Name of the University

Name of the department

PROPOSAL FOR MS THESIS (CT-5002)

MS Degree Title

Fall Semester 2020

An Intelligent System for Enhancing Student Experience

Name of Student: ABC ABC ABC

Roll No: CT-113/2016-17

Contact Number: 03xx-xxxxxxx

Email Address: student@neduet.edu.pk

Proposed Supervisor: Dr. ABC

Designation: Assistant Professor

Table of Contents

Table of Figures

Abstract

The current era of Artificial Intelligence, Machine Learning and Big Data is allowing education institutions to use technology to generate and use a wide variety and volume of data to analyse students' performance and to enhance their learning experience. This in turn can significantly tackle the challenges faced by universities to improve levels of student retention and progression. This research proposal introduces Educational Data Mining and Learning Analytics systems, and a methodology to develop a proof-of-concept prototype that may have the future potential to support higher education institutions in using the data collected in educational settings to better understand and support their students.

Although studies on Educational Data Mining and Learning Analytics have focussed on the application of various data mining algorithms to study educational attributes, their approaches seem to lack a holistic view to explore most of the key factors and service areas that may enhance student experience. This project aims to apply data mining techniques to extract knowledge from student data, IT services and National Student Surveys, with the goal to help educators and administrators gain insight into student opinions, predict student performance and to improve services and student experience.

Keywords: Educational Data Mining (EDM), Learning Analytics (LA), Data Mining, Algorithm

1. Introduction

In recent years, universities around the world are focussing on institutional transformation to improve student's learning experience. Thus, Learning Analytics (LA) and Educational Data Mining (EDM) are gaining increasing momentum as a process for providing insights that could be used to inform decision making process to support university management and to improve student services (Dawson et al., 2008; Vahdat et al., 2016).

Educational Data Mining (EDM) refers to "an emerging discipline, that uses statistical, machine learning and Data Mining (DM) techniques to analyse the data sets from educational institutions, to improve learning experience, academic performances and institutional effectiveness" (Baker & Yacef, 2009, p. 4). Potential Data mining applications in education can range from the predictive modelling of student retention/attrition and student's learning behaviour to student's learning experience (Dawson et al., 2010). Such analysis assists academics in better understanding the specific learning needs of the students. While the potential of Educational Data Mining to improve learning experience and institutional effectiveness is emerging within the literature, current systems still lack to analyse the data without considering other Virtual Learning Environment (VLE) systems and data within the institution (e.g., IT services, student surveys etc.). Analysis on such data may help assess informed decisions relating to the enhancement of the student experience.

Learning Analytics (LA) refers to "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and environments in which it occurs" (Anon, 2011, p. 1). Linan et al. (2015) provides an in-depth comparison between EDM and LA. EDM focuses more on the techniques and methodologies used to analyse individual components of data, while LA deals more with applications by adopting a holistic framework. Although there is a considerable overlap between both these fields of research, the most striking difference is in the techniques and methods applied. EDM employs classification, clustering, Bayesian modelling, relationship mining, and visualization; while LA focuses on social network analysis, text analysis, sentiment analysis etc. Both EDM and LA are potential fields of research that focuses on improving educational experiences by helping stakeholders (instructors, students, administrators and managers) to make better decisions using data. In other words, these two research areas are complementary and hence to derive a holistic insight, it would be beneficial to utilize both approaches (Papamitsiou & Economides, 2014).

Given that universities are facing challenges due to decline in enrolment figures, increased competition and high attrition levels, there is a significant need for a critical review of academic strategies (Lee et al., 2009). Thus, this project will utilise EDM and LA techniques after a thorough critical review of previous research studies for devising an intelligent system to enhance student's experience.

1.1. Background

Universities around the world are facing the challenges to enhance student experience and services which thereby can significantly alleviate the issue of student attrition and retention. Although previous researches have made several attempts to address this issue. But not much progress is made to understand the factors and to identify the best practises from the application areas of Educational Data Mining and Learning Analytics systems to derive a prototype that can help improve student experiences.

1.2. Aims and Objectives

The aim of this project is to design and develop a system to support the enhancement of student experience using Machine Learning capabilities.

The above aim raises the following objectives:

- Conduct a comprehensive literature review covering research over a decade on the application of intelligent systems used to improve student experience and student retention.
- Identify and prioritise from previous research the most significant attributes and factors influencing the student experience.
- Design, develop and test an intelligent system prototype aimed to improve the student experience and levels of student retention.

1.3. Research Questions

As part of this project proposal the following research questions needs to be addressed:

RQ1: What is the State-of-the-Art systems developed in support of student's experience?

RQ2: How can an intelligent prototype support student experience and levels of retention?

1.4. Ethical Consideration

Several ethical codes have been reviewed as per Marczyk et al. (2017) and for this project following ethical considerations will be taken care of at the design stage.

- Data anonymity and confidentiality will be adhered to as per the NDA (Non-Disclosure Agreement)
- Unbiased Data samples will be used for analysis
- Any type of communication in relation to the research will be dealt with transparency.
- Acknowledgement of works of other authors used in any part of the dissertation with the use of Harvard system according to the Dissertation Handbook
- Adherence to Data Protection Act; GDPR: General Data Protection Regulation

1.5. Project Timeline

Projects are time bound. Dvir et al. (2003) examines the relationship between project timings and scheduling to project success. The author finally claims that an effective project planning, scheduling and milestone planning has a positive correlation to the project success. To keep projects on track and set realistic time frames appropriate tools and techniques needs to be employed. For our project planning, the Gantt chart is utilised to display project schedule, tracking and deadline. Geraldi and Lechter (2012) provides a comprehensive description of Gantt chart with a critical review with other techniques like Program Evaluation and Review Technique (PERT) and Critical Path Method (CPM) methods. Finally, the author weighs the Gantt chart as one of the widely used tools for planning and controlling project schedules. On similar lines, Slack et al. (2010) suggests Gantt chart as the universal tool across spectrums of projects like low varied, low volume, high volume etc.

The Gantt chart clearly displays the activities in order of entry, by start date and bar visualisation which depicts the start, duration, finish and overlaps if any. The activities are specified in the chart's two dimensions: the vertical axis indicates the activity, while the horizontal axis depicts the time schedule (Wilson, 2013). The below Figure 7 shows the Gantt chart displaying our project schedule with major phases as Analysis, Design, Development, Testing and Evaluation with appropriate sub categories of activities. The milestones for the project are indicated in the Gantt chart for the “Project Proposal Submission” and “Project Report Submission” events.

