An Early Detection System for Incomplete Application in Master's Degree program at University level

التنبوء المبكرللطلبات الغير مكتملة في برنامج درجة الماجستير على مستوى الجامعة

By

Manju Vishnu Sankar

2013210252

Dissertation submitted in partial fulfilment of the requirements for the degree of MSc Information Technology Management

Faculty of Engineering & Information Technology

Dissertation Supervisor
Dr. Sherief Abdallah

June-2016

DISSERTATION RELEASE FORM

Student Name	**Student ID**	**Programme**	**Date**
Manju Vishnu Sankar	2013210252	MSc Information Technology Management	22. June. 2016

Title: An Early Detection System for Incomplete Application in Master's Degree Program at University level

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

Electronic Submission Copyright Statement
Please choose one of the following two licenses and check appropriate box.

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation worldwide including the users of the repository, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Public access to my dissertation in the Repository shall become effective:

☒ Immediately
☐ 12 months after my submission
☐ 24 months after my submission
☐ 48 months after my submission

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation to students, faculty, staff and walk-in users of BUiD Library, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Signature
Manju Vishnu Sankar

Abstract

Selecting a university to pursue higher education is a very difficult and expensive decision that a student can make. A lot of research into student decision-making in this area has been conducted, but the work has been focusing mostly on students' dropout rate after registrations or before completion of the course using surveys, general market trend and also based on other existing research work on this area.

This paper address the major issue of student incomplete application rate that is faced by a higher education institute based in Dubai. By analyzing the past year student dropout records from the university database, this research intends to build a Incomplete Application Prediction Model (IAPM) and in turn come up with a strategy that can help university to achieve application completion rate.

Early identification of these incomplete applications is as important or in this case can be considered more important than student marketing from the university point of view as retaining prospective students who already applied are easier if detected early as details of the students are already available. Moreover in the university for which the study is undertaken, records show an alarming rate of around 16.25% student incomplete applications after enrolling which urges the need for this study.

Various data classification techniques as well as association rules were applied on all the attributes and also on selective attributes that were obtained from the university's original database for research purpose. An Incomplete Application Prediction Model is developed from these techniques to aid student retention for master's degree courses in a specific university. This model can be further customized for other universities in the region if needed.

The results were positive and indicated that past incomplete students' records can be a valuable resource for mining the near accurate reason for students' incomplete application, which in turn can give the management a clear insight for proactive solving of such issues in future intake. Data visualization in Weka offered interesting insights on these data also. By focusing on antecedents of incomplete students'

records, colleges can restructure their strategies for a better student-supportive system. Smaller sample size and the self-explicated data are some limitations of this research work.

Keywords: master's degree, incomplete applications, Incomplete Application Prediction Model, student retention strategy, data classification, association rule.

الخلاصـــــــــــة

يعدّ اختيار الجامعة لمواصلة التعليم العالي قراراً صعباً جداً ومكلفاً بالنسبة لأيّ طالب. وقد أجريت الكثير من الأبحاث حول قيام الطالب باتخاذ مثل هذا القرار إلا أنها تركزت في معظمها على معدّل تسرّب الطلاب بعد التسجيل أو قبل الانتهاء من الدورة وذلك عن طريق إجراء المسوحات ودراسة التوجّه العام للسوق كما استندت الأبحاث على العديد من الأعمال البحثية المتوفرة في هذا المجال.

أبرز المواضيع التي تتناولها هذه الورقة هي معدّل الطلبات غير المكتملة التي يقدمها الطلاب لدى أيّ معهد للتعليم العالي في دبي. من خلال تحليل سجلات تسرّب الطلاب في السنة الماضية والتي تمّ الحصول عليها من قاعدة بيانات الجامعة، يهدف هذا البحث إلى بناء نموذج للتنبؤ بالطلبات غير المكتملة (IAPM) وبالتالي التوصّل إلى استراتيجية يمكن أن تساعد الجامعة في تحقيق معدّل إتمام الطلبات المستهدف.

إنّ التحديد المبكر للطلبات غير المكتملة لا يقلّ أهميةً أو في هذه الحالة يمكن اعتباره أكثر أهميةً من تسويق الطلاب من وجهة نظر الجامعة إذ أنّ الاحتفاظ بالطلاب المحتملين الذين تقدموا بالفعل يصبح أسهل إذا تمّ اكتشاف ذلك في وقت مبكر وكانت بيانات الطلاب متوفرة أصلاً. وعلاوة على ذلك، ففي الجامعة التي جرت فيها الدراسة، تظهر السجلات أنّ معدّل الطلبات غير المكتملة بعد الالتحاق بلغ حوالي 16.25% وهي نسبة تنذر بالخطر مما دعا إلى ضرورة إعداد هذه الدراسة.

تمّ تطبيق تقنيات تصنيف البيانات المختلفة إلى جانب قواعد تحليل الروابط بين البيانات على جميع الصفات وأيضاً على صفات مختارة تمّ الحصول عليها من قاعدة البيانات الأصلية في الجامعة لأغراض البحث. تمّ وضع نموذج التنبؤ بالطلبات غير المكتملة استناداً إلى هذه التقنيات لتعزيز القدرة على الاحتفاظ بالطلاب لدورات درجة الماجستير في أيّ جامعة معينة. يمكن تكييف هذا النموذج بما يتوافق مع احتياجات الجامعات الأخرى في المنطقة إذا لزم الأمر.

أشارت النتائج الإيجابية إلى أنّ سجلات الطلاب السابقة غير المكتملة يمكن أن تشكّل مصدراً قيماً للتنقيب بدقة عن السبب وراء عدم إتمام تلك الطلبات وهذا بدوره يمكن أن يعطي الإدارة نظرة واضحة عن أيّ حلّ استباقي لمثل هذه المشاكل في المستقبل. قدمت تقنية التمثيل البصري للبيانات في WEKA " أفكاراً مثيرةً للاهتمام حول هذه البيانات " أيضاً. من خلال التركيز على سجلات الطلاب غير المكتملة السابقة، يمكن للكليات إعادة هيكلة استراتيجياتها لتعزيز نظام دعم الطلاب لديها. إلا أنّ صغر حجم العينات والبيانات الواضحة التي لا تحتاج لشرح هي من ضمن القيود التي أثرت على هذا العمل البحثي.

كلمات مفتاحية: درجة الماجستير، الطلبات غير المكتملة، نموذج التنبؤ بالطلبات غير المكتملة، استراتيجية الاحتفاظ بالطلاب، تصنيف البيانات، قواعد تحليل الروابط بين البيانات

Acknowledgements

Foremost, I am so thankful to God for giving me the chance to carry on my study and making me the person who I am.

My sincere thanks to my supervisor and guide Dr. Sherief Abdullah for his valuable support and guidance throughout all the stages of this dissertation.

Also, I owe my deepest gratitude to my family and friends. I owe them all my academic and personal achievements. Their support and love were the main motivates for me to continue this masters study and their belief on my abilities is the key factor of my success.

Declarations

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Manju Vishnu Sankar*)

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Overview about thesis:

The educational system in Dubai shadows the UAE system, at the same time is distinctive in accommodating a range of diverse higher education systems comprising Indian, Canadian, German, British, Pakistani, American, French, Australian, Pilipino, Russian and obviously that of the United Arab Emirates. The mode of instruction in most or all of the private universities in Dubai is in English because of their focus mainly on different expatriate communities. In most of the countries, institutions have independent internal governance. There are deviations amongst universities within Dubai in the types of courses and the level of degree program that is offered. Universities in Dubai can be categorized as either a local university that is founded and based in Dubai or a branch campus founded in another country, with a campus located in Dubai. Dubai at present accommodates over 60 universities. There exists heavy competitions due to which apart from enrollment declines, some of the major problem most of these Universities face are the students' turn down rate either immediately after applying for a specific course or most of the cases, before completion of final thesis work for the course. This situation is predominant in the university we had considered for our case study. Also this is a universal case as indicated by (DeShields et al., 2005). The findings of this paper call in for more attention from the university on student retention strategies as soon as a student apply for a course rather than student enrollment strategies due to the huge number of student incomplete applications. There can be several reasons for a student not selecting a specific university to pursue his or her studies. Some predominant reasons as specified in most of the successful research papers are university accreditation, teaching and learning practices, location of the university, university facilities, environment, current students, social life, tuition fee, cost of living, language etc. Our study focus on the Early Detection of most prominent factors that can help universities to predict students who might decide not to join after applying for a course.

1.2 Aims and Objectives

There may be several reasons due to which a student may not complete the application for joining university. Will following up with some of these students result in successful and completed application? Which of the students with incomplete applications will most likely not continue in the university? How do we target such students?

Our study is formulated to address these issues. The main objective of this study is to explore those factors that can be focused on to retain students with incomplete application at higher educational institutions. In order to decrease the student incomplete application rate especially as the situation faced in masters degree programs in the university for which this research is undertaken, an attempt is made to develop a 'Incomplete Application Prediction Model' based on the past student application records of this particular university. Various classification techniques as well as association rules are used to achieve the best prediction model in this research. This prediction model may allow the university management to develop a focused strategic plan for retaining the students with incomplete applications. A standard strategic plan for student retention is also recommended in this paper for the specific university. Other universities according to their requirements if needed can customize the same.

1.3 Research Questions

The research paper targets to answer the following questions

- What attributes can best contribute for designing the Incomplete Application Prediction Model (IAPM)
- Which classification techniques are suitable for building IAPM
- How to formulate a standard strategy plan for student retention (SPSR) with respect to the prediction model obtained

1.4 Structure of the Thesis

The next section focuses on the literature review of similar prediction models and various classification techniques if any used for this purpose. The section following literature review elaborates on the data used, how the data is refined for experimental purpose and various criteria applied on the data for research purpose. Section 4 is about the experimental setup for the research, which is followed by the analysis, conclusion and future work.

1.5 Uniqueness of this Research

Past researches focuses majorly on retention of students after joining the university and attended classes. In such case, the students get a chance to experience the university facilities and classroom experience. But our research focus on the students have just applied and not joined the course. The information we have on hand and tend to use is about the students who only applied and have not yet joined the university. Here the students' did not have a chance to attend the courses or experience the university campus life at all. This research is unique in the sense that the past year students' incomplete application data records of the university were used as the major resource for predicting students' incomplete application in order to design the Student Incomplete Application Prediction Model (IAPM). Through these results, we intend to develop a standard Strategy Plan for Student Retention (SPSR), which may help the university to retain these students who might not want to join the university.

