ANALYSIS OF STUDENTS' LEARNING OUTCOMES USING EDUCATIONAL DATA MINING APPROACH

[1]JEFFERSON L. LERIOS, [2]JASMIN D. NIGUIDULA

[1]Laguna State Polytechnic University, Laguna, Philippines
[2]Technological Institute of the Philippines, Manila, Philippines
E-mail: [1]jeffersonlerios@yahoo.com, [2]jasniguidula@yahoo.com

Abstract - In the academic institution, analysis of students' academic performance is important as an input in improving the quality of students' education. This can be done by analyzing students' academic records stored in the database using educational data mining. This study aims to explore factors that affect students' performance in their professional courses with prerequisites in a specific program. The prediction of students' performance was made using classification techniques in data mining. The study was able to identify performing and at-risk students in terms of their academic performance based on the pattern extracted from the dataset. This research will serve as guidelines for the administrators and faculty members for the improvement of their teaching methodologies.

Keywords - Educational Data Mining, Classification, Performance, Prediction

I. INTRODUCTION

The main goal of an academic institution is to ensure the delivery of instruction to students is achieved. It is where every student learns the fundamental values and knowledge that define the quality of good education. Determining the success factor of students in their studies will depend on their academic performance. Therefore, an in depth understanding and analysis of students performance becomes a very good area of research. It will help the administrators and faculty members to validate the effectivity and efficiency of the different teaching strategies that are being implemented in a particular course. Normally, these are supported by students records that are stored in large repositories and became voluminous over time confined into different storage devices and can be analyzed through data mining.

Data mining (DM) is an interesting method in the process of knowledge discovery that can be obtained through large scale data from databases and data warehouse of an organization. As these data are kept, different techniques are created for research, meta-analysis and discovering useful information by understanding data patterns, exploring unique behavior of a particular variable and identifying the relationship between these variables. DM contributes to the knowledge building of providing decision makers an alternative solution by means of extracting data pattern for analysis. The analytical components of DM includes classification, clustering, association among others that is designed for statistical, visualization and machine learning as part of information extraction for better understanding [3][4][5].

In DM, exploration of educational data is called Educational Data Mining (EDM). EDM is a technique that could be used to analyze students' records through pattern extraction, application of an algorithm and graphical illustration or visualization. Application of EDM is significant in unveiling discovery of new information that lies on the data set stored in large databases, an interesting field of research to help improve better decision making. EDM has been cited in several research studies to be useful in predicting future students' performance in their learning environment because of its ability to identity at-risk students. Students profile and grades are the common prognostic parameters used to measure students' academic performance in a particular courses or subjects. However, these studies were conducted in general using students profile, academic and course information[4].

In the collegiate level, the curriculum is designed for specific program with general education courses, core courses that are common to all related programs, professional courses that are required for each program and elective courses. In this study, students' academic performance from the professional courses from the Bachelor of Science in Information Technology (BSIT) in a select university will be given emphasis. It will focus more on the courses with prerequisites. The objectives are to a) determine students' performance significant relationship for the courses with prerequisites b) knowledge discovery of the success factors students for learning behaviors and c) production of the best model prediction for student academic performance. Using EDM predictions of students' performance in their professional subjects are significant by understanding and analyzing student educational data that could improve students' performance.

The following sections present the related works in section 2, methodology in section 3, results and discussion in section 4 and section 5 for the conclusion.

II. RELATED WORKS

Application of data mining in analyzing academic data is significant to discover hidden knowledge and pattern extraction. Several studies were conducted to explore the theoretical factors that may affect students' performance. Amjad Abu Saa research study used multiple performance indicators to find a qualitative model which best classifies and predicts the students' performance based on related personal and social factors. Multiple classification technique was used in the study for predicting the students' performance. These were C4.5 decision tree, ID3 decision tree, CART decision Tree and CHAID. The results of the study slightly found out that students' performance was not totally dependent on their academic efforts and there were many other factors that have equal to greater influences as well [3].
A framework in producing a model prediction using selected classification techniques was proposed by Fadhilah Ahmad et al. in their research. It consists of three main stages such as Data Collection and Integration, Data Transformation and Patterns Extraction. The study conducted comparative analysis of three classification techniques such as Decision Tree, Naïve Bayes and Rule Based algorithm. The investigation found out that the classification algorithms can achieve the highest prediction accuracy. The prediction was also successful for students who perform good, average and in poor categories [4].

A similar research on assessing students' performance was conducted by Abeer and Elaraby that is focused on generating classification rules and predicting performance based on recorded students behavior and activities. The result was able to predict those students who needed special attention so that appropriate action could be taken to improve their academic performance [6].

While other research study concentrated on the comparative analysis to test multiple decision tree algorithms using an educational dataset. Yadav, Bhardwaj and Pal used three decision tree classification algorithms and found out that CART method worked better based on produced accuracy and precision using 10-fold cross validations[5].

III. METHODOLOGY

A framework for knowledge discovery in database (KDD) was utilized in the study as illustrated by Dr. Alexander Rieger in Figure 1.

