International Journal of Applied Research 2018; 4(5): 242-256

ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 5.2
IJAR 2018; 4(5): 242-256
www.allresearchjournal.com
Received: 10-03-2018
Accepted: 11-04-2018

Hayat Sahlaoui
Engineering Science Laboratory, Polydisciplinary Faculty, Taza. Sidi Mohammed Ben Abdellah University Fes, Morocco

Abdelouahed Essahlaoui
Engineering Science Laboratory, Polydisciplinary Faculty, Taza. Sidi Mohammed Ben Abdellah University Fes, Morocco

Mohamed Khaldi
École Normale Supérieure, Tetouan. Abdelmalek Essaâdi University, Tetouan, Morocco

Correspondence
Hayat Sahlaoui
Engineering Science Laboratory, Polydisciplinary Faculty, Taza. Sidi Mohammed Ben Abdellah University Fes, Morocco

Data mining and its applications in the education sector

Hayat Sahlaoui, Abdelouahed Essahlaoui and Mohamed Khaldi

Abstract
The growing volumes of data continue to outpace human capabilities to extract and discover valuable information capable of supporting sound decision-making and plausible prediction in any business or industry this can only be addressed using automated methods such as data mining.
Data Mining Technology (DMT) is a robust technology that can help organizations to make optimal informed decisions-instead of guesswork decisions [1], in addressing contemporary challenges facing higher education.
Although data mining technology is progressing, its use leaves much to be desired. its adoption and implementation is expensive and any failure in implementation causes not only financial loss but also dissatisfaction among users, which makes it imperative to study the acceptance and adoption of this technology by its potential users.
previous studies about data mining have always focused either on the technical aspect or the development of DMT application algorithms without considering users' perception of the technology [2]. Little effort has been used to encourage users to appreciate DMT's capabilities and let's DMT obtain the users' acknowledgement; as this could help minimizing underutilization or eventual abandonment despite the prevalent benefit.
This study aims to promote and encourage the use of data mining technologies by its potential users within higher education context:

- This paper aims to contribute to the conceptual and theoretical understanding of Data mining within higher education.
- It introduces the notion of Data mining and outlines its relevance to higher education. From a perspective of e- learning practitioners and data mining practitioners

To achieve these goals, this study aims to design a web portal that targets two types of potential users, e-learning practitioners and data mining practitioners and that can serve as a one-stop knowledge base.
That allows to explore the different participants in the education system and how each of these players can benefit from the data mining system in a virtuous cycle.
And allows data mining practitioners to thoroughly understand the data mining philosophy and apprehended the practice of its process by focusing on data preparation and three key areas in the modeling process; criteria to consider in order to optimize the chances of extracting useful information and to make the right choice of data mining algorithms, Model construction and model evaluation.

Keywords: Data mining technology-e-learning practitioners-data mining practitioners-data mining algorithm selection-model construction

1. Introduction
In this article we describe in brief the structure of the web portal demo, the brief content of each web page and finally the detailed content of each web page and some screen shots.

2. Research Methodology
- Exploratory study of the literature and reports of national and international conferences on data mining.
- Consultation of academicians and data mining experts.
- Present a brief process of development of the data mining system demonstration portal and the techniques used to build the website.

The main process of building the demonstration portal can be divided into five steps outlined in figure below

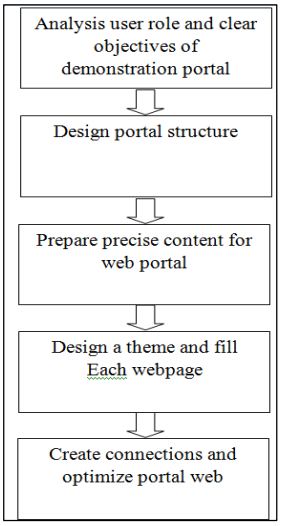


Fig1: procedures of developing the demonstration portal

3. Structure of the demonstration portal

The structure of the demonstration portal is presented graphically below:

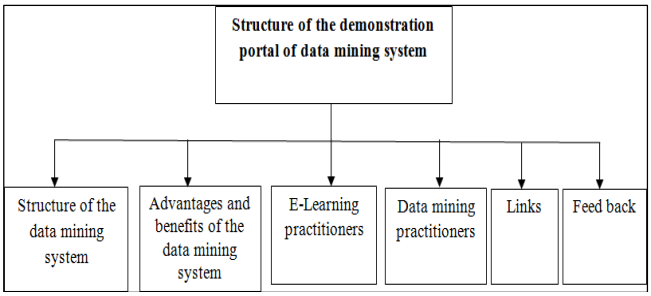


Fig 2: structure of the demonstration portal of data mining system

The portal shows six main parts and each web page is briefly described in Table below

Table 1: Description of the web portal pages

Element	Description
Structure of the data mining system	This page presents the structure and conceptual framework of the data mining system
advantages and benefits of the data mining system page	This page gives a brief introduction to the advantages and benefits of the data mining system
e-Learning practitioners page	This page shows how data mining tools could help e-Learning practitioners to function proactively and efficiently
data mining practitioners page	This page allows data mining practitioners to fully understand and master the data minig process and to optimize the chances of extracting useful information by showing them the criteria to consider when choosing algorithms and data mining software
Links page	This page is the collection of existing websites on data mining applications and solutions in education
Feed back page	This page is an interaction link where users can make suggestions and report problems

4. The detailed content of the web pages of the demonstration portal and some screen shots

4.1 The data mining system structure page

This page presents the structure of the data mining system by showing the main components of the system, namely data sources, data warehouses, data mining, Data mining, optimization and decision. It then presents the conceptual framework with three layers of the system namely the conceptual layer, the technical layer and the application layer, their components, the relations that govern them and their schematic representations

4.1.1 Functional or function oriented structure of the data mining system

Figure 3 illustrates the functional architecture of the data mining system, which is composed of several layers: (appendix)

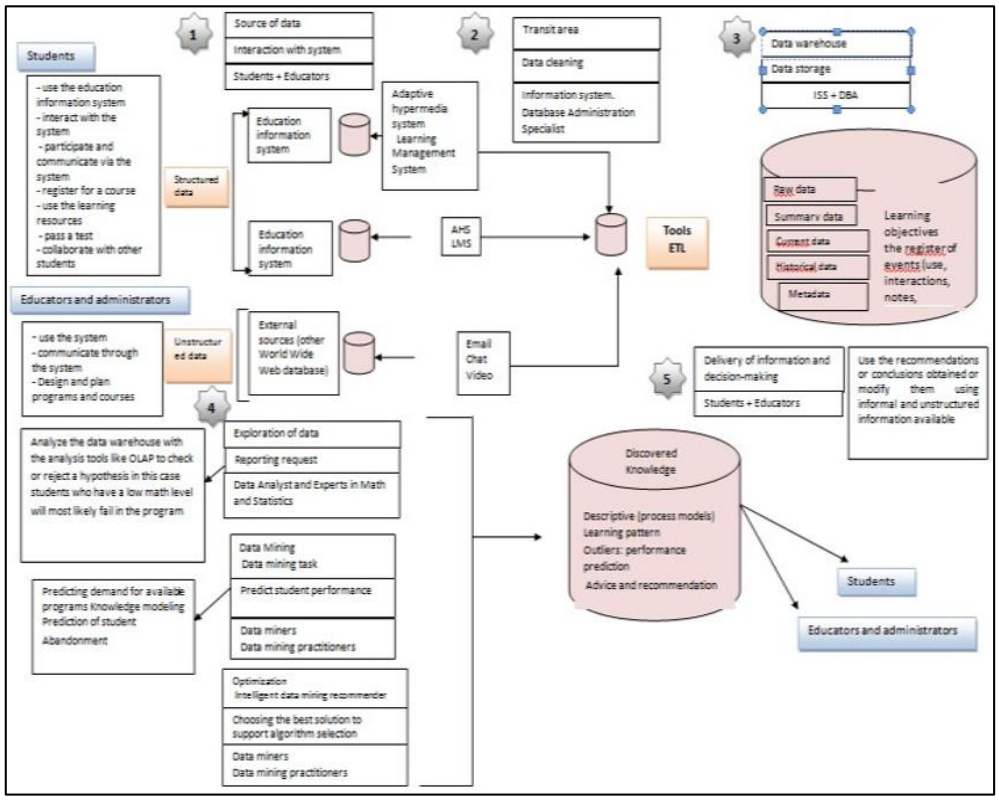


Fig 3: the functional architecture of the data mining system

- **Data sources**

Educational data in the data warehouse is derived from the operational systems that support the basic processes of the educational institution [3].

