Supervised Learning In The Context Of Educational Data mining To Avoid Students Dropout

Niveditha.G [1], Santhanamari.A [2], Ponselvi.P [3] and Pavithra.R[4]
[1, 2, 3, 4] PSR Rengasamy College of Engineering for Women, Sivakasi, India

Abstract- *Educational data mining is a research field that looks for extracting useful information from large educational datasets. This area provides tools for improving student retention rates around the world. In this paper we propose a computational approach using educational data mining and different supervised learning techniques (Decision Trees, K-nearest Neighbor, Neural Networks, Support Vector Machines, Naive Bayes and Random Forests) to evaluate the behavior of different prediction models in order to identify the profile of at-risk university students. The results of this paper indicate that some algorithms can be used as tools for supporting decisions that reduce college dropout. Overall the Educational Data Mining aims to interact the relevant information from any educational data and further transform into a systematic and understandable knowledge for the sake of decision making. Classification techniques can be highly helpful in predicting student's performance. The study found highest accuracy in Naive Bayes among four algorithms as above ninety percent. Naive Bayes was concluded to be the best algorithm and execution of proposed model confirmed the claim.*

Keywords- exam, student, education

I. INTRODUCTION

Data mining is very prominent technique to identify the hidden facts from the huge volume databases; these techniques are applied when data is outsized and less knowledge about data available. Data mining techniques also progressive when extracted making plays an active role to direct the students to upright path. Data mining also is to investigate educational linked data used to solve educational questions . Recently, prediction of student performance is focused area, and key to keep the student on right way. Data mining methods ensure to make the data observable and better way to analyze it. Educational Data Mining is unique area of Data Mining methods to educational data. Its aim is to investigate the data permitted to solve educational study questions

The research curiosity in educational data mining is observed as a necessity of this era. This novel developing area concerns over statistics based approaches that discover knowledge by data gathering from educational locations. Educational Data Mining follows various techniques; Bayes theorem is one of the earliest techniques to find out hidden knowledge. With the passage of time, techniques are expanded such as Decision Trees, Naive Bayes, Neural Networks and Association Rule Mining

Using mentioned statistical techniques, knowledge can be extracted in the form of rules of associations and classifications that are highly recommended techniques of prediction also grouping the educational related items using clustering. These techniques facilitate the observers from almost every possible angle. Prediction in educational data may include course registration of students in best fit course; marks of students in different assessment categories and identification of students needing more help in specific examination categories. The main purpose of this research is to practice the data mining approaches to educational related data and prediction of student enrolment status after one year completion in university like which students likely to be dropped or promoted, used mechanism also is to notify the administration about the students needing more help in specific area. This research also explores the correctness of different classification techniques for prediction student performance.

Education is the important factor for succeeding a long term improvement in any sector of a country. It enlightens the individual and develops his/her capability to the limit. The main purpose of the education system is to build up each and every student's skills and knowledge needed to reach the successful career pathway. This is the main target of most of the educational institutions. But the task is a huge challenge because a large amount of students drops out every year due to various reasons. A vast resource is being wasted due to the dropping out of students. It pulls down the nation backward. As the rate of drop out students' increases, the economic growth and development of a nation decreases. It's very difficult for the educational institutions to analyze and find out the key reasons behind the dropping out of students by looking after each and every student. It's a costly procedure and needs a lot of manpower. In this case, data mining techniques can be used to predict the students who are at risk of dropout.

Data mining aims at discovering previously unknown, potentially useful and non-trivial knowledge from a huge amount of data. Data mining techniques have both predictive and descriptive natures. Usually, supervised learning methods have predictive nature and unsupervised learning methods have descriptive nature. By using the predictive nature of data mining techniques, one can identify the students who are at risk of drop out. In supervised learning method both input and target class are given and an algorithm is used to learn the function so that one can easily predict the target variable when a new input is given. In this paper, a threshold value based approach has been presented where the threshold value is calculated by using the attributes of the datasets and their corresponding information gain value. By using the threshold value it can easily be identified whether the students are at risk of dropping out.