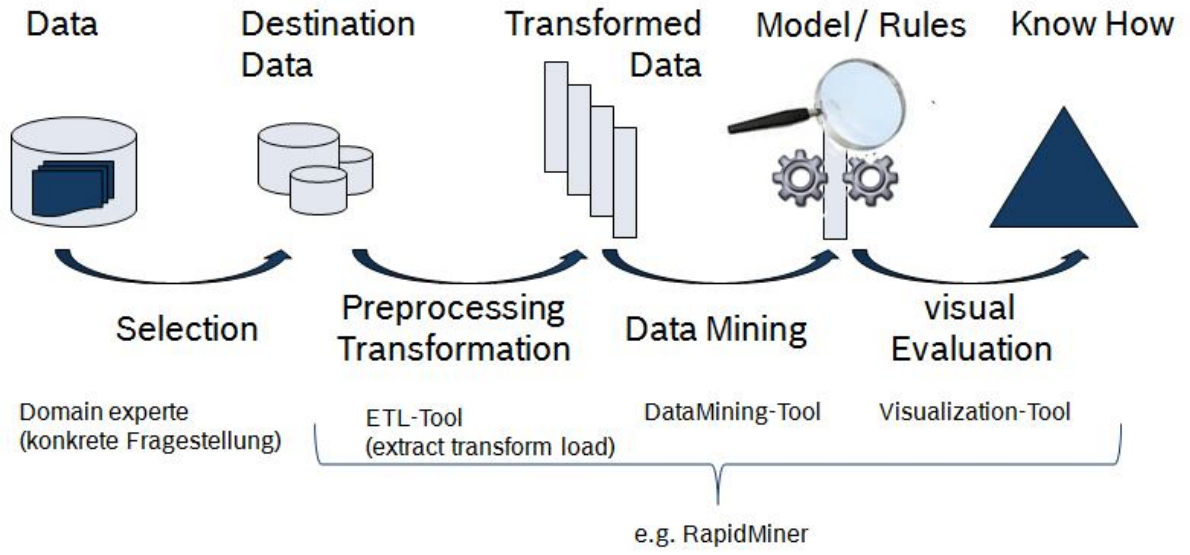


Figure 1 Knowledge Discovery in Database Process

The dataset is extracted from large quantities of data as starting point. Selection of attributes is done for the analysis that is mandatory in all data mining process. A suitable model will be used for preprocessing and transformation of data into structure data for analysis and interpretation [1].

3.1 Data Preparation

The data was collected from a State University in Laguna. It consists of 6,210 students' records from 2013 to 2017 the dataset for the study was extracted from the back up database of the University. It was done using Ubuntu operating system and Xampp as platform for managing the database. Only tables that were directly relative to dataset were used and imported in Comma Separated Value (CSV) format. The attributes of the dataset were described in Table 1 for clarity and easy understanding.

Attributes	Description	Possible Values
Barangay	Name of Barangay where the student live	Mayondon, Malinta, Timugan etc.
Town/City	Name of Town or City	Los Banos, Calamba, Bay etc.
Province	Name of Province	Laguna, Cavite, Batangas etc.
Gender	The gender of the students	Male, Female
Professional Courses	Professional Courses are the required courses of the program	ITEP 101, ITEP 202, ITEP 203, ITEP 204, ITEP 205, ITEP 206, ITEP 207, ITEP 208, ITEP 308, ITEP 309, ITEP 310, ITEP 311, ITEP 312, ITEP 313, ITEP 314, ITEP 315, ITEP 416
Grades	Final grade of the students at the end of the semester	Excellent, Very Satisfactory, Satisfactory, Fairly Satisfactory, Passed, Incomplete, Conditional and Failed

Table 1 Attributes, Description, and Possible Values of the dataset

The attributes presented in Table 1 were the complete address of the students that was important in analyzing the spatial distribution of students' population. The gender, professional courses, and grades were used for the prediction of students' performance.

Course Code	Description	Pre-requisites
ITEP 101	Networking 1	ITEC 101
ITEP 202	Data Structure	ITEC 103
ITEP 203	Networking 2	ITEP 101
ITEP 204	Database Management System 1	ITEP 202
ITEP 205	PC Troubleshooting and Maintenance	ITEC 206
ITEP 206	Operating System Applications	ITEC 101
ITEP 207	Accounting Principles	Math 211
ITEP 208	Service Culture	ENGL 3
ITEP 309	Object Oriented Programming	ITEP 204
ITEP 310	Systems Analysis and Design	ITEP 204
ITEP 311	Database Management System 2	ITEP 204
ITEP 312	Web Development	ITEP 204
ITEP 313	Methods of Research for IT	ENGL 311
ITEP 314	Software Engineering	ITEP 202
ITEP 315	Multimedia Systems	ITEP 310
ITEP 416	Capstone Project	4^{TH} year status

Table 2 BSIT Professional Courses

Professional courses are subjects that belong to a specific program (e.g. BSIT, BSHRM). A different program requires professional courses depending on its design curriculum in accordance with the Commission on Higher Education (CHED) Memorandum. In this case, Professional courses from the Bachelor of Science in Information Technology (BSIT) from a selected University in Laguna were utilized as described in Table 2.

Attribute	Mode	Least
Barangay	Mayondon [450]	Masaya [21]
Town	Los Banos [1638]	San Pablo [2]
Gender	Male [4634]	Female [2172]
Professional Courses	ITEP 101 [739]	ITEP 416 [121]
Grades	Satisfactory [1978]	Dropped [12]

Table 3 Dataset Summary of Statistics

Table 3 shows the dataset summary of statistics that present the mode (the value with the high frequency) and the least (the value with the least frequency) as illustrated in detail in Figure 2 and Figure 3.