- **Data staging**

Three major functions performed in data staging are data extraction, data transformation and preparing it for loading.

- **Data storage**

In this stage data is stored in the data warehouse. The operational system of an enterprise supports only the current data but in data warehouse historical data is also kept.

- **Data exploration**

In this layer, data analysts and mathematical and statistical model experts use tools to perform passive business intelligence analysis.

- **Data mining**

In this layer data miners use active business intelligence methodologies, whose purpose is the extraction of information and knowledge from data.

- **Optimization or intelligent data mining recommender**

In this layer we find optimization models that allow us to determine the best solution out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite. Also, an intelligent data mining assistant can provide support for selecting a model/algorithm, thus suggesting to users the most appropriate data mining techniques for a given problem.

- **Information delivery**

In this stage useful information is provided to the wide community of data warehouse users through various systems like online, intranet, internet and e-mail etc.
Corresponds to the choice and the actual adoption of a specific decision, and in some way represents the natural conclusion of the decision-making process.

4.1.2 The conceptual or concept-oriented structure of the data mining system

A three-layered conceptual framework is proposed by Yao [4] in, consisting of the philosophy layer, the technical layer, and the application layer. The layered framework represents the understanding, discovery, and utilization of knowledge, and is illustrated in Figure.

- **The philosophy layer**

The philosophy layer investigates the essentials of knowledge. One attempts to answer the fundamental question, namely, what is knowledge?
A precursor to technology and application, it generates knowledge and the understanding of our world.
Data is unprocessed and raw facts that hold no meaning, as we move to the next layer, the data is now information, this means that data has undergone some processing to produce meaningful information, the next layer is mining the information for knowledge, and moving to the top layer means we have managed to acquire wisdom or insight that help increases the effectiveness of decision-making and chooses the best solution on a set of alternative measures

- **The technical layer**

The technical layer is the study of knowledge discovery in machine. One attempts to answer the question, how to discover knowledge?
Logical analysis and mathematical modeling are considered to be the foundation of technique layer study of data mining.

- **The application layer**

The ultimate goal of knowledge discovery is to respectively use discovered knowledge. The application layer therefore should focus on the notions of usefulness" and meaningfulness" of discovered knowledge for the specific domain.
These notions cannot be discussed in total isolation with applications, as knowledge in general is domain specific.
Also, the three layers mutually function on each other.
This point is explained by three facts:

- It is expected that the results from philosophy layer will provide guide-line and set the stage for the technique and application layers. It provides the conceptual guidance of knowledge structures, which serves as a pilot lamp for the further research work.
- The technical layer is the systematic pursuit of computer science activities of the framework. The technology development and innovation cannot go far without the conceptual guidance. Notwithstanding, the philosophical study cannot leave the technology either. At the meantime, technical layer is the bridge between philosophical view of knowledge and the application of knowledge.
- The applications of philosophical and technical outcomes give an impetus for the re-examination of philosophical and technical studies too. The application outputs are required an immediate evaluation and assessment.

These feedbacks come from the users and the customers necessitate the researchers work on the other two layers to make respond, either to complete or modify the knowledge structure, the methodology, or innovate the existing technology.

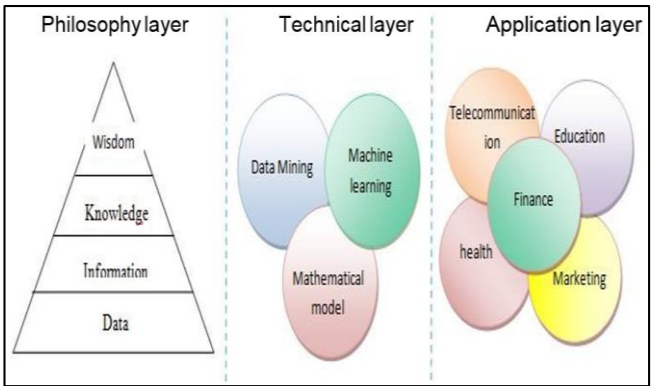


Fig 4: three-layer conceptual framework

4.2 Data mining system benefits page

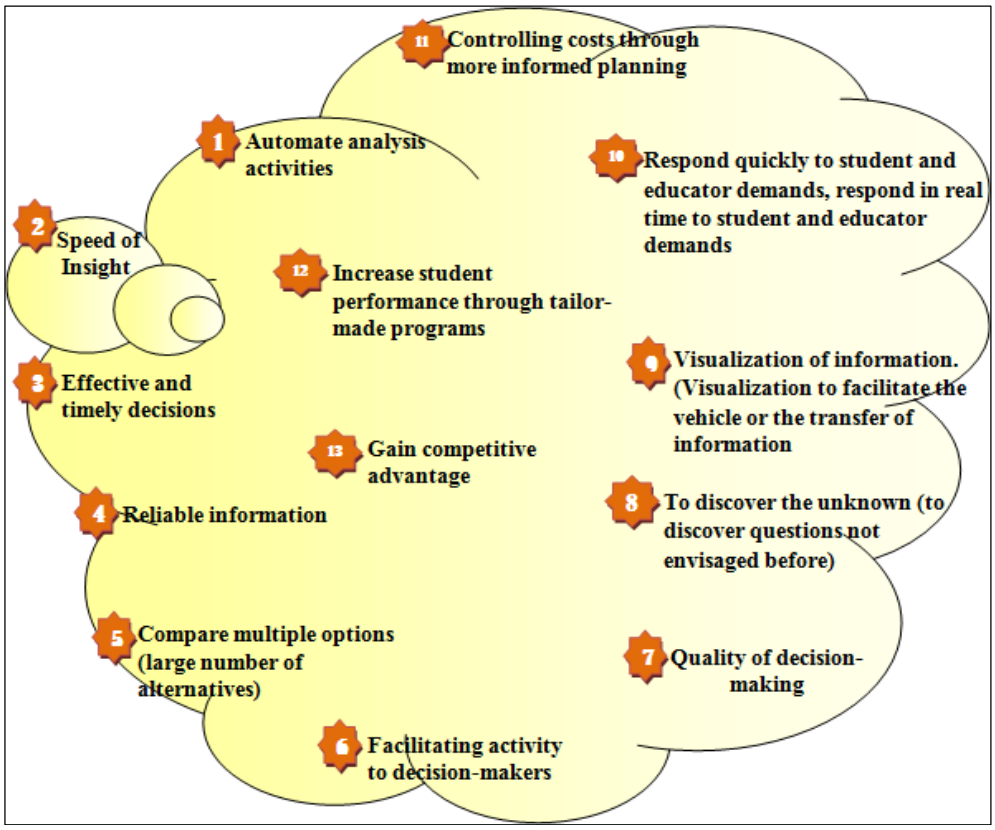


Fig 5: Data Mining Advantages

a) Automate data processing and analysis and provide insight speed.
b) Achieve more accurate conclusions and reach effective and timely decisions.
c) Make better decisions based on reliable information.
d) Facilitate decision-makers activities and improve the overall quality of the decision-making process. With the help of mathematical models and algorithms, it is actually possible to analyze a larger number of alternatives
e) Reveal patterns in data by sifting out nuggets of information from masses of unanalyzed or under-analyzed data, and transforms these mined nuggets into gold.
f) Move beyond simply answering questions posed by educators and researchers and begin to uncover key questions previously not even considered.
g) Move beyond simply answering questions posed by educators and researchers and begin to uncover key questions previously not even considered.
h) The most appealing aspect of data mining is what it produces-complex statistics transformed into usable visualization charts that convey large amounts of information to a user in the most meaningful and transferable form.
i) Manage and control costs through more informed planning and by allowing decision-makers to respond in real-time to
j) Students and educators demand and assure better management to the education institutions [5]
k) Provide management of the education institutions with information to increasing the performance of students through tailoring curriculum (education program) offerings.
l) Achieve competitive advantage by reducing the time required for insightful analysis and decision making from days to minutes and allowing the decision makers to visualize exception conditions requiring immediate action.
m) these predictions tools help reduce decision uncertainty by providing a degree of confidence to those decisions related to education programs

