Research Journal of Computer and Information Technology Sciences ____________________E-ISSN 2320 – 6527
Vol. **4(11),** 1-15, November **(2016)** *Res. J. Computer and IT Sci.*

Mining Student Academic Performance on ITE subjects using Descriptive Model Approach

Mia Villar Villarica
College of Computer Studies, Laguna State Polytechnic University, Brgy. Bubukal Sta. Cruz, Laguna, Philippines
mia.villarica@lspu.edu.ph

Available online at: www.isca.in, www.isca.me
Received 29th August 2016, revised 24th October 2016, accepted 6th November 2016

Abstract

Data mining techniques has been useful to corporate world, however, it is transferable to educational environment. One of the objective of higher educational institution is to acquire potential alumni contributors. Secondary educational level echoes the educational value on college performance, the exact same reason the educational institution recognizes the significance of student admission, monitoring and evaluate potential students with good academic performance. Contemporary research paper on educational data mining focuses on applying data mining techniques in the context of student performance measurement and predictions. In this research, the descriptive model association and descriptive statistics cross tabulation is used to determine the subject area proficiencies of secondary educational institution and generate association rules using Predictive Apriori algorithm to discover hidden patterns. CRISP-DM methodology is used in this study to guide the researcher. The knowledge obtained from the model describes the subject area expertise of the high school as well as the subject ineptness of the school that would decision makers in student admission and academic planning of the educational institution.

Keywords: EDM, DM, CRISP-DM, Association Rules, Apriori and Predictive Apriori Algorithm.

Introduction

Higher Educational Institution (HEI) are positioned in a very high competitive environment all aiming to lead the battle towards optimal superiority among its competitor. In pursuit to educational and institutional excellence HEIs must acquire competitive advantage through acquiring knowledge and using it for a better valuation, preparation, development, management and decision making. According to Baradwaj, one way to achieve upmost level of quality in HEIs is by discovering knowledge[1]. Knowledge which is hidden among the educational data sets or university databases which is extractable and discoverable through data mining techniques and tools. There are increasing research interests in using data mining in education. Educational Data Mining (EDM) concerns with developing methods that discover knowledge from data originating from educational environments[2]. The father of EDM Ryan Baker, defined it as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, knowledge transfer, learning situations and the organization which they learn in[3].

Application of Data Mining (DM) with educational data sources is a recent research domain which has gained a larger recognition among the research society. Measuring of academic performance of students is challenging since students' academic performance hinges on diverse factors like personal, socio-economic, psychological and other environmental variables[4]. Proper understanding on the learning behavior of the students helps the educational foundations to manage their educational programs, to a much more improved level, which can increase the learning capabilities of the student, who followed their educational programs.

The university database containing records of students from admission to graduation is stored and waiting to be mined and be dig up for knowledge. Using the academic performance datasets gathered from the university database.

The research aims to discover knowledge in the university academic database specifically analyze the student grades in Information Technology Education major subjects associating the high school institutions attended to guide decision makers and academicians specifically the College Dean in College admission. Also, to utilized data mining tasks and techniques to uncover hidden patterns and discover tacit knowledge from large set of students' database record following the CRISP-DM methodology to complete the data mining process.

Educational Data Mining (EDM) is said to be an emerging discipline and it is a current trend in the field of educational and social research. Educational data mining is a wide term used recently to describe the use of data mining in educational data sources[5]. According to Romero there is a pressure in HEIs to provide up-to-date information on institutional effectiveness[6]. One response to this pressure based on the study conducted by Ranjan is finding new ways to apply analytical and data mining methods to educationally related data. Even though data mining has been applied in numerous industries and sectors, the application of DM to educational contexts is limited[7].

Data mining can assist in order to guide decision-making[8]. According to Luan on his article entitled "Data Mining Application in Higher Education" he defined data mining as a series of tools and techniques for uncovering hidden patterns and relationships among data[9]. Also he mentioned that data mining has presented advantages in predicting students' possibility of student attrition for a proactive intervention. His study also recommends that the institution should look at the student categories and attributes who are most at risk of not passing on a higher level which can be resolved early[10]. Based on the output of the research universities enables to predict the likelihood of student transferability or student attrition. Also, Luan mentioned that data mining can also link student's academic behaviors with their final grade. Lastly his study found out that these identified variables such as student types and student academic behaviors can affect on the final outcome of the students retention and attrition that can help the universities to focus on the student that needs more academic assistance by organizing extra classes, continuous consultation with the university's academicians, administration, counselors and psychologies[8].

In the study conducted by Waiyamai entitled "Improving Quality of Graduate Students by Data Mining" states that to improve the quality of graduate students' academic researchers can usedata mining techniques. His studyaims to understand and determine important information that can be turned into knowledge from a huge set of engineering student's databases records. Based on the study the discovered knowledge is valuable and beneficial in the development of new curricula, refining of existing curricula and most importantly serving the students in helping them in major selection[11].

In the research by Delavari entitled "Data Mining Application in Higher Learning Institutions" which presents the importance of data mining in the higher educational system. In particular the study by presents the application of data mining techniques to discover new explicit knowledge which could be useful for the decision making processes. Also the research output recommends an analytical guideline for higher education institutions to improve and augment their current decision procedures[12]. However in study conducted by Goyal suggests the use of data mining techniques to improve the efficiency of higher education institution. Stated in the study, data mining techniques such as clustering, decision tree and association are applied to higher education processes, it would help to improve students' performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution. The study's outcome discovers an approach to examine the effect of using data mining techniques in higher education[13].

As mentioned in the above literature, educational data mining has been recognized and utilized by different educations institutions in support with development of curriculum, student retention and attrition, learning preference, tracing students' academic performance and educational institution strategic management decision making.

Methodology

The CRISP-DM methodology suggested by Chapman *et al*, was utilized in this study. CRISP-DM remains the most popular methodology for analytics, data mining, and data science projects[14]. CRISP-DM: Cross-Industry Standard Process for Data Mining is a standardized approach to data mining and process model that describes commonly used approaches that data mining experts use to resolve problems and challenges during the process. The CRISP-DM methodology is the life cycle of a data mining project consists of six phases, shown in Figure-1. The order of the phases is not rigid. Moving back and forth between different phases is always essential and the outcome of each phase regulates which phase, or particular task of a phase, has to be achieve next. The arrows indicate the most important and frequent dependencies between phases. The outer circle in the figure symbolizes the cyclical nature of data mining itself. Data mining does not end once a solution is deployed. The knowledge erudite during the procedure and from the organized solution can trigger new, often more-focused questions. The subsequent data mining processes will benefit from the experiences of previous ones.