Company Profile

Ascox Techno Soft is a leading software development and web designing company which born on 2014n in Madurai, Tamilnadu, India. We are the young and energetic team committed to the permit of excellence. Our successful projects with client requirements have represented our reputation as superior providers. Ascox Techno Soft has established multi-branches at Chennai and Coimbatore for Continuous and better serve its Clients. Ascox Techno Soft's differentiation point meets with three philosophies.

True Participation
Perfect Understanding
Patience in completing the job

In the past three years Ascox Techno Soft travelled more than 35 projects and has a large client's base of more than 20 clients all over India.

Our Mission and Vision

- ✓ Our Mission is to enriching the business growth of our clients with creative design and development to deliver high qualified solutions.
- ✓ Our Vision is to develop efficient software solutions to the most complex requirements with the highest levels of integrity. Professionalism and technological capabilities. When the project is specific and the result cannot fulfill your requirements-you need efficiently developed solutions for your software, being Ascox Techno Soft's clients, you will receive a perfect and expected solution.

II. EXISTING SYSTEM

Starting from the previous models (rules and decision trees) generated by the DM algorithms, a system is used to alert the teacher and their parents about students who are potentially at risk of failing or dropout can be implemented.

DISADVANTAGE:

- The problem of imbalanced data classification occurs when the number of instances in one class is much smaller than the number of instances in another class or other classes.
- It does not include complete academics of any student as it covers only till higher secondary school.

III. PROPOSED SYSTEM

We propose that Data mining techniques are applied to Engineering colleges and once students were found at risk, they would be assigned to a tutor in order to provide them with both academic support and guidance for motivating and trying to prevent student failure. We have shown that classification algorithms can be used successfully in order to predict a student's academic performance and, in particular, to model the difference between Fail and Pass students.

ADVANTAGES:

- ➢ As we have seen classification task has been used on student database to predict the students' performance on the basis of previous database record.
- ➢ There are many approaches that are basically used for the data classification. Information like Class test, Attendance, Seminar, innovative activities and Assignment marks were collected from the students' previous database record, to predict the performance at the end of the each semester.
- ➢ This study will definitely help for the students and the teachers to improve the performance of the student. This study will also help full to identify those students which needed special attention to reduce failure ration and taking appropriate action for the next academic examination.

IV. SYSTEM DESIGN

ARCHITECTURE DESIGN

MODULES DESIGN:

- Data acquisition and noise removal.

- Data pre-processing.
- Feature selection.
- Feature extraction algorithms.
- Evaluation and analysis of results.

DATA ACQUISITION AND NOISE REMOVAL:

Data Acquisition- This is the process where data is gathered/collected from various primary sources such as earlier records, social media platforms, data centres, and manual collection from paper records.It is very much important to track the origin of the database and check whether that data is up to date or not, as it is very much important to match the real-time results.Since each data is very important, so it is important that the data should be uploaded on the server so that there will be enough space to hold that data for an accurate results.

Data Preparation- To structure or prepare a database for a specific purpose is called Data Preparation. It primarily means manipulation of data into a form suitable for further analysis and processing Data preparation is a fundamental stage of data analysis. While a lot of low- is available in various data sources and on the Web, many organizations or companies are interested in how to transform the data into cleaned forms which can be used for high-profit purposes.

Data preparation is a fundamental stage of data analysis. While a lot of low-quality information is available in various data sources and on the Web, many organizations or companies are interested in how to transform the data into cleaned forms which can be used for high-profit purposes.

Data preparation comprises those techniques concerned with analyzin graw data so as to yield quality data, mainly including data collecting, data integration, data transformation, data cleaning, data reduction, and data discretization.

DATA PREPROCESSING

In the knowledge discovery process, data preprocessing is an essential part, because excellence decision making is grounded on quality of data. Raw data may contains missing values, abnormal values as outliers and inconsistencies, data preprocessing deals with all such kind of data to clean it and make it available in tuned form. Data cleaning refers to clean the data by filling missing values, smoothing the data by removing outliers and noise and make it consistent. Data transformation refers normalization and aggregation of data is the next stage to make the data open to participate toward formation of data mining process.

DATA CLEANING:

Investigating a data record for incorrect entries and fields and thus appending/updating it with correct info is known as Data Cleaning. This process is also referred to as Data Cleansing/Scrubbing. It may include removing typographical errors, standardizing database fields so that all the entries in the database have the same format.