4.3 The e-learning practitioners page

We will explore the different participants in the education system and how each of these players can benefit from the data mining system in a virtuous cycle [6]. Figure 5
For this we have developed the following scenario

4.3.1 on the students' side

Individual students at the college would perhaps be the most impacted by data mining system.
New kind of data generated by a student as the way they interact with their university becomes increasingly digital. This is what we call the =digital footprint'-the data that is left behind as a student interacts with their university through online systems and on- campus technology.
For example, a university which requires swipe card access to its buildings will have a data set on how often each student is visiting campus, which buildings they are most frequently visiting, and which days and times of day they are most often on campus.
Each time a student logs into their institution's Virtual learning environment, they create a set of data including log in times, page clicks, downloads, length of time visited and comments made, etc. Video and audio lectures (if available) will also generate data, such as how long a student spends listening/watching to a single file, how often they

rewind/fast-forward, and any points at which they close the file and stop listening.

In the library there will be a record of how many books each student is borrowing, and libraries which provide e-textbooks may also be able to collect data about how students are using these. For example, when a student uses an e-textbook, they will be generating data on page clicks, the speed at which they read, any highlights and notes made in the text, and potentially even tracking data on where students‘ eyes are falling on the page.

As colleges and universities move to a more student centered learning environment, the most important use of data mining system may be in the hands of students. Students will be able to plan their academic experience and track their progress in each course, and be able to compare their efforts and results to those of their peers.

This powerful tool could aid students ‘engagement in coursework and with their engagement at the college.

Students could analyze their own progress, and they could benchmark their progress against other students ‘progress and course goals and behaviors.

The use of Data mining system could help students select the correct courses and levels based on their past performance and prior courses taken.

And provides activity recommendations based on their activity to date, and shows the student their progress so far students access to their own data encourage greater self-reflection on their performance and the factors that particularly affected this. It also encourage competition, in that students will try to beat their own scores of the previous week or that of the class average, or even comparing with their peers For students, receiving information about their performance in relations to their peers or about their progress in relation to their personal goals can be motivating and encouraging.

Students will also be able to uncover potential career and employment prospects and design future educational and life goals. When students have access to their own data and are able to relate to it in an applicable manner, they can shape a more meaningful and real future for themselves Students could use data mining system to better plan their educational experience, search transfer locations, seek financial aid prospects, and plan, research, and discover future employment opportunities, including areas of which they would not otherwise be aware. Data mining system could be a liaison for current students and alumni to share like goals, employment possibilities, and mutual interests. Data mining system could provide a richer college experience that keeps students engaged through their entire school life.

4.3.2 on the teacher side

Data mining system help teachers in various areas of Teaching and Learning

- **Student Performance Effectiveness**.

Data mining system could help teacher Monitor ongoing student test performance and compare to previous tests results as well as clusters of similar students. Integrate social media data and teacher notes to create a more detailed profile on the student‘s behaviors and propensities. Develop student- and class-specific recommendations, such as individual or small group tutoring, supplemental learning materials in-problem subject areas, or even changes in classes or majors

- **Student Work Groups**

Data mining system could help teacher Leverage cohort analysis for groups of students that can collaborate inside and outside class to improve individual performance. The data mining system allows the teacher to see cohort assignment, factors, and reasons for the assignment, and allow teacher overrides. The data mining system could-reshuffle‖ cohort assignment based on planned design elements and/or random factor; teacher to record observations (to be blended with objective cohort performance) after reshuffle to update the data set.

- **students feedback**

Data mining system could help teacher Provide students with better feedback on their progress

Feedback is traditionally one of the poorer areas of performance for universities, with

_assessment and feedback ‘generally being the two areas where students are the least satisfied.

Data mining system can provide a way for tutors to better understand how each individual student is progressing on their course, and in turn provide better-and more immediate-feedback for students. In many courses feedback is only provided when a student submits an assignment or sits the final exam, meaning that it comes too late to have any impact on their learning experience. A data mining system that combines data from all of the systems a student uses in the course of their study can provide a highly accurate, instant picture of student performance and engagement, and allow a tutor to provide high-quality, specific feedback more quickly.

- **Students' engagement levels**

data mining system help also detecting students' engagement levels, emotional states, and creating Heat maps‖ of activity in classrooms by applying advanced facial recognition, computer-vision algorithms and motion-tracking algorithms to video footage (digital data) from constantly running cameras in classroom

Data flowing from the interaction of student with the digital content would then be searched by system algorithms for patterns in each student's engagement level, moods, use of classroom resources, social habits, language and vocabulary use, attention span, academic performance, The resulting insights say, would be fed to teachers, parents, and students via school digital learning platform and mobile app, which are currently being tested. The information would be accompanied by scheduling tips, recommendations for more, and a playlist of assignments customized to each student.

How those suggestions are used, and whether they make a difference in how well each student learns, would also be tracked, creating a never-ending feedback loop of insights, experiments, recommendations, and product tweaks.

- **Enhancing teaching**

Information from data mining system can also provide powerful feedback to tutors on how effective their teaching strategies are for particular cohorts of students. For example, if a tutor can see that no students are downloading a particular resource, or that all students are heavily relying on one, then this becomes a powerful piece of feedback for the tutor to take on board when devising teaching strategies or designing modules and lessons.

At a time when teaching quality in higher education institutions is increasingly coming under scrutiny, data about students 'learning behaviors are potentially a powerful measure for how well tutors are performing. Better teaching should produce students who are more engaged with their studies and perform better.

On the one hand, if students are paying a high tuition fees for a three-year course, there may be an obligation on institutions to ensure that the teaching students are receiving is of high quality.

Data on how a student is interacting with their course and their institution can be an indicator as to how engaged the student is, and subsequently how likely they might be to drop out. For example, a student who isn't logging into the VLE or going to campus is likely not engaged with their studies, and might be at risk of completely disengaging and dropping out. Crucially, data mining system give tutors the ability to identify students who are disengaging at a much earlier stage, meaning that tutors can intervene before the situation escalates, improving the likely success of their students.

This data is used to measures engagement in order to target early interventions, and to improve retention over time.

data mining system could help teacher identifying at risk students and then recommending appropriate intervention resources or strategies, the larger promise of data mining system, is that it will-enable faculty to more precisely understand students' learning needs and to tailor instruction appropriately far more accurately and far sooner than is possible today.

The availability of real-time insight into the performance of students has positive implications for both teachers and students. For teachers, real-time data means they can take immediate steps to adjust and customize their teaching styles to better meet the needs of students.

4.3.3 On the Administrator side

Data mining system help administrator in various areas of educational institution

- **Enrollment Management**

Data mining system help administrator to identify, recruit, and retain a specific cohort of students based on certain pre-determined variables and criteria. It help them identify-ideall students and decide which students to accept, deny, or put on hold, based on historical data of previously enrolled students.

Data mining system can also help administrators tailor recruitment packages and follow-up efforts. -Just as Amazon.com knows when to send someone an e-mail notice of a new book that he/she might be interested in buying, so does an admissions office know whether to invest in printing and postage necessary to send a high school junior a glossy campus view book.

Data mining system can help target which prospective students are most likely to matriculate, based on past behaviors and characteristics of successful applicants.

-The history behind who enrolled at your school, -tells a powerful story about who you can expect to enroll in future terms.