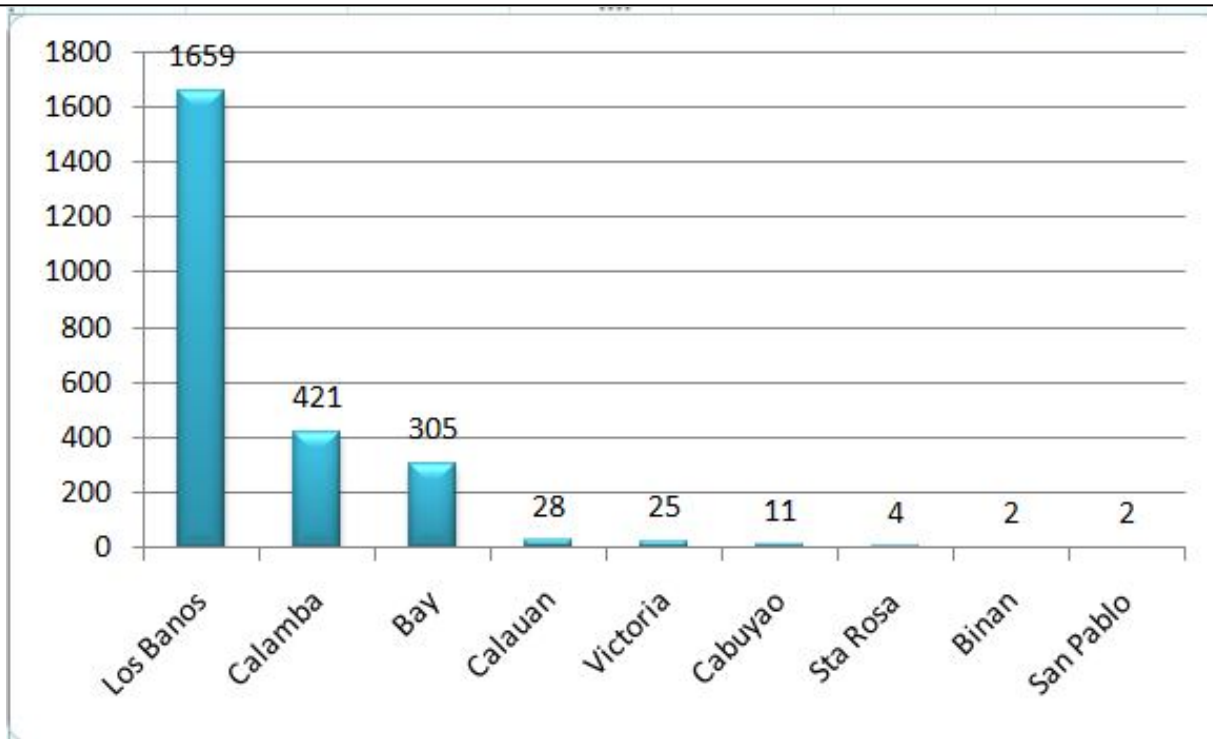


Figure 2 Distribution of Students Population by City/Town

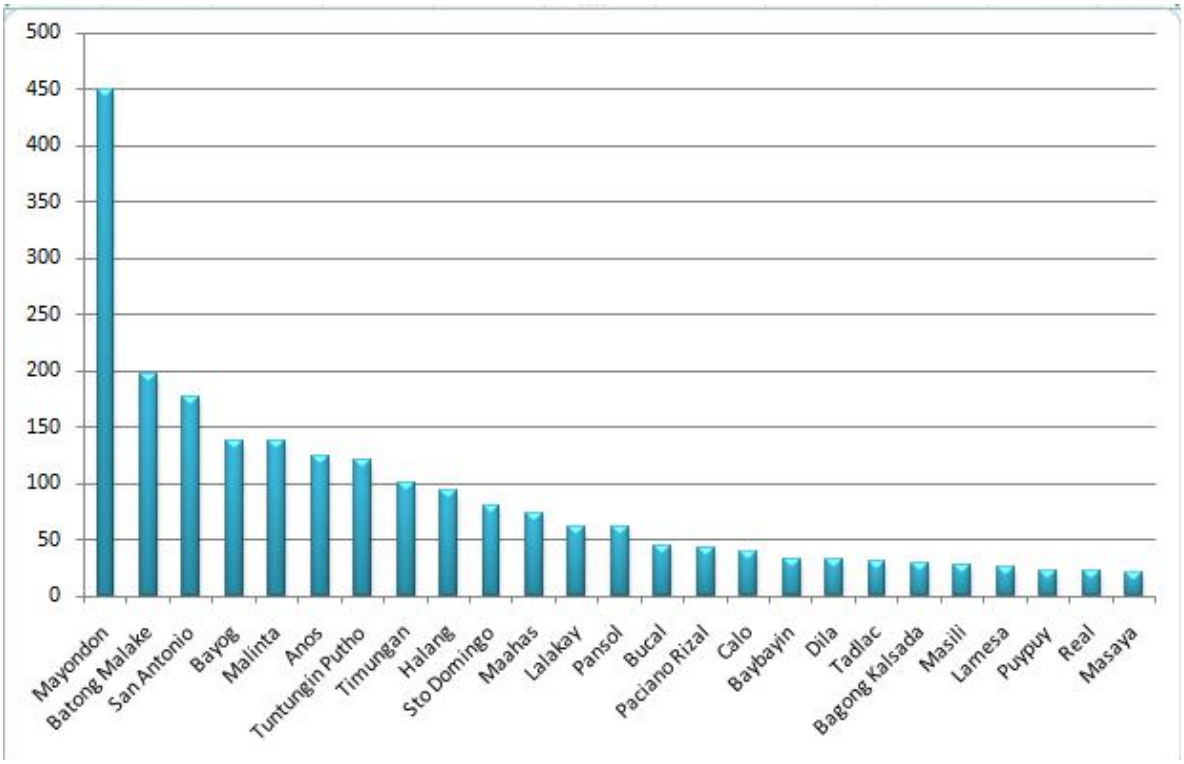


Figure 3 Distribution of Students Population by Barangay

Figure 2 shows that most of the students were residence of Los Banos, Calamba and Bay. This indicates the majority of the University's clientele was from the nearby town/city area. While at the Barangay level in Figure 3 shows that Mayondon, Batong Malake, and San Antonio were the highest in terms of students' population.

