CS 773/873: Data Mining and Security Summer 2020

Course project Points: 100

Due: August 1, 2020

Team Members:

Pavan Galagali – UIN: 01125293

Divya Uppala – UIN: 01155746

Analyzing at-risk students in the Open University Learning Analytics Dataset (OULAD)

Executive Summary:

The clickstream information of students which is captured can be exploited in obtaining their performance in online education platforms whereas in higher education, the academic performance of students is linked with organizing optimal educational policies that largely impact economic and financial development. In the current study, the given problem of student's performance prediction is analyzed by using different algorithms for the Open University Learning Analytics dataset. We demonstrated that the clickstream data generated due to the online learning platforms which helped for students' interaction can be evaluated at a granular level to improve the early prediction of at-risk students. Our model can predict pass/fail class with around 90% accuracy for student interaction in a virtual learning environment

(VLE). Of the given data, 60% of data was taken as training dataset and 30% of data was taken as test dataset and rest of the 10% is ignored as it has incorrect values.

Let us consider the five algorithms which we took for accurate results. The **Bayesian classifier** has a precision of 99.5% and recall of 99.5% with more accurate results when compared with other algorithms like **Simple CART** has precision of 95.3% and recall of 95.3%, **Decision tree** has precision and recall of 78.6% and **Logistic regression** with precision of 61.2% and recall of 61.3%. We have also considered **K-Means clustering** where we got clusters with 53% at-risk, remaining not-at-risk. We have also tried **LSTM** and **SVM** with 60 epochs for which we the models took a lot of time to train providing fair results. We are considering the best classification algorithms to proceed with our analysis. Let us go through about each algorithm in detail further below and discuss its relevancy.

A contribution of our research is an informed approach to advanced higher education decision-making towards sustainable education. It is a bold effort for student-centric policies, promoting the trust and the loyalty of students in courses and programs.

Introduction:

The available large amount of educational data provides opportunities to utilize it for various purposes, such as tapping the learning behaviors of the stakeholders involved, improving these behaviors by addressing the issues, and optimizing the learning environment. With the readily available educational data that is accessible, several research communities have exhibited noticeable interest in predicting students' patterns and extracting meaningful insights from these patterns. Such information extraction is not only bound to the data mining community. However, new communities have also emerged that focus not only on the goal to improve students’ performance but on optimizing the learning environment, referred to as learning analytics. Educational data, accumulated due to the interactional activity between learners and instructors, has been concluded as a multidisciplinary field of study, involving researchers from various research communities, which has yielded to the inclusion of numerous terms associated with the exploration of the educational data, such as academic analytics, predictive analytics, and learning analytics.

Various data analytic techniques and machine learning practices are administered for the prediction of several measures and events, with deep artificial neural networks (ANNs) being a prominent practice due to their learning abilities. The paradigm of deep learning is defined as hierarchical representational learning, encompassing various layers of computation and enabling the system to learn from prevailing examples, intervening in the traditional feature engineering methods.

To intervene early with the students for optimal performance, various forms of classification algorithms were implemented to predict the probability of the student being at risk. Classification can be performed on structured or unstructured data. Classification is a technique where we categorize data into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under. These practices correspond to the course duration as sequential data by analyzing each learner’s daily, weekly, or monthly performance. Overall, in this discipline, the research community is embarking towards the adoption of these sequential practices to predict early on the students at risk of poor performance and intervene with them on time for optimal results.

Problem statement:

Open University Learning Analytics Dataset contains data about courses, students and their interactions with the Virtual Learning Environment for seven selected courses. Our aim is to deliver an understanding of the behavior of students at risk of failure, contributing to the decision-making policies using classification, clustering and regression algorithms, to devise suitable interventions to help students succeed.

Solution methodology:

The model that we used is a classification model. The Bayesian classification was the most useful among all the models experimented. The prediction capability of the model is used in further analysis for this study. The **comparison metrics** in this study is based on **precision, recall and F-measure**. Higher values indicate better accuracy.