Clearly, any tool that will help colleges and universities make better decisions about which students are likely to matriculate can have direct cost benefits. Knowing where to target that money and how to spend it can help administrators optimize both effectiveness and efficiency.

- **Student Acquisition**.

data mining system can also help administrators Use historical performance and demographics data of current and former students to create profiles of applicants most likely to enroll-then augment with social media data to score the institution's sentiment scores. Employ graphic analysis to examine current and prospective students 'social networks to identify first-level friends that may be potential new students.

This Data constitute a significant asset for institutions administrators, and it's used to inform their day-to-day operational decisions as well as longer-term business and strategic decisions. To take a common example, creating the timetable for each semester requires drawing on a range of different types of data across the institution. To work out how many lectures, tutorials and labs to schedule per module, the time table will need to bring together information about student enrollments in the module, staff numbers in each faculty (including staff with the relevant expertise/qualifications to lecture or lead labs) and estates data on rooms available with the necessary capacity and any equipment required.

- **Student Retention**

Today, most colleges and universities are grappling with the problem of student retention-and with good cause. Among those students who begin their college career as full-time freshmen in four-year colleges and universities, only a small portion complete their degree. This means that each year a lot students fail to achieve a college degree.

While there may be numerous reasons why a student doesn't continue to pursue his or her college degree after freshman year, research shows if a college or university can identify at-risk students in their critical first year, and intervene with appropriate resources or university support programs, they can potentially increase retention and help students persist to graduation. This is important not only for the student, but for university when education stakeholders are now measuring the success of an institution in terms of its graduation rates.

A number of schools administrators could turn to data mining system as a way to identify at-risk first-year students and to recommend the appropriate intervention, for outreach efforts, counseling, or other action.

Schools administrators could also turn to data mining system to combine previous metrics and scores including Student Performance Effectiveness and Student Work Groups, coupled with individual demographic, financial and social data to 1) score the likelihood of attrition, and 2) deliver recommendations that allow the institution to make a decision on whether to try to retain this student. Deliver and measure the effectiveness of specific recommendations-based upon the success of previous interventions. Empower teachers to make their own recommendations, which can be monitored for results and applied in futureretention intervention recommendations.

- **students' performance**

data mining system help administrators to Tracking students' performance across cohort, departments and courses and creating clusters based on different characteristics enables targeted strategies for specific segments of students. Such as Students pursuing a particular course and performing exceptionally well or average or below average students finding the course very tough. For

the below average cluster, the university administration can initiate structures intervention and provide them some special training to ensure retention and improved performance.
Analyzing the attendance data and focusing on students who missing the assigned course credit can help identify likely dropouts. Specific actions or retention programs for such students can have a significant impact on dropout rates.

- **Capture attendance data**

data gathered by swiping access card smart phone apps, portable card readers or proximity cards can provide administrators a good measure of engagement the system collects data from all timetabled activities and generates responses to patterns of attendance based on analysis of previous cohort demographics, behaviors, and outcomes. Interventions range from automated emails and text messages to a sophisticated targeting of _support priority students 'in the highest risk categories, who can be provided with personal support tailored to their circumstances and risks. Support is offered well before the usual signs of significant problems are evident, ensuring that it tackles issues before they become deep-seated and hard to address.
Data mining system help administrators provide support to particular student groups, such as underachieving students, students from minority groups, and other widening participation groups. Data mining system can be a powerful way to identify students who are struggling, and when linking this with demographic data, it can provide insights to particular issues faced by certain groups. And put in place additional support for these groups
Data mining system can help administrators understand the reasons behind a student's decision to leave the course midway. E.g. Insufficient or no financial-aid, high cost of education, poor grades, choice of subjects, distance from home, good job opportunity or better choice of college etc.
Data mining system can help administrators predict a particular student's probability of drop-out, right at the time of application, based on certain characteristics they possess. As a result, the university administrators took several measures to fine tune their short listing process and make necessary interventions to ensure the student stays back and completes the course.

- **Student Course Major Selection.**

Data mining system can help administrators compare a first-year college student's high school performance and aptitude tests to former student profiles to recommend a possible curriculum and major. Create detailed profiles based upon high school performance, areas of interest captured in both survey and social media, and aptitude test results. Compare those profiles to profiles on courses and majors to find the right match. Integrate external data regarding future workforce skills demands and salaries to help students make informed decisions on a major and minor.

- **Teacher Effectiveness**.

Data mining system can help administrators Measure and fine-tune teacher performance. While some institutions may be limited here, those that have the freedom to measure performance can benefit from insight into an individual teacher's effectiveness when compared similar teachers. Performance can be measured by subject matter, number of students, student demographics, student behavioral classifications, student aspirations, and a number of other variables to ensure that the teacher is matched to the right classes and students to ensure the best experience for teachers and students alike

- **Student Lifetime Value/Booster Effectiveness**.

Data mining system can help administrators Plan ahead with respect to potential giving levels for both current students and alumni. Understanding the likelihood to recommend and current or future earnings/wealth potential can all be major factors in profiling, targeting, and messaging to optimize alumni giving. Take advantage of these insights for the early identification of future boosters and uncovering Booster Top Performance predictors.

- **Student Advocacy**.

Data mining system can help administrators Leverage graphic analysis to examine a student's social network and score/monitor the likelihood to recommend (LTR) and triage specific areas of the college experience that are better or worse for a particular student. Use this data to create a Student Advocacy score that can be leveraged in the Student Acquisition (targeting a happy student's friends), Student Retention (flagging changes that can be a precursor to retention problems), Student Performance Effectiveness (flagging changes that can be a precursor to classroom performance problems), and Student Lifetime Value apps.

- **Bookstore Effectiveness**.

Data mining system can help administrators to use retail industry best practices to improve bookstore profitability using data-driven applications like merchandising effectiveness and textbook inventory optimization.

Data mining system can better empower career and academic advisers as they search for student employment opportunities, and help merge curriculum to industry needs.
Organizationally, the adoption of data mining system allows academic administrator to track student success and student needs in a timely manner. Academic administrator may instantly see when a student experiences a gap in success, attendance, resource management, or other retention factors.
Data mining system can help College capitalize on community partnerships and alumni contributions, making a positive impact in serving the community and alumni using better data to strategize where there is greater need for workforce development activities.

4.3.4 on the top manager side

Data mining system can help top manager analyze academic, financial and operational data to identify specific patterns and trends. This insight helps better decision making around planning budgeting and forecasting.
Data mining system can help top manager to improve access to university and develop more sophisticated recruitment strategies.
Data mining system can help top manager to use geo and demographics factors data to improve the way to identify market to and recruit students to institutions.
Data mining system can help top manager to use data to target potential students at different points of the decision making cycle which is particularly helpful in reducing information overload and assisting potential students by providing them with different packages of information when they need them. Data also helps us to target mature learners

and potential students who may have specific subject interests or professional needs.

Data mining system can help top manager to develop a client relationship management system (CRM) to enhance the service to applicants/students across the student life cycle-i.e. across the inquiry, application, admission and on-course stages

Data mining system dashboard delivers real- time monitoring of a variety of data sources and draws together student, HR and financial planning information in a user friendly way so that all managers across the university have access to real-time data. Managers can access information on cash balance, financial performance, space utilization as well as staff and student number data. The dashboard is central to the university‘s ability to manage assets, set targets and benchmarks, and produce detailed forecasts

On a strategic scale, data is used to inform senior management‘s business planning and overall strategy for their institutions. Student enrollment data, both historical and projected, as well as estates data, will influence the plans institutions make to build new buildings or refit current buildings to meet projected need. Financial data influences strategic decisions on expanding or reducing particular faculties or services provided.

Data mining system can help top manager to analyze the trend.

Analyze the curriculum and instructor development effectively on a regular basis to keep up with latest trends Getting insights on how to stay competitive and maximize profits Measure spend effectiveness against the results Manage fundraising, advancements and alumni relations Linking student information with administrative data enables better capacity planning Survey Analytics for understanding student sentiment.