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn't the fanciest part of machine learning and at the same time, there aren't any hidden tricks or secrets to uncover. However, proper data cleaning can make or break your project.

FEATURE EXTRACTION:

Deep learning feature extraction, we treat the pre-trained network as an arbitrary feature extractor, allowing the input image to propagate forward, stopping at pre-specified layer, and taking the outputs of that layer as our features.

Feature extraction is data manipulation. In data manipulation, the task is to modify the data to make it easier to read and more organized. We manipulate the data for data analysis and data visualization. Data manipulation is also used with the term 'data exploration' which involves organizing data using the available sets of variables.

FEATURE SELECTION:

There are two main types of feature selection techniques: wrapper and filter methods. Filter-based feature selection methods use statistical measures to score the correlation or dependence between input variables that can be filtered to choose the most relevant features.

Feature Selection methods in Data Mining and Data Analysis problems aim at selecting a subset of the variables, or features that describe the data in order to obtain a more essential and compact representation of the available information.

EVALUATION AND ANALYSIS OF RESULTS:

In our study, the experimental results were segregated into two parts. In first and second part the Decision Tree results and Bayes results were observed respectively. Finally all techniques were ranked and declared best fit technique for our research.

V. LITERATURE SURVEY

C. Romero and S. Ventura, "Educational data mining: Educational data mining: a survey from 1995 to 2018,

The quality of empirical research on student attrition depends on the availability of good data. There are two types of data that have been exploited in the literature: administrative data and survey data. Due to the lack of available administrative data, information on student attrition in German universities is largely gathered from surveys (Larsen t al., 2013). However, student surveys have significant limitations when investigating the causes of attrition. In ex-ante interviews, the dependent variable, student attrition, must be replaced with the intention of dropping out. Using the intention to drop out as a predictor for actually dropping out is, however, controversial in the literature as it assumes that the intention is not exaggerated or otherwise subjected to self-adjustment (Brandstätter et al., 2006). But clearly, one advantage of survey data as compared to administrative data is that survey data allows for learning more about determinants of the dropout decision. Tinto's (1975) "student integration model" established the central importance of the social and academic integration of the student. Pascarella and Terenzini (1979) adopt the idea of integration and extend the model by distinguishing between forced and voluntary attrition. Bean (1983), on the other hand, presents the importance of integration as a main predictor of attrition and adds student satisfaction as a central variable. The importance of academic performance and informational frictions for explaining attrition has been stressed in recent literature (Stinebrickner & Stinebrickner, 2008; 2012; 2013; 2014; Arcidiacono et al., 2016). But despite the importance of the topic, there is still much that is unknown about the underlying determinants of attrition and the effective means for reducing it. But all of the aforementioned determinants, namely, integration, identification, and satisfaction of students are not components of student administrative data. Analyzing administrative data implies knowingly using incomplete data and making the best of it.

Aside from the acknowledged shortcomings of administrative data, they are much better suited for studying the extent of dropout occurrences and, more importantly, the use of administrative data allows predicting student dropout and analyzing student study behavior.

cement of a voice-based e-assessment master framework for the outwardly debilitated understudies in ODL. The examination utilizes a blend of advances, for example, framework structure, server-side scripting, voice-based framework improvement, information the board and rule-based thinking in building up the framework. The framework was assessed to decide the degree of convenience. The consequences of the ease of use assessment demonstrated that the created application has a 'normal ease of use' rating of 3.48 out of 5 scales. The discoveries show that the voice-based e-assessment framework won't just be of tremendous advantage to the outwardly debilitated understudies in ODL in particular of separation, however will likewise supplement the current electronic strategy for online assessment.

VI. PROBLEM DEFINITION

Identification of drop out students is a major task because without educating people building up a developed country is not possible. For this, the educational department and the government of a country are aware of this. Researchers have tried to invent new models for identifying students who are at risk of dropping out. Predictive, Descriptive and also Subgroup Discovery algorithms have been utilized in this field to identify drop out students and the key factors which are responsible for drop out students. They have used Decision Tree, Logistic Regression, Stepwise Regression, and Cox Regression model for their way of task. Subgroup discovery is a significant data mining technique which has been largely used in educational data mining. The subgroup discovery technique discovers interesting associations among

PROBLEM DESCRIPTION

In this paper, to predict about the drop out students, we have proposed a novel approach based on threshold value calculation. Threshold values are calculated using important features. These features are selected using Information Gain [14]. The more is the information gain, the feature is more important. According to the information gain, we have taken some features to calculate the threshold value.