The steps taken to build a model are: (Solution Steps)

1) Preprocessing
2) Cleaning and structuring the data.
3) Train the model with datasets obtained.
4) Use the model to find patterns in the data.
5) Provide suitable intervention techniques on the pattern to avoid students being at-risk.

Classification differs from regression in that the former predicts categorical (discrete, unordered) labels while the latter predicts missing or unavailable, and often numerical, data values. This pair of tasks is similar in that they both are tools for prediction. Challenges to data mining regarding data mining methodology and user interaction issues include the following: mining different kinds of knowledge in databases, interactive mining of knowledge at multiple levels of abstraction, incorporation of background knowledge, data mining query languages and ad hoc data mining, presentation and visualization of data mining results, handling noisy or incomplete data, and pattern evaluation.

The final step in understanding the patterns is visualizing the data.

This primitive refers to the form in which discovered patterns are to be displayed. In order for data mining to be effective in conveying knowledge to users, data mining systems should be able to display the discovered patterns in multiple forms such as rules, tables, cross tabs (cross-tabulations), pie or bar charts, decision trees, cubes or other visual representations. Our most suited method was using the bar-chart. The Illustration is explained with the patterns.

Experimental setup and data used:

The procured Open University Learning Analytics dataset (OULAD) constituted student demographics, clickstream history, and assessment submission information of students over a course, from 2013 to 2014. The data were composed of several (7) courses, with each course being taught at different intervals in a year. Four distinct performance classes were defined: distinction, pass, fail, and withdrawal. Our assumption for students being at risk are “Fail” and “Withdrawal”.

The OULAD comprised students’ information regarding their interaction with the VLE—their assessments, quizzes, and course performances. The interaction with the VLE was further categorized into 20 different activity types with each activity referring to a specific action, such as downloading or viewing lectures, course content, or quizzes. The names of each of these activity types are as follows: data plus, dual pane, external quiz, folder, forum-ng, glossary, homepage, HTML activity, oucollaborate, oucontent, ouelluminate, ouwiki, page, questionnaire, quiz, repeat activity, resource, shared subpage, subpage, and URL.

The aggregated average clicks per student were processed to visualize the students' interactions and performance. The number of attempts taken by the students to pass the exam also plays a crucial role in determining the students probability of being successful.

Data Preprocessing:

The OULAD was organized in a raw structured format with numerous data files. The log-file data were computed to obtain features implying to the various actions indicating students' interactions with the VLE. These features were formulated by processing the provided data tables in the database. From the given datasets we have combined "**student info", "student VLE", "VLE "csv data** into one dataset and rest of the files as is for processing/mining using different algorithms. We have cleaned the data with '**0's** wherever we have '**?'s** or unknown values for the variables in above csv files. The data was computed in a weekly fashion, with each week constituting the same activity features, and each week comprising a homogenous set of students, that is, students in the week 'I' were also present in the week (i-1) and so on. The result overview is described as an overlook at the data comprising of all weeks combined. Each student was identified with a unique ID in the data. Only to small ratio in the data took more than one course and hence were repeated; however, the highlighting part of this study is that we also intend to find academic performances of a student in multiple courses over multiple attempts. Therefore, each student was identified by a unique ID for each course. The unique student IDs were computed by a combination of their older IDs, the study was intended to analyze pass or fail instances with at-risk parameters.

Approach:

We have used different algorithms for mining the given data to analyze students at-risk. Outcomes of each of the training models were later used to perform analysis on the data provided. The models are listed in the decreasing order of obtained accuracy. Relevancy of each of these models are as described before for "classification model" selection.

1. Bayesian network:

Bayesian networks (BNs) are an increasingly popular method of modelling uncertain and complex domains such as ecosystems and environmental management. At best, they provide a robust and mathematically coherent framework for the analysis of this kind of problems.

Bayesian methods interpolate to this extreme because the Bayesian prior can be a delta function on one model of the world. Bayesian and near-Bayesian methods have an associated language for specifying priors and posteriors. Computing a posterior may be extremely difficult. This difficulty implies that computational approximation is required. The "think harder" part of the Bayesian research program is a "Bayesian employment" act.