Conducting satisfaction surveys and analyzing survey data is crucial for understanding the students ‘sentiments. Surveys help identify key areas of focus needed thereby enabling focused investments and action plans. Periodic surveys help track improvement in key areas of concern and thereby the effectiveness of the actions taken.

Data mining system can help top manager to predict the future based on the past. It involves digging into historical data, finding key patterns and predicting future trends on the basis of those patterns.

Data mining system can help top education managers to consume more meaningful information from big data, generate actionable insights from complex education problems and make data driven decisions across pan-educational institution processes to create sustainable education impact.

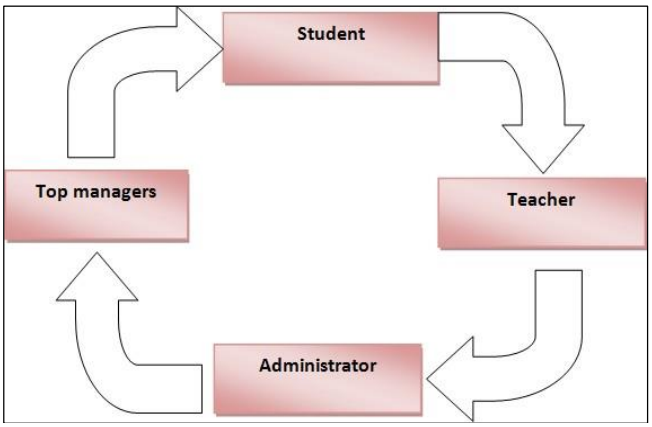


Fig 5: E-learning practitioners’ virtual cycle

4.4 The data mining practitioners page

This page allows data mining practitioners to thoroughly understand the data mining philosophy and apprehended the practice of its process by focusing on data preparation and three key areas in the modeling process; criteria to consider in order to optimize the chances of extracting useful information and to make the right choice of data mining algorithms, Model construction and model evaluation. Figure [6]

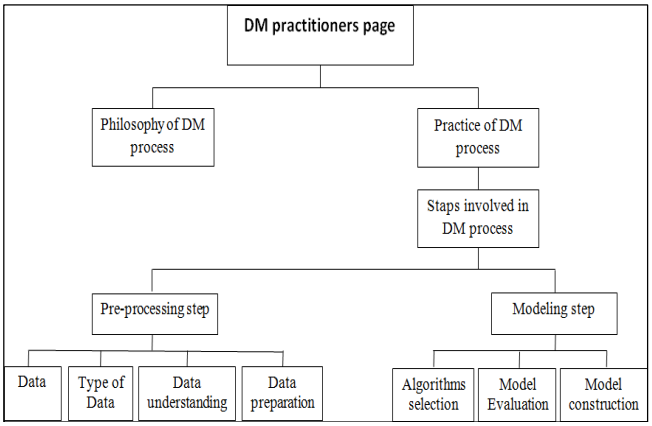


Fig 6: Philosophy and practice of DM process

4.4.1 Philosophy of the data mining process

Although CRISP-DM describes how data mining is performed, it does not explain what data mining is or why the process has the properties that it does. In this paper I propose nine maxims or -laws‖ of data mining (most of which are well-known to practitioners), together with explanations where known. This provides the start of a theory to explain (and not merely describe) the data mining process.

Here is an overview of Tom Khabaza’s “9 Laws of Data Mining.”

- **1st Law of Data Mining**, or -Business Goals Law‖: **Business objectives are the origin of every data mining solution**

We explore data to find information that helps us run the business better. Shouldn‘t this be the mantra of all business data analysis?

It‘s significant that this law comes first. Everyone should understand that data mining is a process with a purpose. Real miners don‘t play in dirt, they follow a methodical process to uncover specific valuable material. Data miners also follow methodical processes in search of what‘s valuable to them.

- **Quoting Tom Khabaza:**

Data mining is not primarily the technology, it is the process, which has one or more business objectives at its heart. Without a business objective … there is no data mining.‖

- **2nd Law of Data Mining**, or "Business Knowledge Law": **Business Knowledge is central to every step of the data mining process**

There‘s a horrible misconception floating around – that data mining doesn‘t require the investigator to know anything. This is a misinterpretation of the true philosophy of data

mining, that discovery of useful patterns in data can and should be put in the hands of business people who are not formally trained statisticians. Data mining is meant to bring power to the people-business people-who use their business knowledge, experience and insight, along with data mining methods, to find meaning in data.

- **3rd Law of Data Mining or-Data Preparation Law: Data preparation is more than half of every data mining process**

This should come as no surprise to anyone with experience dealing with data, whether as a data miner, a traditional analyst, or in another role. However, this is another area where there is mythology surrounding data mining, implying that data mining overcomes all issues of data quality and completeness. This myth was propagated by some long-forgotten vendors of data mining products, but the data mining community is still working to set the record straight. Data mining calls for good data.

But there‘s more to it than just needing good data. Manipulation of the data is an important part of the data miner‘s process. Here‘s how Tom Khabaza explains it: -The reason is deeper than the state of the data: during data preparation, the data miner customizes the problem space.

There are two aspects to this -problem space shaping‖. First the data miner must put the data in a suitable form for the algorithms to use at all-for many algorithms this means one row per example. Secondly, the data miner makes it easier for the algorithm to find a solution by enhancing the data with useful information or by putting the information into a helpful form. Examples include calculated fields, binning and calculating date and time differences.‖

- **4th Law of Data Mining, -NFL-DM‖: The right model for a given application can only be discovered by experiment**

(NFL-DM= -There is No Free Lunch for the Data Miner‖)

Here we could begin some very colorful discussion. At the end of this article, I‘ll direct you to some places where you can read and participate in such discussions. For now, it is important that you simply understand that experimentation is central to data mining philosophy and practice.

- **5th Law of Data Mining, or -Watkins‘ Law‖: There are always patterns**

The practical experience of data miners is that useful patterns are consistently found when data is explored.

[The -Watkins‖ mentioned here refers to David Watkins, also a well-known data miner and one of the developers of Clementine.]

- **6th Law of Data Mining: Data mining amplifies perception in the business domain**

This law speaks to the benefits of data mining algorithms and processes-they bring to light patterns in the data that would otherwise have gone undiscovered.

- **7th Law of Data Mining or -Prediction Law‖: Prediction increases information locally by generalization**

This is the law that I have found most challenging to clarify in my own mind, but here goes:

Data mining offers us ways to look at a case whose outcome is unknown, and find similarities to past cases where the outcome is known. By understanding those similarities, we gain information about likely outcomes for new cases.

- **8th Law of Data Mining, or -Value Law‖: The value of data mining results is not determined by the accuracy or stability of predictive models**

The real value of the process is in filling a business need. Accuracy or stability in a model are good, of course, but may be less important than issues such as the importance of predicted values to a business, meaningful insights, or the ease of putting the predictions to use.

- **9th Law of Data Mining, or -Law of Change‖: All patterns are subject to change**

A model that has great business value today may be just another old model tomorrow. Business does not sit still, neither can data miners.