Data mining process starts with data collection then preprocessing, transformation and finally development of prediction model. To predict the student performance, many factors needed to be measured. Prediction model must include all environmental attributes that are imposed to effective prediction of student performance. Data concerning to student's previous academics, core subjects studied, student aptitude in specific assessment type, favorite subjects of students etc., Outlier detection and removal is needed as an important part of preprocessing before transformation of data.

VII. IMPLEMENTATION

IMPLEMENTATION PLAN

To develop prediction model, many techniques are used such as Neural Networks, Decision Trees, Association Rule Mining, Nearest Neighbor Method, Clustering and Classification. Classification is one of the most commonly used techniques for prediction especially when more accuracy needed. Classification prediction based on class label having successful and failure scenarios. In this study, we used the two Decision Tree algorithms, Random Forest and J48graft versus two Bayes classification algorithms Naive Bayes and Bayesian Logistic Regression. Some details of proposed classification algorithms are following

Decision Tree

Approach Decision Tree is all about splitting data into segments called branches, shaped as parent child relation. These branches form an upward down Tree that originates with top of the node called root node. Decision Tree also deals with both continuous and categorical data. Decision Tree is influential and wide spread method for prediction. The charm of Decision Trees is because of its comparison with Neural Networks, rules are drawn in Decision Trees that are very easy to understand and interpret. In our research, Random Forest and J48graft algorithms are used to calculate the accuracy of our model.

Bayes Approach

Bayes rule is basically core of Bayesian inference method, used to update the probability estimation for a hypothesis as additional evidence. Bayesian updating is dynamic throughout statistics, and in mathematical statistics, its role is very important. Especially to predict a target class, Naive Bayes classification technique is preferred. It works with Bayesian theorem to calculate the probabilities. Bayesian results are often proved more accurate comparatively. Naive Bayes method is good for a number of causes. Its narrowing is easy, complex iterations are not needed, that shows that it can be applied to balky data. Its interpretation is very easy, that's why a non-technical users can understand. In our research, Naive Bayes and Bayesian Logistic Regression algorithms were used to calculate the accuracy of our model.

TESTING

White Box Testing

By using this technique it was tested that all the logical paths were executed at least once, all the logical decisions were tested on both their true and false sides.

Black Box Testing

By using this technique, the missing functions were identified and placed in their positions. The errors in the interfaces were identified and corrected. This technique was also used to identify the initialization and termination errors and correct them.

Unit Testing

This is the first level of testing. In this different modules are tested against the specifications produced during the coding of the simple program module in an isolation environment. Unit testing first focuses on the modules independently of one another to locate errors. After coding each dialog is tested and run individually. All unnecessarily coding were removed and it as ensured that all the modules worked, as the programmer would expect. Logical errors found were corrected. So, by working all the modules independently and verifying the outputs of each module in the presence of staff was conducted that the program was functioning as expected.

Integration Testing

Data can be lost access an interface, one module can have as adverse effort on another sub function when functions combined, may not produce the desired major functions. Integration testing is a systematic testing for constructing tests to uncover errors associated within the interface. The objectives are to take unit tested as a whole. Here correction is difficult because the vast expenses of the entire program complicated the isolation of causes. Thus in the integration testing step, all the errors uncovered are corrected for the next testing steps.

Validation Testing

This provides the final assurance that the software meets all functional, behavioral and performance requirements. The software is completely assembled as a package. Validation succeeds when software functions in a manner in which the user expects. Validation refers to the process of using software in a live environment in order to find errors. During the course of validating the system, failures may occur and sometimes the coding has to be changed according to the requirement. Thus the feedback from the validation phase generally produces changes in the software.