2. Literature Review

2.1 Overview of research findings:

Due to the vast number of colleges and universities in the UAE, there exists a heavy competition of attracting students, which ultimately forces these universities and colleges to focus on various student enrollment strategies. Even though there is no doubt that the development of students' education along with successful completion of his/her course are the major reason for survival of higher educational organizations, the university management still tend to pay more emphasis on strategies for attracting and admitting students instead of developing strategies for retaining existing enrolled students. Satisfying enrolled students is as important for any university as companies that strive to satisfy their customers in order to retain them (DeShields et al., 2005). As indicated by (Rojas-Méndez et al., 2009), from the measurements led in American schools, about 40 percent of undergrads leave advanced education without acquiring a degree, around 75 percent of undergrads leave university in the initial two years. (Reisberg, 1999) Shows that over 26.4 percent of fresh year students are not returning for the following fall semester and 46.2 percent of freshman does not pass out from college. (Kemerer et al., 1982) Also demonstrates that in universities that have large number of students who travel have greater dropout rates contrasted with organizations accommodating students who stay in the university premises and study, which clarifies the fact that discontented undergrads may slash down on the courses or even leave college completely. The management should thus assess retaining students in higher education more sensibly.

Apart from the focus on components of student satisfaction, (DeShields et al., 2005) also looks in to the retention aspect in a college or university that may influence students' college experience. Findings analyzing student contentment in higher educational organizations from client point of view add an extra element to the educational strategies of colleges and universities. According to (DeShields et al.,

2005) drawing students, administering their applications, and directing students who are admitted are exceedingly important tasks. While (Kotler and Fox, 1995) contends that undergrads must be treated as accomplices starting from day one of enrolling until the student graduates so as to enhance their experience at the university. Thus he affirms that performance of academic staff, counselors along with students' classroom experience are some of the vital factors that impact students' experience and overall contentment in college life. This fulfillment at long last impacts undergrads goals whether to stay at or leave the organization. (Greenhaus and Parasuraman, 1986) States that satisfactory level of the client is dictated by the distinction between the client's desires and administration execution as experienced by the client. As pointed out by (Aarinen, 2012), the overall competition in the market of higher education institutes has prompted a circumstance where establishments go up against different foundations all around the globe. This has constrained establishments to perceive that they may need to market themselves diversely in this atmosphere. Just like the significance of fulfilling clients to hold them in revenue making establishments, fulfilling accepted undergrads is likewise critical for any university as indicated by (Anderson and Sullivan, 1993). However, retaining undergrads is one of the pointers that higher educational finance associations and other educational institutes believe is a measure of the quality offered by educational organizations as remarked by (Lykourentzou et al., 2009).

After a widespread literature study by (Tharp, 1998), just the background information taken for predicting the dropout cases did not have a positive result in the situation of students who are enrolled as regular and full-time. Where as the background information was meaningful for students who opt to study through distance education or open education where social integration and organizational commitment are not crucial in the student experience.

(Nandeshwar and Chaudhari, 2009) Affirms that higher education organizations today confront the problem of retaining students that is related to graduation rates.

Those colleges that have higher student freshman retention rate also manage to have higher graduation rates. The average nationwide retention rate is around 55% and for some cases it is less than 20% of incoming student graduates and over 50% of students enrolled in engineering course leave even before they graduate. The paper also indicates the significant amount of loss of income by some of the universities due to the high rate of students who do not continue their university education.

(Kovacic, 2010) Explores the socio-demographic factors like gender, age, education, disability, work status and ethnicity along with study atmosphere, which are courses taken, and the course block that may impact persistence or students turnover at the Open Polytechnic of New Zealand. The results from this paper show that the most of the crucial factors separating successful from unsuccessful students are ethnicity and course block. Classification and Regression Tree (CART) used for study gave the best successful result with 60.5 percentage of correct classification. The paper also emphases that early identification of susceptible students who are prone to drop out from their enrolled courses is vital for any retention strategy success. This can permit educational organizations to take timely and pre-emptive actions. When the 'at-risk' students are identified, they can be targeted with the help of academic and also administrative strategies to boost their chance of continuing the program.

In our scope of study, attributes such as disability and work status is not available since these fields are not mentioned in the application form while a few attributes focused on the demographic information is available though. The only noteworthy factors were ethnic origin and other features such as the program and course. But these features were found not quite successful in classifying the students 'at-risk'. The paper states that background characteristics could be significant initially, but when factors related to the academic performance were included in the classification model they dived down on the rank list of significant factors while detecting study outcome from students who dropped out of the course. Our study is

solely depending on these background information as this is the only source of information the university has on the candidates who have applied that is going to be used for making early decisions based on the results from our study.

(Vergolini and Zanini, 2015) Points out that working part-time while doing higher education is usual in many OECD (Organization for Economic Co-operation and Development) countries and permits students to increase their incomes. The paper also states that average OECD student in 2003 was employed about 27% of the time during their higher education. It is clear from the studies that although the chance to work during university studies may improve their standard of living, there are also possibility of increased risks of students' dropping out of the courses in between or elongation of their studies. Henceforth, education policy makers should consider evaluating the association between the financial aid system and students' higher education as an important factor. Here in our paper we are unable to focus on the financial part as the study is done purely based on the information available on the application form. The current design of the application of the university under study does not have any provision to know the financial status of the enrolling student. If this attribute had been available, it could have been a very strong attribute for prediction as most of the research in this area have proven that financial situation is one of the main attribute that aid in students' decision making.

Since our research data does not have this particular information, the study tries to extract the best available attribute from the background information for decision making.

2.2 Classification Algorithms:

Classification is a data mining technique that represents data into pre-defined classes. Classification can be categorized as supervised type of learning. Classification is a two-fold step in which model construction is the first step where there is a set of predetermined classes and each record is assumed to fit in to a pre-defined class. The record sets that are used for model construction is called as the training set. In the second step, this constructed model is used for classifying unknown entities. Accurateness is determined from the percentage of test set samples that are correctly classified by the model. Here the test set and training set are completely different entities used to predict the maximum accuracy. The classification algorithms basically follow the process below for prediction.

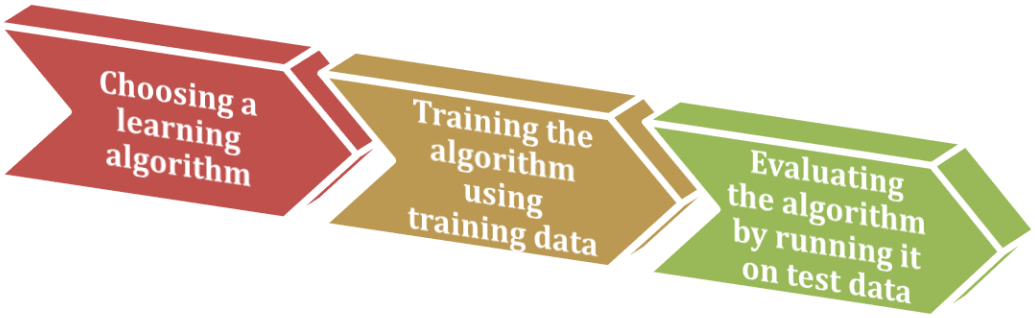

Figure 1: Machine Learning Task process (Sabnani, 2008)

Following is the brief introduction of each classification algorithm used for our experiment purpose.

2.2.1 Simple CART algorithm:

Simple CART classification technique creates a binary decision tree. The output for this technique is binary tree that generates two child nodes. Here entropy is used to select the best splitting attribute. The benefit of Simple CART is that, it can also handle missing data by overlooking that record. The Simple CART algorithm is said to be best suitable for training data (Dunham, 2006).

2.2.2 Naïve Bayes:

Naïve Bayes classification is also categorized as supervised type of learning and is based on Bayes rule of conditional probability. This algorithm makes use of all the attributes contained in the data file and individually analyses each and every attribute as if they are equally significant and independent of each other.

2.2.3 K Nearest Neighbor:

K Nearest Neighbors is the easiest compared to all other machine-learning algorithms used to classify data. In this algorithm all available cases stored and new cases are classified based on the measure of similarity. Here, an object is classified by a majority vote of its neighbors. The object is then assigned to the class that is most common among its k nearest neighbors. K is usually an insignificant positive integer. If the value of K is 1 then the object is assigned directly to the class of that single nearest neighbor.

2.2.4 J48 algorithm:

A decision tree is a machine-learning model used for prediction that decides the target value of sample test data based on several attribute standards of the existing data used as training set. Inner nodes of the decision tree signify the various features while the branches amid the nodes denotes potential values that these

attributes can contain in the perceived sample data. The terminal nodes denote the target value of the variable that is dependent. The predicted attribute is known as the dependent variable, because its value depends on the values of all the other attributes. The other attributes that help in prediction of the value of the dependent variable are identified as variables that are independent in the given dataset.

2.2.5 NBTree:

The NBTree system is a mixture of Naive Bayes and Decision-tree classifiers. This algorithm denotes the trained information in the visual form of a tree, which is built recursively. Here, the end nodes are Naive Bayes categorizers rather than nodes identifying a single class as denoted by (Sabnani, 2008). A maximum limit is chosen for continuous attributes so as to limit the entropy measure. Discretizing the data and computing the fivefold cross-validation accuracy estimation using Naive Bayes gauge the effectiveness of a node. The effectiveness of the split is the weighted sum of effectiveness of the nodes and this depends on the number of occurrences that go past that node. The NBTree algorithm attempts to estimate if the generalization accurateness of Naive Bayes at each leaf is greater than a single Naive Bayes classifier at the present node. A split is assumed to be important if the comparative decrease in error is larger that 5% and there are minimum of 30 occurrences in the node (Sabnani, 2008).

2.3 Apriori- Association Rule Learning:

In data mining as well as computer science, Apriori is considered as a classic algorithm for association rule learning. Apriori is intended to function on databases that contains transactions. According to (Kavšek and Lavrač, 2006), Apriori algorithm is widely studied amended to other areas of data mining as well as machine learning and is also efficaciously applied in several problem areas. An association rule has the form A -> B, where A and B are item sets and are also subsets of I, which is the set of all items containing of a set of transactions. In the

standard machine-learning lexicon, transactions resemble to the training examples, an item is a binary feature, and item sets are conjunctions of features.

2.4 About WEKA Software:

For our research purpose we are using WEKA software. WEKA software has a wide collection of machine learning algorithms for classifying and clustering tasks in data mining. The algorithms in WEKA can either be applied directly to the dataset or can be called using Java code. WEKA comprises of tools for data pre-processing, classification, regression, clustering, association rules and visualization. WEKA is also suitable for developing new machine learning schemes. The software reads data form ".arff" files as input. The native file format for WEKA is ARFF data file format but WEKA can also read ".csv" file formats. This is an advantage because many databases applications and spreadsheet applications can save or export data into flat files in .csv format (Hall et al., 2009).

