigned through this admin module.

CART (Classification And Regression Tree)

The term Classification And Regression Tree (CART) analysis is used to refer both of the above procedures. Trees used for both regression and for classification have some similarities and also have some differences. In classification the procedure used to determine. In regression procedure used to split.

CART algorithm was able to calculate the class of 108 substance out of 270, which gives it an Accuracy value of 40%.

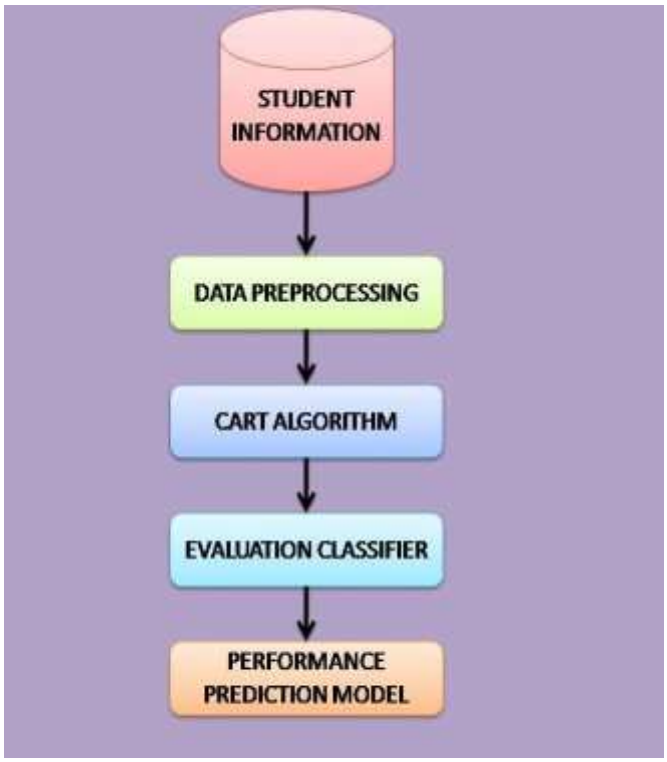

Fig 2.2 Student Performance Analysis using CART Algorithm

The main elements of CART algorithm are:

1. Rules for splitting data at a node based on the importance of variable;
2. Stopping regulations for deciding while a subdivision is terminal and can be divided and
3. Finally, a forecast for the target variable in each terminal node.

Fig 2.3 Histogram of Class Precision and Class Recall of CART

In detail, CART had the best accuracy of 40%, CART algorithms worked better with the dataset than others.

USES OF CART ALGORITHM

Uses Gini impurity(not to be confused with Gini coefficient).

A good discussion of the differences between the impurity and coefficient is available on Stack Overflow.

Uses the cost-complexity method of pruning. Starting at the bottom of the tree.

CART evaluates the misclassification expenditure with the node or without the node.

The decision nodes have exactly 2 branches.

Uses surrogates to distribute the missing values to children.