The trained dataset is split into **60% and test data is 30%.** We got **18** attributes in the output. The remaining **10%** is ignored as it has misclassified data.

From the output, we observed the **average precision of 99.5%** which comprises of both students **at-risk** and students **not-at-risk**. We have got more accurate results using this classifier out of the others used.

2. Simple CART:

The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.

CART is nonparametric and therefore does not rely on data belonging to a distribution. CART is not significantly impacted by outliers in the input variables. Your belief on the builders to do all the things they promise to make sure the perfect expertise. It could be barely tough to make use of it at first and you might need to spend a while getting used to the system.

The trained dataset is split into **60% and test data is 30%.** The remaining **10%** is ignored as it has misclassified data.

From the output, we observed the **average precision of 95.3%** which comprises of both students **at-risk** and students **not-at-risk**. We have got the next most accurate results using this classification algorithm of the other used algorithms.

3. Decision Tree:

Decision trees are tree-like diagrams that attempt to display the range of possible outcomes and subsequent decisions made after an initial decision. It helps us to weigh the likely consequences of one decision against another. It’s also used to estimate expected payoffs of decisions.

The drawback is outcomes of decisions, payoffs and subsequent decisions may be based primarily on expectations.

Here also the **trained dataset** is split into **60% and test data is 30%.** From the output, we observed the **average precision of 78.6%** which comprises of both students **at-risk** and students **not-at-risk**. We have got the next most accurate results using this classification algorithm of the other used algorithms.

4. Logistic Regression:

Logistic Regression is a classification algorithm that uses a linear and additive combination of the predictor variables to predict the binary output of the response variable. Logistic Regression is one of the simplest predictive algorithms. It is transparent, means we can see the process and understand what is going on at each step, contrasted to the more complex ones.

Even with a small number of big-influential, the model can be damaged sharply. It is essential to pre-process the data carefully before giving the data to the Logistic model. It also requires more data for processing.

We consider the **trained dataset** into **60% and test data** into **30%.** From the output, we observed the **average precision of 78.6%** which comprises of both students **at-risk** and students **not-at-risk**. We have got the next most accurate results using this classification algorithm of the other used algorithms.

5. K-Means clustering:

K-means clustering is used to simplify large datasets into smaller and simple datasets. Distinct patterns are evaluated, and similar data sets are grouped together. The variable K represents the number of groups in the data. It’s Simple, flexible and best suitable to use when we have large datasets. It’s computed much faster than the smaller dataset. It can also produce higher clusters.

K-means does not allow development of an optimal set of clusters and for effective results, you should decide on the clusters before. When we deal with a large dataset, using a dendrogram technique will crash the computer du0e to a lot of computational load and Ram limits.

From the given datasets, **trained dataset** is categorized into **60% and test data** into **30%.** The remaining **10%** is ignored as it has incorrect data. From the output, we got **3** iterations and observed two clusters **cluster-0** with group percentage of 47% and **cluster-1** with group percentage of 53% which are **not-at-risk** and **at-risk** respectively.

LSTM and SVM took longer time to train providing an accuracy of around **60%.** Hence, we moved forward with our analysis with classification algorithms. Although these algorithms have potential in future analysis of the OULAD.

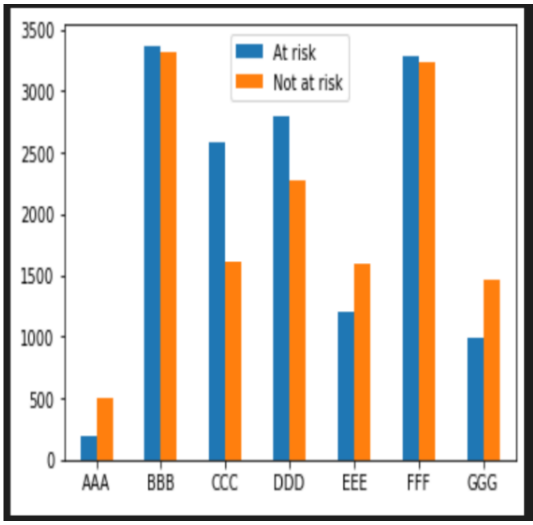
Experimentation and Evaluation:

Evaluation with Baseline:

To evaluate the deployed different datasets the early prediction of at-risk students, several machine learning algorithms were deployed as baselines. Logistic Regression, Bayesian Network, K-Means clustering, Decision tree, simple cart, ANN, Support Vector Machine (SVM), and Logistic Regression (LR) have been frequently adopted in the educational research community for evaluating other proposed models.

Bayesian Network works more efficiently for the given dataset, and then **simple cart, decision tree, Logistic regression and K-Means** is the **decreasing** order of accurate results based on taken based on **value of average precision** we got by processing the data using **WEKA.** Hence behaviors of students are analyzed in a weekly manner and produced optimal results compared to baseline techniques such as pandas in python. The model with a good fit captured the learning behavior of students and efficiently predicted the at-risk students on the basis of their interactions and engagement patterns.

Let us take the consideration of data with number of attempts. Here we had taken 7 attempts to see the number of students at-risk and not-at-risk.



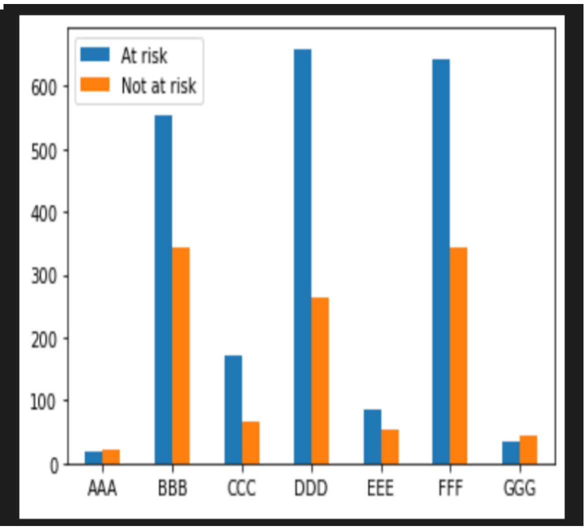
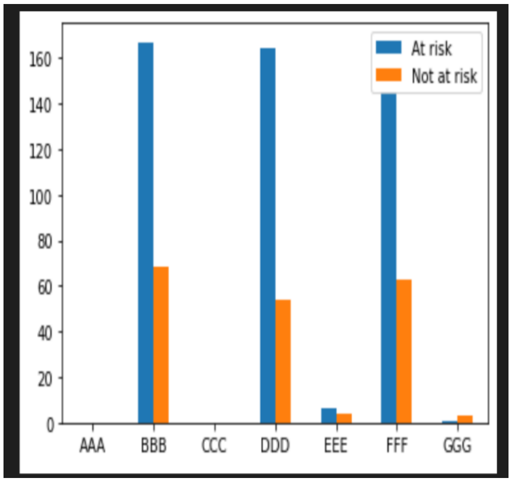