3. Research Design:

The research is split into a six-phase design model starting from dataset collection, data analysis, and data preparation, acquiring the predictive model and its implementation, result evaluation and recommendation. The details as depicted below:

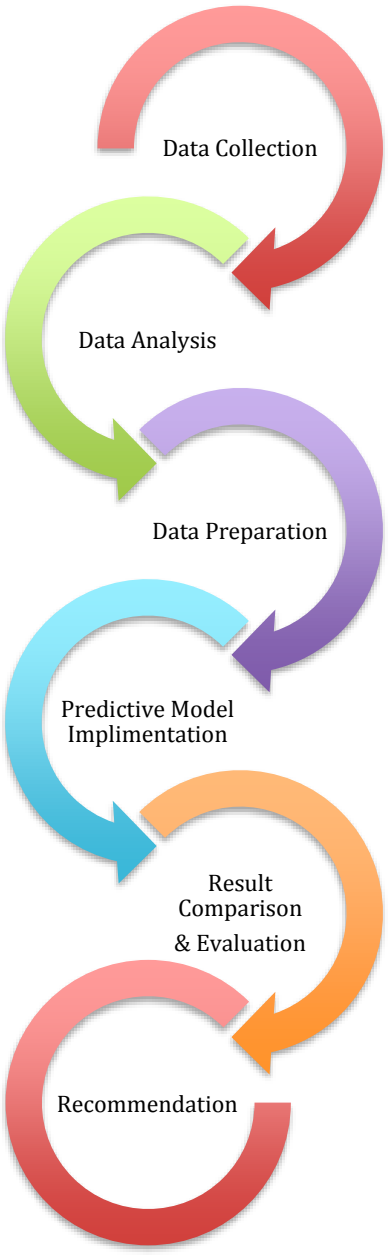

Figure 2: Design Phases

3.1 Dataset Collection

Two files in which one file had the details of 'registered students' and the other with details of drop out students with 'incomplete' status were received from the university for which the research is undertaken. The total number of data records we received originally from the university after removal of duplicate records is 1009. After the preliminary cleaning process, the number of records available for research is 997. Out of which 161 records are details of students who dropped out of the course classified as 'incomplete' and the remaining classified as 'registered'. This shows an alarming rate of around 16.25% student incomplete application after enrolling. The crucial information such as enrollment year and names of the students are not revealed on the original file due to data restrictions and university policy.

3.1.1 Dataset Description

Attributes: ID, Gender, Emirate Name, Nationality, Resident Country, Degree Acquired, Awarding Institute, Subject Major, Final Percentage, TOFEL Taken, TOFEL Scores, IELTS Taken, IELTS Score, Program Code, Stream Name, Study Mode, Semester Name, Status Description.

3.1.2 Details of each attribute is as follows:

i. ID: Nine-digit integer number generated by the university software.

ii. Gender: Male or Female. Total of 464 female and 532 male candidates out of which 66 female students and 95 male students are of incomplete status.

iii. Emirate Name: Emirate of residence whether Dubai, Abu Dhabi, Ras Al khaimah, Umm Alquin, Fujairah, Sharjah, Ajman. Here majority of the students in the record file shows are residing in Dubai. Following table shows the detailed information of students from different emirate.

Place of Residence	Total No. Of Students	No. Of Students With Incomplete Status
Dubai	423	33
Abu Dhabi	263	41
Sharjah	99	8
Others	68	28
Fujairah	46	22
Ajman	35	0
Ras Al Khaimah	32	4
Not filled	26	0
Umm Al Quwain	4	0

Table 1: Comparison Of Total No. Of Enrolled Students With That Of Incomplete Students

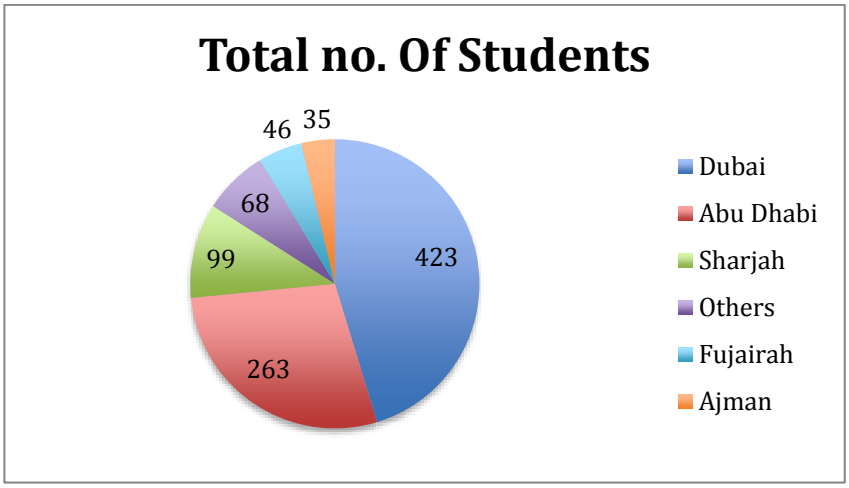

Figure 3: Majority Residence of Enrolled Students

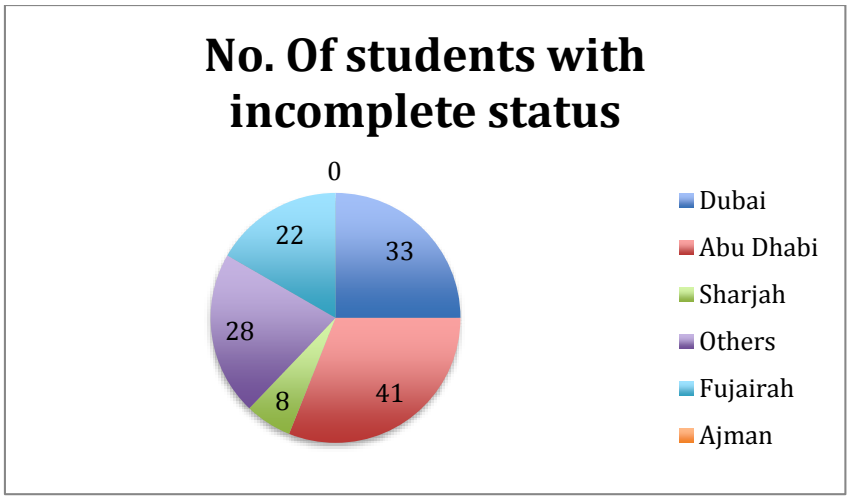

Figure 4: Majority Residence Of Students With Incomplete

iv. Nationality: Denotes the nationality of student

Nationality	**Total No. Of Students**	**No. Of Students With Incomplete Status**
Emirati	273	20
Jordanian	149	10
Egyptian	122	15
Indian	51	11
Syrian	45	9
Palestinian	43	6
Iraqi	36	5
Pakistani	33	16

Lebanese	28	6
Not filled	25	25
Omani	23	3
Iranian	22	3
Sudanese	18	1
British	14	2
American	11	2
Canadian	12	1
Algerian	8	2
Nigerian	7	3
Saudi	6	1
Tunisian	6	2
Yemeni	5	0
Bahraini	4	1
Bangladeshi	3	1
Australian	2	1
Cameroonian	2	2
Moroccan	3	2
Zimbabwean	2	1

Kuwaiti	2	1
Filipino	3	1
Singaporean	3	2
Indonesian	2	1
French	3	1
Azeri	1	1
Chinese	3	1
Turkish	2	1
Sri lankan	1	1
Uzbek	2	0
Somalian	3	0
Kazakh	1	0
Libian	2	0
New Zealander	1	0
Russian	2	0
Spanish	1	0
Albaninan	1	0
Ghanaian	1	0
Irish	3	0

Braziallian	2	0
Comoros	2	0
Seychellian	1	0
Polish	1	0
Romanian	1	0
Mexican	1	0

Table 2: Comparison Of Total Number Of Students With That Of Incomplete Students From Different Nationality