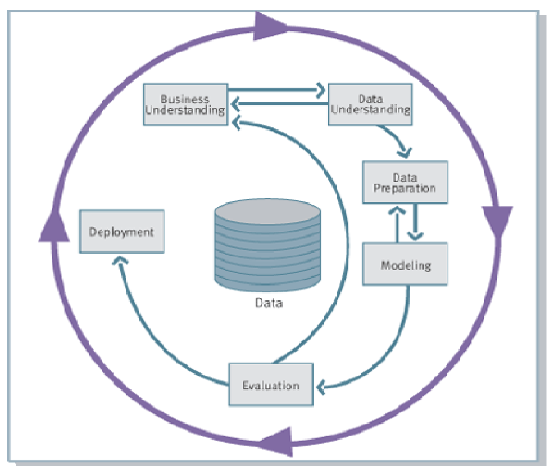


Figure-1
Steps of CRISP-DM Methodology[14]

The researcher utilized the university database approved by the University President and Campus Director the researcher utilized the operational database of the institution. Necessary legalities and assuring data confidentiality was recognized and valued. Data available was analyzed and literatures and related works were referred to and evaluated, converting this information into problem definition as well as designing the preliminary plan to achieve the objectives. CRISP-DM methodology consists of six phases:

Business Understanding: The initial phase of CRISP-DM is business understanding which focuses on understanding the project objectives and requirements from a business perspective. Organization and university process assessment was conducted specifically college admission process. Literature consultation and relevant research was carried out to translate the information into knowledge and converting this knowledge into a data mining problem definition, and formulate a preliminary plan designed to achieve the objectives. In this study, research related to data mining for educational environment were sought and trace the appropriate data mining techniques the previous study utilized to be adopted on this research.

The project objective is to identify the potential student for the College of Computer Studies based on the academic performance of previous students, associating the grade and high school institution for academic decision making purposes.

Data Understanding: After understanding, the subsequent phase is data understanding and it starts with an initial data collection. This phase proceeds with familiarization with the data, identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses for hidden information. This phase was performed in parallel with the initial and the subsequent phase. The university database contains student information from admission to graduation. The database is in My SQL has been used for almost 10 years, outsourced and not well maintained. The database possesses flaws and few bugs.

Data Preparation: This phase covers all activities to construct the final dataset from the initial raw data. These tasks are likely to be completed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools. In this study, data preparation was performed parallel to business and data understanding.

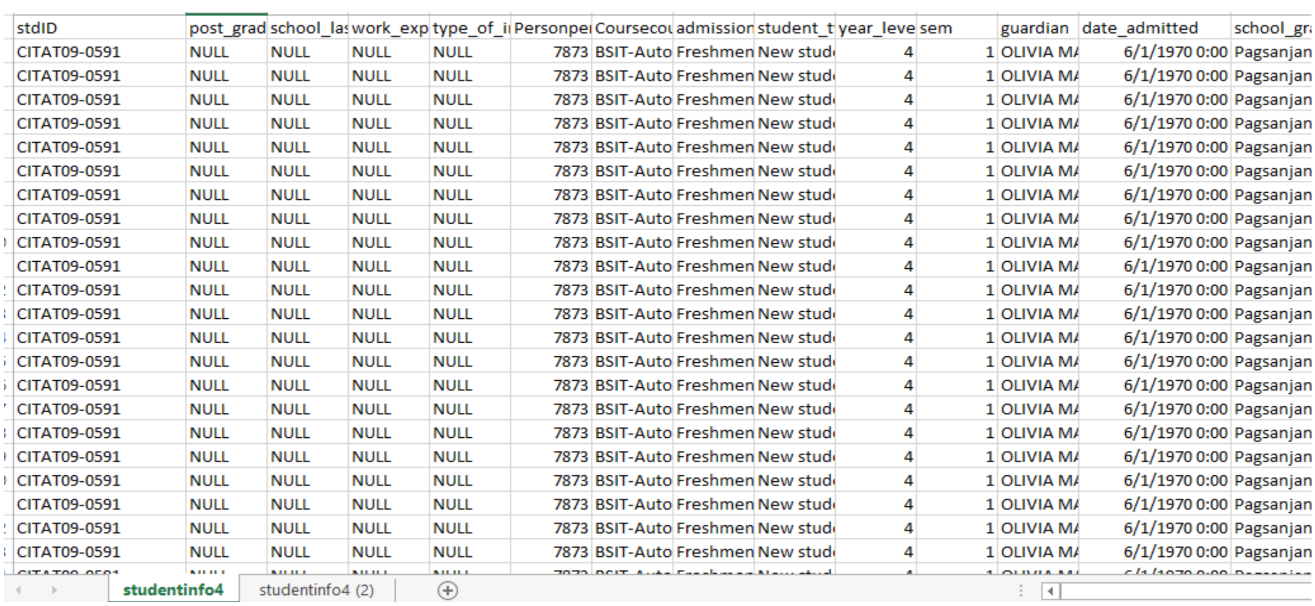
In accomplishing the preprocessing, *data integration* was performed combining data from multiple sources to form a coherent data store. Cross Tabulation Query was used to extract data and join operations using SQL Code, linking admission information and student grade from the registrar generating 84,880 instances from 2009-2015 records with 30 attributes.

The data quality problem has been identified, SQL code was utilized to parse attributes, and data was further process on the next phase.

Data cleaning was accomplished to check misplaced values, smooth out noise while classifying outliers, and correct discrepancies in the data. It was performed as an iterative two-step process consisting of discrepancy detection and data transformation.

Data transformation routines convert the data into appropriate forms for mining. High school code was assigned to High School institutions. Figure-2 demonstrates the sample data converted to excel file, converted from its original state which is SQL format to Comma Separated Value at Attribute-Relation File Format (ARFF) for WEKA Tool. Column of values are transformed into many columns in the new view.

Data reduction is used to obtain a reduced representation of the data while minimizing the loss of information content. In this study, various data reduction was done while maintaining its integrity and preserving the quality of information.



stdID	post_grad	school_las	work_exp	type_of_i	Personpe	Coursecou	admission	student_t	year_leve	sem	guardian	date_admitted	school_gra
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan
CITAT09-0591	NULL	NULL	NULL	NULL	7873	BSIT-Auto	Freshmen	New stud	4	1	OLIVIA MA	6/1/1970 0:00	Pagsanjan

Figure-2
Sample Student Data

Datasets tend to expose new issues and challenges while in the process of transformation. With the goal in mind, it is important to choose the right data mining algorithms, techniques and tools which are expected to give best results with data mining.

Descriptive data summarization techniques was used to identify the typical properties of the data and highlight which data values should be treated as noise or outliers and it provides the analytical foundation for data preprocessing. The basic statistical measures for data summarization was utilized which includes mean, weighted mean, median, and mode for measuring the central tendency of data, range, quartiles, and standard deviation for measuring the dispersion of data.