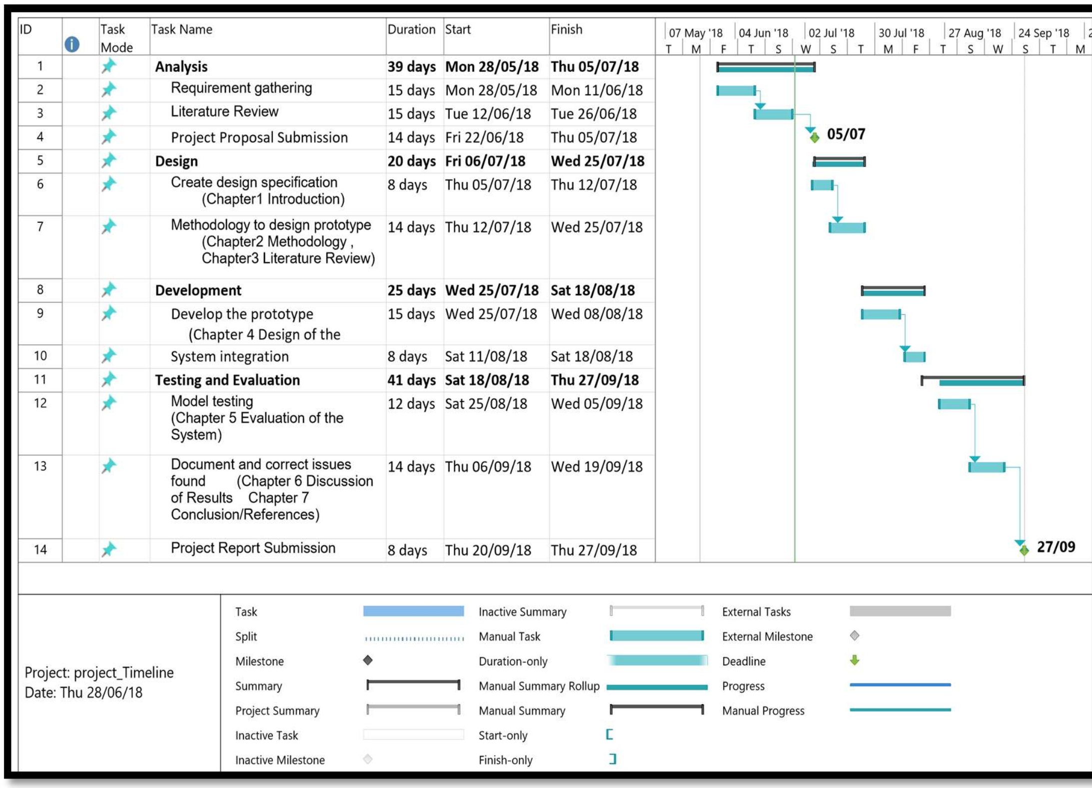

Figure 1: Gantt chart for the project showing the phases of the project as Analysis, Design, Development, Testing and Evaluation

To manage the project effectively, the schedule will be adhered with utmost diligence with special focus on the deliverables, identified milestones and the quality.

1.6. Skills Audit

The initial skills and competency assessment and the strategy for assessing the skills and competency portfolio are essential for improving the skill levels (Gibbings and Brodie, 2006). As shown in the below Table 5, the major skill audit parameters are provided along with the competency.

Table 5: Skill audit describing skill indicators, competency and ways to improve.

Skills	Skill Indicators	Initial competence level	Approach to improve competency

Project management	Apply effective project management strategy to plan, organise, implement and assess the project deliverables within the timelines and quality.	* Competent * **Average** * Needs Improvement	Adhere to the timelines and review deliverables for each phase.
IT Skills	*Use of state of art techniques from data mining/machine learning to provide insights to data. *Evaluating and Optimising the Results. *Effective interpretation and visualisation.	* Competent * **Average** * Needs Improvement	Explore Sentiment Analysis and DM techniques with special focus on accuracy checks.
Academic writing	Deliver a concise well-structured report incorporating the use of academic research method.	* Competent * **Average** * Needs Improvement	*Incorporate all sections *Adhere to the Harvard Style citation
Personal effectiveness	*Persistent effort to deliver projects with accuracy. *Demonstrate self-discipline and motivation to explore new methodologies to solve difficult problems.	* Competent * **Average** * Needs Improvement	Plan and achieve the milestones with proper review.

2. Methodology

Methodology is a systematic approach to efficiently execute and manage the research study. It is essential to utilize appropriate methodology to ensure quality deliverables (Kothari, 2004). This section provided the details of the methodologies used in this project mainly: Literature Review Methodology and Design Methodology.

2.1. Literature Review Methodology

For undertaking a critical assessment of previous research and to identify any research gaps in EDM and LA, a Literature Review (LR) will be carried out to analyse relevant articles. A LR allows for the search and appraisal of significant research in the areas of EDM and LA. A LR provides and includes the identification, selection and analysis of evidence from the literature (Boell and Cecez, 2015; Salleh et al., 2017).

For this project, the methodology includes:

- Definition of the LR protocol (research questions, sources to be searched, search term, search strategy and Inclusion/Exclusion criteria).
- Selection criteria.
- Analysis strategy.

Rigorous review of previous research in the following databases will be undertaken: articles from IEEE, Springer, Google Scholar, Science Direct, JEDM (Journal of Educational Data Mining), ACM Digital Library and Summon. The search terms to be included are: *Education Data Mining, Learning Analytics, Student Experience, Student Attrition and Retention*. The search process will comprise from 2000 to 2018 and the inclusion and exclusion criteria for the papers was devised as shown in the Table1.

Table 1: Inclusion and Exclusion Criteria for the Research

Inclusion Criteria	Exclusion Criteria
English articles only	Data from blogs, Wikipedia etc,
Published Research Articles and journals	
Only full access articles	
Book chapters	
Date from 2000 to 2018	

As shown in Figure 1, rigorous search will be undertaken, duplicate records will be deleted, the quality of the literature/research work will be assessed by means of citations and interpretation of the results. Finally, after reading abstracts, available articles will be downloaded for full review.

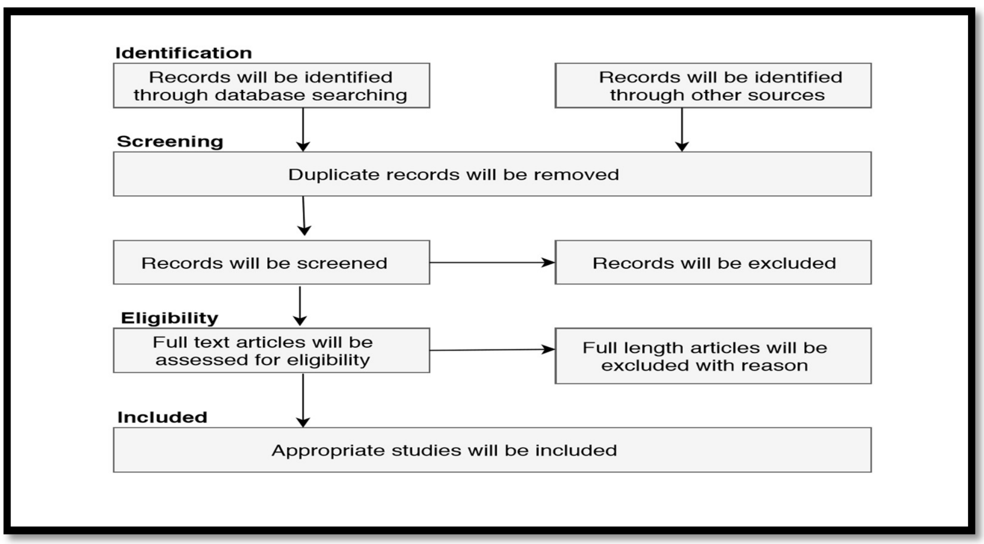

Figure 2: PRISMA Statement showing flow of information through the different phases of a literature review.

In this project, SWOT (Strengths, Weakness, Opportunities and Threats) is used as the analysis framework. Helms and Nixon (2010) provides an in-depth benefit of using SWOT analysis as a tool for planning purposes. SWOT analysis as a framework has been widely used due to its simplicity and practicality (Jackson et al., 2008). The overall SWOT analysis for this project proposal has been provided in detail in Section 6.

2.2. Design Methodology

An efficient project should be executed in an iterative way to ensure flexibility for requirement and implementation changes amidst the project lifecycle within the agreed time lines and quality. Project failures can be greatly reduced by executing the project using this approach (Nilsson and Wilson, 2012).

Spiral Model is a risk driven model that can be characterized by iterative processes that helps in mitigating risks. As shown in the below figure, Spiral Model consists of four main software development life cycle (SDLC) phases. The whole development process repeatedly passes through these stages. Each iteration is called Spiral (Gurendo, 2018).
As shown in the Figure 2, the four main phases are:

1. **Determine Objectives:** In this project, objectives will be defined with appropriate feasibility study and milestones.