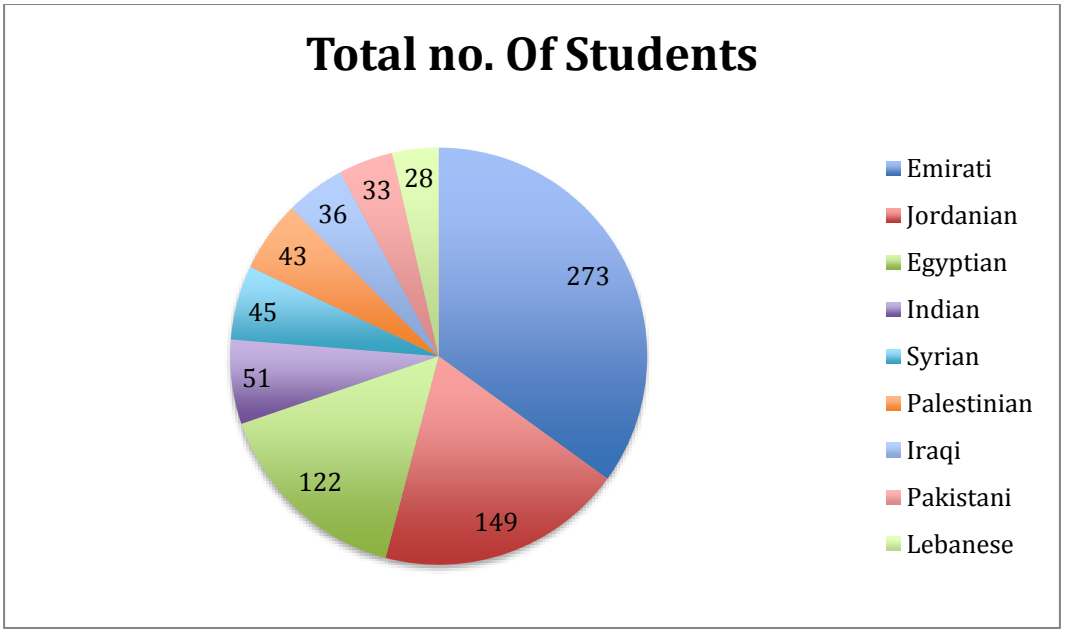

Figure 5: Majority Of Students From Different Nationality

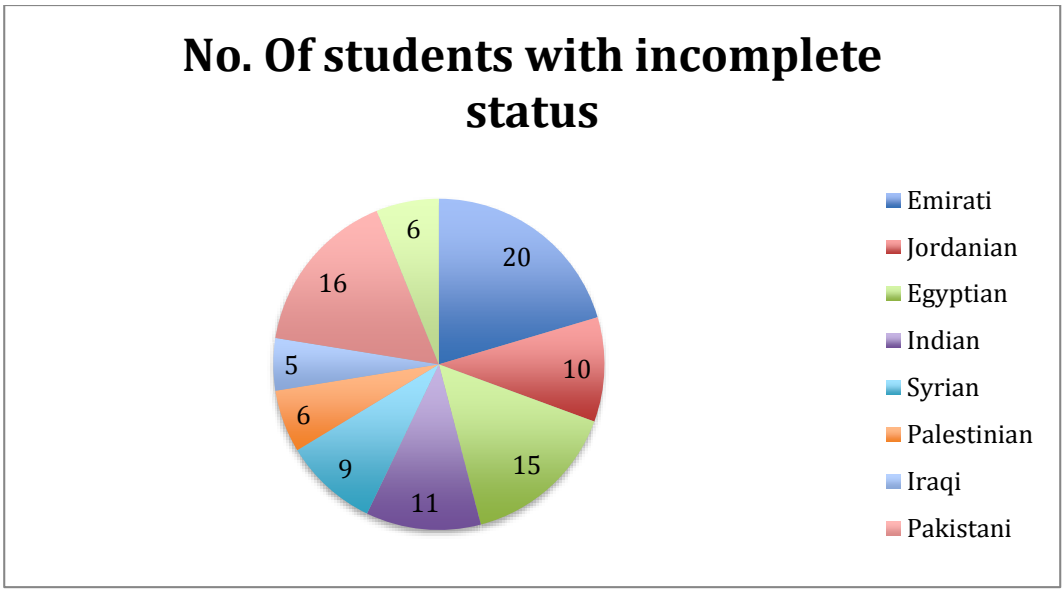

Figure 6: Majority Of Students With 'Incomplete Status' From Different Nationality

v. Resident Country: Country of student's residence

Residence Country	**Total No. Of Students**	**No. Of Students With Incomplete Status**
United Arab Emirates	770	82
Algeria	2	2
Australia	1	1
Azerbaijan	1	1
Bahrain	4	1

Canada	3	1
Egypt	8	3
India	7	3
Iran	5	3
Jordan	4	0
Libya	2	2
Malaysia	1	1
Morocco	1	1
Nigeria	6	3
Not Filled	126	26
Oman	19	4
Other	1	1
Pakistan	10	10
Qatar	1	1
Saudi Arabia	9	5
Sudan	1	1
Tunisia	2	2
Turkey	1	1
UK	4	3

US	1	1
Cameroon	1	1
Indonesia	1	0
Kuwait	1	1
Lebanon	1	0
Panama	1	0
Uruguay	1	0

Table 3: Comparison Of Total Number Of Students With That Of Incomplete Students From Different Country Of Residence

vi. Degree Acquired: Student's last degree - Bachelor degree or Master degree

vii. Awarding Institute: Name of awarding institute for student's last degree

viii. Subject Major: Major subject taken by the student in previous degree

ix. Final Percentage: Final GPA secured by the student in his/her bachelor's degree

x. TOFEL Taken: Indicated by '1' if taken and '0' if not taken.

xi. TOFEL Scores: Numeric value.

xii. IELTS Taken: Indicated by '1' if taken and '0' if not taken.

xiii. IELTS Score: Score in scale of 1 to 10

xiv. Program Code: Code issued by BUiD for the enrolled Master's program scheme

Program Code	**Total No. Of Students**	**No. Of Students With Incomplete Status**
MEd	192	30
MInf	14	12

MPM	146	24
MFB	38	18
MITM	41	10
SDBE	86	6
MCLD	79	9
EdD	90	19
MSYS	21	3
MHRM	24	7
IBDA	12	1
PMD	62	9
CPDEDU	2	1
PGDHRM	2	2
MINFO	20	1
MCM	17	2
MBA	94	5
PHD ASBE	5	1
PHD CS	12	1
CPD – INFORMATICS	2	0
PMFB	5	0

CPDPM	5	0
CPDF&B	2	0
PGDPM	1	0
CPD-PHD	1	0
ASU	1	0
CPDSYS	1	0
CPD-CLDR	1	0
MSTRE	7	0
CPD CM	1	0
CPD-STRE	3	0
MEM	4	0
CPDHRM	1	0
CPD MBA	2	0
CPD ITM	1	0
PGD-E	1	0

Table 4: Comparison Of Total Number Of Students With That Of Incomplete Students Enrolled In Different Courses

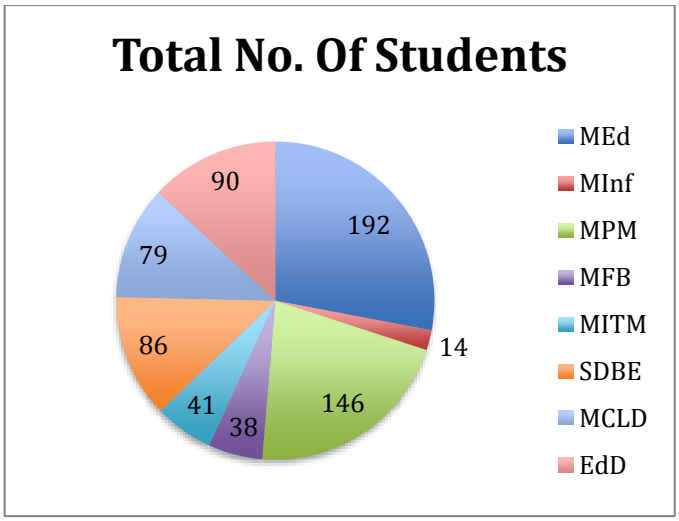

Figure 7: Majority of Courses Enrolled By Students

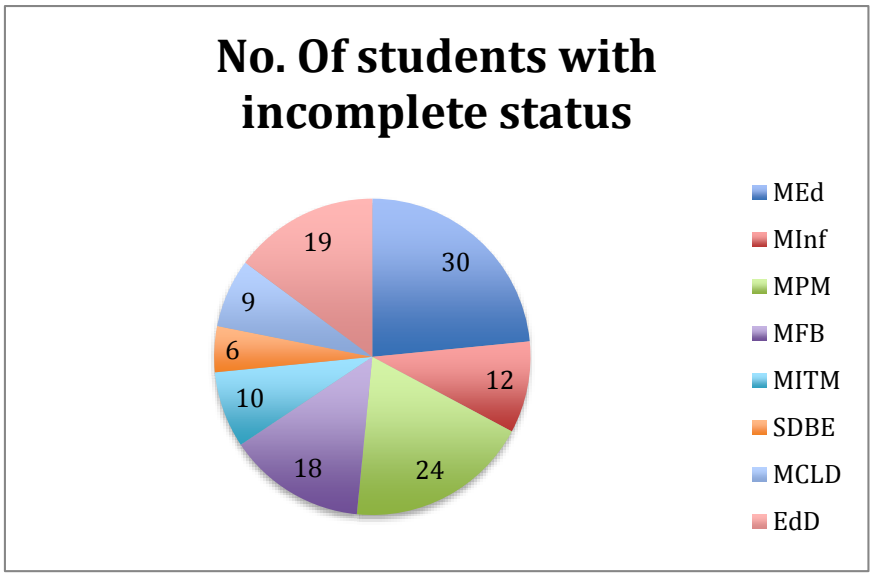

Figure 8: Major Courses Showing Number Of Students With 'Incomplete Status'

xv. Stream Name: Specific Course name

xvi. Study Mode: Full-Time or Part-Time. A total of 195 students had enrolled for full-time and 801 students enrolled for part-time. From the full-time students data, a total of 85 students are incomplete and 76 students enrolled are registered. Out of the 801 students enrolled, 76 students are incomplete and the remaining is registered.

xvii. Semester Name: January/ April /June /September.

Semester	Total No. Of Students	No. Of Students With Incomplete Status
January	473	70
April	87	0
June	14	0
September	421	91

Table 5: Comparison Of Total Number Of Students With That Of Incomplete Students Enrolled In Different Semesters

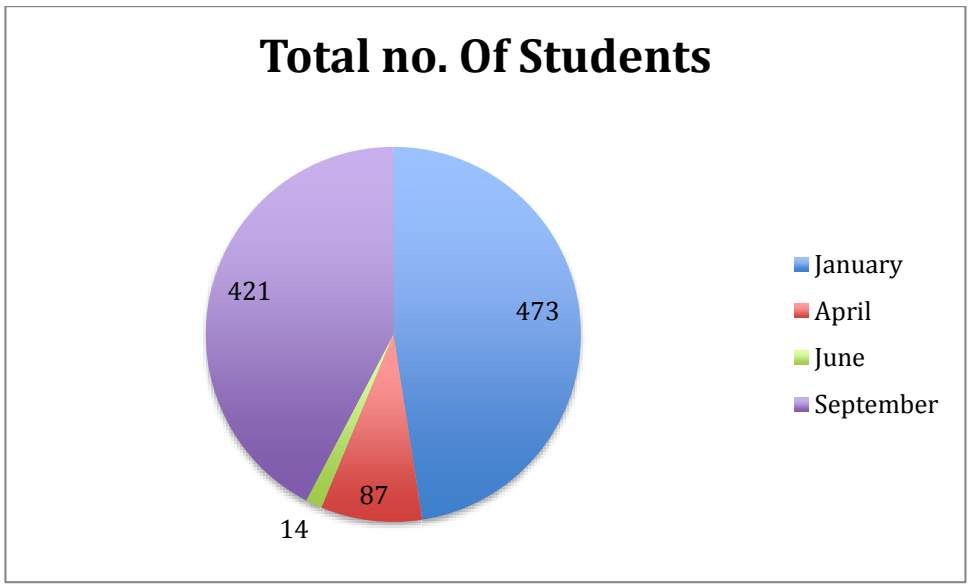

Figure 9: Students Enrollment Per Semester

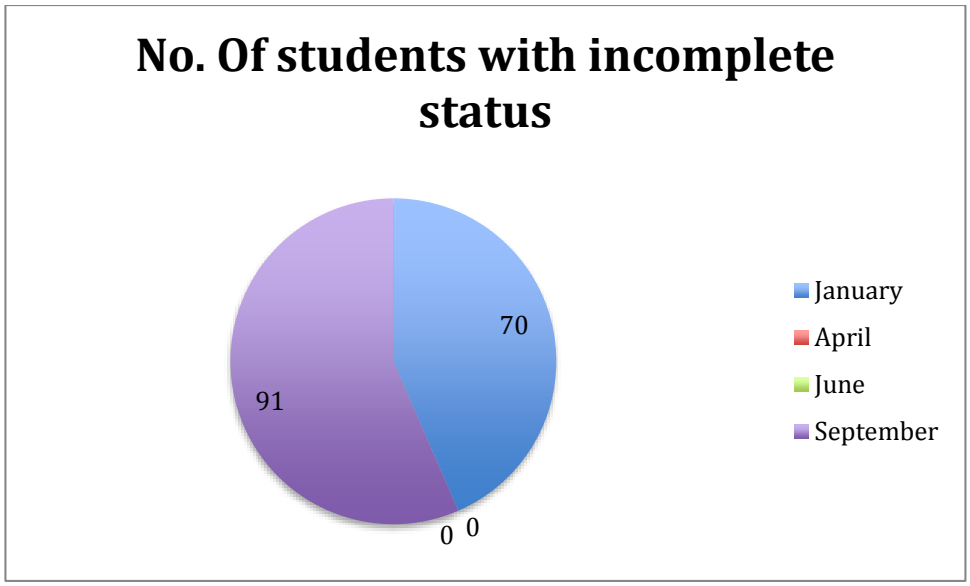

Figure 10: Incomplete Students' Enrollment Per Semester

xviii. Status Description: Also the prediction class, labeled 'incomplete' or 'registered'. Total of 161 records have the status 'incomplete' and 835 students have the status 'Registered'.