Table-1
High School Code

HS Code	High School Institution
1	Banahaw Institute
2	Banca-Banca Nat'l High School
3	Calumpang National High School
4	Cavinti National High School
5	Dayap National High School
6	Don Manuel Rivera Memorial National High School
7	Liceo De Cavinti
8	Liceo De Luisiana
9	Liceo DE Majayjay
10	Liceo DE Pagsanjan
11	Liceo DE PILA
12	Linga National High School
13	Pedro Guevarra Memorial National High School
14	Pedro Guevarra Memorial National High School –Annex

Table-2
The Symbolic Attribute Description

Attribute	Description	Position Values
HS Code	High School Code	[1-14]
Software Engineering	ITE Major Subject Software Engineering – student grade	[1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4,5]
Systems Analysis and Design	ITE Major Subject Systems Analysis and Design - student grade	[1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4,5]
Dbase1	ITE Major Subject Database 1 - student grade	[1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4,5]
Dbase2	ITE Major Subject Database 2 - student grade	[1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4,5]
WD	ITE Major Subject Web Development - student grade	[1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4,5]
MulDev	ITE Major Subject Multimedia Development - student grade	[1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4,5]

As a result of preprocessing phase, the total number of data was reduced to 4552 instances after extracting BSIT students and 596 instances after narrowing it to 13 High Schools and taking the IT Education major subjects Software Engineering, Systems Analysis and Design, Database 1, Database 2 Web Development and Multimedia Development. As shown in Table-1 High School Code was assigned to High School Institutions, which was chose based on a result of top 14 High School with most number of enrollees.

Data preprocessing was completed, the overall representation of your data was scrutinized using IBM SPSS Statistics as well as Rapid Miner and WEKA Tool.

Modeling: During this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. DM is an interdisciplinary field of astronomy, business, computer science, economics and others to discover new patterns from large data sets[1]. The actual data mining task is to analyze large quantities of data in order to extract previously unknown patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). Data mining is an iterative process even after the data preparation was completed and modeling is being performed problems and issues arise.

According to Han, descriptive model defines the data set in a summarized and concise way and presents the interesting general properties of the data. It explains the patterns in existing data, which may be used to guide decisions. Also according to him, it describes data in a concise and summative manner and presents interesting general properties of the data. This is entirely different from predictive data mining, which analyzes data in order to construct one or a set of models, and attempts to predict the behavior of new data sets[9].

Based from the review of related literature and studies, it was found out that the research should use association rule aligned on the objective. In this experiment, the researcher used two different methods for rule mining—the Apriori algorithm that has become a standard approach and the Predictive Apriori algorithm published by Scheffer[15]. Apriori Algorithm is the most popular and useful algorithm of Association Rule Mining of Data Mining, and one of the useful techniques is association rule it was first introduced by Agarwal[16]. On the other hand, predictive apriori implements to also mine association rules and it examines with an increasing support threshold for the best 'n' rules. The most significant difference between the two relates to how the interestingness of an association rule is measured. Although both are confidence-based, the confidence is estimated differently. Both algorithms start the same way by building frequent item sets. An item set is called frequent when its support is above a predefined minimum support.

Also according to Agarwal, these are the difference of Apriori Algorithm and Predictive Apriori Algorith:

Apriori Algorithm is an association rule mining algorithm is used to discover recurrent item sets in transaction database. Quality and eminence of association rules hinged upon support and confidence[17].

Support: Support of an association rule $A \cup B$ is the number of transactions that contains item set $A \cup$[17].

$$support A \Rightarrow B = P(A \cup B)$$

Confidence: Confidence of an association rule $A \cup B$ is the ratio of number of transactions that contains $A \cup B$ to the number of transaction that contains A [18].

$$Confidence\ A \Rightarrow B = P\ B\ A = \frac{support\ (A\ \cup B)}{support\ (A)}$$

Predictive Apriori Algorithm is an algorithm produces association rule based on predictive accuracy. Predictive accuracy is derived from support and confidence. The algorithm searches with an increasing support for the best “n‟ rules[19].

The researcher applied Apriori algorithm to the preprocessed datasets to extract association rules in order to find frequent item sets specifically in associating ITE major subjects with student grade and high school attended and to uncover the hidden information. For this experiment, WEKA tool 3.7 was used for extracting the results. In search for more hidden patterns and unknown knowledge the researcher also applied Predictive Apriori Algorithm. In Predictive Apriori association rule algorithm, support & confidence is combined into a single measure called “Accuracy‟. {Support, Confidence}=> Accuracy[20].

At this stage, a model or models that appears to have high quality, from a data analysis perspective, has been built. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the objectives. A key objective is to determine if there is some important issue that has not been sufficiently considered.

At the end of this phase, a decision on the use of the data mining results should be reached. The knowledge obtained from data mining techniques can provide managerial decision makers knowledge for decision making.

Deployment: Generation of the model is generally not the end of the research. Even if the purpose of the model is to upsurge knowledge of the data, the knowledge gained will need to be prepared and presented in a way that the client in this case the HEI administrator can use it. Depending on the necessities, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process.