Figure 1: Attempt 1

Figure 2: Attempt 2



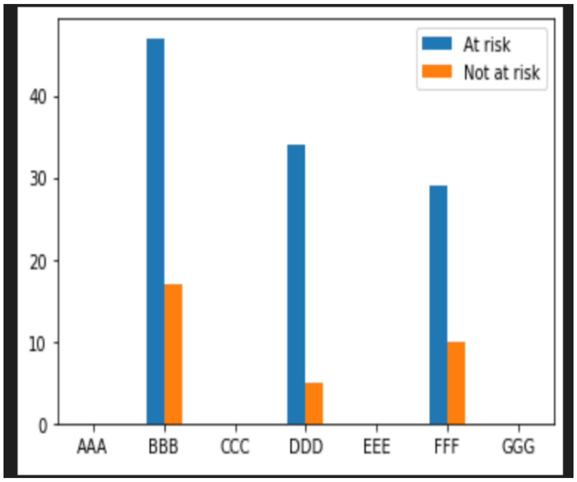


Figure 3: Attempt 3

Figure 4: Attempt 4

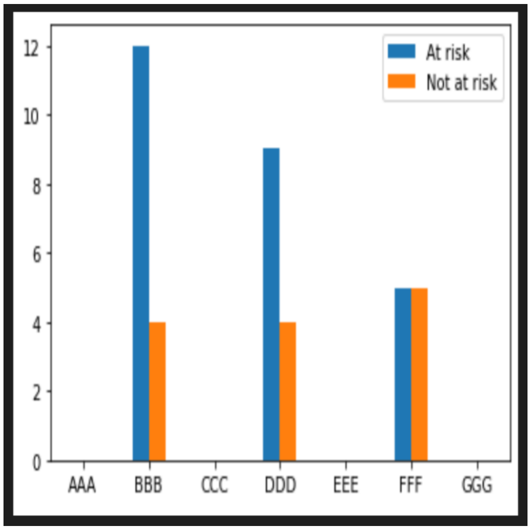


Figure 5: Attempt 5

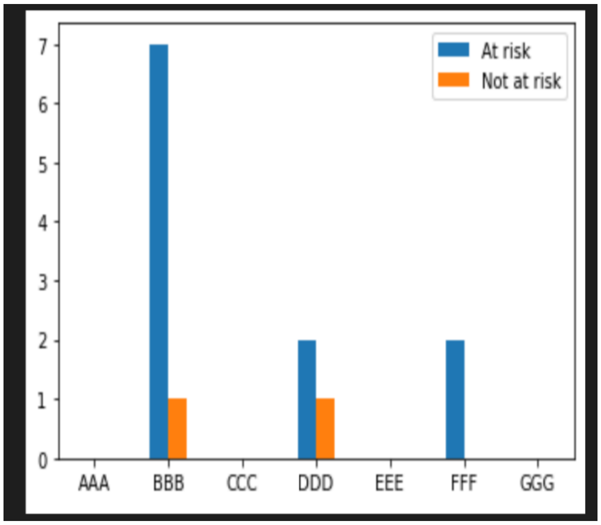


Figure 6: Attempt 6

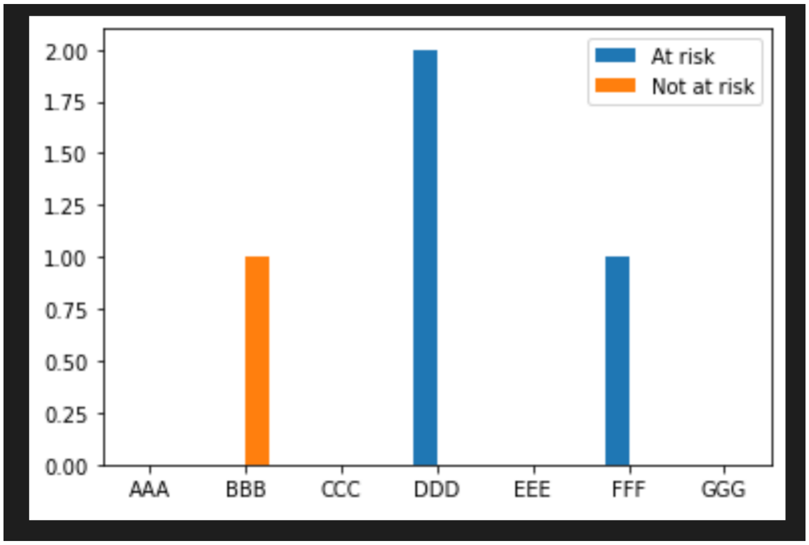


Figure 7: Attempt 7

From the above 7 attempts we analyzed the graphs as follows. We had taken "courses" on X-axis and number of students on "Y-axis". At-risk is represented in blue bar and not-at-risk is represented in orange bar. Let us consider and analyze each course and student’s performance. Here are the observations.