3.2 Data Preparation:

The original files obtained from University were in the form of spreadsheet with Microsoft Excel (XLS) format. WEKA software required these original files to be converted to Comma-Separated Values CSV format for analyzing. CSV files store plain text as a series of values separated by commas in form of rows. CSV files can be opened in a text editor in an easily readable form.

Two files of data were obtained from the university. One file is with information about students who registered for the course and completing the studies. The other file is with details about students who enrolled but did not join the course. Both the files initially had different attributes. For experimental purpose, the files have to be

merged that required deletion of dissimilar attributes. Final merged file contains 18 attributes. In some of the fields, null values were dominant. These fields were kept for experiment without removing since we wanted to analyze the result for the original data without much modification. Many fields with date were removed as not much prediction was expected with this field. Since our main intention is to predict whether the student will continue or not, the last column is designated as the field for classification purpose. Original file contained hundreds of duplicate records, which were all removed as the first step of pre-processing. Some of the attributes had extreme values such as Final Percentage where some students had entered their average marks as percentage and some had entered their GPA. This had to be normalized so as to provide a meaningful input for the software.

3.2.1 Adding Filters:

Fields with 'null' values were replaced with an arbitrary value for some cases using WEKA's 'AddValues' filter. Another simple way is to remove the records with 'null' values so as to get meaningful results from WEKA but we replaced the a uniform arbitrary value in order to get the results without much modification of the original file acquired from the university itself. Alternate plan was also to test the same set of data by removing the records with 'null' value and compare the prediction results. Also original data file contained symbols like apostrophe, which were sometimes entered as a part of data when students entered the names of the universities awarded their bachelors degree. Care was taken to remove these special symbols since it is considered invalid data by WEKA software. Some classification technique works only with nominal attributes. By using WEKA's Numeric to Nominal attribute filter, numeric attributes in the original file is converted to nominal form in order to be classified with that particular technique. This process not only enables the use of very useful classification techniques but also decreases the time taken to classify.

3.2.2 Feature Selection:

Feature Selection is mainly done to improve accuracy and decrease training time taken by the respective classification techniques. The WEKA attribute selection through filter and wrapper proved futile for this data file since the software couldn't display the most relevant attribute. Manual feature selection didn't contribute either. So in this case original values were kept as such.

3.3 Predictive Model Design:

As a basic thumb rule, 80% of continuing students and students' incomplete application records should be allocated for training purpose and 20% of the same should be allocated for testing purpose. This requires around 200 records of continuing students and 33 records of incomplete students to be used for testing purpose. Remaining 797 records of continuing students and 128 records of incomplete students to be used for training purpose. So 33 random records were selected from incomplete students' records and 200 random records were selected from continuing students' records for testing file.

The classification algorithms used in our experiment for prediction are CART, Naïve Bayes, KNN, J48 and NBTree.

Naïve Bayes algorithm has the capacity to also analyze files with missing values. Since the original file obtained from the university contained many missing values in various attributes since most of the students chose not to answer most of the fields in the enrollment form for the university. Naïve Bayes is popular for its accuracy and can handle files with missing values.

All the above-mentioned classification and prediction algorithms can be used with a variety of test options in WEKA. For our research purpose we have used three of the following test options.

4. Model Evaluation

4.1 Test Options:

- Training Set: In this option, one part of the file loaded in preprocess is used as training set and other part of it is used for testing purpose.
- Cross Validation: In this option, the classification results will be evaluated by cross validation with 10 folds sampling technique. This mode allows us to change the number of folds. Most of the research on prediction relies on cross-validation techniques for accuracy on prediction performance. Cross validation is a resampling technique, which makes use of several random training and subsamples. Advantages of cross validation technique are that all records in the given sample data are used for testing and most of them are also used for training model. The cross validation analysis will generate valuable perceptions on the reliability of machine learning with respect to variation in the samples.
- Percentage Split: While using the built-in percentage split function, the default settings recommends around 66% of the total records for training purpose and the remaining for testing purpose for getting closer to accurate in predicting results. So from 997 student records, 339 random records were selected for testing purpose and the machine learning software utilized the remaining 658 records for training purpose.

4.2 Performance of Predictive Models:

Performance of the classification techniques used is depicted in the form of confusion matrix by WEKA. Confusion matrix is a predictive analytics table with dual rows and columns that gives the number of false positives, false negatives, true positives, and true negatives. False negatives (under-prediction) and false positives (over-prediction) are the types of error possible in any predictive models. The relative percentages of these errors are conveyed in the form of a confusion matrix, or error matrix (Fielding and Bell, 1997). Confusion matrix contains four elements

as shown in table 5 below. Element 'a' represents known 'incomplete status' where the student did not join the course that is correctly predicted as 'incomplete', and 'd' reflects 'registered status' where the student is a continuing student and that are classified by the model as 'registered'. Thus, 'a' and 'd' are considered as correct classifications; while 'b' and 'c' are inferred as errors. Element 'b' signifies incorrectly predicted 'Registered' by the model. Conversely, c is a measure of incorrectly predicted 'incomplete'. Thus, confusion matrix shows thorough analysis than just proportion of correct guesses. Accurateness is not a trustworthy metric for the real performance of a classifier, since it will generate ambiguous outputs if the sample size varies. The table of confusion matrix contains the average values for all of the above-mentioned classes in a combined manner (Anderson et al., 2003).

Predicted values	**Actual values**	
	Incomplete	**Registered**
Incomplete	a	b
Registered	c	d

Table 6: Elements Of A Confusion Matrix

5. Experimental results Comparison & Evaluation:

5.1 Comparative Analysis:

The following classification algorithms were the most popular for predicting and so being used for predicting the student turnover rate in our experiment.

- CART
- Naïve Bayes
- K Nearest Neighbor
- J48
- NBTree
- Logistic Regression

All these algorithms are tested using three different testing options available in WEKA software for obtaining accurate results. The prediction results are given in the form of confusion matrix. The results are then compiled for each of the classification technique and shown in the comparison table below.

Details of table 7 given below,

Section A lists the details of all the classification techniques mentioned above applied using Training Set option on both original data file with empty fields and also for the file that had been replaced with arbitrary values. Results shows that DTNB has the best prediction results of over 98.4% but the time taken is pretty high. The next best accuracy is given by Naïve Bayes of around 97% accuracy with zero time taken.

Section B lists the details of all the classification techniques applied using Percentage Split option on both original data file with empty fields and also for the file that had been replaced with arbitrary values. The best results are shown by Naïve Bayes classification technique with accuracy of about 94% in just 0.03 seconds.

Section C lists the details of all the classification techniques mentioned above applied using Cross Validation option on both original data file with empty fields and also for the file that had been replaced with arbitrary values. The best results are shown by DTNB classification technique with accuracy of about 95.5% in 1.91 seconds.

Section D lists the details of all the classification techniques mentioned above applied using Cross Validation option on both original data file with empty fields and also for the file that had been replaced with arbitrary values with the field 'Student ID' removed from both the files. The best results are shown by DTNB classification technique with accuracy of about 90% in 1.26 seconds.

S.No	Sample Size	Algorithm / Test Option	Values	Correctly identified Instances	Incorrectly identified Instances	Accuracy	Time in sec
A. CLASSIFICATION TECHNIQUES USING TRAININ SET							
1	996	CART Training Set	Empty fields replaced with arbitrary values	835(83.8%)	161(16.2%)	83.8%	42.7
2	997	NaiveBayes Training Set	Original file with empty fields	967 (97%)	29(3%)	97%	0
3	997	NaiveBayes Training Set	Empty fields replaced with arbitrary values	969 (97.3%)	27(2.7%)	97.3%	0
4	996	K nearest neighbor Training Set	Empty fields replaced with arbitrary values	939(94.3%)	57(5.7%)	94.6%	0

5	996	Decision Tree. J48 Training Set	Empty fields replaced with arbitrary values	959(96.3%)	37(3.7%)	96.4%	0.01
6	996	DTNB Training Set	Empty fields replaced with arbitrary values	980(98.4%)	16(1.6%)	98.4%	7.58
B. CLASSIFICATION TECHNIQUES USING PERCENTAGE SPLIT							
7	996	Decision Tree. J48 Percentage split (66%)	Empty fields replaced with arbitrary values	306(90.3%)	33(9.7%)	90%	0.01
8	997	NaiveBayes Percentage split (66%)	Original file with empty fields	318 (93.8%)	21(6.2%)	93.8%	0
9	997	NaiveBayes Percentage split (66%)	Empty fields replaced with arbitrary values	319 (94%)	20(6%)	94%	0.03
10	996	K nearest neighbor Percentage split (66%)	Empty fields replaced with arbitrary values	313(92.3%)	26(7.7%)	92.3%	0
C. CLASSIFICATION TECHNIQUES USING CROSS VALIDATION							
11	996	CART Cross Validation	Empty fields replaced with arbitrary values	No Results	No Results	No Results	41.6
12	997	NaiveBayes Cross validation	Original file with empty fields	942 (94.5%)	54 (5.4%)	94.5%	0

13	997	NaiveBayes Cross validation	Empty fields replaced with arbitrary values	947(95%)	49(5%)	95%	0.01
14	996	K nearest neighbor Cross validation	Empty fields replaced with arbitrary values	931(93.5%)	65(6.5%)	93.5%	0
15	996	Decision Tree. J48 Cross validation	Empty fields replaced with arbitrary values	930(93.4%)	66(6.6%)	93.2%	0.02
16	996	DTNB Cross validation	Empty fields replaced with arbitrary values	952(95.5%)	44(4.5%)	95.5%	1.91
D. CLASSIFICATION TECHNIQUES USING CROSS VALIDATION WITHOUT 'Student ID' FIELD							
17	996	Logistic Cross validation	File without StudentID	876(88%)	120(12%)	88%	11.52
18	996	Decision Tree. J48 Cross validation	File without StudentID	877(88%)	119(12%)	88%	0.15
19	996	DTNB Cross validation	File without StudentID	897(90%)	99(10%)	90%	1.26

Table 7: Comparison Table On Accuracy Rates Of Different Prediction Techniques With Variation In Data Files

5.2 Confusion Matrix for various Classification Models used in this experiment:

Total sample size: 997 student records
Sample size used for machine learning as training set: 658, which is (66%)
Sample size used as testing set: 339, which is (34%) of the total sample size.