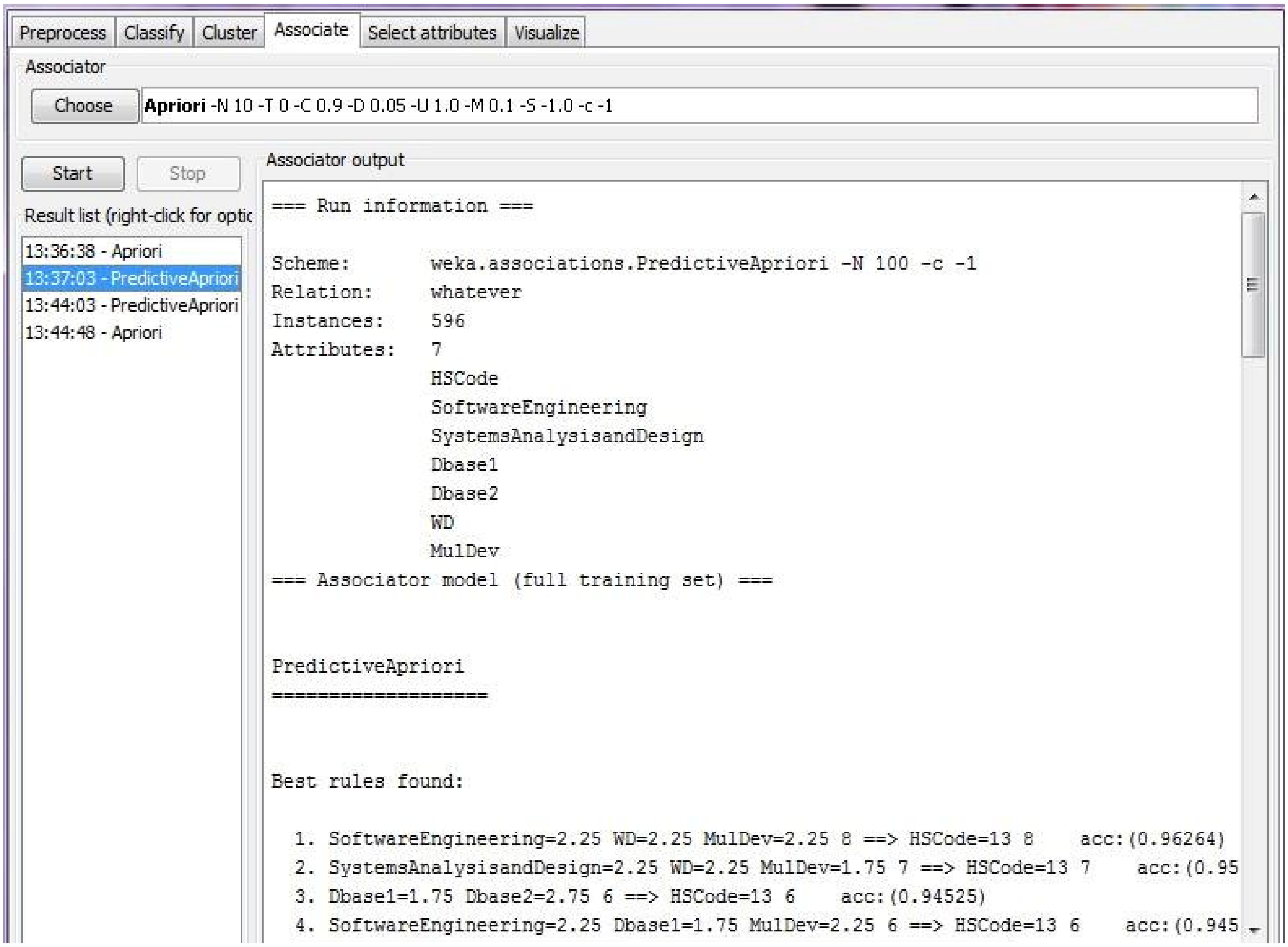


Figure-3
WEKA tool 3.7

Results and Discussion

Series of iterative process such as statistical testing using IBM SPSS Statistics and Weka Tool. The Descriptive statistics was performed, the frequency per attribute was generated with process time 1.75 minutes and elapsed time 2.57, percentile of 4 using statistics mean, median and mode. Frequencies per attribute were represented by pie charts as shown in Figure-5. As shown in the succeeding figures and tables majority of the students come from HS Code 13,out of 596 instances 261 instances comes from this HS Code which pertains to Pedro Guevarra Memorial National High School.

The Descriptive Statistics of the attributes having 596 records valid with no missing record, graphically presented as shown in Figures-5, the frequency distribution of attributes. Figure-5A shows the frequency distribution of High School with the highest frequency HSCode 13 - Pedro Guevarra Memorial National High School. Figure-5B illustrates the frequency distribution of ITE Major Subject Software Engineering with the highest number of students who gets 2.00 and 2.25 respectively. As illustrated in Figure-5C the frequency distribution of Systems Analysis and Design highest number of students who gets 2.00. Figure-5D displays the frequency distribution of ITE Major Subject with the highest frequency Database 1 highest number of students who gets 2.00. As presented in respective figures the frequency distribution of ITE Major Subject with the highest frequency Database 2, Web Development and Multimedia Development with highest number of students who gets 2.00.

Cross Tabulation Analysis: Cross tabulations enable you to examine relationships within the data that might not be readily apparent when analyzing the total data sets. IBM SPSS Statistics was used to perform Cross Tabulation Analysis. As represented by Tables-2 to 7, having the High School Code as the row and columns represented the student grade categorized per subjects. Results exhibits the relationship between attributes, ITE major Subject are analyze in relations with the High School.

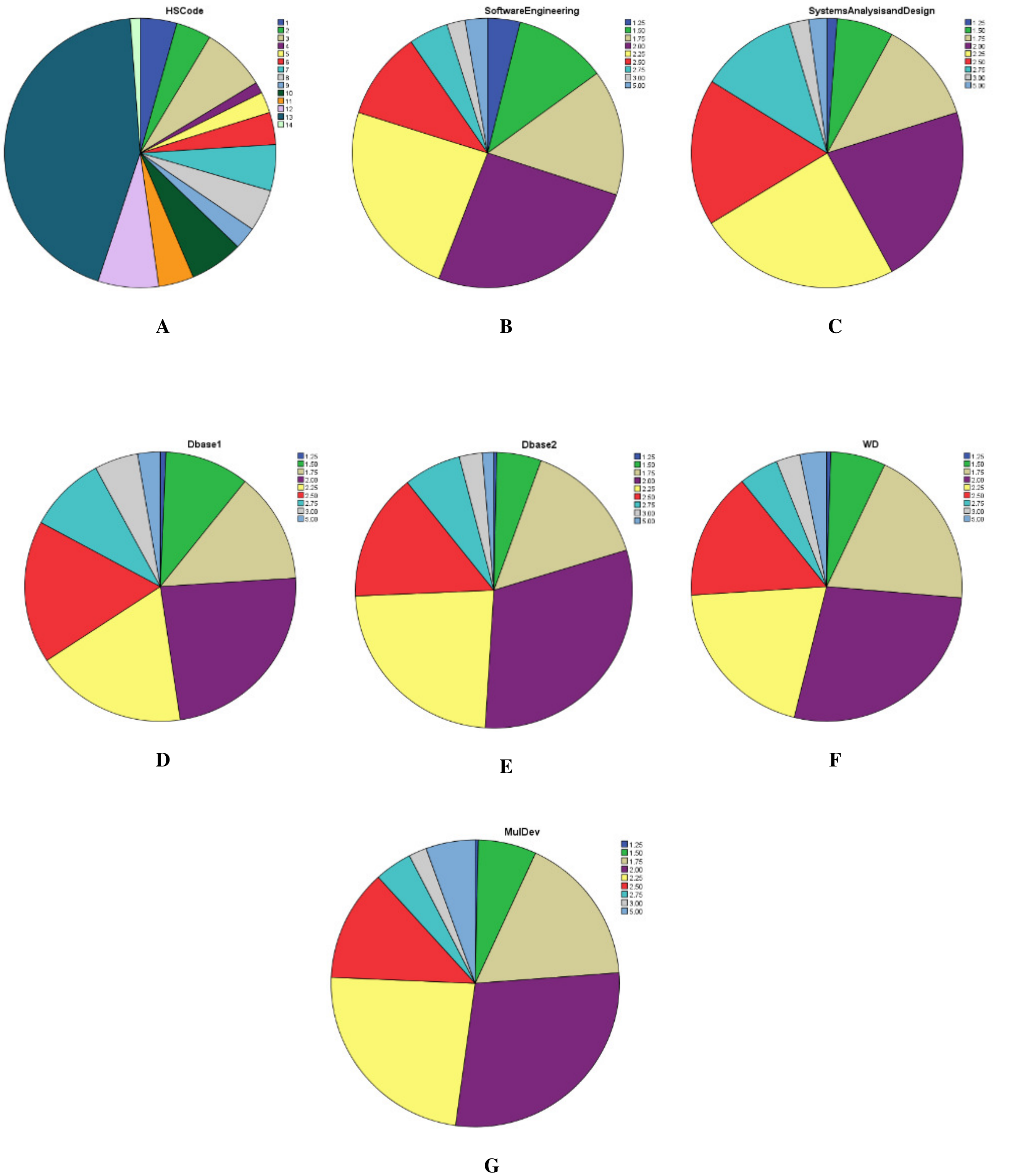


Figure-4
Frequency Distribution

Table-3
Cross Tabulation Analysis between High School and Software Engineering

HSCode * SoftwareEngineering

Crosstab

Count

		SoftwareEngineering									
		1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	5.00	Total
HSCode	1	1	1	2	5	5	4	6	2	0	26
	2	0	1	2	8	6	2	3	3	0	25
	3	2	5	6	8	13	5	3	2	2	46
	4	0	0	2	0	5	1	0	0	0	8
	5	0	0	5	2	3	2	2	1	0	15
	6	0	2	3	9	5	3	0	1	0	23
	7	1	3	5	8	10	5	1	0	0	33
	8	0	3	4	6	7	4	3	2	1	30
	9	1	0	3	5	3	4	0	0	0	16
	10	1	3	2	5	8	4	6	0	9	38
	11	1	3	4	8	2	6	1	0	0	25
	12	0	4	11	16	8	1	0	2	1	43
	13	15	41	41	72	65	22	2	0	3	261
	14	1	0	0	2	3	0	1	0	0	7
Total		23	66	90	154	143	63	28	13	16	596