The common observation from the data is, there is a gradual decrease in the number of attempts made by the students in every course. Initial assumption is that the students tend to pass the subject with minimum attempts. Analyzing the data reveals good patterns.

Course wise performance evaluation:

Course AAA: The probability level of at-risk is less when compared with not-at-risk for this course and found no attempts after second attempt. This course is simple and easier and level of risk is less. The probability that a student will pass in this exam is higher.

Course BBB: Majority of the students are at-risk in every attempt. In the 6th and 7th attempt, none of the students reconsidered this course. So, we can understand the level of difficulty of the course and its complexity. But the course might be very important or mandatory as attempts count is more. So, we can consider the students at-risk as more when choosing course BBB as the number of attempts gets increased.

Course CCC: Course has probability of at-risk more than two attempts and none of the students took after that. We analyze from the observation like the course is a bit difficult and attempts were stopped.

Course DDD: This course, the level of complexity is more than Course BBB. In all the attempts the level of at-risk is more and none of the students have passed in last attempt. Majority of students taking attempts got reduced but the at-risk is not reduced.

Course EEE: Of all the students only, few have taken this course from attempt 1 to attempt 7. In first attempt a smaller number of students were at-risk and in second attempt the number of students at-risk was very low but more than not-at-risk and from the attempt 3 none of them took this course. More number of student pass the exam.

Course FFF: This is also one of the complex and major course where probability of at-risk is large in all the attempts and not-at-risk percentage decreased to 0 in the attempt number 7. The course is very difficult and major subject to be taken by the students.

Course GGG: The probability level of at-risk is very low. Initial three attempts were the last considered by the students and none of them have taken after that. It probably not that difficult and majority of the students were able to pass this course.

Intervention suggestions:

1) The student should interact more with VLE.
2) Week wise interactions should be improved.
3) Analyze the difficulty of the course before taking it.
4) More learning is needed with subsequent interactions.

To summarize, the **courses BBB, DDD, FFF** were attempted by students many number of times and their probability of at-risk is more when compared to rest of the three courses **AAA, CCC, GGG** and the number of students taken that course is more when compared to other courses in initial attempts.

Let’s analyze the number of clicks done by students for courses.