Result 1: Confusion Matrix for SimpleCART using Numeric To Nominal filter on full Data with Training Set test option

a b <-- classified as

0 161 | a = Incomplete

0 835 | b = Registered

Result 2: Confusion Matrix for Naïve Bayes using original file with empty fields using Percentage Split (66%) test option

a b <-- classified as

46 12 | a = Incomplete

9 272 | b = Registered

Result 3: Confusion Matrix for Naïve Bayes using original file after replacing empty fields with arbitrary values with Percentage Split (66%) test option

a b <-- classified as

46 12 | a = Incomplete

8 273 | b = Registered

Result 4: Confusion Matrix for Naïve Bayes using original file after replacing empty fields with arbitrary values with Training Set test option

a b <-- classified as

143 18 | a = Incomplete

9 826 | b = Registered

Result 5: Confusion Matrix for Naïve Bayes using original file with empty fields values using Training Set test option

a b <-- classified as

144 17 | a = Incomplete

12 823 | b = Registered

Result 6: Confusion Matrix for Naïve Bayes using original file after replacing empty fields with arbitrary values with cross validation test option

a b <-- classified as

132 29 | a = Incomplete

20 815 | b = Registered

Result 7: Confusion Matrix for Naïve Bayes using original file with empty fields values using cross validation test option

a b <-- classified as

131 30 | a = Incomplete

24 811 | b = Registered

Result 8: Confusion Matrix for KNN using Numeric To Nominal filter on full Data with cross-validation test option

a b <-- classified as

102 59 | a = Incomplete

6 829 | b = Registered

Result 9: Confusion Matrix for KNN using Numeric To Nominal filter on full Data with percentage split test option

a b <-- classified as

35 23 | a = Incomplete

3 278 | b = Registered

Result 10: Confusion Matrix for KNN using Numeric To Nominal filter on full Data with training set test option

a b <-- classified as

105 56 | a = Incomplete

1 834 | b = Registered

Result 11: Confusion Matrix for J48 using Numeric To Nominal filter on full Data with Cross Validation test option

a b <-- classified as

108 53 | a = Incomplete

13 822 | b = Registered

Result 12: Confusion Matrix for J48 using Numeric To Nominal filter on full Data with percentage split test option

a b <-- classified as

38 20 | a = Incomplete

13 268 | b = Registered

Result 13: Confusion Matrix for J48 using Numeric To Nominal filter on full Data with training set test option

a b <-- classified as

125 36 | a = Incomplete

1 834 | b = Registered

Result 14: Confusion Matrix for DTNB using Numeric To Nominal filter on full Data with cross validation test option

a b <-- classified as

133 28 | a = Incomplete

16 819 | b = Registered

Result 15: Confusion Matrix for DTNB using Numeric To Nominal filter on full Data with training set test option

a b <-- classified as

45 13 | a = Incomplete

5 276 | b = Registered

Result 16: Confusion Matrix for Logistic Regression using file without 'StudentID' field with Cross Validation test option

```
  a   b   <-- classified as

124  37 |   a = Incomplete

 36 799 |   b = Registered
```

Result 17: Confusion Matrix for J48 using file without 'StudentID' field with Cross Validation test option

```
  a   b   <-- classified as

108  53 |   a = Incomplete

 13 822 |   b = Registered
```

Result 18: Confusion Matrix for DTNB using file without 'StudentID' field with Cross Validation test option

```
  a   b   <-- classified as

115  46 |   a = Incomplete

 18 817 |   b = Registered
```

5.3 Result Analysis:

5.3.1 Graphical representation of results from various classification techniques:

5.3.1.1 CART Result Analysis: Table and graph below showing the result of application of CART classification technique using both Training Set option and Cross Validation option on the data file with empty fields replaced with arbitrary values. The results show that time taken to classify is quite high with quite low accuracy rate. In case of Cross Validation, no results were obtained.

CART		
Algorithm / Test Option	**Accuracy**	**Time in sec**
Training Set	83.80%	42.7
Cross Validation	No Results	41.6

Table 8: CART Result Analysis with Training Set option and Cross Validation option

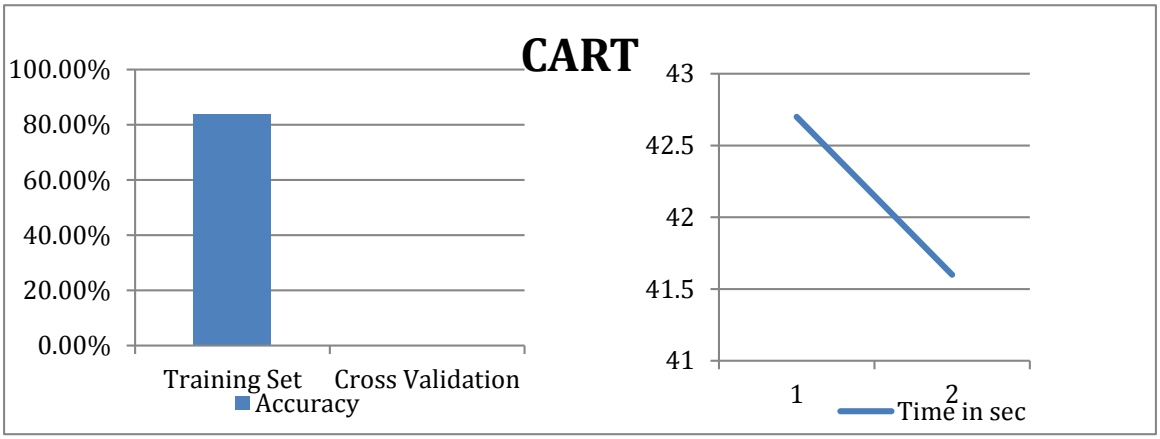

Figure 11: CART Result Analysis displaying Accuracy and Time Taken with Training Set and Cross Validation options

5.3.1.2 Naïve Bayes Result Analysis: Table and graph below showing the result of application of Naïve Bayes classification technique using Percentage Split, Training Set and Cross Validation options on the data file with empty fields replaced with arbitrary values. The results show that Training Set option has given highest accuracy rate of 97.3% in 0.01 seconds.

Naïve Bayes		
Algorithm / Test Option	**Accuracy**	**Time in sec**
Percentage Split (66%)	94%	0.03
Training Set	97.30%	0
Cross Validation	95%	0.01

Table 9: Naive Bayes Result Analysis with Percentage Split, Training Set option and Cross Validation option

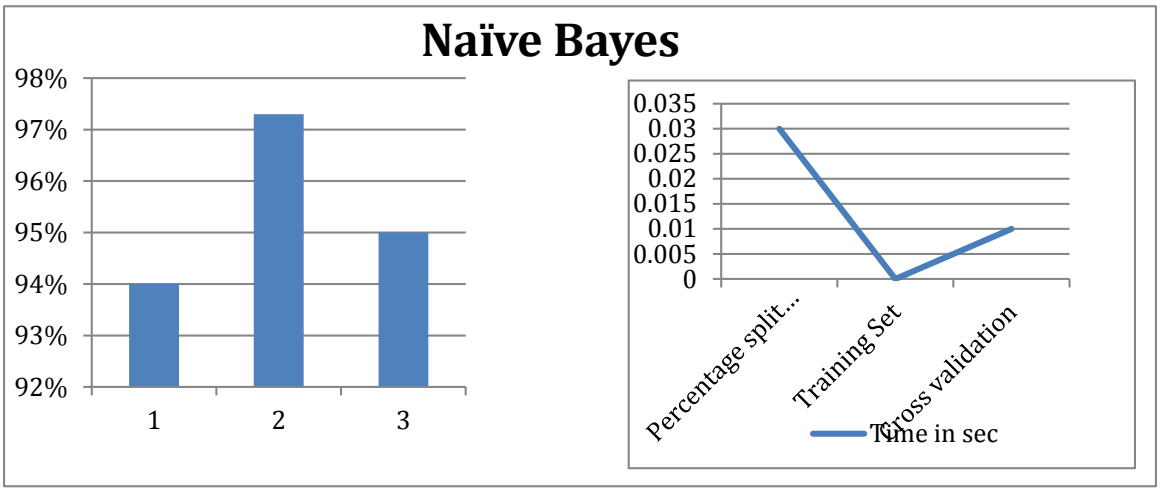

Figure 12: Naive Bayes Result Analysis displaying Accuracy and Time Taken with Percentage Split, Training Set and Cross Validation options

5.3.1.3 K Nearest Neighbor Result Analysis: Table and graph below showing the result of application of K Nearest Neighbor classification technique using Percentage Split, Training Set and Cross Validation options on the data file with empty fields replaced with arbitrary values. The results show that Training Set option has given highest accuracy rate of 94.60% in 0 seconds.