Table-4
Cross Tabulation Analysis between High School and Systems Analysis and Design

Crosstab

Count

		SystemsAnalysisandDesign									
		1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	5.00	Total
HSCode	1	1	3	7	4	5	4	2	0	0	26
	2	0	2	2	4	4	7	1	0	5	25
	3	0	1	4	11	10	13	4	3	0	46
	4	0	1	3	3	0	0	1	0	0	8
	5	0	1	0	3	4	7	0	0	0	15
	6	1	1	3	8	4	1	2	0	3	23
	7	0	0	6	15	6	2	3	0	1	33
	8	0	3	2	8	8	6	2	1	0	30
	9	0	0	1	2	7	4	1	1	0	16
	10	0	7	10	5	9	3	2	2	0	38
	11	0	1	8	7	1	3	5	0	0	25
	12	0	4	4	10	10	10	3	1	1	43
	13	5	16	22	49	75	44	41	6	3	261
	14	0	0	1	2	1	1	2	0	0	7
Total		7	40	73	131	144	105	69	14	13	596

Table-5
Cross Tabulation Analysis between High School and Database1

HSCode * Dbase1

Crosstab

Count

		Dbase1									
		1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	5.00	Total
HSCode	1	0	2	3	8	7	2	3	1	0	26
	2	0	0	0	9	7	7	0	1	1	25
	3	0	5	5	8	4	11	6	4	3	46
	4	0	0	1	1	2	1	0	2	1	8
	5	0	1	1	4	3	2	2	2	0	15
	6	1	0	2	6	4	2	6	2	0	23
	7	0	2	2	11	5	5	3	5	0	33
	8	1	1	2	8	6	5	2	4	1	30
	9	0	0	5	4	3	4	0	0	0	16
	10	0	6	8	5	7	6	4	1	1	38
	11	0	2	2	7	8	2	4	0	0	25
	12	0	6	6	8	9	8	1	1	4	43
	13	2	35	42	61	40	46	23	8	4	261
	14	0	0	0	1	3	1	1	0	1	7
Total		4	60	79	141	108	102	55	31	16	596

Table-6
Cross Tabulation Analysis between High School and Database2

HSCode * Dbase2

Crosstab

Count

		Dbase2									
		1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	5.00	Total
HSCode	1	0	1	2	11	10	1	1	0	0	26
	2	0	0	6	6	5	3	2	3	0	25
	3	0	5	9	15	6	6	3	2	0	46
	4	0	3	3	1	1	0	0	0	0	8
	5	0	0	2	4	4	4	1	0	0	15
	6	0	1	3	7	6	5	0	1	0	23
	7	0	3	3	15	8	3	1	0	0	33
	8	0	1	3	10	9	5	1	1	0	30
	9	0	0	0	6	9	1	0	0	0	16
	10	0	3	8	12	5	8	0	2	0	38
	11	0	2	4	8	7	2	0	2	0	25
	12	0	3	8	16	10	3	0	0	3	43
	13	2	9	37	71	55	46	31	5	5	261
	14	0	0	0	1	4	2	0	0	0	7
Total		2	31	88	183	139	89	40	16	8	596

Table-7
Cross Tabulation Analysis between High School and Web Development

HSCode * WD

Crosstab

Count

		WD									
		1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	5.00	Total
HSCode	1	0	2	4	8	7	4	1	0	0	26
	2	1	2	2	9	4	4	0	2	1	25
	3	0	6	15	8	5	4	4	3	1	46
	4	0	1	2	2	0	0	0	0	3	8
	5	0	2	3	4	4	1	1	0	0	15
	6	0	1	4	8	6	3	1	0	0	23
	7	0	4	8	13	4	3	1	0	0	33
	8	0	2	8	6	7	7	0	0	0	30
	9	0	0	3	2	4	7	0	0	0	16
	10	0	3	11	12	6	5	1	0	0	38
	11	0	2	8	9	4	0	2	0	0	25
	12	0	1	3	18	8	9	0	0	4	43
	13	2	13	44	65	60	43	15	11	8	261
	14	0	0	0	0	1	1	2	1	2	7
Total		3	39	115	164	120	91	28	17	19	596

Table-8
Cross Tabulation Analysis between HSCode and Multimedia Development

HSCode * MulDev

Crosstab

Count

		MulDev									
		1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	5.00	Total
HSCode	1	0	2	0	8	6	3	1	4	2	26
	2	0	0	4	5	5	3	0	2	6	25
	3	0	3	10	10	7	8	3	0	5	46
	4	0	2	2	4	0	0	0	0	0	8
	5	0	2	1	6	2	3	1	0	0	15
	6	0	2	1	10	9	0	0	1	0	23
	7	0	2	4	10	8	7	2	0	0	33
	8	0	5	7	5	6	4	1	1	1	30
	9	0	2	2	9	2	1	0	0	0	16
	10	1	3	4	14	7	2	1	0	6	38
	11	0	3	7	6	7	1	1	0	0	25
	12	0	0	2	15	11	7	3	0	5	43
	13	1	13	56	64	68	36	11	4	8	261
	14	0	0	1	3	2	0	1	0	0	7
Total		2	39	101	169	140	75	25	12	33	596

Graphical Representation of Cross Tabulation Analysis: As clearly stated on Figure-6 the Graphical Representation of Cross Tabulation of High School and ITE Subjects HS Code 13 has the highest frequency and because of that HS Code 13 – Pedro Guevarra National High School has the most number of high ranking grade in respect with ITE Major Subjects.