2. **Identifying and resolving risks:** Risks will be identified, prioritized and resolved according to importance.

3. **Development and Testing:** At this stage, a high-quality working prototype in priority order, will be developed in accordance to the requirements. Later, in subsequent spirals, working version of the product will be delivered.

4. **Planning Phase:** At this stage, the output of the project will be evaluated prior to the next spiral.

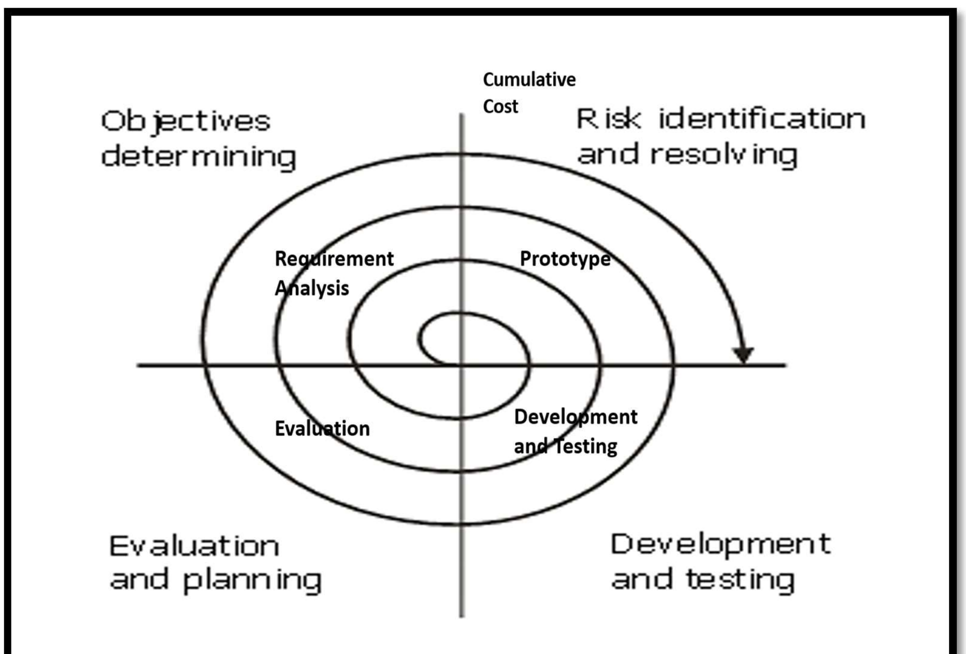

Figure 3: Spiral model (Adapted from Source: Boehm's model Gurendo,2018)

The key benefit of implementing project using spiral model is that its range of options accommodates the good features of existing software process models, while its risk-driven approach avoids many of their difficulties (Nilsson and Wilson, 2012). The Spiral Model provides control, flexibility and helps to keep the project on schedule

(Oriogun,1999). For this project, Spiral model is more suitable than Waterfall model and Human centric design as it is flexible with critical focus on risk assessments. Spiral model also incorporates prototype development which makes it the most suitable methodology for this project (Heim, 2007).

For this project, the spiral model will be applied in four cycles:

- Cycle 0: Determine the feasibility of the project. Project guideline negotiation with the stakeholders.
- Cycle 1: Develop objectives, prototype, plans and specifications.
- Cycle 2: Establish a specific and detailed design and identify any potential risk.
- Cycle 3: Achieve a workable prototype marking the identified milestone.

At a high level, the project workflow traverses through these three stages: Define, Design and Implement. As shown in the Figure 3, the below steps will be carried out for each phase for design and development of the prototype. As part of the Define stage, the aims and objective will be determined and after data pre-processing the features and the model will be selected. During the Design stage, the model will be trained and evaluated using proper accuracy matrices. Finally, as part of the implementation stage, the model will be deployed and iterated with new data.

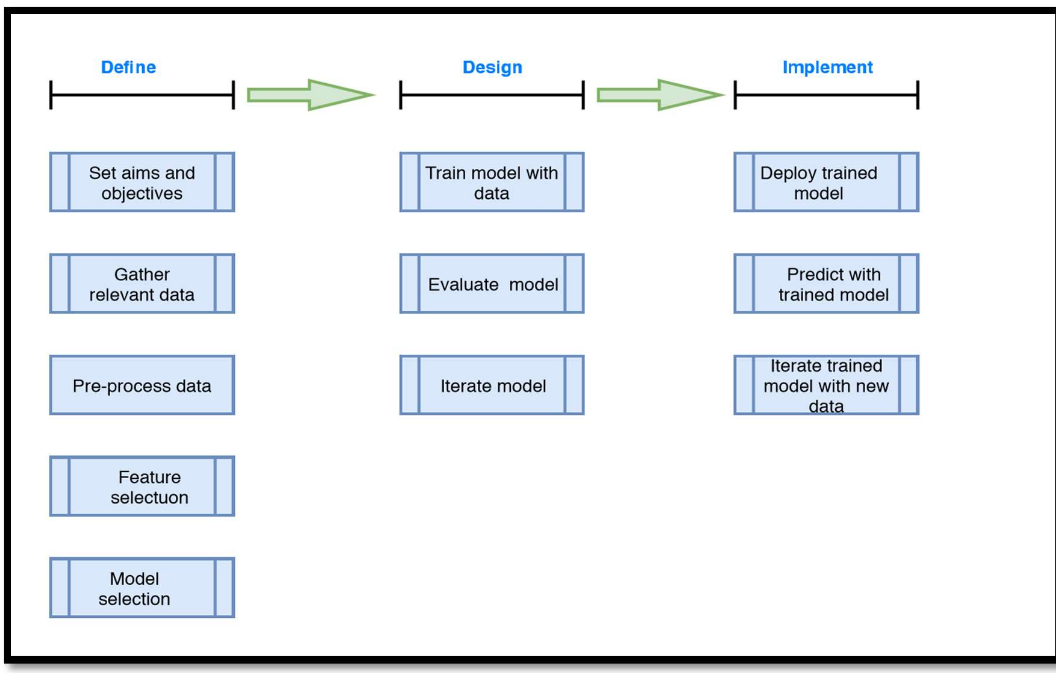

Figure 4: Project stages and the corresponding project workflow

2.3. Theoretical Background

Data Mining, Machine Learning and Sentiment Analysis is the core of the Knowledge Discovery in Databases (KDD) process, involving the inferences from the algorithms and techniques that explore the data, develop the model, discover patterns and provide insights.

Data Mining(DM) is a computational method of processing data in any domain that aims to obtain meaningful patterns and insights (Zyt et al., 2012). The application of data mining in educational systems is an iterative process and can be applied to data coming from varied sources and the techniques tends to be linked to supervised learning.

Machine Learning(ML) provides the technical basis of data mining that uses algorithms that learn from and make predictions on data (Witten et al., 2016). ML utilizes techniques for solving several problems: classification, regression, clustering, supervised learning, reinforcement learning, etc.

Besides the data mining techniques, there is a text analysis which deals with detection of sentiments from the text through specialized Semantic Analysis for varied purposes ranging from feedbacks to surveys and comments (Khadjeh et al., 2014).

As shown in the Figure 4, KDD consists of the following phases (Ltifi et al., 2013; Kurgan and Musilek, 2006). Although any of these three methods (DM, ML and Sentiment Analysis) may be used in Phase 3 of the KDD process. The techniques need to be different for each system due to varied data sources and objectives (Maimon and Rokach, 2009). The most widely used data mining and machine learning techniques in the field of education are classification, prediction, clustering, relationship mining and distillation of data for human judgment (Nithya et al., 2016). The KDD process is an interactive and iterative (with many decisions made by the users) and different stakeholders are involved at each phase.