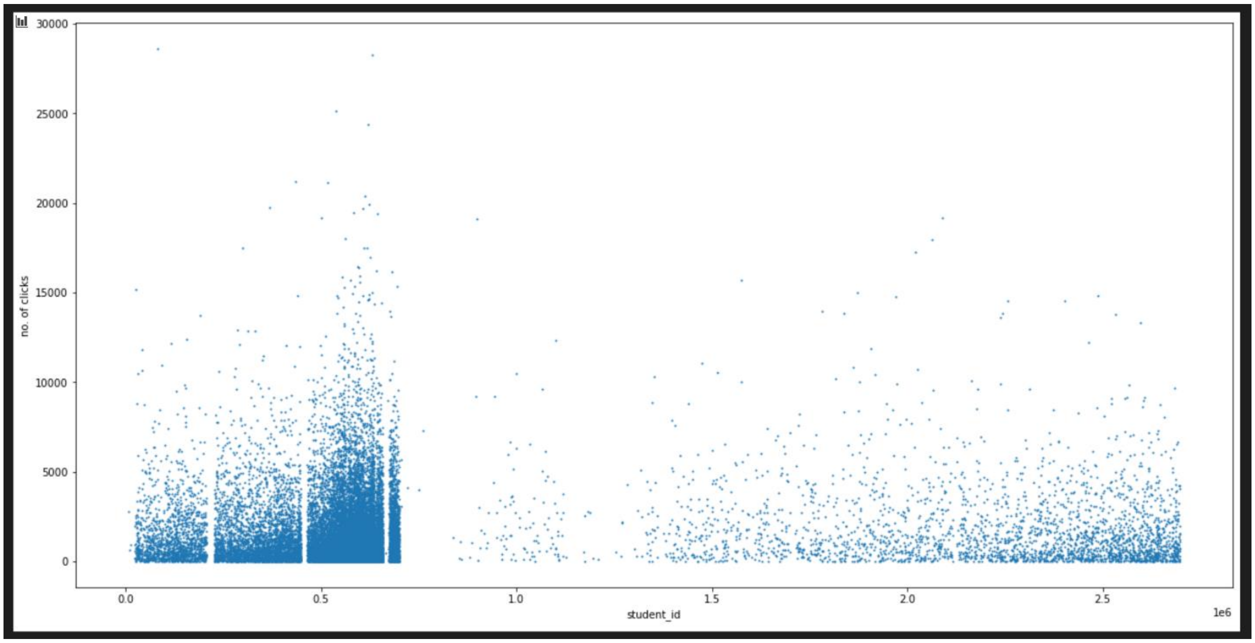


Figure 8: Student interaction with VLE in terms number of clicks

Here we had taken **scatter-plot graph** for the students with number of clicks on y-axis and student_id’s on x-axis with dots presented in blue representing number of clicks made over the student’s period of time with the university.

From the graph we can predict if the number of clicks made by a student is more then the probability of his visits to VLE might be more and his at-risk level is low and if click count is less then probability of his visit to VLE might be low where his at-risk might be more. If we look deeper in the pattern, the student interaction with the system is more than 60%. We can also predict that if the number of clicks made by student on VLE visit is more then, his probability of number of attempts to the courses might be more from previous attempts graphs description which could also mean the student has interacted with the system efficiently. The average number of clicks made by each student is approximately more than 10000. Although fewer students attempted no click which either means they never visited VLE or they have withdrawn from the course which makes their at-risk probability above 95%.

Intervention suggestions:

1) Any section in the VLE should not be overlooked.
2) Improve interaction with VLE.

Let us analyze the below graph for all the classification algorithms we used for prediction. (Comparison)

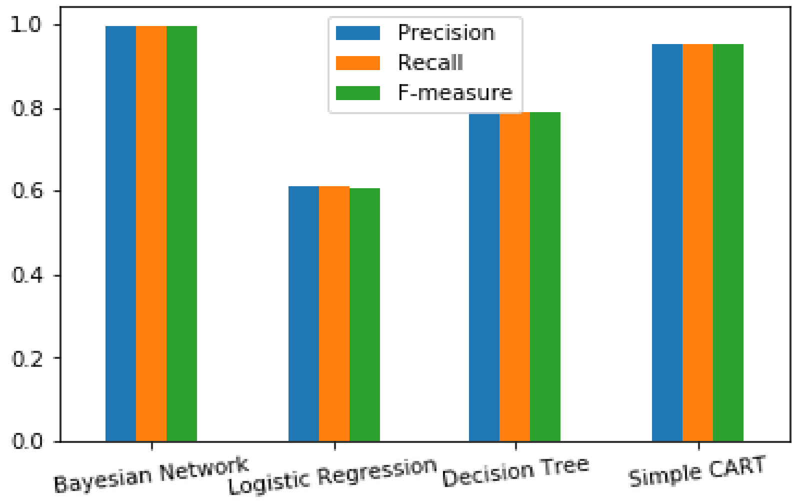


Figure 9: Model performance comparison

From the above graph, we have taken classification algorithms on x-axis and accuracy values from 0 to 1 (indicating 0-100%) on y-axis. Precision is represented in blue bar, Recall in orange bar and f-measure in green bar.

When we analyze the graph almost all the three (**precision, recall, f-measure**) values are approximately same or equivalent for all the classifications. The **Bayesian Network** has more accurate results of all the three values with 0.995 when compared to other classifications. The Simple CART goes second with a value of 0.953 then the decision tree goes third accurate with a value of 0.786 and the least accurate result is for Logistic Regression with a value of 0.612, 0.613, 0.608 respectively for precision,recall and f-measure.