4.4.2 The Practice of the Data Mining Process

Data mining involves the following steps show in figure 7, which must be executed in sequence until the results are documented and reported [7]

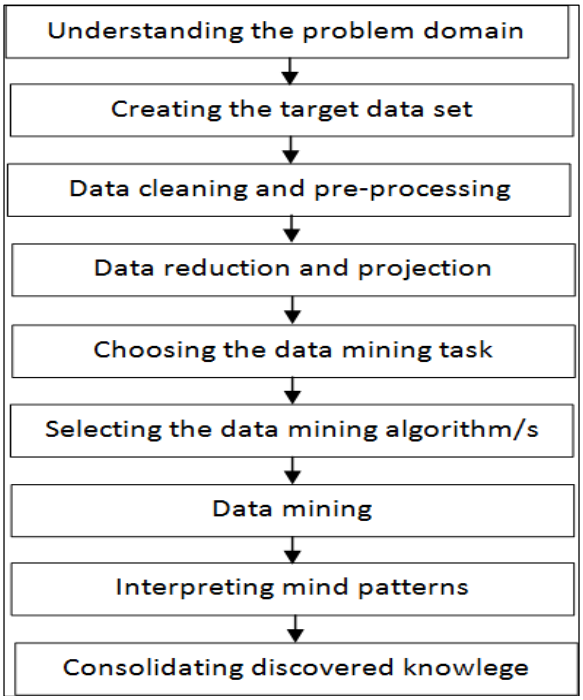


Fig 7: Steps Involved in the Data Mining Process

- **Pre-processing of data**
- **What is Data?**

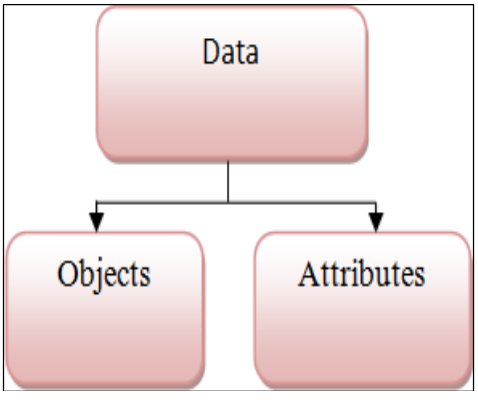


Fig 8: What is Data Collection of data objects and their attributes

- **An attribute is a property or characteristic of an object**
- Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
- Object is also known as record, point, case, sample, entity, or instance

Types of data sets

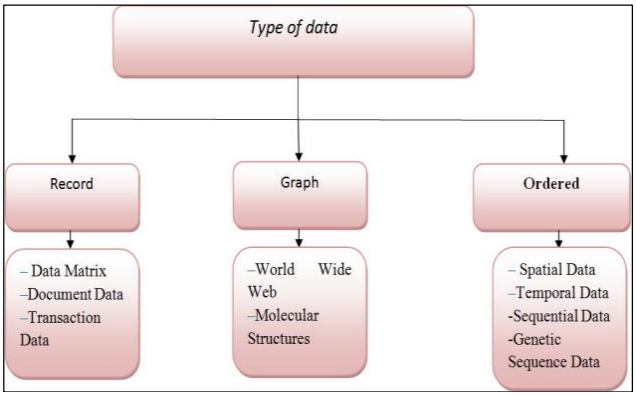


Fig 9: type of Data

- **Record Data**

Data that consists of a collection of records, each of which consists of a fixed set of attributes [8]

- **Data Matrix**

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.
- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

- **Document Data**

Each document becomes a `term' vector,

- Each term is a component (attribute) of the vector,
- The value of each component is the number of times the corresponding term occurs in the document.

- **Transaction Data**

A special type of record data, where

- Each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

- **Graph Data**
  - Examples: Generic graph and HTML Links Chemical Data
  - Benzene Molecule: C6H6

- **Ordered Data**

- Sequences of transactions
- Genomic sequence data
- Spatio-Temporal Data
- Average Monthly
- Temperature of
- land and ocean

- **Data Understanding**

The data understanding phase starts with an initial data collection. It proceeds with activities

- to get familiar with the data,
- to identify data quality problems,
- to discover first insights into the data, or to
- Detect interesting subsets to form hypotheses for hidden information.

- **Data Preparation**

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools [9].

- **Select Data**

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality and technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

- **Clean Data**

Raise the data quality to the level required by the selected analysis techniques.

Problems that can occur with-dirty data‖ include missing data, empty values, non- existent values, and incomplete data. Data cleaning may involve selection of clean subsets of the data, the insertion of suitable defaults or more ambitious techniques such as replacing the dirty data with derived values, or building separate models for those entities that possess dirty data. However, these approaches can introduce additional problems. Specifically, filtering the problematic data can introduce sample bias into the data and using data overlays could introduce missing values

- **Construct Data**

This task includes constructive data preparation operations such as the production of derived attributes, entire new records, or transformed values for existing attributes.

- **Integrate Data**

Two methods used for integrating data are merging data and generating aggregate values. In these methods information is combined from multiple tables or other information sources to create new records or values. For example, merging tables refers to joining together two or more tables that have different information about the same objects; generating aggregate values refers to computing new values computed by summarizing information from multiple records, tables or other information sources.

- **Format Data**

Formatting transformations refer to primarily syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.

- **Modeling**

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, several techniques can be applied to the

same data mining problem type. Some techniques require a specific form of data. Therefore, stepping back to the data preparation phase is often needed.

- **Algorithm and software selection**
- **Criteria for selection of data mining algorithms**

This study revealed that there are factors that researchers may consider when selecting a data mining algorithm [10].

- **Main goal of the problem to be solved**

This factor considers the reason why we are mining the data as well as the nature of the problem we are trying to address. A loan company may use a statistical decision procedure to determine whether to accept or reject cases for loan applications. The company may need to use classification rules to predict the number of customers who are likely to default their loan repayments based on information such as age, years with current employer, years with the bank and other credit cards possessed. So depending on the problem we are trying to solve data miners must select an appropriate algorithm. Also multiple algorithms may be used in a single solution to perform multiple tasks, for example using regression to obtain financial forecasts and then use neural network algorithm to perform an analysis of factors that influence sales. Table2 below shows how the nature of problems to be solved may be matched to possible algorithms.

Table 2: Example Problems and the Possible Algorithms that may be adopted to solve them to find anomalies in data, one may choose to use the decision trees algorithm.

Problem to be solved	Data mining Technique	Possible algorithm/s
identify anomalies in data, to find outliers	Classification	Decision trees One Class Support Vector Machine
Find items that tend to co-occur in the data and the rules that govern their occurrence	Association rules	Apriori
Find groupings in data	Clustering	K-Means
Create new features using linear combinations of the original attribute	Feature extraction	Non- Negative Matrix factorization

- **Structure of the available data set**

The data you provide is first analyzed to identify specific types of patterns or trends before defining the mining model. The results of the analysis are used to define the parameters for the mining model hence the need to consider the data set. The relationships between the objects/data, relationships between variables and the way that the data is stored influences the choice of an algorithm. Table 3 (below) shows examples of the data that will be required if a data miner is to implement a specified algorithm.

Table 3: Examples of the Required Structure of Data Sets for Some Algorithms

Algorithm	Structure of data set
Association algorithm	Single key column Single predictive column Input columns contained in two tables
Linear Regression	A key column Input column At least one predictable column
Clustering algorithms	Single key column At least one input column with values that are to be used to build the clusters Optional predictable column
Naïve Bayes	Single key column At least one predictable attribute At least one input attribute None of the attributes can be continuous numeric data as it will be ignored.
Neural networks	One input column One output column Data can be continuous, cyclical, discrete, key table or ordered
Time Series	Key time column that contains unique values Input columns At least one predictable column

The table above shows the recommended structure of the data if a researcher is to adopt a specific technique. The single key column is the column that uniquely identifies each record, in other words it is the primary key in a table.
The predictive column is the key column in the nested table, the foreign key. Therefore based on these examples, to implement a time series algorithm, there must be one column with time data, input columns and at least one predictable column.

- **Familiarity with an Algorithm**

Having experience in implementing an algorithm may make the selection much easier. Data miners may adopt those algorithms that they are familiar with although there is a risk that the chosen algorithms may not be suitable for the task to be performed. This factor however becomes useful when the same type of tasks is to be performed. A case base where all the experiences with different algorithms are documented may be useful when adopting this factor in selecting an appropriate algorithm. Researchers may get an idea of which algorithms work best in specific domains as well as the challenges they are likely to face, well before the data mining resumes.

- **Configuration Parameters**

When considering this factor, select an algorithm that may be integrated to the data source at minimal costs. An algorithm that may be fully integrated to the organization‘s database would be more ideal, where extracting, importing and exporting the data will be easy to automate without incurring extra costs. However, it does not mean that an algorithm must be selected if it can only be integrated without satisfying other criteria. To make an informed choice, researchers may consider more than one of the factors in selecting a data mining algorithm to use.
Considering only one aspect may increase chances of yielding undesired results, rendering the data mining process meaningless.