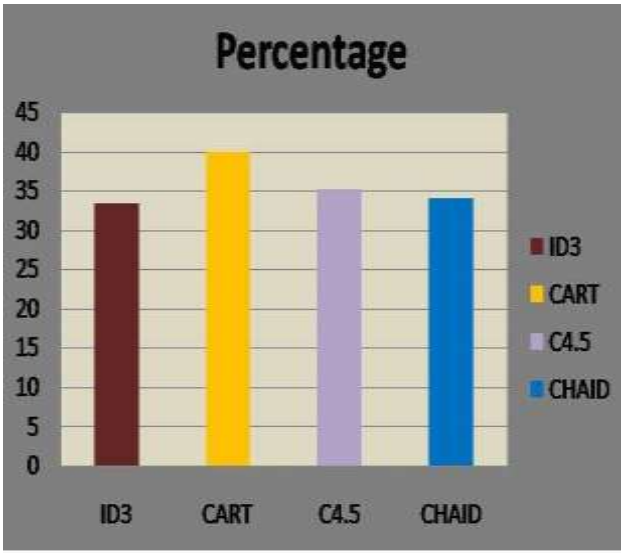

Fig 2.4 Histogram of Accuracy

3. CONCLUSION

This model is mainly focused on analyzing the prediction accuracy of the academic performance of the students. This research work is about how data mining techniques can be effectively used to analyse and identify the weaker performance of the student and improve their performance by giving special coaching to them. This research work developed a new data mining algorithm namely CART algorithm. It is used to predict better performance of the student. Also the effect of parameter tuning for optimization of decision trees was also analysed.

4. REFERENCES

[1]. Xiaofeng Ma and Zhurong Zhou.," Student Pass Rates Prediction Using Optimized Support Vector Machine and Decision Tree"., 2018 IEEE.
[2]. Amjad Abu Saa., et al.,"Educational Data Mining & Students' Performance Prediction"., (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5,2016., Pp no.212-220.
[3]. P. Cortez and A. Silva..," Using Data Mining to Predict Secondary School Student Performance." In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
[4]. Baradwaj, B.K. and Pal. S., "Mining Educational Data to Analyze Students' Performance "., (IJACSA) International Journal of Advanced Computer Science and Applications., Vol. 2, No. 6, 2011.
[5]. Ahmed, A.B.E.D. and Elaraby., I.S., 2014." Data Mining: A prediction for Student's Performance Using Classification Method"., World Journal of Computer Application and Technology, 2(2)., Pp.43-47.
[6]. P.V.Praveen Sundar., " A Comparative Study For Predicting Student's Academic Performance Using Bayesian Network Classifiers"., IOSR Journal of Engineering (IOSRJEN) e-ISSN: 2250-3021., p-ISSN: 2278-8719., Vol. 3, Issue 2 (Feb. 2013)., Pp 37-42.
[7]. G. Jayanthi and Dr.V.Ramesh., "Design of Academic Performance Prediction System Using Multi-Layer Perceptron"., International Journal of Computer Science and Software Engineering ., Volume 1, Number 1 (2015)., Pp. 9-15.
[8]. B.K. Bharadwaj and S. Pal. Data Mining: "A prediction for performance improvement using classification". International Journal of Computer science and Information Security (IJCSIS), Pp.136-140,2011.
[9]. Surjeet Kumar Yadav and Saurabh Pal. Data Mining: "A Prediction for Performance Improvement of Engineering Students using Classification". World of Computer Science and InformationTechnology Journal (WCSIT),Pp.1309-1314, 2012.
[10]. Georgios Kostopoulos, et al ., Predicting Student Performance in Distance Higher Education Using Active Learning". Engineering Applications of Neural Networks, Pp. 75-86, 2017.
[11]. Xiaofeng Ma and Zhurong Zhou.," Student Pass Rates Prediction Using Optimized Support Vector Machine and Decision Tree"., 2018 IEEE.
[12]. Amjad Abu Saa., et al.,"Educational Data Mining & Students' Performance Prediction"., (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5,2016., Pp no.212-220.
[13]. P. Cortez and A. Silva..," Using Data Mining to Predict Secondary School Student Performance." In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
[14]. Baradwaj, B.K. and Pal. S., "Mining Educational Data to Analyze Students' Performance "., (IJACSA) International Journal of Advanced Computer Science and Applications., Vol. 2, No. 6, 2011.
[15]. Ahmed, A.B.E.D. and Elaraby., I.S., 2014." Data Mining: A prediction for Student's Performance Using Classification Method"., World Journal of Computer Application and Technology, 2(2)., Pp.43-47.
[16]. P.V.Praveen Sundar., " A Comparative Study For Predicting Student's Academic Performance Using Bayesian Network Classifiers"., IOSR Journal of Engineering (IOSRJEN) e-ISSN: 2250-3021., p-ISSN: 2278-8719., Vol. 3, Issue 2 (Feb. 2013)., Pp 37-42.