Comparison with models used:

Models	Precision	Recall	F-measure	Goodness
Bayesian Network	0.995	0.995	0.995	They can easily and in a mathematically coherent manner, incorporate knowledge of different accuracies and from different sources.
CART	0.953	0.953	0.953	CART is nonparametric and therefore does not rely on data belonging to a particular type of distribution. Implicitly perform

				feature selection
Logistic Regression	0.612	0.613	0.608	Better than a Decision tree when data is distributed such that it can be linearly classified with lower variance.
Decision Tree	0.786	0.786	0.786	It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis.

Table 1: Comparison of efficiency with different models used.

From these analysis, we can predict or probably say more accuracy rate is for bayesian network and it might be the best suitable algorithm for our trained and test datasets.

Results:

This research presented the critical concern of the identification of students at risk of failure. Different computation methods were deployed to predict the students at risk of a failure based on their behavior and engagement patterns with the VLE. **Bayesian Network** improved the predictability of the students' decisions and assisted the educational community to develop guidelines for helping the at-risk students. Such classification models design a path for the administrative authorities to contribute to formulating policies and strategies to implement timely interventions, regulate the decision-making process, and ultimately assist students through the provision of support systems. Next **simple cart, decision tree,** are the **two classification algorithms, Logistic regression (Regression) and K-Means clustering algorithm** follows in descending order. We have compared average precision value of algorithms to order their accuracy. Moreover, such settings will also help establish guidance committees and regular student counseling sessions for maintaining a motivational infrastructure and tapping the behavior of students for data-driven decision-making processes.

Let us describe more in detail by using below graph.

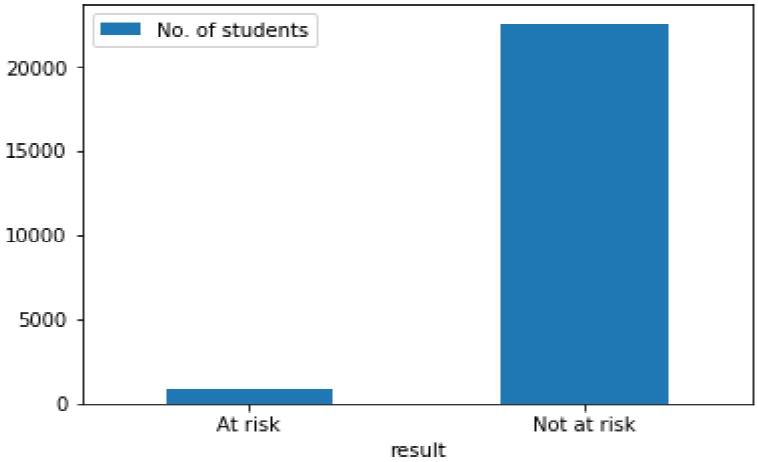


Figure 10: Assessment wise comparison

The graph represents number of students on y-axis and result at-risk and not-at-risk on x-axis. This graph is generated using the assessment result which was given to students. When we observe the graph, the prediction or probability of failing in the initial assessment is low and pass count is more in number. The probability of at-risk in future assessment is more if at-risk is more in present assessment and the probability of not-at-risk is more in future assessment if it is more now. In this scenario, Bayesian classification provides accurate results as to if the student can "Pass" in the next assessement based on the previous attempts.

Conclusions:

The study examined students at-risk by converting problem into course wise performance data format and measured the effectiveness of classification models. The **Bayesian Network** tended to observe the sequential week-wise pattern of students and their activities performing better compared to other classifiers dealing with students' interactions in collective and aggregated way. Such early predictions give an idea to timely advise the students at-risk with counseling and alert mails.

However, A course-level analysis helped to capture the behavioral differences of the same students in different courses and identify the elements to tap their academic performance. We plan further analysis to enrich our model prediction by including students' assessment scores, analyzing the association between their assessment submission pattern and performance. A framework catering to the prominent attributes—extrinsic and intrinsic—associated with students' performance may enable the learning analytics community to adhere towards more effective decision-making systems.