- **Criteria for selection of data mining software**

Software selection depends on various factors such as software performance, functionality, accessory tasks, hardware and software requirements to run the software effectively, vendor responsibility, and Quality or software capability to manage data discrepancies [11]

- **Performance Criteria**

Software performance criteria is taken into consideration which depends mainly on factors like reliability, efficiency and output of the software. Any data mining software which is being implemented in any company should give consistent results for business development.

- **Functionality Criteria**

Software functionality criteria is taken into Consideration which depends on adaptability and capability of the software to customize according to the requirement of the particular organization

- **Auxiliary task support**

Ancillary task support criteria is taken into consideration to find software capability to handle data with discrepancies and faults like blanks

- **Software Quality Characteristics**

Quality criteria is taken into consideration which includes software capability to modify history of actions being carried out, portability and easily understandable GUI.

- **Criteria related to vendor Criteria**

Vender criteria is taken into consideration which includes software upgrading, training and technical support from the vendor for improved customer service.

- **Criteria related to hardware and Software**

System requirements, including supported computer platforms are often particular to a company or project.
Hardware and software criteria is taken into consideration which includes the resources that are available with the organization and the investment required for software implementation and smooth functioning of the software package.

- **Data mining algorithms and their applications in the field of education**

Here is a brief description of the most popular techniques of data mining and their real applications in the Education domain. Simply to provide the minimum information to make the final choice [12]

Table 4: Examples of applications of data mining algorithm in the field education

technique	algorithm	Supervised unsupervised	application
Classification	Decision tree	✓	- Predicting Failure - Predicting Dropout - Alumni Fund Raising - Classify the students as Bad Average Very Good and Excellent Prediction of Placements - Student Retention
	Bayesian	✓	- Classify the students as Bad Average Very Good and Excellent - Academics & Recruitment
association	aprioi	✗	- Finding Adept Teacher Dealing with Students Failure - pattern extraction
clustering	K-Means	✗	- Persistent and Non-Persistent student Comprehension - Groupin Similar Students
Outlier detection		✓	- Rare Events Analysis and understanding - Understanding irregular events

Model construction

The purpose of building models is to use the predictions to make more informed business decisions. The most important goal when building a model is stability, which means that the model should make predictions that will hold true when it's applied to yet unseen data [9].
Regardless of the data mining technique being used, the basic steps used for building predictive models are the same. The model set first needs to be split into three components: (1) the training set, (2) the test set, and (3) the evaluation set. A fourth dataset, the score set, is not part of the model set.
Each of these components should be totally separate; that is, they should not have any records that are in common since each set performs a distinct purpose.
Models are created using data from the past in order for the model to make predictions about the future. This process is called training the model. In this step, the data mining algorithms find patterns that are of predictive value. Next, the model is refined using the test set.
The model needs to be refined to prevent it from memorizing the training set. This step ensures that the model is more general (i.e. stable) and will perform well on unseen data. Next, the performance of the model is estimated using the evaluation set. The evaluation set is entirely separate and distinct from the training and test sets. The evaluation set (or hold out set) is used to assess the expected accuracy of the model when it is applied to data outside the model set.
Finally, the model is applied to the score set. The score set is not pre-classified and is not part of the model set used to create the data model. The outcomes for the score set are not known in advance. The final model is applied to the score set to make predictions. The predictive sores will, presumably, be used to make more informed business decisions. The process is summarized in Figure 10.

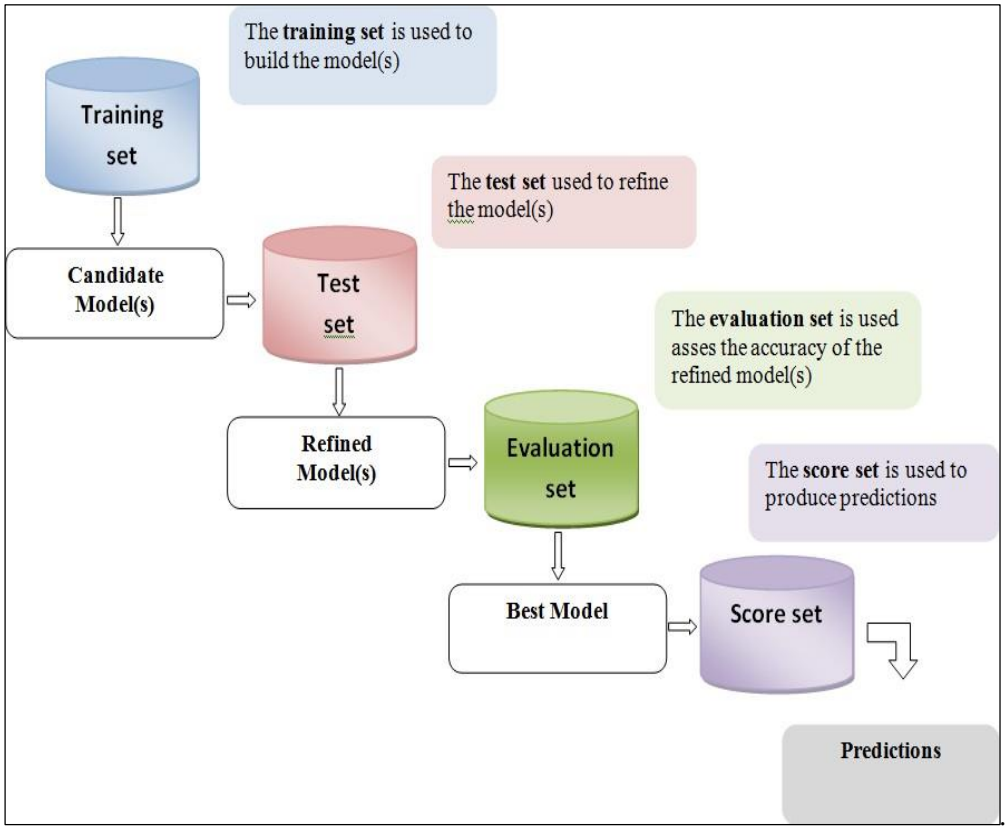


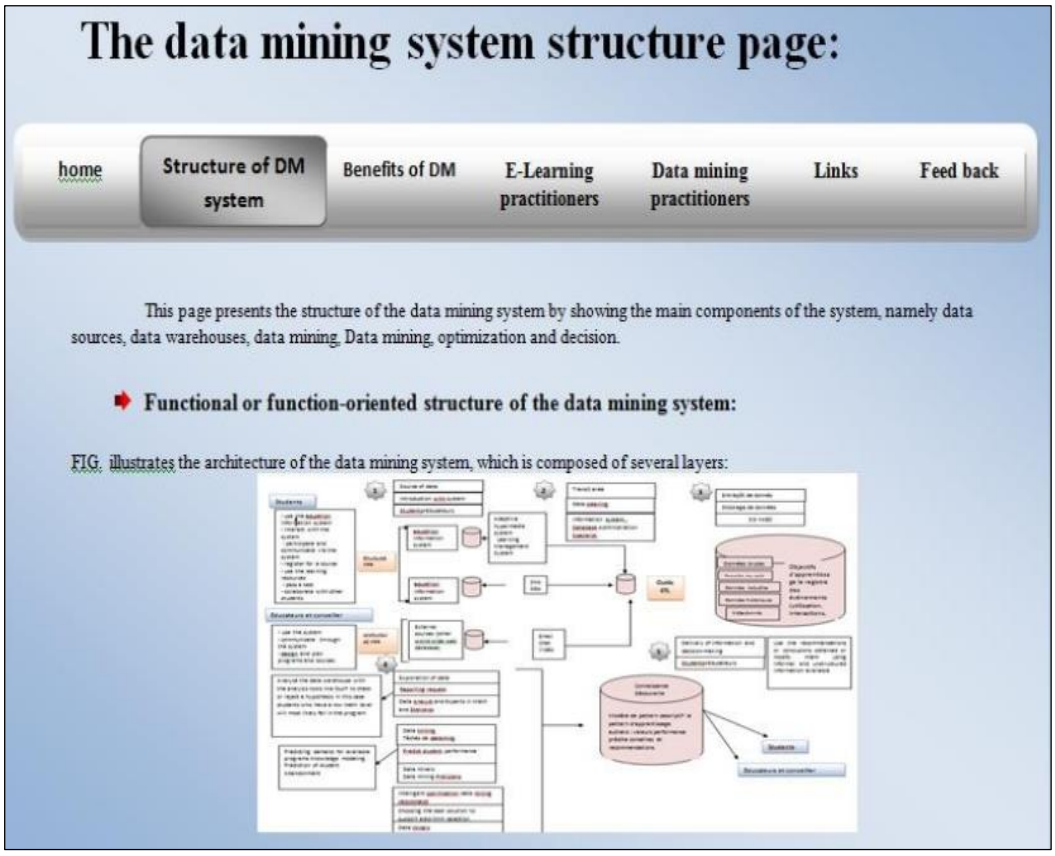
Fig 10: The Process of Building a Predictive Model