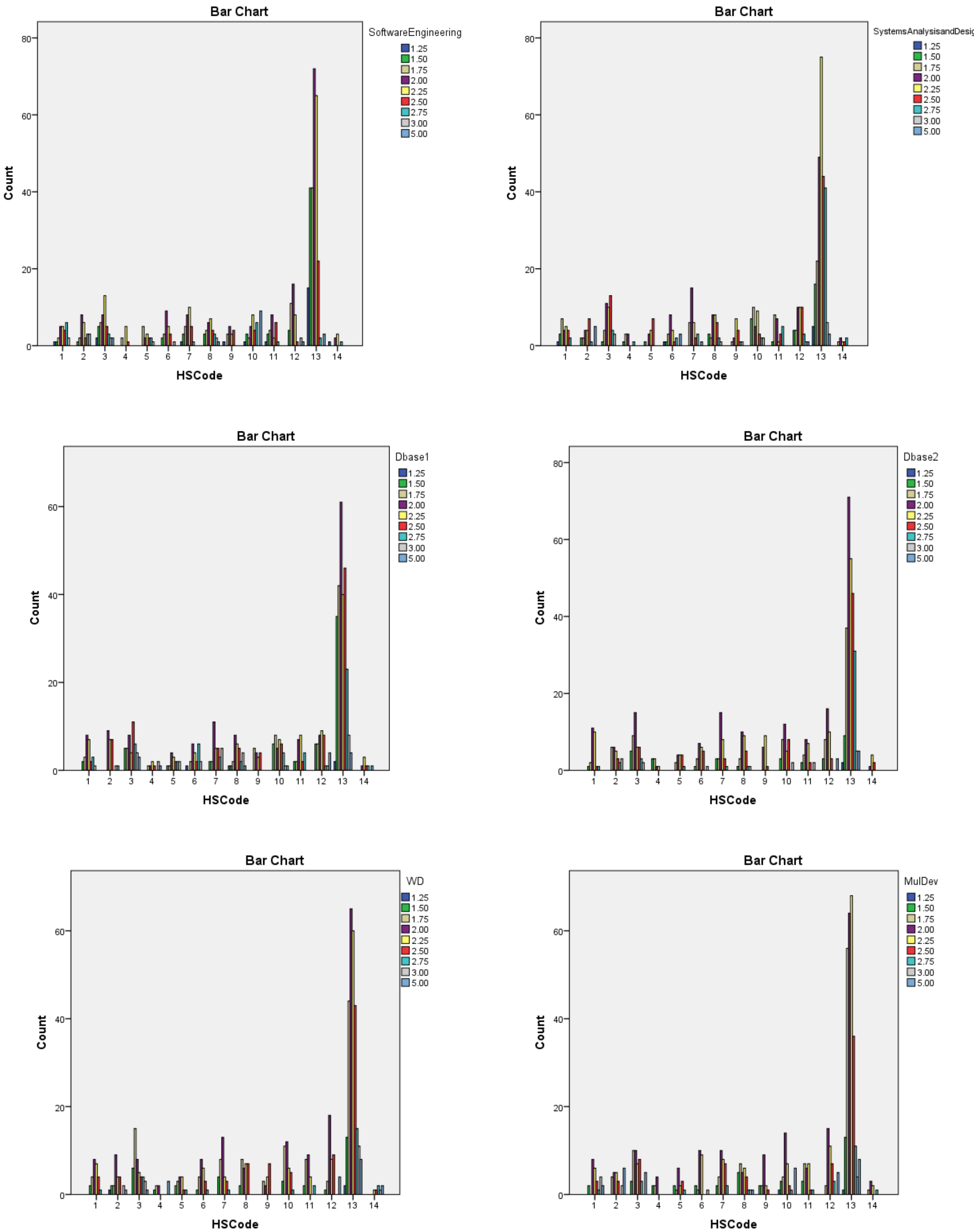


Figure-6
Graphical Representation of Cross Tabulation of High School and ITE Subjects

Apriori and Predictive Apriori Algorithm: For further analysis, the Association rule mining was performed using both Apriori and Predictive Apriori using Weka Tool. As a result, Apriori algorithm has generated 10 best association rules as shown in Table-8 while Predictive Apriori generated 100 association rules as shown in Table-9 (*25 best results indicated*). As presented in Table-9, with 17 cycles performed with minimum support 0.15 and minimum confidence of 0.9 unknown patterns were discovered:

Rule 1: DBASE2=2 9 ==> MULDEV=5 9
<conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.51)
with the top highest level of support and confidence result, which suggests that students with grade of 2.0 in Database2 has 5.0 grade in Multimedia Development, with an instance of 9.

Rule 2: SYSTEMSANALYSISANDDESIGN=1.75 8 ==> MULDEV=5 8
<conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.46)
Rule 2 can be interpreted as students with grade of 1.75 in Systems Analysis and Design has 5.0 grade in Multimedia Development, with an instance of 8.

Rule 6: HSCODE=10 6 ==> MULDEV=5 6 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34)
Rule 6 can be interpreted as students from the High School 10LICEO DE PAGSANJANhas grade of 5.0 grade in Multimedia Development, with an instance of 6.

Rule 7. HSCODE=2 6 ==> MULDEV=5 6
<conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34)
Rule 7 can be interpreted as students from the High School 2BANCA-BANCA NAT'L HIGH SCHOOLhas grade of 5.0 grade in Multimedia Development, with an instance of 6.

Table-9
RULE SET Generated by WEKA TOOL using Apriori Algorithm

```
=== Run information ===
Scheme:       weka. associations. Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     whatever
Instances:    35
Attributes:   7
              HSCODE
              SOFTWAREENGINEERING
              SYSTEMSANALYSISANDDESIGN
              DBASE1
              DBASE2
              WD
              MULDEV
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.15 (5 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17
Generated sets of large itemsets:
Size of set of large itemsetsL(1): 25
Size of set of large itemsetsL(2): 24
Size of set of large itemsetsL(3): 1
Best rules found:
 1. DBASE2=2 9 ==> MULDEV=5 9    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.51)
 2. SYSTEMSANALYSISANDDESIGN=1.75 8 ==> MULDEV=5 8    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.46)
 3. DBASE2=2.5 8 ==> MULDEV=5 8    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.46)
 4. WD=2.5 8 ==> MULDEV=5 8    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.46)
 5. DBASE1=2.25 7 ==> MULDEV=5 7    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.4)
 6. HSCODE=10 6 ==> MULDEV=5 6    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34)
 7. HSCODE=2 6 ==> MULDEV=5 6    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34)
 8. DBASE1=2.5 6 ==> MULDEV=5 6    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34)
 9. HSCODE=3 5 ==> MULDEV=5 5    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.29)
10. HSCODE=12 5 ==> MULDEV=5 5    <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.29)
```