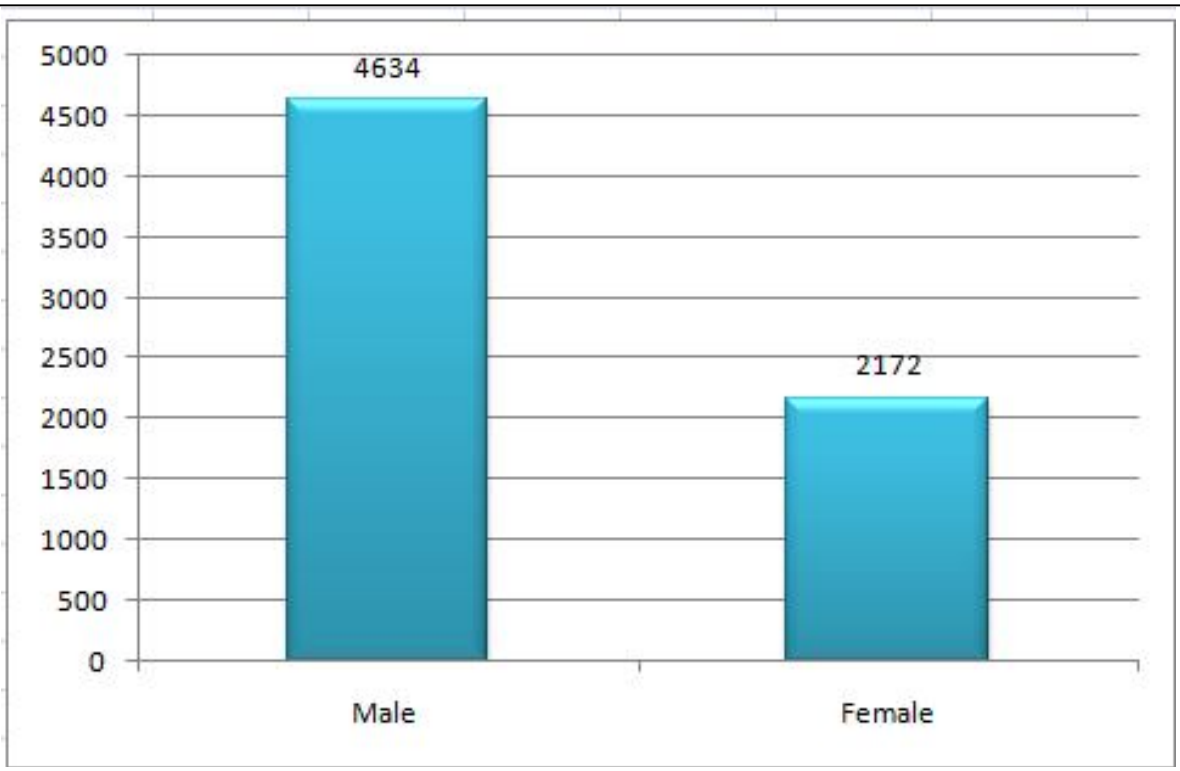


Figure 4 Distributions of Students by Gender

In terms of gender distribution, the male students were dominant than the female students in the BSIT Program as shown in Figure 4.

3.2 Data Transformation

The process of data transformation consists of data selection, data cleaning, and normalization. The data selection phase includes the extraction of students' information needed in the database. Data cleaning followed to check if there were incorrect/missing values to avoid the noise of the results. The final stage was the normalization process on which by means of SQL queries dataset were produced.

3.3 Pattern Extraction

In order to fully understand the data mining process, a model prediction framework in Figure 5 was adopted from Fadhilah Ahmad et.al a step by step process in the analysis of this research study. It shows the structure of the stages involved in Patterns Extraction process [2].

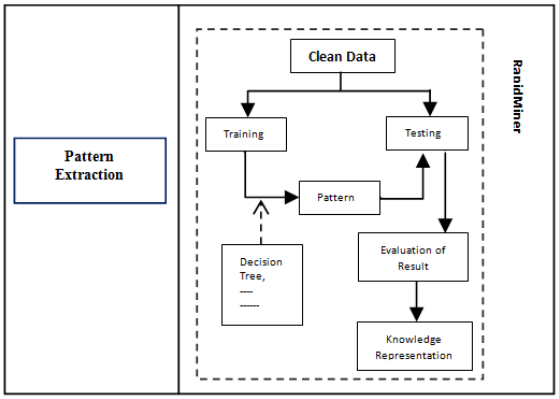


Figure 5 Model Prediction Framework (Fadhilah Ahmad et.al)

3.4 Data Mining Implementation and Results.

There were several techniques in data mining for identifying pattern and data processing. By examining data set attributes deducing of information can be structured to be helpful for decision making. Among these techniques were Classification, Association, Clustering and Sequential Patterns [2]. Classification was the mostly used and studied because it was simple and easy to understand. A method for predicting data object class or category based on learned classes from training dataset [3]. Multiple classification techniques are K-Nearest Neighbor (K-NN), Neural Networks, Decision Trees and Naïve Bayes. In this study, different classification techniques were used such as Naïve Bayes, Rule Induction, and Decision Trees to validate and verify the algorithm outcomes. The results consist of acceptable percentage of accuracy and precision measurement using Rapid Miner as data mining tool.

RESULTS AND DISCUSSIONS

This section presents the results of study after application of data mining process using the different classification algorithms. An interesting fact has been extracted in the process of data mining after generation of probability distribution matrix using the three classification algorithms. Studying students' performance in their professional courses was important as an input to curriculum development and

improvement of academic related activities. The focal point of the study was to analyzed students' performance based on the BSIT professional courses in Table 2 co- existed with its pre-requisites.