K Nearest Neighbor		
Algorithm / Test Option	**Accuracy**	**Time in sec**
Cross validation	93.50%	0
Percentage split (66%)	92.30%	0
Training Set	94.60%	0

Table 10: K Nearest Neighbor Result Analysis with Percentage Split, Training Set option and Cross Validation option

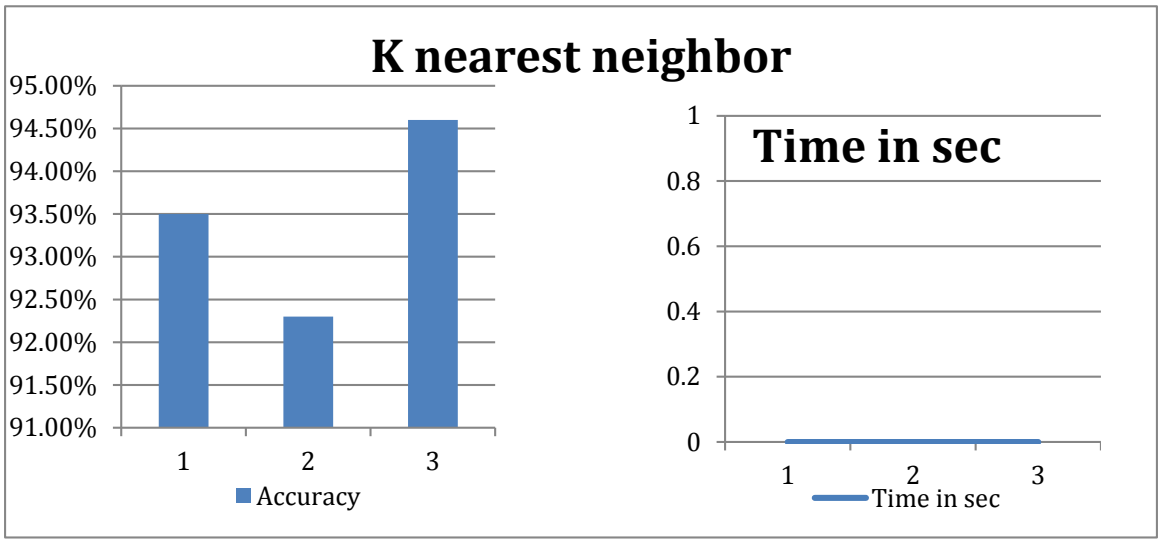

Figure 13: K Nearest Neighbor Result Analysis displaying Accuracy and Time Taken with Percentage Split, Training Set and Cross Validation options

5.3.1.4 Decision Tree. J48 Result Analysis: Table and graph below showing the result of application of Decision Tree. J48 classification technique using Percentage Split, Training Set and Cross Validation options on the data file with empty fields replaced with arbitrary values. The results show that Training Set option has given highest accuracy rate of 96.40% in 0.01 seconds.

Decision Tree. J48		
Algorithm / Test Option	**Accuracy**	**Time in sec**
Cross validation	93.20%	0.02
Percentage split (66%)	90%	0.01
Training Set	96.40%	0.01

Table 11: Decision Tree. J48 Result Analysis with Percentage Split, Training Set option and Cross Validation option

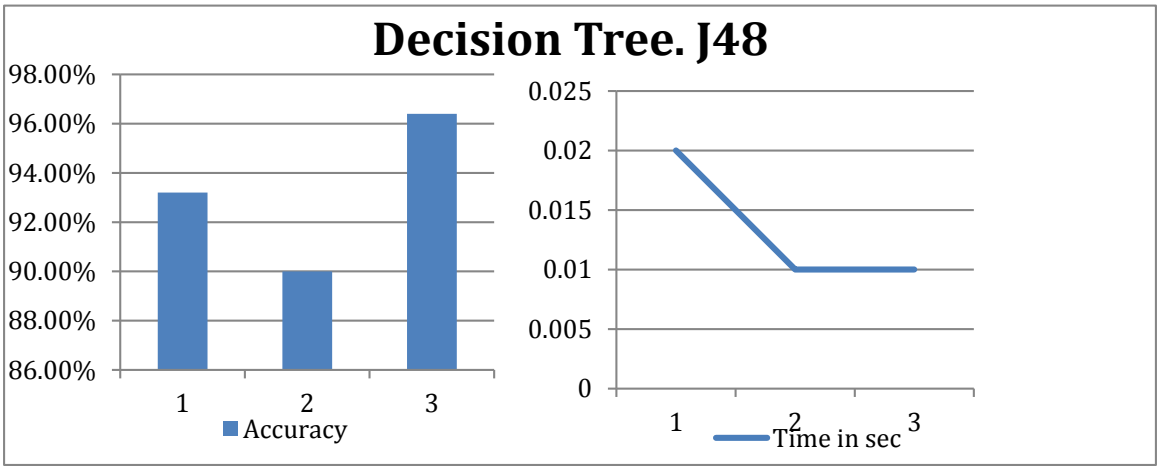

Figure 14: Decision Tree. J48 Analysis displaying Accuracy and Time Taken with Percentage Split, Training Set and Cross Validation options

5.3.1.5 Decision Tree NB Result Analysis: Table and graph below showing the result of application of DTNB classification technique using Training Set and Cross Validation options on the data file with empty fields replaced with arbitrary values. The results show that Training Set option has given highest accuracy rate of 98.40% but the time taken is quite high 7.58 seconds.

DTNB		
Algorithm / Test Option	**Accuracy**	**Time in sec**
Cross validation	95.50%	1.91
Training Set	98.40%	7.58

Table 12: Decision Tree NB Result Analysis with Training Set option and Cross Validation option

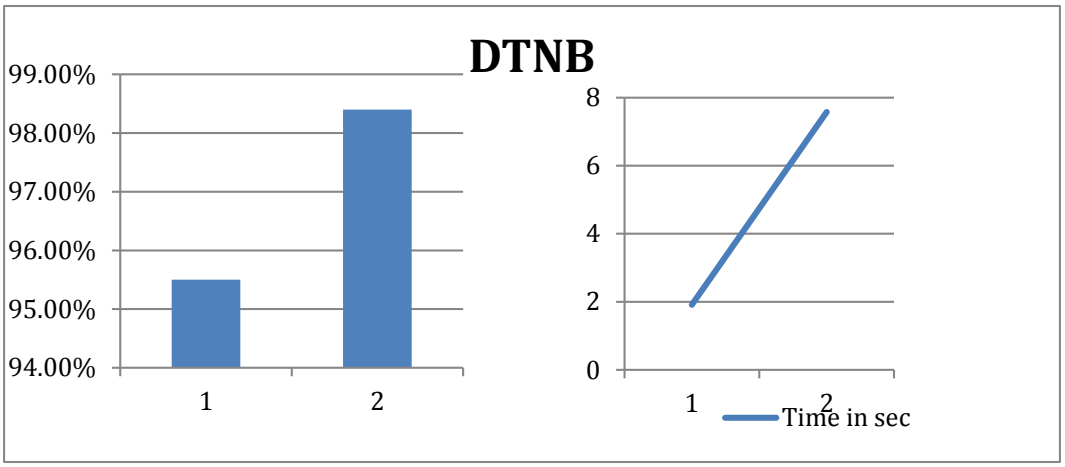

Figure 15: Decision Tree NB Result Analysis displaying Accuracy and Time Taken with Cross Validation and Training Set options

5.3.1.6 Results comparison between Naïve Bayes and DTNB: Table and graph below shows the comparison of results from Naïve Bayes and DNTB classification techniques using Training Set option on the data file. The results show that DNTB has given highest accuracy rate of 98.40% but the time taken is quite high 7.58 seconds.

Best Accuracy using Training Set		
Algorithm / Test Option	**Accuracy**	**Time in sec**
Naïve Bayes	97.30%	0
DTNB	98.40%	7.58

Table 13: Results Comparison Between Naive Bayes & DTNB using Training set option

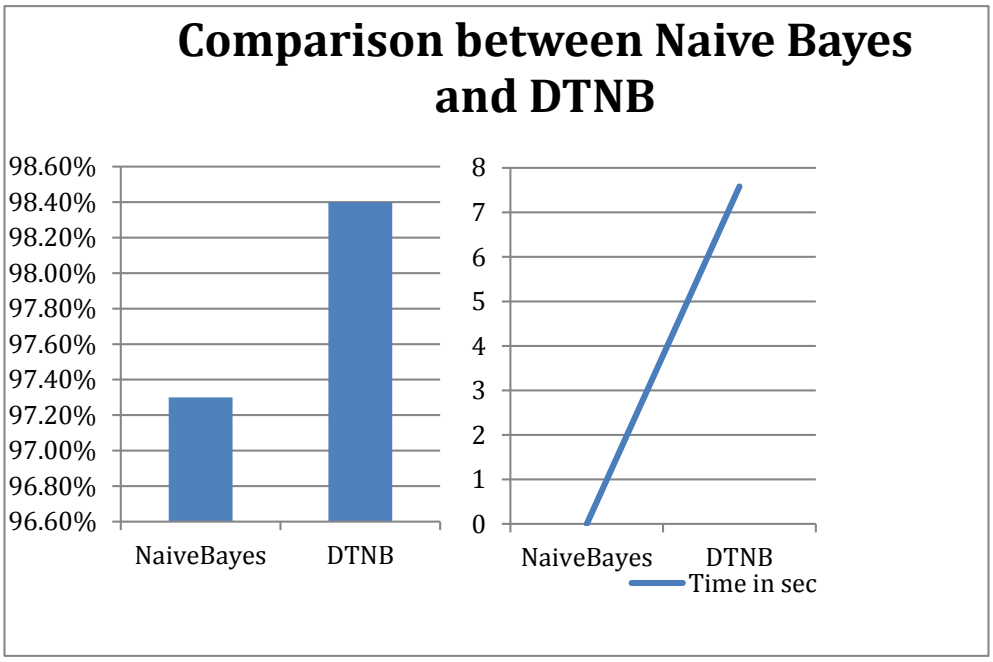

Figure 16: Results Comparison Between Naive Bayes & DTNB displaying Accuracy rate and Time Taken using Training set option

5.3.1.7 Results comparison between Logistic, Decision Tree J48 and DTNB:

Table and graph below shows the comparison of results from Logistic, Decision Tree J48 and DNTB classification techniques using Cross Validation option on the data file without 'Student ID' attribute. The results show that DNTB has given highest accuracy rate of 90% in 1.26 seconds.