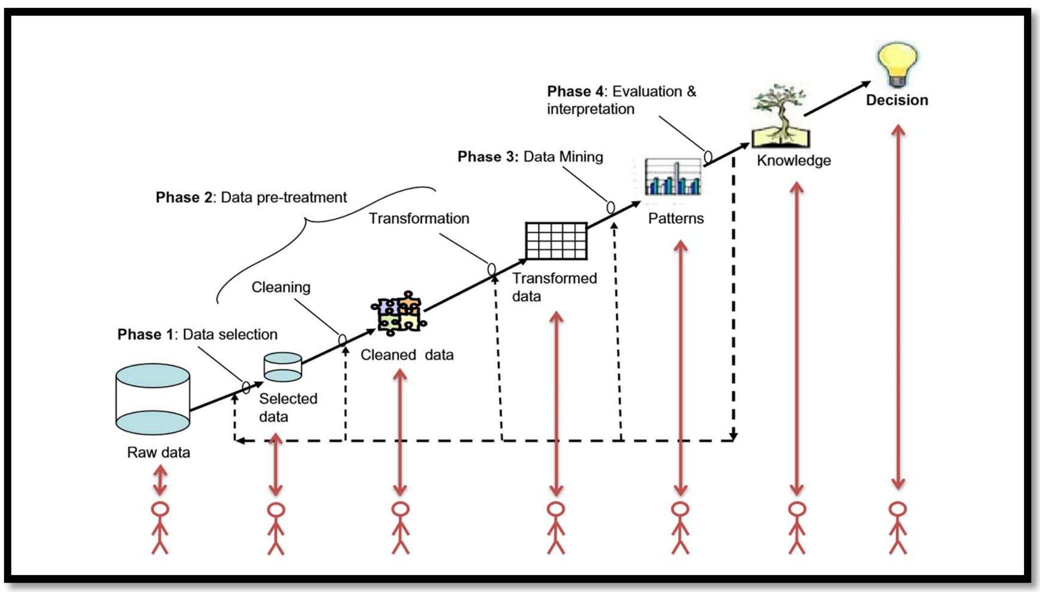

Figure 5: The KDD process (Source: Fayyad et al., 1996)

The different phases of the process as shown in Figure 4 are as follows:

• **Data selection**: Creating the target dataset.

• **Data cleaning and pre-processing**: Data cleansing and pre-processing for noise removal, outlier detection and handling missing and unknown value.

• **Data Transformation**: dimension reduction (feature selection and extraction) and attribute transformation (discretisation of the numerical features).

• **Data mining task**: Selecting appropriate data mining task depending on the KDD goals (e.g., summarization, classification, regression and clustering) to identify meaningful insights.

• **Choosing the Data Mining algorithm**: Selecting appropriate one or more machine learning algorithms method for searching patterns, insights and predictions. Wu et al., 2008 provided a list of widely used data mining algorithms mainly Decision Tree, K-Means, KNN, Naïve Bayes and Sentiment Analysis etc. These algorithms cover classification, clustering, association analysis, and text mining, which are all among the most significant areas in data mining research and development that can be utilised in this project.

• **Interpretation and Evaluation**: Evaluate the algorithms using accuracy matrix and interpret the extracted patterns using visualisation techniques.

• **Knowledge Management**: incorporate the discovered knowledge in an information processing system.

As seen in Figure 5, the application of data mining in educational systems is an iterative process and can be applied to data coming from two types of educational systems: traditional educational processes and virtual learning environment. In traditional classrooms, educators can monitor student's progress and performance by student's behaviour, analysing historical data, student's attendance etc. However, for students using web-based learning environment, educators need to use a different strategy to monitor student's performance (Romero & Ventura, 2007). Recently, there is a growing need for analysis of learner interaction data with web-based learning environments by means of Data mining techniques (Rabbany et al., 2012). This has been emphasized by Zafra et al. (2011) in the paper on classifying students in a web-based learning environment. The author provides an insightful analysis of the students' massive volumes of records representing clickstream or click-flow data from VLE's log data.

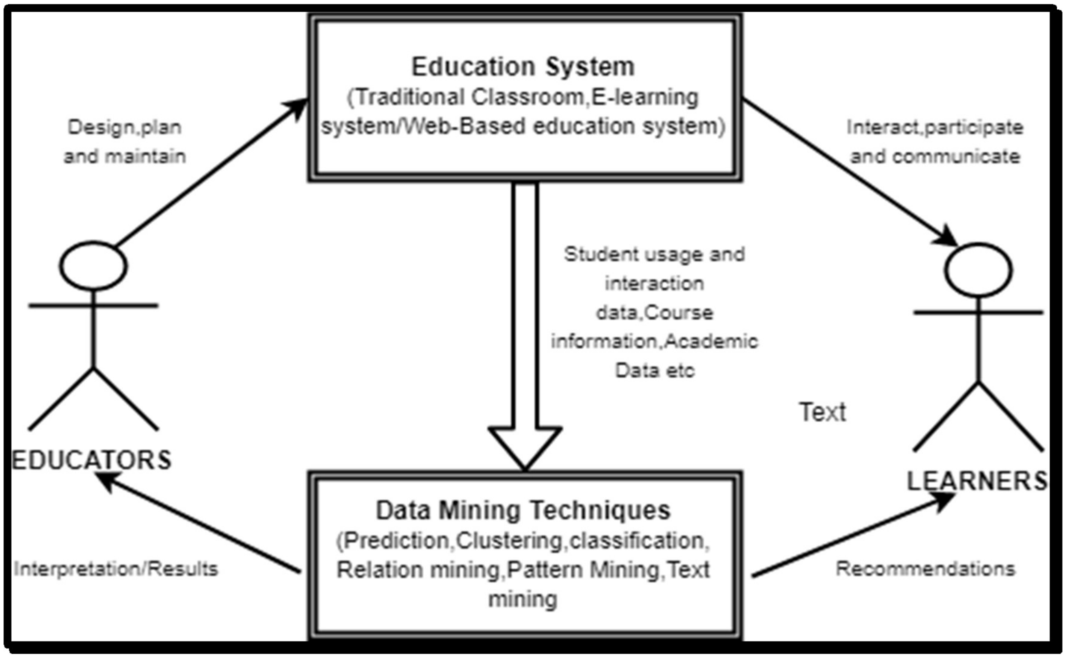

Figure 6: EDM/LA Process (Adapted from Source: Romero and Ventura, 2007, pp. 136)

2.4. Evaluation of the Prototype

Evaluation is essential in determining the project quality and development, which in turn has potential to stimulate development of future research (Stige et al., 2009). Quantitative evaluation and functional testing of the techniques and the prototype would be beneficial in assessing the efficient functioning of the project outcome. Specialised performance measures can be used for evaluating the algorithms for classification, regression, and clustering. Sokolova and Lapalme (2009) has provided an analysis of performance metric for the classification tasks which can be utilized for the evaluation of this project. The author has provided an in-depth comparison and strengths between the metrics like Accuracy, ROC-AUC, precision and recall which will provide a base for the evaluation of this project. For regression and forecasting, Bergmeir and Benítez (2012) provides an accuracy measure description using RMSE (root mean squared error) and MAPE (mean absolute percentage error). Most of the researches in the field of evaluation of text analysis are highly subjective and thus a qualitative evaluation approach needs to be followed for the textual analysis of the prototype. Thus, to summarize, the below Table 2 provides the performance measurement metrics that will be used for the evaluation of the prototype.