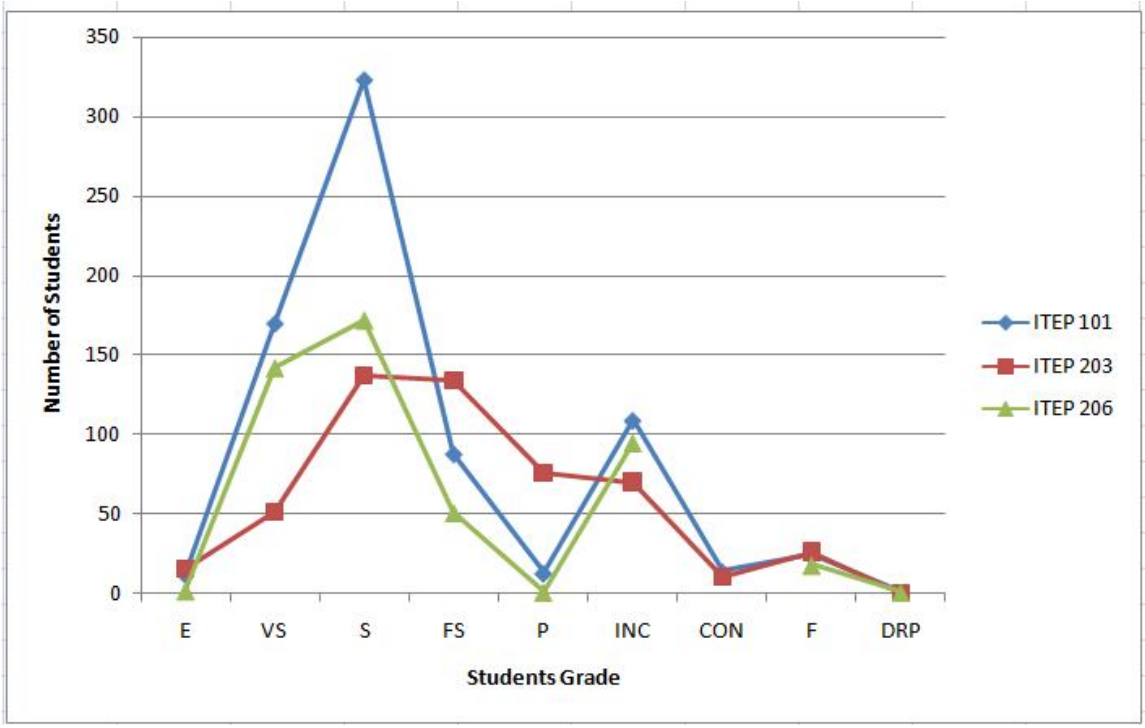


Figure 6 Students rating for ITEP 203, ITEP 206 and ITEP 101 as pre-requisite.

For the students to be allowed to enroll in ITEP 203 Networking 2 and ITEP 206 Operating system a passing grade must be achieved in ITEP 101 Networking 1 as pre-requisite. Figure 6 shows that most of the students perform Satisfactory with 52.87% and Very Satisfactory with 52.09% class recall in these courses meaning students who passed ITEP 101 most likely to passed ITEP 203 and ITEP 206. However, there were higher probabilities that those students who obtained an INC rating in ITEP 101 could have the same result in ITEP 203 and ITEP 206 with 40.78% of class prediction.

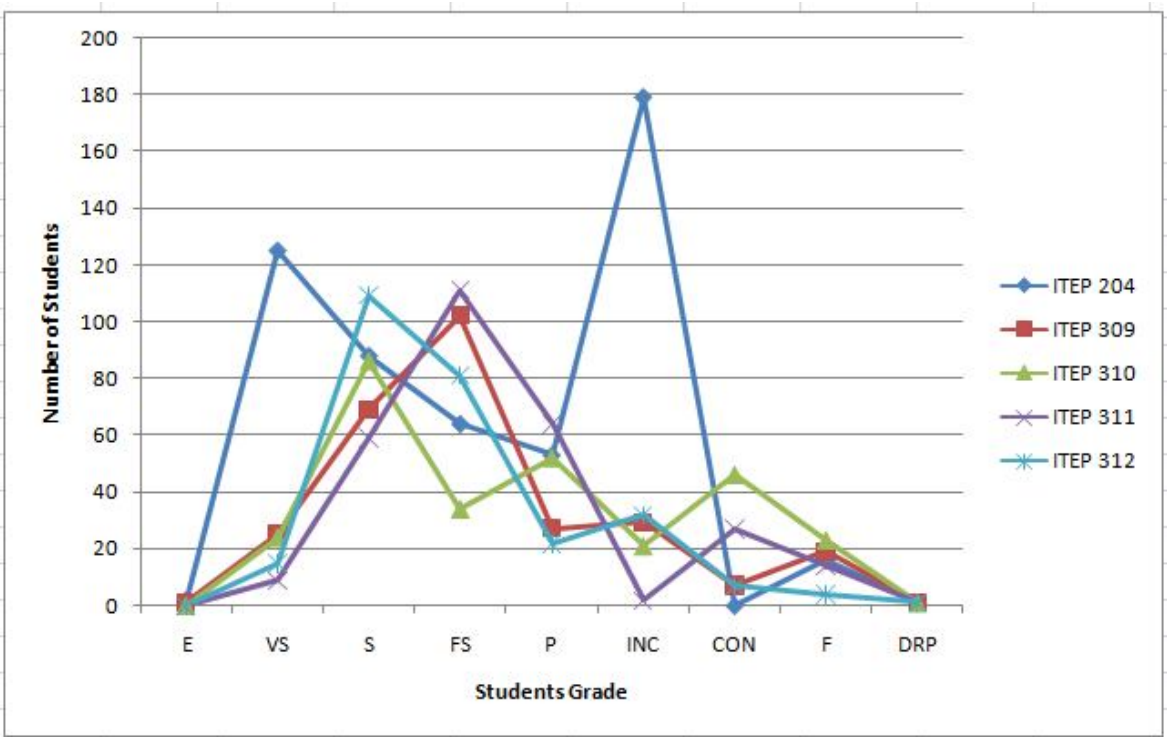


Figure 7 Students rating for ITEP 309, ITEP 310, ITEP 311, ITEP 312 and ITEP 204 as pre-requisite.