INSTALATION

SOFTWARE SPECIFICATION:

- Operating system : Windows 7,8,10

- Front End : Java
- Backend : SQL Server

HARDWARE SPECIFICATION:

- Processor - Pentium –IV
- Speed - 1.1 GHz
- RAM - 256 MB(min)
- Hard Disk - 20 GB
- Key Board - StandardWindows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

SOFTWARE CHARACTERISTICS

ABOUT THE SOFTWARE

JAVA

Java is a widely used programming language expressly designed for use in the distributed environment of the internet. It is the most popular programming language for Android Smartphone applications and is also among the most favored for the development of edge devices and the internet of things.

Java was designed to have the look and feel of the C++ programming language, but is simpler to use and enforces an object-oriented programming model. Java can be used to create complete applications that may run on a single computer or be distributed among servers and clients in a network. It can also be used to build a small application module or applet for use as part of a webpage

HISTORY OF JAVA

The internet and the World Wide Web were starting to emerge in 1996 and Java was not originally designed with the internet in mind. Instead, Sun Microsystems engineers envisioned small, appliance-sized, interconnected devices that could communicate with each other.

As a result, the Java programming language paid more attention to the task of network programming than other competing languages. Through the java.net APIs, the Java programming language took large strides in simplifying the traditionally difficult task of programming across a network.

The first full increment of Java occurred on Jan. 23, 1996. The well-known JavaBeans interface was introduced in Java 1.1 in February 1997.

Later versions of Java releases have received nicknames, such as JDK 1.2 being referred to as Java 2. Java 2 saw considerable improvements to API collections, while Java 5 included significant changes to Java syntax through a new feature called Generics.

In October 2009, Google released the Android software developer's kit (SDK), a standard development kit that made it possible for mobile device developers to write applications for Android-based devices using Java APIs.

Oracle Corp. took over the Java platform when it acquired Sun Microsystems in January 2010. The acquisition delayed the release of Java 7, and Oracle scaled back some of the more ambitious plans for it.

Java 8 was released in March 2014. It included Lambda expressions, which are common features in many competing languages but had been absent in Java. With Lambda expressions, developers can write applications using a functional approach, as opposed to an object-oriented one.

March of 2018 saw the release of Java 10 followed by Java 11 in September 2018. Java 12 was released in March of 2019.

WHY JAVA IS POPULAR

It is difficult to provide a single reason as to why the Java programming language has become so ubiquitous. However, the language's major characteristics have all played a part in its success, including the following components:

Programs created in Java offer portability in a network. Source code is compiled into what Java calls bytecode, which can run anywhere in a network, on a server or on a client that has a Java virtual machine (JVM). The JVM interprets the bytecode into code that will run on computer hardware. In contrast, most programming languages, such as COBOL or C++, will compile code into a binary file. Binary files are platform-specific, so a program written for an Intel-based Windows machine cannot on run a Mac, a Linux-based device or an IBM mainframe. As an alternative to interpreting one bytecode instruction at a time, the JVM includes an optional just-in-time (JIT) compiler which dynamically compiles bytecode into executable code. In many cases, the dynamic JIT compilation is faster than the virtual machine interpretation.

Java is object-oriented. An object is made up of data as fields or attributes and code as procedures or methods. An object can be a part of a class of objects to inherit code common to the class. Objects can be thought of as "nouns" that a user can relate to "verbs." A method is the object's capabilities or behaviors. Because Java's design was influenced by C++, Java was mainly built as an object-orientated language. Java also uses an automatic garbage collector to manage object lifecycles. A programmer will create objects, but the automatic garbage collector will recover memory once the object is no longer in use. However, memory leaks may occur when an object which is no longer being used is stored in a container.

The code is robust. Unlike programs written in C++, Java objects contain no references to data external to themselves or other known objects. This ensures that an instruction cannot include the address of data stored in another application or in the operating system itself, either of which would cause the program and perhaps the operating system to terminate or crash. The JVM makes a number of checks on each object to ensure integrity.

Data is secure. Unlike C++, Java does not use pointers, which can be unsecured. Data converted to bytecode by Java is also not readable to humans. Additionally, Java will run programs inside a sandbox to prevent changes from unknown sources.

Applets offer flexibility. In addition to being executed on the client rather than the server, a Java applet has other characteristics designed to make it run fast.