- **Model evaluation**

Previous evaluation steps dealt with factors such as the accuracy and generality of the model. This step assesses the degree to which the model meets the business objectives and seeks to determine if there is some business reason why this chosen model is deficient. Another option of evaluation is to test the model(s) on test applications in the real application if time and budget permits.

- **some screen shots of the demonstration portal are given below**

Figure 11 is a view of the first page it show the structure of data mining system



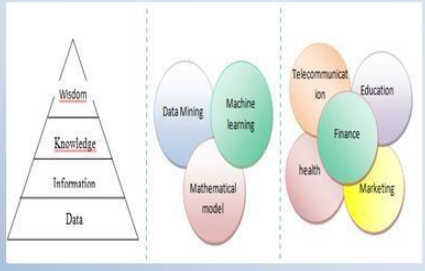
The data mining system structure page:

Home | Structure of DM system | Benefits of DM | E-Learning practitioners | Data mining practitioners | Links | Feed back

A three-layered conceptual framework is proposed by Yao in, consisting of the philosophy layer, the technical layer, and the application layer. The layered framework represents the understanding, discovery, and utilization of knowledge, and is illustrated in Figure.

- **The philosophy layer**

The philosophy layer investigates the essentials of knowledge. One attempts to answer the fundamental question, namely, what is knowledge?



Data mining system benefits page

home | Structure of DM system | Benefits of DM | E-Learning practitioners | Data mining practitioners | Links | Feed back

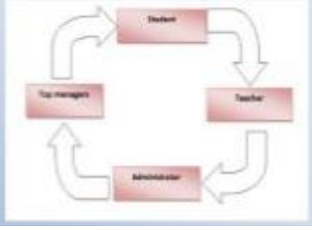
The e-learning practitioners page:

Home | Structure of DM system | Benefits of DM | E-Learning practitioners | Data mining practitioners | Links | Feed back

We will explore the different participants in the education system and how each of these players can benefit from the data mining system in a virtuous cycle. For this we have developed the following scenario

➔ **on the students' side:**

Individual students at the college would perhaps be the most impacted by data mining system. new kind of data generated by a student as the way they interact with their university becomes



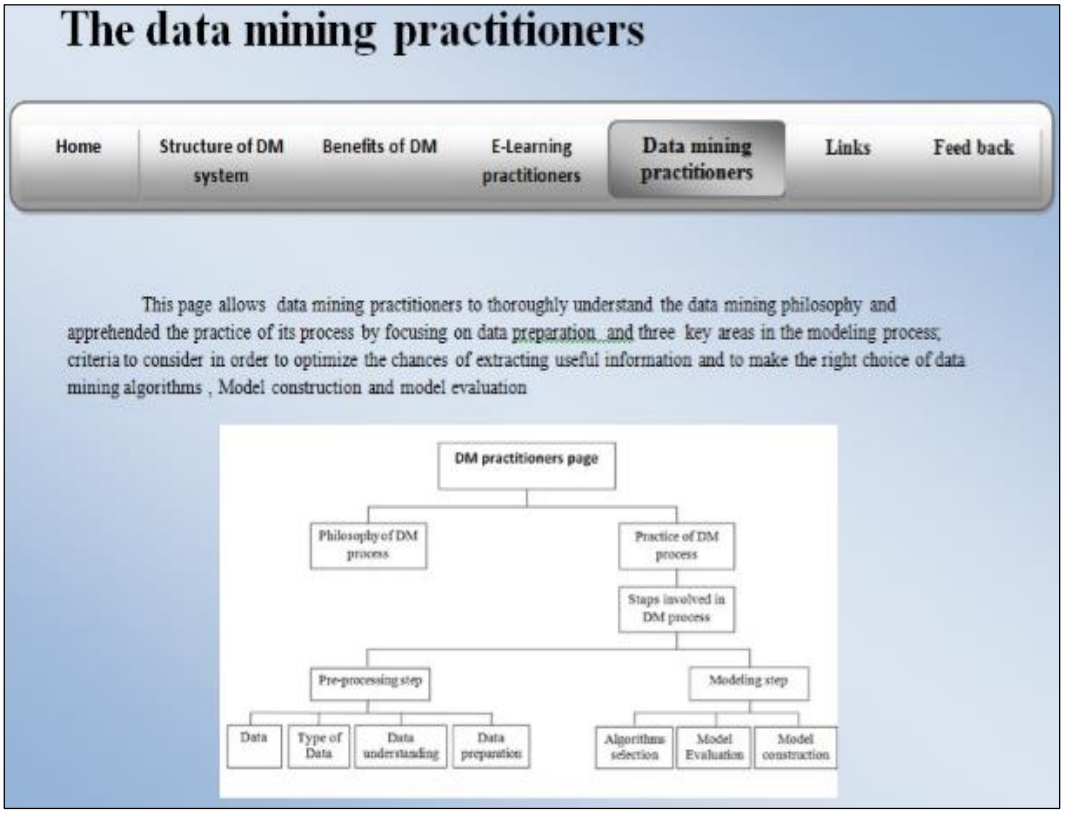


Fig 11: screen shots of the demonstration portal

5. Reference

1. Nemati HR, Barko CD. Organizational Data Mining. In: Data Mining and knowledge Discovery Handbook, Maimon O. and L Rokach (Eds). Springer, Boston, MA, 2010, 1041-1048
2. Romero C, Ventura S. Educational data mining: A review of the state of the art. IEEE trans. Syst. Man Cybernet Part C: Appl. Rev. 2010; 40:601-618
3. Vercellis C. Business Intelligence: Data Mining and Optimization for Decision Making. John Wiley & Sons Ltd, 2009.
4. Yao Y, Zhong N. Three-layered Conceptual Framework of Data Mining, 2004
5. Heinrichs J. Integrating web- based data mining tools with business models for knowledge management. Decision Support Systems. 2003; 35:103-112
6. AlHammadi D, Aksoy M. Data Mining in Higher Education. Periodicals of engineering and natural sciences, 2013, 1:2
7. Gibert K, Rodríguez-Silva G, Rodríguez-Roda I. Knowledge Discovery with Clustering based on rules by States: A water treatment application. Environmental Modelling & Software. 2010; 25:712-723
8. Gibert K, Sànchez-Marrè M, Codina V. Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation, 2010.
9. Kumar T. Introduction to Data Mining. Lecture Notes for Chapter 2, 2004.
10. Jackson J. data mining: a conceptual overview, 2002.
11. Communications of the Association for Information Systems, 2002, 8:267-296
12. Chikohora T. A Study of the Factors Considered when Choosing an Appropriate Data Mining Algorithm International Journal of Soft Computing and Engineering (IJSCE), 2014, 4(3).
13. Bhargava N, Arya R. Selection Criteria for Data Mining Software: A Study IJCSI International Journal of Computer Science, 2013, 10(3).
14. Walte D, Reddy H. Overview of algorithms in Educational Data Mining for Higher Education: An Application Perspective. International Journal of Engineering Research & Technology (IJERT), 2014, 3(2).