In case of Figure 7 ITEP 204 Database Management System 1 was all the pre-requisite of ITEP 309 Object Oriented Programming, ITEP 310 System Analysis and Design, ITEP 311 Database Management System 2 and ITEP 312 Web Development. Failing with ITEP 204 would deeply affect students' academic standings. Furthermore, these 4 courses were taken by students during their 3rd year first semester together with 3 more courses. It will be more challenging since these were major courses with laboratory. The result indicates that most of the students achieved a Satisfactory rating with 72.67% of class recall, 56.64% Fairly Satisfactory and 3.12% for Very Satisfactory. This shows that despite the circumstances and their academic load they still excel in most of the professional courses for 3rd year students.

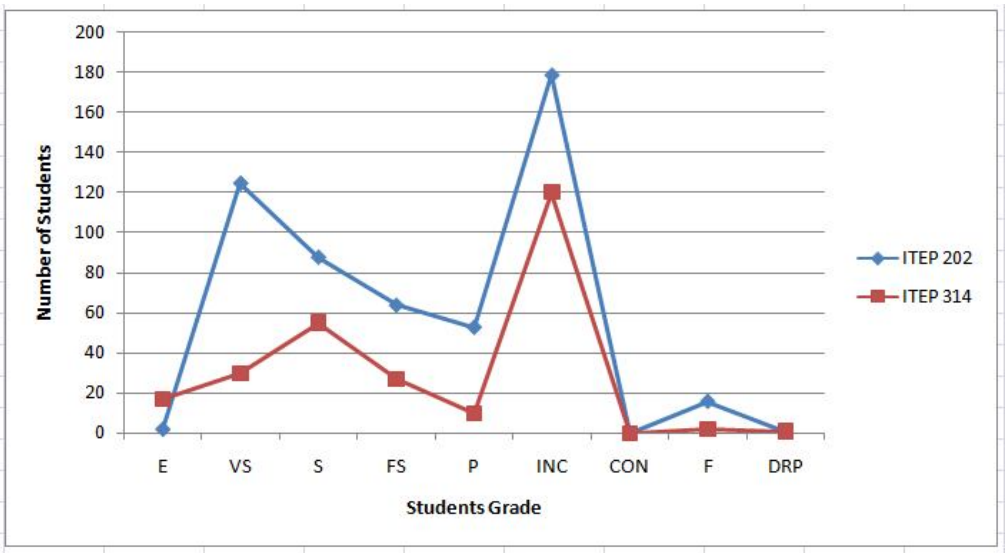


Figure 8 Students rating for ITEP 314 and ITEP 202 as pre-requisite.

The findings with ITEP 202 Data Structure and ITEP 314 Software Engineering needed an in depth research because of the 100% class recall in both courses for an Incomplete marks and with 37.85% class prediction. It shows significant relationship but with ambiguity of the reason. A more detailed dataset that includes attendance, quizzes, major examination, projects and the like must be included in the dataset since; an INC rating was given to students who passed the course but with lacking requirements.

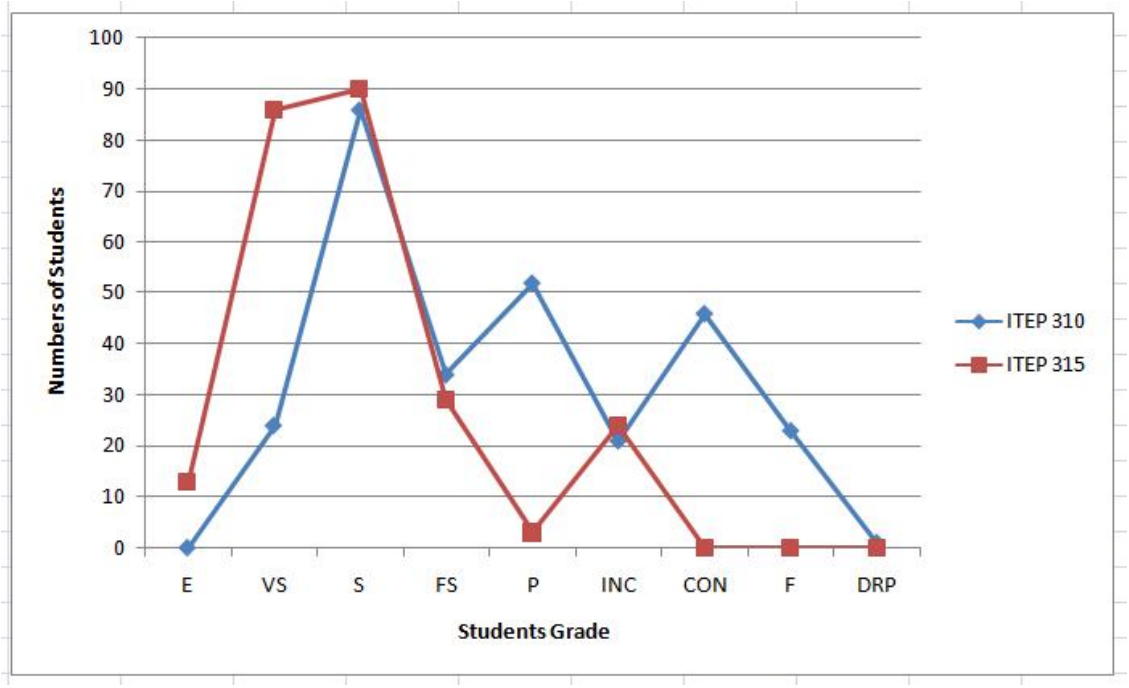


Figure 9 Students rating for ITEP 315 and ITEP 310 as pre-requisite.