Developers can learn Java quickly. With syntax similar to C++, Java is relatively easy to learn, especially for those with a background in C.

JAVA PLATFORMS

The three key platforms upon which programmers can develop Java applications are:

Java SE- Simple, stand-alone applications are developed using Java Standard Edition. Formerly known as J2SE, Java SE provides all of the APIs needed to develop traditional desktop applications.

Java EE- The Java Enterprise Edition, formerly known as J2EE, provides the ability to create server-side components that can respond to a web-based request-response cycle. This arrangement allows the creation of Java programs that can interact with Internet-based clients, including web browsers, CORBA-based clients and even REST- and SOAP-based web services.

Java ME- Java also provides a lightweight platform for mobile development known as Java Micro Edition, formerly known as J2ME. Java ME has proved a prevalent platform for embedded device development, but it struggled to gain traction in the smartphone development arena.

MAIN USES OF JAVA

It is easy for developers to write programs which employ popular software design patterns and best practices using the various components found in Java EE. For example, frameworks such as Struts and JavaServer Faces all use a Java servlet to implement the front controller design pattern for centralizing requests.

A big part of the Java ecosystem is the large variety of open source and community built projects, software platforms and APIs. For example, the Apache Foundation hosts a variety of projects written using Java, including simple logging frameworks for Java (SLF4J), both Yarn and Hadoop processing frameworks, Microservices development platforms and integration platforms.

Java EE environments can be used in the cloud as well. Developers can build, deploy, debug and monitor Java applications on Google Cloud at a scalable level.

In terms of mobile development, Java is commonly used as the programming language for Android applications. Java tends to be preferred by Android developers because of Java's security, object-oriented paradigms, regularly updated and maintained feature sets, use of JVM and frameworks for networking, IO and threading.

Although Java is widely used, it still has fair criticisms. Java syntax is often criticized for being too verbose. In response, several peripheral languages have emerged to address these issues, including Groovy. Due to the way Java references objects internally, complex and concurrent list-based operations slow the JVM. The Scala language addresses many of the shortcomings of the Java language that reduce its ability to scale.

MySQL

MySQL is (as of July 2013) the world's most widely used open-source relational database management system (RDBMS), enabling the cost-effective delivery of reliable, and high-performance and scalable Web-based and embedded

database applications. It is widely-used as the database component of LAMP (Linux, Apache, MySQL, and Perl/PHP/Python) web application software stack.

MySQL was developed by Michael Widenius and David Axmark in 1994. Presently MySQL is maintained by Oracle (formerly Sun, formerly MySQL AB).

MySQL is very commonly used in conjunction with PHP scripts to create powerful and dynamic server-side applications. MySQL is used for many small and big businesses. It is developed, marketed and supported by MySQL AB, a Swedish company. It is written in C and C++.

VIII. REASONS OF POPULARITY

MySQL is becoming so popular because of these following reasons:

- MySQL is an open-source database so you don't have to pay a single penny to use it.
- MySQL is a very powerful program so it can handle a large set of functionality of the most expensive and powerful database packages.
- MySQL is customizable because it is an open source database and the open-source GPL license facilitates programmers to modify the SQL software according to their own specific environment.
- MySQL is quicker than other databases so it can work well even with the large data set.
- MySQL supports many operating systems with many languages like PHP, PERL, C, C++, JAVA, etc.
- MySQL uses a standard form of the well-known SQL data language.
- MySQL is very friendly with PHP, the most popular language for web development.
- MySQL supports large databases, up to 50 million rows or more in a table. The default file size limit for a table is 4GB, but you can increase this (if your operating system can handle it) to a theoretical limit of 8 million terabytes (TB).

MySQL EDITIONS

There are five types MySQL editions.