Table 2: Performance Measure metrics

Measure	Evaluation Focus
Average Accuracy	Overall effectiveness of a classifier
Precision	Proportion of positive outcomes that are truly positive
Recall	Measure of completeness of result. Calculated from the confusion matrix
AUC	Area Under the Receiver Operating Characteristic Curve from positive scores
F-Score	Harmonic Mean of Precision and Recall
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error

In addition to the quantitative evaluation, in this project, functional testing and analysis will be done to ensure the prototype has all the required functionality as per the requirement specification.

3. Literature Review

Several researches have been done in the field of EDM and LA. In general, both fields offer valuable potential in gaining insights from learning environment, predicting student behaviour and attrition but have significant differences in their origin and techniques (Baker and Inventado, 2014). A closer look into the literature, however, reveals that the differences between EDM and LA are focussed on the research questions, and the use of models rather than on the methods used (Siemens and Baker, 2012).Based on the literature analysis, there are a wide variety of research studies and applications of EDM and LA. Romero et al. (2010) provided a comprehensive review of data mining applications in the field of education and identified 11 categories(Analysis and Visualization of Data, Feedback Analysis, Student Recommendation, Student's Performance Prediction, Student Modelling, Detecting Undesirable Student Behaviours, Grouping Students, Social Network Analysis, Developing Concept Maps and Constructing Courseware) based on 300 research studies completed before 2010. Although the literature covers a wide variety of analysis, the review for this project will focus on five major application areas which emerge repeatedly in the literature. These are: student's behaviour, student's retention and attrition, infrastructure services and feedback/survey analysis. Although the literature presents these applications in a variety of contexts, this research proposal will primarily focus on their application using data mining and machine learning algorithms.

Such applications from EDM and LA can help stakeholders to understand special needs of the students, help make better decisions using data and enhance student experience.

For this research proposal, a literature review was undertaken to understand the background, methodologies of research carried out by previous researchers and to identify research gaps in this identified field.

3.1. Student's Behaviour

An educational institution maintains varied types of student data, ranging from student's academic records to their personal data. For that reason, the major task of Educational Data Mining (EDM) is filtering out information that can be used to model a student's behaviour and performance (Baker & Yacef, 2009; Romero & Ventura, 2010). A series of recent studies has indicated that personal factors, followed by school-related information and social factors are some of the significant attributes in understanding and predicting student's learning behaviour (Martínez et al., 2016; Dutt et al., 2017).

The research study by Baker and Yacef (2009) evaluates one of the goals of EDM as predicting student's future learning behaviour. The study proposes a student model that incorporates student's characteristics, including detailed information such as academic level, behaviour and motivation to learn. A different approach by Kock and Paramythis (2010) utilises clustering algorithm to discover different kinds of patterns in learners' problem-solving behaviour. The author used multi-level clustering approach at three different levels, each tailored to different pattern detection purposes to identify behavioural patterns of individual learners. On similar lines, Talavera and Gaudioso (2004) employs clustering approach on student's data to characterise similar behaviour groups and discover patterns reflecting user behaviours.

One of the pioneer research study that has inspired a great deal of later EDM work with ability to predict is given in paper by Beck and Woolf (2000). The study uses a variety of variables to predict student's learning behaviour. Furthermore, a recent study by Athani et al. (2017) proposed an automated system to predict student's academic performance and social behaviour. Although the study provides meaningful insights but was limited to utilising Naïve Bayes algorithm for classification and prediction and lacks critical comparison with other classification methods for higher accuracy.

Furthermore, from a different perspective out of student's online learning data, Lee et al. (2009) suggested that cognitive style is an important factor that determines student's learning behaviour. Meanwhile, Jie et al. (2017) points out the most significant factor from student's online study behaviour as student level as compared to demographic details. Munoz-Organero et al. (2010), utilised data from the Virtual Learning Environment (VLE) to identify relationships between student's motivation index and performance. The study was successful in establishing a positive correlation between student's e-learning usage to their motivation index. One of the tough challenges for all researchers in this domain of analysing from online data is the handling of the large amount of data. All the listed researches have been done only from sample data thus lacking accuracy and requiring a more systematic and robust approach.

To summarise, as reported above, the previous researches have a special focus on determining the significant student attributes and applying appropriate algorithm to analyse student's behaviour. However, further research work needs to be done to evaluate the accuracy levels and to utilise other sources of online student log data to provide further insights. Thus, in this project, an appropriate feature selection will be carried out, followed by application of the algorithms with critical focus on checking the accuracy using different techniques like confusion matrix, ROC-AUC, precision and recall.

3.2. Student's Retention and Attrition

Retention and attrition rates in educational institutions have long been a focus of research study. High level of student attrition has a critical impact on university's reputation and even financial implications to the government. In fact, one of the major goals of academic institutions is to make sure student's complete the course within the course time. Hence, the concern of efficiently enhancing the high skill level of university students has become critically important in many countries, including EU States, where reducing attrition rates is one of the Europe 2020 strategy goals (Ec.europa.eu., 2018).

Several distinct factors contribute to student attrition. Several authors have recognized personal, family-related, school-related, and social variables as some of the most important determinant of student attrition (Martínez et al., 2016). On the same line, a recent research study by Sarra et al. (2018) used Bayesian Profile Regression(BPR) and devised a covariate model comprising of professional experience, student satisfaction index, motivation factor and student's resilience skills. The author demonstrates BPR in identifying clusters to develop profiles of drop-out prone students which, in turn, provides university's decision support system with relevant information to improve retention.

Previous studies have almost exclusively focused on predictive models for identifying students at risk of academic failure and attrition. Hoffait and Schyns (2017) presents an early detection of potential failure using student's enrolment data at registration. Along similar lines, Vandamme et al. (2007) emphasizes the need of analyses at an early phase by using classification data mining tools, such as decision trees, neural networks and linear discriminant analysis to identify students at risk of dropout. Majority of the research study deals with majorly three categories of features mainly personal characteristics (gender, nationality, year of enrolment etc.), previous academic data (grades, type of school attended etc.) and socio-economic factor (parental education level, income etc.)

One of the early researches by Dekker et al. (2009) presented a data mining case study demonstrating the effectiveness of several classification techniques and the cost-effective learning approach from student's pre-university dataset. The author established that simple classifier (Decision Tree) had higher accuracies as compared to Random Forest, BayesNet etc. A more comprehensive research by Quadri and Kalyankar (2010) incorporated several machine learning methods for automatically analysing the student performance from cumulative grade point average (CGPA). In addition to deriving meaningful insight from the underlying relationship in the data, the study also utilised Decision tree to identify the factors that influence dropouts. Based on the inferences from the tree analysis, logistic regression was used to quantify the dropouts and understand the effect of each risk factor. The study found that the most significant risk factor affecting the dropout frequency is the parental income as compared to gender and parent education. Another effort towards student performance analysis which in turn could be used for predicting attrition was undertaken by Angeline and James (2012). The study uses Apriori algorithm to predict students under three categories: good, average and poor and finally accuracy was assessed by the value of the confidence value.

Virtual learning environments (VLE) in addition to the traditional classroom teaching has emerged as an important aspect of academic learning. Romero et al. (2008) reviews and recognises the strength of log data from VLE in student tracking and assessment management which in turn helps to enable early detection of students with potential difficulties. One of the most commonly used VLE is the Moodle (modular object oriented developmental learning environment), a free learning management system to help educators create effective learning communities (Rice, 2006). The analysis of usage logs can help understand student's interaction with the online system which can be one of the ways to improve performance and avoid attritions (Psaromiligkos et al., 2011). Romero et al., 2008 employs K-means clustering to identify student groups and thereafter utilizes a Decision Tree to classify students with similar grades depending on the activity carried out in Moodle. On similar lines, Lykourentzou, et al. (2009), proposed using combination of machine learning techniques like Artificial Neural Network (ANN) and Support Vector Machines (SVMs) to identify dropout prone students by analysing the student's usage logs from VLE. The author successfully established that the combinations of data mining techniques provided more accurate prediction of the dropout-prone students.