Table-10
RULE SET Generated by WEKA TOOL using Predictive Apriori Algorithm

```
=== Run information ===
Scheme: weka.associations. Predictive Apriori -N 100 -c -1
Relation: whatever Instances: 35
Attributes: 7
HSCODE SOFTWAREENGINEERING SYSTEMSANALYSISANDDESIGN DBASE1 DBASE2 WD MULDEV
=== Associator model (full training set) ===
PredictiveApriori ===================
Best rules found:
1. DBASE2=2 9 ==> MULDEV=5 9 acc:(0.9875)
2. SYSTEMSANALYSISANDDESIGN=1.75 8 ==> MULDEV=5 8 acc:(0.98575)
3. DBASE2=2.5 8 ==> MULDEV=5 8 acc:(0.98575)
4. WD=2.5 8 ==> MULDEV=5 8 acc:(0.98575)
5. DBASE1=2.25 7 ==> MULDEV=5 7 acc:(0.98327)
 6. HSCODE=10 6 ==> MULDEV=5 6 acc:(0.97956)
7. HSCODE=2 6 ==> MULDEV=5 6 acc:(0.97956)
8. DBASE1=2.5 6 ==> MULDEV=5 6 acc:(0.97956)
9. HSCODE=3 5 ==> MULDEV=5 5 acc:(0.97374)
10. HSCODE=12 5 ==> MULDEV=5 5 acc:(0.97374)
11. SOFTWAREENGINEERING=2.5 5 ==> MULDEV=5 5 acc:(0.97374)
12. SYSTEMSANALYSISANDDESIGN=2.5 5 ==> MULDEV=5 5 acc:(0.97374)
13. DBASE1=1.5 5 ==> MULDEV=5 5 acc:(0.97374)
14. WD=2 5 ==> MULDEV=5 5 acc:(0.97374)
15. WD=2.25 5 ==> MULDEV=5 5 acc:(0.97374)
16. SOFTWAREENGINEERING=1.5 4 ==> MULDEV=5 4 acc:(0.96412)
17. SOFTWAREENGINEERING=1.75 4 ==> MULDEV=5 4 acc:(0.96412)
 18. SYSTEMSANALYSISANDDESIGN=2.75 3 ==> MULDEV=5 3 acc:(0.94717)
19. DBASE2=5 3 ==> MULDEV=5 3 acc:(0.94717)
20. DBASE2=1.75 3 ==> DBASE1=2.25 MULDEV=5 3 acc:(0.94717)
21. WD=3 3 ==> MULDEV=5 3 acc:(0.94717)
22. WD=5 3 ==> MULDEV=5 3 acc:(0.94717)
23. HSCODE=13 SOFTWAREENGINEERING=2 3 ==> MULDEV=5 3 acc:(0.94717)
24. SYSTEMSANALYSISANDDESIGN=1.75 WD=1.75 3 ==> HSCODE=10 MULDEV=5 3 acc:(0.94717)
25. SYSTEMSANALYSISANDDESIGN=2 WD=1.75 3 ==> MULDEV=5 3 acc:(0.94717)
```

Table-10 shows association rules generated by WEKA TOOL using Predictive Apriori Algorithm generated 100 results in the table top 25 best association rules are presented.

Rule 1: DBASE2=2 9 ==> MULDEV=5 9
acc:(0.9875)
with the top highest level of accuracy result, which suggests that students with grade of 2.0 in Database2 has 5.0 grade in Multimedia Development, with an instance of 9.

Rule 2: SYSTEMSANALYSISANDDESIGN=1.75 8 ==> MULDEV=5 8
acc:(0.98575)
Rule 2 can be interpreted as students with grade of 1.75 in Systems Analysis and Design has 5.0 grade in Multimedia Development, with an instance of 8.

Rule 6: HSCODE=10 6 ==> MULDEV=5 6
acc:(0.97956)
Rule 6 can be interpreted as students from the High School 10 LICEO DE PAGSANJAN has grade of 5.0 grade in Multimedia Development, with an instance of 6.

Rule 7. HSCODE=2 6 ==> MULDEV=5 6
acc:(0.97956)
Rule 7 can be interpreted as students from the High School 2 BANCA-BANCA NAT'L HIGH SCHOOL has grade of 5.0 grade in Multimedia Development, with an instance of 6.

In addition to this interesting association rules were generated in performing Predictive Apriori Algorithm such as:

Rule 23: HSCODE=13 SOFTWAREENGINEERING=2 3 ==> MULDEV=5 3 acc:(0.94717)
Rule 23 can be interpreted as students from the High School 13PEDRO GUEVARRA NATIONAL HIGH SCHOOL has grade of 2.0 grade in Software Engineering has a grade 5.0in Multimedia Development, with an instance of 3.

Rule 24: SYSTEMSANALYSISANDDESIGN=1.75 WD=1.75 3 ==> HSCODE=10 MULDEV=5 3
acc: (0.94717)
Rule 24 can be interpreted as students with the grade of 1.75 grade in Systems Analysis and Design and Web Development

has a grade 5.0 in Multimedia Development, with an instance of 3.

Rule 25: SYSTEMSANALYSISANDDESIGN=2 WD=1.75 3 ==> MULDEV=5 3
acc:(0.94717)
Rule 25 can be interpreted as students with the grade of 2.0 grade in Systems Analysis and Design and grade of 1.75 in Web Development has a grade 5.0 in Multimedia Development, with an instance of 3.

The researcher performed random algorithm experiments both in WEKA Tool and Rapid Miner. Apriori Algorithm at first gave an interesting rule but as the experiment continues in an iterative process the researcher executed the mining procedures using Predictive Apriori Alorithm it was found out to generate more association rules because it generate association rules even if it is low level of accuracy. As shown in Table-10 the comparison of the two Association Data Mining.

Table-11
Comparison of RULE SET Generated by WEKA TOOL using Apriori and Predictive Apriori Algorithm