ITEP 310 Web Development and ITEP 315 Multimedia System demonstrate a relative result of Satisfactory with 100% class recall and Class Prediction of 33.08%. It means that greater percentage of students would do well if they pass with ITEP 310.

In general, Naïve Bayes classification is a classification technique that assumes that all given attributes in a dataset are independent from each other and estimate the likelihood of a property given the dataset as evidence. To summarize the result of the BSIT Professional Courses the probability distribution matrix was generated by the Bayes model in Table 4 and the graphical illustration is presented in Figure 10.

Probability Distribution Matrix									
Professional Courses	E	VS	FS	S	P	INC	CON	F	DRP
ITEP 101	0.016	0.225	0.117	0.428	0.017	0.145	0.019	0.032	0.001
ITEP 202	0.004	0.237	0.121	0.167	0.100	0.339	0.000	0.030	0.002
ITEP 203	0.029	0.098	0.258	0.264	0.146	0.135	0.019	0.050	0.000
ITEP 204	0.027	0.308	0.171	0.355	0.058	0.067	0.002	0.011	0.002
ITEP 205	0.026	0.234	0.083	0.526	0.026	0.056	0.004	0.042	0.002
ITEP 206	0.004	0.295	0.106	0.357	0.002	0.197	0.000	0.037	0.002
ITEP 207	0.065	0.173	0.185	0.400	0.060	0.085	0.023	0.006	0.002
ITEP 208	0.020	0.317	0.157	0.292	0.031	0.102	0.074	0.004	0.002
ITEP 309	0.004	0.089	0.364	0.246	0.096	0.104	0.025	0.068	0.000
ITEP 310	0.000	0.084	0.118	0.300	0.181	0.073	0.160	0.080	0.003
ITEP 311	0.000	0.031	0.387	0.206	0.223	0.007	0.094	0.049	0.003
ITEP 312	0.000	0.055	0.299	0.042	0.081	0.118	0.026	0.015	0.004
ITEP 313	0.016	0.506	0.031	0.230	0.000	0.218	0.000	0.000	0.000
ITEP 314	0.065	0.115	0.103	0.210	0.038	0.458	0.000	0.008	0.004
ITEP 315	0.053	0.351	0.118	0.367	0.012	0.098	0.000	0.000	0.000
ITEP 416	0.139	0.836	0.000	0.000	0.000	0.025	0.000	0.000	0.000

Table 4 Summary of Bayes Probability Distribution Matrix

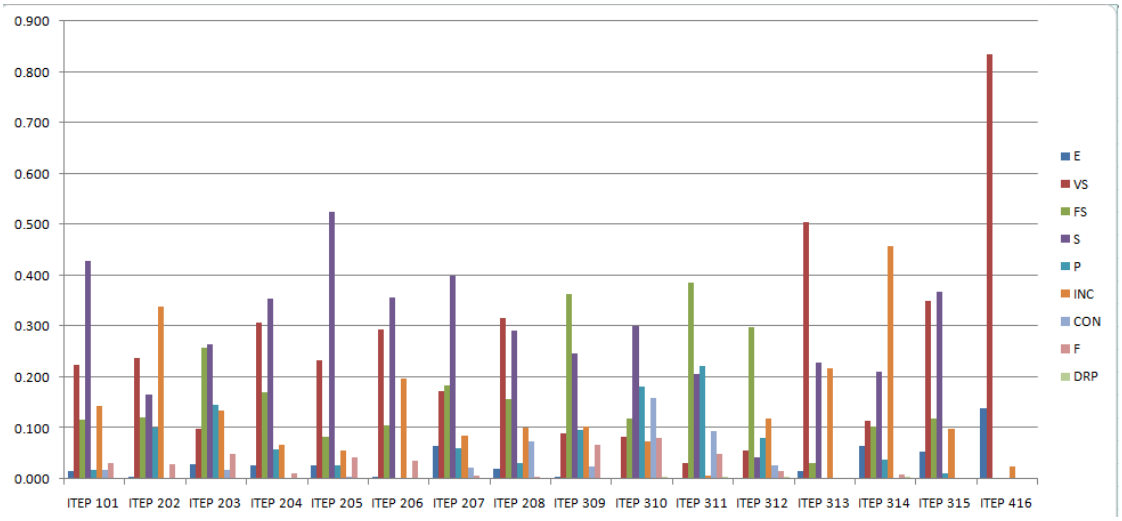


Figure 10 Graphical Illustration of Bayes Probability Distribution Matrix

Naïve Bayes Distribution matrix presents the 16 professional courses and the grades that the students obtained in fraction. ITEP 416 Capstone Project obtained with the highest Excellent and Very Satisfactory rating, in the BSIT curriculum this course was the culminating activity for Fourth Year Graduating Students in the form of system application or an enterprise resource plan. ITEP 311 Database Management System 2 achieved with the utmost Fairly Satisfactory and Passed rating. It emphasizes the importance of database analysis, design and implementation. ITEP 205 PC Troubleshooting and Maintenance was dominant with the Satisfactory rating, the course was designed to provide basic understanding of how personal computers work that also covers the basic of PC repair with an emphasis on troubleshooting and maintenance. ITEP 314 Software Engineering got the highest Incomplete rating. It focused on the basic of software engineering principles including the different software development techniques. ITEP 310 System Analysis and Design has the biggest fraction for the Conditional and Failed rating, the course is focused on the study of the different phase of information systems development. It can be simplified that in all courses, Conditional, Failed and Dropped ratings were at the least fraction. This shows that performance of BSIT Students in their professional courses was a lot better. To fully illustrate the summary of grades of students in all courses, the data is presented in Figure 11.