Finally, having reviewed the previous literature, another promising line of research to be included in this project would be to incorporate a feature selection strategy that will help to reduce a total number of features and usage of combination of machine learning techniques/ensemble methods by using multiple learning algorithms for optimum results.

3.3. Improving Student Infrastructure and Student Services

Previous research has pointed out vigorous debate about the effect of quality of infrastructure and services on the student's performance. Some researchers have underlined that the impact of infrastructure and the quality of services from the administrative staff should not be underestimated when trying to improve student satisfaction and opportunity for learning (Wiers-Jenssen et al., 2002). Hing and Zimmer (2016) provides some positive evidence of improved infrastructure on student proficiency levels. On the same line, Ramsden (2010) provides evidence that students are more dependent on high quality support services, easily available technology and better infrastructure. The paper points out details about academic institution's increase in investments to enhance learning spaces and upgrade of infrastructure. On the other hand, the author also points out the need for streamlining and further improving the services' effectiveness from student's perspective.

One of the earliest research study by Garfield and Wermter (2006) attempted call classification from corpus of IT helpdesk and compared the performance with different machine learning algorithm. The author employed neural network, support vector machine and finite state automata. The performance of each technique was evaluated using evaluation metrics recall and precision and the F-score and finally recurrent neural network was found to perform better in particular, when factors such as irregularities in the utterance and the number of call classes are considered. On the other hand, Povoda et al. (2015) attempted an emotion recognition method using machine learning techniques to classify helpdesk messages. In this paper, the text messages were obtained from helpdesk system and 5 emotional classes (afraid, angry, sad, satisfied and surprised) were identified from the tokens of the text and finally the accuracy of the model was compared with the state of art algorithms. Although the research identifies an emotional classification model but lacks an in-depth phase of pre-processing of data which can be very critical in determining model accuracy.

Previous researches point out infrastructure services as a neglected area in the field of student experience enhancement. This suggests opportunities for future research in this project to improve infrastructure services by means of call classification and helpdesk text analysis.

3.4. Student Feedback and Survey Analysis

Academic institutions have introduced a range of improved initiatives for student engagement to understand their experiences and to provide better strategies for responding to student feedback. There is a critical need to review quality assurance arrangements for students, including information for prospective students and analyses of student's experiences (including data from internal surveys and the National Student Survey (NSS) which are linked to the academic goals (Ramsden, 2010).

The emerging field of text mining has the potential to transform natural language into critical results, acquiring new insight and help in management's decision support system (Froelich and Ananyan, 2008). Chen et al. (2014) presented an analysis of social media data for understanding of student's learning experience. In the study, student's digital footprints helped inform institutional decision-making that helped improving education quality. They utilised text mining for assigning prominent categories and based on the categories Naïve Bayes classification algorithm was used to identify at-risk students. Furthermore, research study by Abd-Elrahman et al. (2010) uses text mining techniques for understanding free-style question answers in course evaluation forms. The author initially identified the polarity of the text and finally used co-occurrence-based analysis to compute the Teaching Evaluation Index(TEI) for the course. Li and Xu (2014) provided another perspective by looking for meaningful features to understand emotions instead of simply choosing words with high co-occurrence degree and thereby create a text-based emotion classification. Even though they managed to obtain fair results, the work was far from accuracy perspective with major factors unexplored and high dependency on human judgement.

Thus, as described, although there are many studies, the research in the field of in-depth sentiments analysis from survey and feedback data in EDM/LA remains limited. Further exploration on this area of text analysis can help in understanding challenging tasks.

Below Table 3 shows an in-depth comparison and evaluation of the research work done on the focus areas of EDM and LA.

Table 3: Comparison and Evaluation of the LR for the Project Proposal

Research Objectives	Algorithm and DM approach	Critical Results (Advantages and Disadvantages)	Author(s)
Student Behaviours	Clustering to discover pattern learning behaviour.	Exploratory Analysis and characterization using Clustering is beneficial prior to applying the machine learning techniques.	Talavera and Gaudioso, 2004; Kock and Paramythis, 2010
	Naïve Bayes algorithm for classification and prediction of academic and social behaviour	Significant lesson learnt is to be more selective on the type of data needed and to relate the selection to specific goal.	Athani et al., 2017
	Co-relational Statistics for motivation prediction from student behaviour and interaction pattern with VLE.	Evaluation and Interpretation of the results from the algorithm requires further tuning.	Munoz-Organero, et al., 2010
Student Retention and Attrition	Early detection of students with potential difficulties using Decision Tree, Random Forest, Logistic Regression and Artificial Neural Network Algorithms.	Combination of techniques known as ensembles provides more accurate and robust predictio.n	Hoffait & Schyns, 2017
	Identified the classifier that is most accurate to predict student's academic performance.	Feature selection is significant but adding more predictor variable does not necessarily improve the prediction accuracy	
	Association Rule Generation for Student Performance Analysis using Apriori Algorithm.	Discovered rules from the algorithm must be pruned to avoid redundancy.	Angeline and James, 2012
	Dropout prediction from VLE log data through the combination of machine learning techniques.	Visualisation techniques can help provide general view of the student usage data. Higher amount of training data and removal improves the predictive model	Psaromiligkos et al., 2011; Romero et al., 2008
Improve Infrastructure Services	Call classification using neural network, support vector machine and finite state automata.	Positive correlation between improved infrastructure and student performance Evaluation metrics like recall, precision and the F-score recommended to assure accuracy	Garfield and Wermter, 2006
	Emotion recognition method using machine learning techniques to classify helpdesk messages.	In-depth data pre-processing needs to be done to avoid issues in classification.	Ramsden, 2010; Povoda et al., 2015

		Sentiment Analysis has high potential for providing further insights.	
Feedback and Survey Analysis	Text mining techniques for understanding free-style question answers.	Focus on pre-processing textual data/corpus significant for valid inferences	Abd-Elrahman et al., 2010; Froelich and Ananyan, 2008
	Social media analytics for understanding student experience.	Sentiment Analysis can be explored to provide further insights.	Chen et al., 2014

In addition to the applications of EDM, Shahiri et al. (2015) reviews most widely used data mining algorithms in this field and the important student attributes contributing to the accuracy of the respective algorithm. The Figure 6 shows the most important attributes as per prior literature study.
The majority of prior studies have emphasized on the student grades which is a concrete indication of academic performance as one of the most important determinant attribute of student behaviour/performance followed by student's demographic and high school background (Angeline et al., 2012; Quadri and Kalyankar, 2010; Osmanbegovic and Suljic, 2012).However some research studies have utilised qualitative attribute such as psychometric factors like student behaviour and family background for predicting student performance (Gray et al., 2014).