=== Run information === Scheme: weka.associations. PredictiveApriori -N 100 -c -1 Relation: whatever Instances: 35 Attributes: 7 HSCODE SOFTWAREENGINEERING SYSTEMSANALYSISANDDESIGN DBASE1 DBASE2 WD MULDEV === Associator model (full training set) === PredictiveApriori =================== Best rules found: 1. DBASE2=2 9 ==> MULDEV=5 9 acc:(0.9875) 2. SYSTEMSANALYSISANDDESIGN=1.75 8 ==> MULDEV=5 8 acc:(0.98575) 3. DBASE2=2.5 8 ==> MULDEV=5 8 acc:(0.98575) 4. WD=2.5 8 ==> MULDEV=5 8 acc:(0.98575) 5. DBASE1=2.25 7 ==> MULDEV=5 7 acc:(0.98327) 6. HSCODE=10 6 ==> MULDEV=5 6 acc:(0.97956) 7. HSCODE=2 6 ==> MULDEV=5 6 acc:(0.97956) 8. DBASE1=2.5 6 ==> MULDEV=5 6 acc:(0.97956) 9. HSCODE=3 5 ==> MULDEV=5 5 acc:(0.97374) 10. HSCODE=12 5 ==> MULDEV=5 5 acc:(0.97374) 11. SOFTWAREENGINEERING=2.5 5 ==> MULDEV=5 5 acc:(0.97374) 12. SYSTEMSANALYSISANDDESIGN=2.5 5 ==> MULDEV=5 5 acc:(0.97374) 13. DBASE1=1.5 5 ==> MULDEV=5 5 acc:(0.97374) 14. WD=2 5 ==> MULDEV=5 5 acc:(0.97374) 15. WD=2.25 5 ==> MULDEV=5 5 acc:(0.97374) 16. SOFTWAREENGINEERING=1.5 4 ==> MULDEV=5 4 acc:(0.96412) 17. SOFTWAREENGINEERING=1.75 4 ==> MULDEV=5 4 acc:(0.96412) 18. SYSTEMSANALYSISANDDESIGN=2.75 3 ==> MULDEV=5 3 acc:(0.94717) 19. DBASE2=5 3 ==> MULDEV=5 3 acc:(0.94717) 20. DBASE2=1.75 3 ==> DBASE1=2.25 MULDEV=5 3 acc:(0.94717) 21. WD=3 3 ==> MULDEV=5 3 acc:(0.94717) 22. WD=5 3 ==> MULDEV=5 3 acc:(0.94717) 23. HSCODE=13 SOFTWAREENGINEERING=2 3 ==> MULDEV=5 3 acc:(0.94717) 24. SYSTEMSANALYSISANDDESIGN=1.75 WD=1.75 3 ==> HSCODE=10 MULDEV=5 3 acc:(0.94717) 25. SYSTEMSANALYSISANDDESIGN=2 WD=1.75 3 ==> MULDEV=5 3 acc:(0.94717)	=== Run information === Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1 Relation: whatever Instances: 35 Attributes: 7 HSCODE SOFTWAREENGINEERING SYSTEMSANALYSISANDDESIGN DBASE1 DBASE2 WD MULDEV === Associator model (full training set) === Apriori ======= Minimum support: 0.15 (5 instances) Minimum metric <confidence>: 0.9 Number of cycles performed: 17 Generated sets of large itemsets: Size of set of large itemsetsL(1): 25 Size of set of large itemsetsL(2): 24 Size of set of large itemsetsL(3): 1 Best rules found: 1. DBASE2=2 9 ==> MULDEV=5 9 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.51) 2. SYSTEMSANALYSISANDDESIGN=1.75 8 ==> MULDEV=5 8 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.46) 3. DBASE2=2.5 8 ==> MULDEV=5 8 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.46) 4. WD=2.5 8 ==> MULDEV=5 8 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.46) 5. DBASE1=2.25 7 ==> MULDEV=5 7 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.4) 6. HSCODE=10 6 ==> MULDEV=5 6 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34) 7. HSCODE=2 6 ==> MULDEV=5 6 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34) 8. DBASE1=2.5 6 ==> MULDEV=5 6 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.34) 9. HSCODE=3 5 ==> MULDEV=5 5 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.29) 10. HSCODE=12 5 ==> MULDEV=5 5 <conf:(1)> lift:(1.06) lev:(0.01) [0] conv:(0.29)

Conclusion

The research aims to acquire knowledge in the university academic database which is hidden and which has not been discovered yet. Specifically examine the student grades in Information Technology Education major subjects associating the high school institutions attended by the students to guide decision makers and academicians particularly the College Dean in College admission.

During the conduct of the research CRISP-DM methodology to have properly guided the researcher to complete and accomplished iterative the data mining process. Also, data mining tasks and techniques was performed to uncover hidden patterns and discover tacit knowledge from a large set of students' data. The researcher conduct experiment with different approaches such as Cross Tabular Query using SQL, descriptive statistics such as Frequencies, Mean, Median, Mode and Cross Tabulation Analysis. K-Means Clustering was also tested as well as Decision Tree using IBM SPSS. After tedious testing and experiments WEKA Tool 3.7 generated an interesting association rules using Apriori and Predictive Apriori.

Association rule algorithms are functional in generating rules for unidentified patterns. The results show that the algorithms Apriori and Predictive Apriori performed well based on the intended objective. Top ten rules generated by the Algorithms are identical and important. Most importantly Predictive Apriori generated an interesting and remarkable rules. It will help decision makers to classify the students based on the High School with respect to the academic performance of the students in ITE major subjects. Furthermore, the parameters that determine the subject area proficiency of high school are identified. This can be used for another facet of research.

Acknowledgement

The author would like to thank the University President Dr. Nestor M. De Vera, Campus Director Dr. Roberto C. Miranda and Vice President of Academic Affairs Mario C. Pasion for the support and confidence that has been given to university researchers.

References

1. Baradwaj B. and Pal S. (2011). Mining Educational Data to Analyze Student s' Performance. *International Journal of Advanced Computer Science and Applications*, 2(6), 63-69.

2. Han J. and Kamber M. et. al. (2006). Data Mining: Concepts and Techniques. 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

3. Baker R.S.J.d. and Yacef K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1), 3-17.

4. Ramesh V., Parkavi P. and Ramar K. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. *International Journal of Computer Applications*, 63(8), 0975–8887.

5. Bienkowski M., Feng M. and Means B. (2012). Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief. Office of Educational Technology, U.S. Department of Education, 1-57.

6. Romero C. and Ventura S. (2010). Educational Data Mining: A Review of the State of the Art. Systems*, Man, and Cybernetics,* 40(6), 601-618. doi: 10.1109/tsmcc.2010.2053532.

7. Ranjan J. and Malik K. (2007). Effective educational process: a data-mining approach. *The Journal of Information and Knowledge Management Systems*, 37(4), 502-515.

8. Luan J. (2004). Data Mining Applications in Higher Education. Executive Report SPSS, http://www.pse.pt/Documentos/Data%20mining%20in%20higher%20education.pdf.

9. Luan J. (2002). Data mining knowledge management in higher education potential applications. Proc. Workshop Assoc. Inst. Res. Int. Conf., 1-18.

10. Luan J., Zhao C.M. and Hayek J. (2004). Exploring a new frontier in higher education research: A case study analysis of using data mining techniques to develop NSSE National Survey of Student Engagement institutional typologies. California Association for Instituional Research, California, USA.

11. Waiyamai K. (2003). Improving Quality of Graduate Students by Data Mining. Dept. of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand http://slideplayer.com/slide/8285125/.

12. Delavari N., Amnuaisuk S.P. and Beikzadeh M.R. (2008). Data Mining Application in Higher Learning Institutions. *Informatics in Education - International Journal*, 7(1), 31-54.

13. Goyal M. and Vohra R. (2012). Applications of Data Mining in Higher Education. *IJCSI International Journal of Computer Science*, 9(2), 1,

14. Chapman et al, (2000). CRISP-DM 1.0 - Step-by-step Data Mining Guide. https://www.the-modeling-agency.com/crisp-dm.pdf.

15. Scheffer T. (2001). Finding Association Rules That Trade Support Optimally against Confidence. Proc. of the 5th European Conf. on Principles and Practice of Knowledge Discovery in Databases, 424-435, Springer-Verlag.

16. Agrawal R., Imieliński T. and Swami A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93, pp. 207-216.

17. Agrawal R. and Srikant R. (1994). Fast Algorithms for Mining Association Rules. In Proc. of the 20th Int. Conf. on Very Large Data Bases, 475–486, Morgan Kaufmann.

18. Witten I.H. and Frank E. (2005). Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.

19. Khattak A.M., Khan A.M., Lee S. and Young Koo (2010). Analyzing Association Rule Mining and Clustering on Sales Day Data with XLMiner and Weka. *International Journal of Database Theory and Application*, 3(1), 13-22.