- **MySQL Enterprise Edition:** This edition includes the most comprehensive set of advanced features, management tools and technical support to achieve the highest levels of MySQL scalability, security, reliability, and uptime.
- **MySQL Standard Edition:** This edition enables you to deliver high-performance and scalable Online Transaction Processing (OLTP) applications. It provides the ease of use that has made MySQL famous along with industrial-strength performance and reliability.
- **MySQL Classic Edition:** This edition is the ideal embedded database for ISVs, OEMs, and VARs developing read-intensive applications using the MyISAM storage engine.
- **MySQL Cluster CGE:** MySQL Cluster is a scalable, real-time, ACID-compliant database, combining 5 x 9s availability and open source technology. With a distributed, multi-master architecture and no single point of failure, MySQL Cluster scales horizontally on commodity hardware accessed via SQL and NoSQL APIs.
- **MySQL Embedded (OEM/ISV):** MySQL Database is a full-featured, zero-administration databases that more than 3000 ISVs, OEMs, and VARs rely on to bring their products to market faster and make them more competitive.

MySQL SUPPORTED PLATFORMS

MySQL runs on

- Linux (RedHat, SUSE, Mandrake, Debian)
- Embedded Linux (MontaVista, LynuxWorks BlueCat)
- Unix (Solaris, HP-UX, AIX)
- BSD (Mac OS X, FreeBSD)
- Windows (Windows 2000, Windows NT)
- RTOS (QNX)

MySQL FEATURES

- **Relational Database Management System (RDBMS):** MySQL is a relational database management system.
- **Easy to use:** MySQL is easy to use. You have to get only the basic knowledge of SQL. You can build and interact with MySQL with only a few simple SQL statements.
- **It is secure:** MySQL consist of a solid data security layer that protects sensitive data from intruders. Passwords are encrypted in MySQL.
- **Client/ Server Architecture:** MySQL follows a client /server architecture. There is a database server (MySQL) and arbitrarily many clients (application

programs), which communicate with the server; that is, they query data, save changes, etc.

- **Free to download:** MySQL is free to use and you can download it from MySQL official website.
- **It is scalable:** MySQL can handle almost any amount of data, up to as much as 50 million rows or more. The default file size limit is about 4 GB. However, you can increase this number to a theoretical limit of 8 TB of data.
- **Compatible on many operating systems:** MySQL is compatible to run on many operating systems, like Novell NetWare, Windows* Linux*, many varieties of UNIX* (such as Sun* Solaris*, AIX, and DEC* UNIX), OS/2, FreeBSD*, and others. MySQL also provides a facility that the clients can run on the same computer as the server or on another computer (communication via a local network or the Internet).
- **Allows roll-back:** MySQL allows transactions to be rolled back, commit and crash recovery.
- **High Performance:** MySQL is faster, more reliable and cheaper because of its unique storage engine architecture.
- **High Flexibility:** MySQL supports a large number of embedded applications which makes MySQL very flexible.
- **High Productivity:** MySQL uses Triggers, Stored procedures and views which allows the developer to give a higher productivity.

IX. CONCLUSION

To extract the important features, one only needs the attribute values and their corresponding information gain. From the extracted features the threshold value can be calculated. Once the threshold value is calculated, no classifier is needed for classifying new pattern. Only the threshold value of the new pattern has to be calculated and comparing with the previously calculated threshold value, it can be classified. If the threshold value is less than the calculated threshold value, then it can be said that the student is in risk of drop out. In the section of performance analysis, performance has been shown in two stages: for original datasets and after detecting outliers. Both for the original datasets and the datasets after detecting outlier our proposed approach works better. The work in this paper is limited to applying the method for original datasets and for the datasets of after detecting outliers. In future, the imbalance of the datasets can be considered and removal of the imbalance can enhance the performance of the proposed method.

EVALUATION PROCESS:

This paper presented the analysis of different supervised learning techniques in the context of educational data mining for university student's dropout avoidance. It was found that students dropped from their respective programs more frequently in the 4o semester for the CE and IS programs, and in the 6o semester for the CS one since those were the semesters where the algorithms had their best result.

X. FUTURE WORK

As for future work, we are scheduling to enlarge our dataset with other departments within our institution and to point out more factors as useful attributes that influence our results in more accurate way.

REFERENCES

[1] I. Teixeira, *Instituto Nacional de Estudos e P. E. A. Censo da educao superior 2019*, 2019.

[2] L. Manhaes, S. Cruz, and G. Silva, "Wave: An architecture for predicting dropout in undergraduate courses using edm," *Proceedings of the ACM Symposium on Applied Computing*, 03 2018.

[3] C. Marquez-Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 8, no. 1, pp. 7–14, Feb 2019.