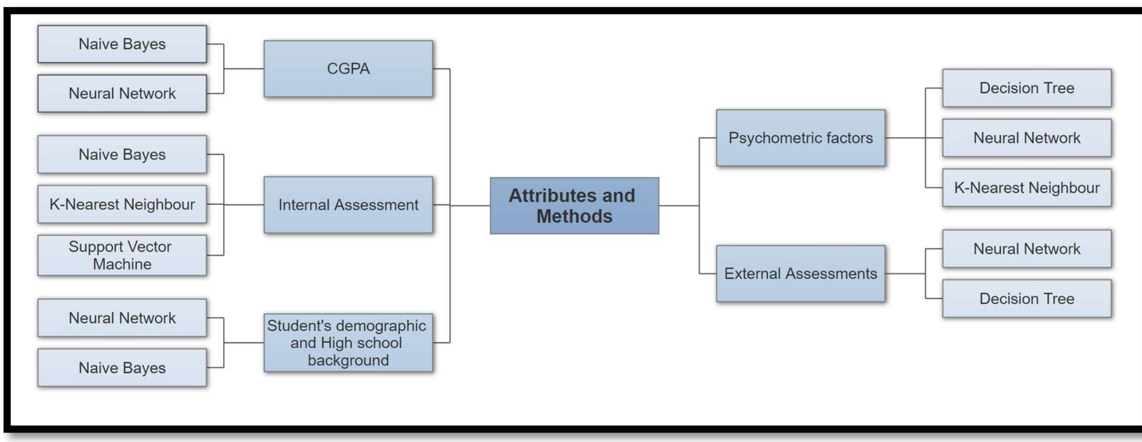

Figure 7: List of common attributes and data mining techniques used in predicting student performance (Adapted from Source: Shahiri et al., 2015)

Finally, Table 4 shows the SWOT Analysis which identifies the Strengths, Weakness, Opportunities and Threats from previous research. The critical issues provided in the four quadrants of the SWOT analysis grid, helps to understand the way strengths can be leveraged for identifying new opportunities and how weaknesses can affect the threats.

Table 4: Overall SWOT Analysis from previous research

SWOT Analysis	
Strengths	**Weaknesses**

State of Art machine learning algorithms available for detail analysis and pattern recognition. Large volume of educational data available for analysis and feature selection. Cutting edge visualisation tools available for result interpretation.	Cross validation needs to be done at multiple level to assure quality for the research study. Must address needs of different stakeholders with insights drawn from the educational data. Limited focus on qualitative analysis. High dependency on human judgement for result interpretation.
Opportunities	**Threats**
Autonomous intelligent system for all important application areas for the EDM/LA. Usage of interpretable results for accurate decision making.	Ethical issue-Privacy and Transparency of the Analysis. Misinterpretation of results leading to erroneous decision making.

4. Conclusion

In conclusion, this project proposal has given an account of methodology and the literature review of the state of art systems to support student enhancement. The evidence from previous reviews suggest that there is immense potential in improving previous studies by incorporating accuracy at multiple levels to assure quality. The previous studies have helped to understand the benefits of expanding dataset in the number of courses, services and surveys and to further explore cutting edge techniques like sentiment analysis and machine learning algorithm that can provide more insights. Another promising line of research for this project would be to develop a consolidated prototype for all-important application areas of EDM/LA that can improve student experience and will be beneficial for all stakeholders.

5. References

Abdous, M. & He, W. (2011), "Using text mining to uncover students' technology-related problems in live video streaming", British Journal of Educational Technology, vol. 42, no. 1, pp. 40-49.

Abd-Elrahman, A., Andreu, M. and Abbott, T. (2010). Using text data mining techniques for understanding free-style question answers in course evaluation forms. Research in Higher Education Journal, 9, p.1.

Angeline, D.M.D. & James, I.S.P. (2012), "Association rule generation using Apriori Mend algorithm for student's placement", International Journal of Emerging Sciences, vol. 2, no. 1, pp. 78

Anon, (2011). Learning Analytics & Knowledge. [online] Available at: https://tekri.athabascau.ca/analytics/ [Accessed 25 May 2018].

Athani, S.S., Kodli, S.A., Banavasi, M.N. & Hiremath, P.G.S. (2017), "Student academic performance and social behavior predictor using data mining techniques", IEEE, pp. 170.

Baker, R.S. and Inventado, P.S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer New York.

Baker, R.S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. JEDM| Journal of Educational Data Mining, 1(1), pp.3-17.

Beck, J.E. and Woolf, B.P. (2000), June. High-level student modeling with machine learning. In International Conference on Intelligent Tutoring Systems (pp. 584-593). Springer, Berlin, Heidelberg.

Bergmeir, C. & Benítez, J.M. (2012), "On the use of cross-validation for time series predictor evaluation", Information Sciences, vol. 191, pp. 192-213.

Boell, S.K. & Cecez-Kecmanovic, D. (2015), "On being ‘systematic’ in literature reviews in IS", Journal of Information Technology, vol. 30, no. 2, pp. 161-173.

Chen, X., Vorvoreanu, M. & Madhavan, K.P.C. (2014), "Mining Social Media Data for Understanding Students' Learning Experiences", IEEE Transactions on Learning Technologies, vol. 7, no. 3, pp. 246-259.

Dawson, S., Heathcote, L. and Poole, G. (2010). Harnessing ICT potential: The adoption and analysis of ICT systems for enhancing the student learning experience. International Journal of Educational Management, 24(2), pp.116-128.

Dawson, S.P., McWilliam, E. and Tan, J.P.L. (2008). Teaching smarter: How mining ICT data can inform and improve learning and teaching practice.

Dekker, G.W., Pechenizkiy, M. and Vleeshouwers, J.M. (2009). Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining.

Dutt, A., Ismail, M.A. & Herawan, T. (2017), "A Systematic Review on Educational Data Mining", IEEE Access, vol. 5, pp. 15991-16005.

Dvir, D., Raz, T. & Shenhar, A.J. (2003), "An empirical analysis of the relationship between project planning and project success", International Journal of Project Management, vol. 21, no. 2, pp. 89-95.

Ec.europa.eu. (2018). [online] Available at: http://ec.europa.eu/dgs/education_culture/repository/education/library/study/2015/dropout-completion-he_en.pdf [Accessed 26 May 2018].

Fayyad, U.M. (1996), Advances in knowledge discovery and data mining, AAAI Press, London;Menlo Park, Calif;Cambridge, Mass;.

Froelich, J. and Ananyan, S. (2008). Decision support via text mining. In Handbook on Decision Support Systems 1 (pp. 609-635). Springer, Berlin, Heidelberg.

Garfield, S. & Wermter, S. (2006), "Call classification using recurrent neural networks, support vector machines and finite state automata", Knowledge and Information Systems, vol. 9, no. 2, pp. 131-156.

Geraldi, J. & Lechter, T. (2012), "Gantt charts revisited", International Journal of Managing Projects in Business, vol. 5, no. 4, pp. 578-594.

Gibbings, P. and Brodie, L. (2006). Skills audit and competency assessment for engineering problem solving courses. In Proceedings of the 2006 International Conference on Innovation, Good Practice and Research in Engineering Education (pp. 266-273). Higher Education Academy Subject Centres for Materials and Engineering.

Gray, G., McGuinness, C. and Owende, P. (2014), February. An application of classification models to predict learner progression in tertiary education. In Advance Computing Conference (IACC), 2014 IEEE International (pp. 549-554). IEEE.

Gurendo, D. (2018). Spiral Model in Software Development Life Cycle (SDLC): Phases, Explanations, Methodology. [online] XB Software. Available at: https://xbsoftware.com/blog/software-development-life-cycle-spiral-model/ [Accessed 30 May 2018].

Helms, M.M. & Nixon, J. (2010), "Exploring SWOT analysis - where are we now?: A review of academic research from the last decade", Journal of Strategy and Management, vol. 3, no. 3, pp. 215.

Hoffait, A. & Schyns, M. (2017), "Early detection of university students with potential difficulties", Decision Support Systems, vol. 101, pp. 1-11.

Hong, K. & Zimmer, R. (2016), "Does Investing in School Capital Infrastructure Improve Student Achievement?", Economics of Education Review, vol. 53, pp. 143-158.