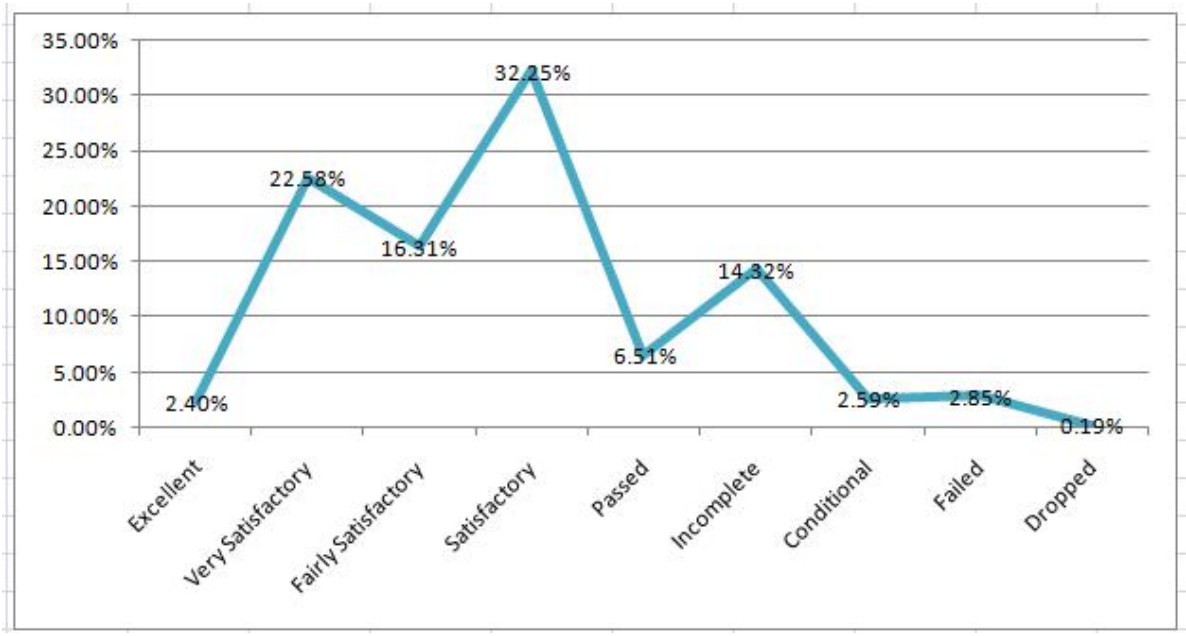


Figure 11 Summary of Grades of Students in all Professional Courses.

The summary of grades shows that students with Satisfactory rating rank the highest in all professional subjects with 32.25%. It shows that majority of students performance was in average level. It was followed by Very Satisfactory with 22.58% that demonstrate students acquired the necessary knowledge above average on the different subject matter. Fairly Satisfactory was in 16.31% that portray students' performance in below average level. Students were considered passed but have lacking requirements were given an Incomplete rating. This obtained 14.32%. Passed rating was given to students that have not acquired the necessary abilities to work proficiently but in the range of passing grades got 6.51%. In terms of students who performed beyond standards and has done exceptional work or Excellent got 2.40%. This was lower than the ratings of Conditional 2.59% and Failed with 2.85% while students who were dropping a course obtained .19% at the very lowest rating.

CONCLUSIONS

Research in Educational Data Mining (EDM) has significant contributions in the process of knowledge discovery. Understanding patterns extracted from the dataset on students' records was an interesting area of research. An academic institution should focus on students' performance and curriculum development. Since students' data was just confined into a large data bank, performing data analytics using EDM has the biggest potential of knowledge discovery.

In this study, performance of students in the BSIT professional courses in a select State University was given emphasis. It shows that all most majority of the students performed Satisfactory and Very Satisfactory. However, the results also show that there was higher probability that students who received an incomplete grade with the course pre-requisites may also be achieved in the subsequent courses. A follow up research on this aspect should

be conducted to determine the reason with a detailed dataset composition. A small fraction of excellent rating was also observed higher than conditional and failing students.

The outcome of the study may serve as guidelines for the administrators and faculty members in the making of possible solution from the identified predicament dealing with students' performance. It will also encourage other academic institution to conduct research on students' performance using data mining.

REFERENCES

[1] Rieger, Alexander (2014, May 18). Big Data Analysis | Predictive Modeling in Data Mining. Retrieved October 06, 2017, from http://blog.bosch-si.com/categories/technology/2011/11/large-scale-data-analysis-and-predictive-modeling-in-data-mining/

[2] Brown,M. (2012, December 11). Data mining techniques . Retrieved October 06, 2017, from https://www.ibm.com/developerworks/library/ba-data-mining-techniques/

[3] Saa, Amjad A. S., 2016. Educational Data Mining & Students Performance Prediction. (IJACSA) International Journal of Advanced Computer Science and Application, Vol. 7 No. 5. 2016.

[4] Ahmad, Fadhilah, et.al S., 2015. The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. Applied Mathematics Sciences, Vol. 9, 2015 no. 129, 6415-6426.

[5] Yadav, S.K. and Pal, S., 2012. Data mining: A Prediction for Performance Improvement of Engineering Students Using Classification. World of Computer Science and Information Technology Journal (WCSIT). Vol.2 No. 2, 51-56, 2012.

[6] Ahmed, A.B.E.D. and Elaraby, I.S., 2014. Data mining: A Prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), pp.43-47.

★★★