Hussain, M., Zhu, W., Zhang, W., Syed Muhammad Raza Abidi & Ali, S. (2018), "Using machine learning to predict student difficulties from learning session data", The Artificial Intelligence Review, pp. 1-27.

Jie, W., Hai-yan, L., Biao, C. & Yuan, Z. (2017), "Application of educational data mining on analysis of students' online learning behavior", IEEE, , pp. 1011.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T. & Ngo, D.C.L. (2014), "Text mining for market prediction: A systematic review", Expert Systems with Applications, vol. 41, no. 16, pp. 7653-7670.

Kock, M. and Paramythis, A. (2010), November. Towards adaptive learning support on the basis of behavioural patterns in learning activity sequences. In Intelligent Networking and Collaborative Systems (INCOS), 2010 2nd International Conference on (pp. 100-107). IEEE.

KURGAN, L.A. & MUSILEK, P. (2006), "A survey of Knowledge Discovery and Data Mining process models", The Knowledge Engineering Review, vol. 21, no. 1, pp. 1-24.

Lee, M.W., Chen, S.Y., Chrysostomou, K. & Liu, X. (2009), "Mining students' behavior in web-based learning programs", Expert Systems with Applications, vol. 36, no. 2, pp. 3459-3464.

Li, W. & Xu, H. (2014), "Text-based emotion classification using emotion cause extraction", Expert Systems with Applications, vol. 41, no. 4, pp. 1742-1749.

Linan, L.C. and Perez, A.A.J. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. International Journal of Educational Technology in Higher Education, 12(3), pp.98-112.

Ltifi, H., Kolski, C., Ayed, M.B. and Alimi, A.M. (2013). A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases. Journal of Decision Systems, 22(2), pp.69-96.

Maimon, O. and Rokach, L. (2009). Introduction to knowledge discovery and data mining. In Data Mining and Knowledge Discovery Handbook (pp. 1-15). Springer, Boston, MA.

Marczyk, G., DeMatteo, D. and Festinger, D. (2017). Essentials of research design and methodology. John Wiley.

Martínez Abad, F. & Chaparro Caso López, Alicia A. (2016), "Data-mining techniques in detecting factors linked to academic achievement", School Effectiveness and School Improvement, vol. 28, no. 1, pp. 39-17.

Munoz-Organero, M., Munoz-Merino, P.J. & Kloos, C.D. (2010), "Student Behavior and Interaction Patterns with an LMS as Motivation Predictors in E-Learning Settings", IEEE Transactions on Education, vol. 53, no. 3, pp. 463-470.

Muslim, A., Chatti, M., Mahapatra, T. & Schroeder, U. (2016), "A rule-based indicator definition tool for personalized learning analytics", ACM, pp. 264.

Nilsson, A. & Wilson, T.L. (2012), "Reflections on Barry W. Boehm's "A spiral model of software development and enhancement"", International Journal of Managing Projects in Business, vol. 5, no. 4, pp. 737-756.

Nithya, P., Umamaheswari, B. and Umadevi, A. (2016). A survey on educational data mining in field of education. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 5(1).

Oriogun, P.K. (1999), "A Survey of Boehm's Work on the Spiral Models and CO-COMO II—Towards Software Development Process Quality Improvement", Software Quality Journal, vol. 8, no. 1, pp. 53-62.

Osmanbegovic, E. and Suljic, M. (2012). Data mining approach for predicting student performance. Economic Review, 10(1), pp.3-12.

oufakou, A., Gosselin, J. & Guo, D. (2016), "Using data mining to extract knowledge from student evaluation comments in undergraduate courses", IEEE, pp. 3138.

Papamitsiou, Z. & Economides, A.A. (2014), "Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence", Journal of Educational Technology & Society, vol. 17, no. 4, pp. 49-64.

Povoda, L., Arora, A., Singh, S., Burget, R. & Dutta, M.K. (2015), "Emotion recognition from helpdesk messages", IEEE, pp. 310.

Quadri, M.M. and Kalyankar, N.V. (2010). Drop out feature of student data for academic performance using decision tree techniques. Global Journal of Computer Science and Technology.

Ramsden, P. (2010). The future of higher education teaching and the student experience. The Higher Education Academy. Retrieved March, 30, p.2010.

Romero, C., Ventura, S. & García, E. (2008), "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, vol. 51, no. 1, pp. 368-384.

Romero, C. & Ventura, S. (2010), "Educational Data Mining: A Review of the State of the Art", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 6, pp. 601-618.

Salleh, M.A., Bahari, M. & Zakaria, N.H. (2017), "An Overview of Software Functionality Service: A Systematic Literature Review", Procedia Computer Science, vol. 124, pp. 337-344.

Sarra, A., Fontanella, L. & Zio, S.D. (2018), "Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework", Social Indicators Research, , pp. 1-20.

Slack, N., Chambers, S. and Johnston, R. (2010), Operations Management, Prentice-Hall, New York, NY

Siemens, G. & Baker, R. (2012), "Learning analytics and educational data mining: towards communication and collaboration", ACM, pp. 252

Shahiri, A.M., Husain, W. & Rashid, N.A. (2015), "A Review on Predicting Student's Performance Using Data Mining Techniques", Procedia Computer Science, vol. 72, pp. 414-422.

Sokolova, M. & Lapalme, G. (2009), "A systematic analysis of performance measures for classification tasks", Information Processing and Management, vol. 45, no. 4, pp. 427-437.

Stige, B., Malterud, K. & Midtgarden, T. (2009), "Toward an Agenda for Evaluation of Qualitative Research", Qualitative Health Research, vol. 19, no. 10, pp. 1504-1516.

Stukalina, Y. (2016), "Modelling student satisfaction and motivation in the integrated educational environment: An empirical study", International Journal of Educational Management, vol. 30, no. 6, pp. 1072-1087.

Talavera, L. and Gaudioso, E. (2004), February. Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence (pp. 17-23).

Vahdat, M., Oneto, L., Anguita, D., Funk, M. & Rauterberg, M. (2016), "Can machine learning explain human learning?", Neurocomputing, vol. 192, pp. 14-28.

Vandamme, J.P., Meskens, N. and Superby, J.F. (2007). Predicting academic performance by data mining methods. Education Economics, 15(4), pp.405-419.

Wiers-Jenssen, J., Stensaker, B.R. and Grgaard, J.B. (2002). Student satisfaction: Towards an empirical deconstruction of the concept. Quality in higher education, 8(2), pp.183-195.

Wilson, J.M. (2003), "Gantt charts: A centenary appreciation", European Journal of Operational Research, vol. 149, no. 2, pp. 430-437.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Zafra, A., Romero, C. & Ventura, S. (2011), "Multiple instance learning for classifying students in learning management systems", Expert Systems with Applications, vol. 38, no. 12, pp. 15020-15031.

Zyt, J., Klosgen, W. and Zytkow, J. (2002). Handbook of data mining and knowledge discovery. Oxford university press.

6. Appendices

7. Glossary

ANN	Artificial Neural Network
BPR	Bayesian Profile Regression
CGPA	Cumulative Grade Points Average
CPM	Critical Path Method
DM	Data Mining
EDM	Educational Data Mining
GDPR	General Data Protection Regulation
HCD	Human Centric Design
ISO	International Organisation for Standardisation
KDD	Knowledge Discovery in Databases
LA	Learning Analytics
NDA	Non-Disclosure Agreement
NSS	National Student Survey
PERT	Program Evaluation and Review Technique
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
SDLC	Software Development Life Cycle
CLR	Critical Literature Review
SVM	Support Vector Machine
SWOT	Strength, Weakness, Opportunities and Threats
TEI	Teacher Evaluation Index
VLE	Virtual Learning Environment