Accuracy using Cross validation		
Algorithm / Test Option	**Accuracy**	**Time in sec**
Logistic	88.00%	11.52
Decision Tree. J48	88.00%	0.15
DTNB	90.00%	1.26

Table 14: Comparison Between Logistic, Decision Tree J48 & DTNB using Cross Validation option on data file without 'Student ID' attribute

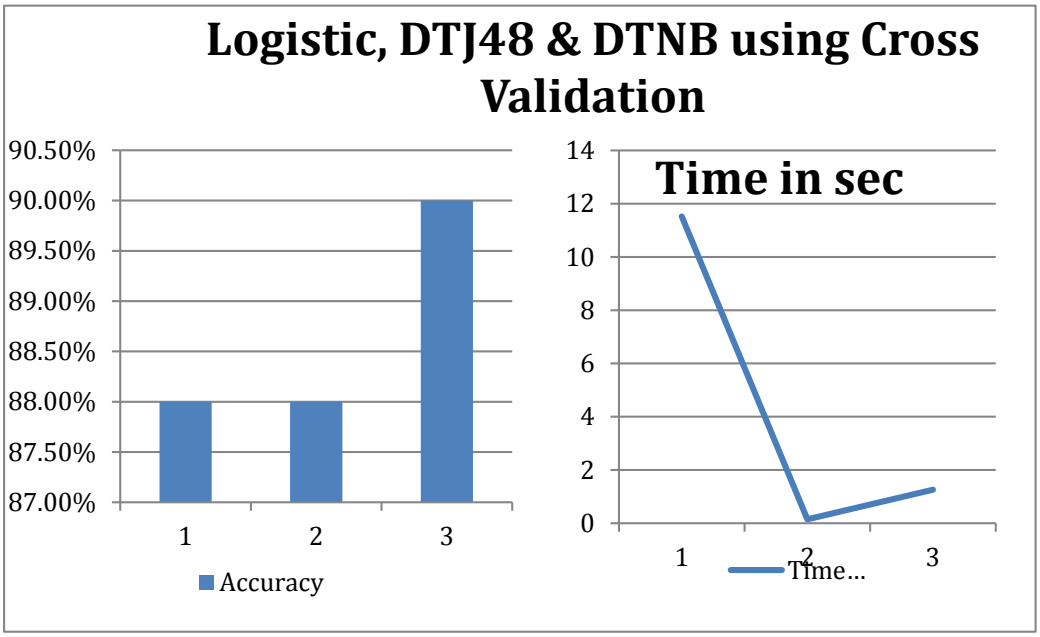

Figure 17: Results Comparison Between Logistic, Decision Tree J48 & DTNB displaying Accuracy Rate and Time Taken in seconds using Cross Validation option on data file without 'Student ID' attribute

5.3.2 Incomplete Application Prediction Model (IAPM):

From the above statistics, it is obvious that the best results are obtained form NBTree and Naïve Bayes classification techniques with Training set option, achieved the maximum accuracy and time taken. Also we notice that in all the above classification techniques, there were no unclassified instances.

In most of the cases, accuracy of results for file with empty fields was less than that of the accuracy of results yielded with the same techniques when applied for file with empty fields replaced with arbitrary fields.

According to (NB tree, Ron), Naïve Bayes accuracy prediction is near accurate for smaller databases. He proves that the accuracy is much better in larger databases using Decision Tree algorithms. Combining the best qualities of both these classifiers, a new hybrid classifier called NBTree is introduced.

From Table 7 results, with files with all attributes inclusive of 'StudentID' field, NBTree classifier has maximum time taken to evaluate when used with cross validation technique but resulted in the best accuracy rate of 98.4% when used with Training Set technique but with 7.58 seconds time taken to evaluate compared with Naïve Bayes that gave us the accuracy result of 97.3% in zero seconds calculation time. Since as per the literature review, cross validation is a more reliable option. After elimination of 'StudentID' field, NBTree has resulted in the best accuracy of 93.6% in 4.05 seconds, Decision Tree J48 resulted in 93.4% accuracy in 0.01 seconds and logistic resulted in accuracy of 88% in 11.52 seconds.

5.3.3 Visualization of Decision Tree J48:

From the visualize tree option; we derive to conclusion that students who have taken IELTS are all registered and continuing. Those who have not taken IELTS but have taken at least TOEFL are also likely to be registered and continuing. But those who have taken neither IELTS nor TOEFL are the most likely not joining the course.

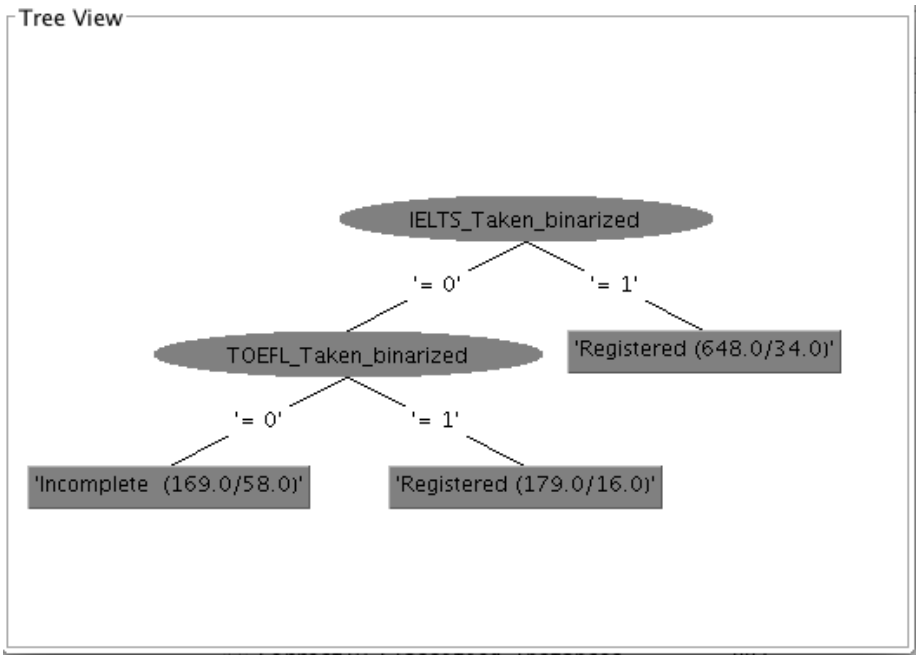

Figure 18: Visual Form of J48 Tree Classification

On removing the fields IELTS and TOEFL, the next strongest attributes as identified by Decision Tree J48 classification technique are Study Mode and Final Percentage. Students registered for Part Time scores are more likely to continue their studies in the University than students registered as Full Time. While the students who has taken Study Mode as Full Time and secured less in the Final Percentage have been categorized as Incomplete Status. The visual tree is as follows:

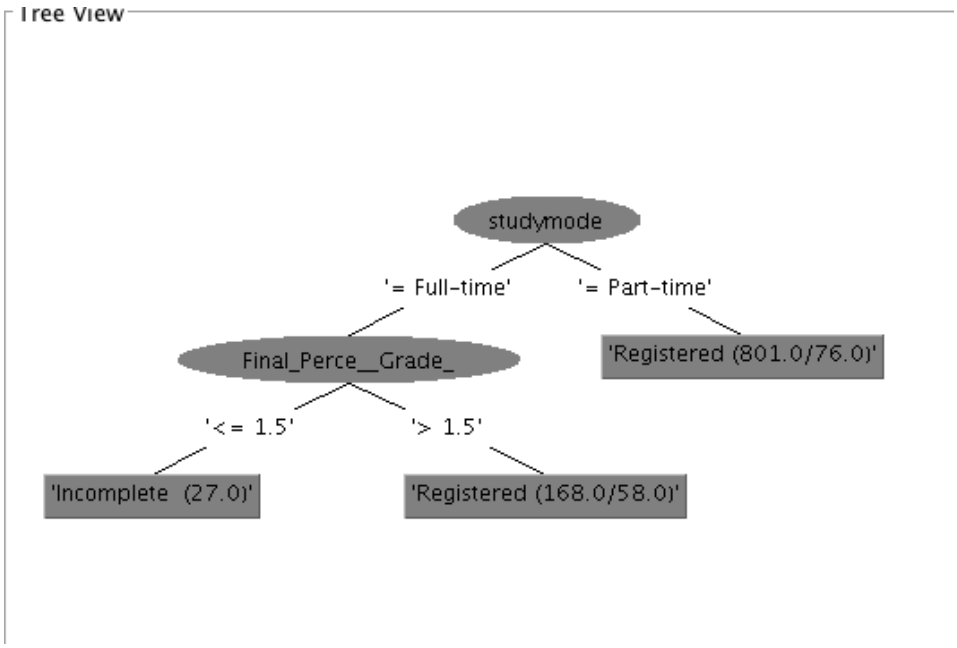

Figure 19: Visual Form Of J48 Tree Classification without IELTS & TOEFL

5.3.4 Analysis of some of the critical results from Apriori Association:

Analyzing the rules derived from Apriori Association given in table 15 below,

Apriori rule 2 associates a confidence rate of 95% if the student had opted Residence Country as United Arab Emirates, Degree Name as Bachelor, Study Mode as Part-time and Stream Name was not filled which gives us the liberty of not focusing on the input for Stream Name while if the other three criteria were satisfied then there is a 95% probability that the student will register and continue.

Apriori rule 3 associates a confidence rate of 95% that the student will register and continue if the student had opted Residence Country as United Arab Emirates, Final Percentage in the previous degree is high, Degree Name as Bachelor, Study Mode as Part-time and Stream Name was not filled.

Apriori rule 4 associates a confidence rate of 94% that the student will register and continue if the student had opted Degree Name as Bachelor, Study Mode as Part-

time and Stream Name was not filled which is a slight variation from the previous rule.

Apriori rule 13 associates a confidence rate of 92% that the student will register and continue if the resident Emirate Name is Dubai.

Apriori rule 16 associates a confidence rate of 92% that the student will register and continue if the student had opted Study Mode as part-time and Semester as January.

When we analyze the data file by semester, for September semester intake, over 421 students are recorded as registered and 90 incomplete in which 46 students registered as full-time and 44 as part-time. For January semester intake, over 473 students are recorded as registered and 70 incomplete in which 41 registered as full-time and 29 as part-time

For analysis the data file by study-mode, a total of 195 students had enrolled for full-time and 801 students enrolled for part-time. From the full-time students data, a total of 85 students are incomplete and 76 students enrolled are registered. Out of the 801 students enrolled for part-time, 76 students are incomplete and remaining 725 students are registered.

5.3.5 Apriori Association result analysis for file without IELTS & TOEFL scores:

Rule No.	Apriori Rule	No. of rules	Confidence rate for continuing students
1	Final_Perce__Grade__binarized=1, studymode=Part-time -> Status = Registered	2	91%
2	Residcountry = United Arab Emirates, Degre_Name=Bachelor, Streamname=Not	3	95%

	filled, studymode=Part-time -> Status = Registered		
3	Residcountry = United Arab Emirates, Degre_Name=Bachelor, Final_Perce__Grade__binarized=1, Streamname=Not filled, studymode=Part-time -> Status = Registered	5	95%
4	Degre_Name=Bachelor, Streamname=Not filled, studymode=Part-time -> Status = Registered	3	95%
5	Degre_Name=Bachelor, Final_Perce__Grade__binarized=1, Streamname=Not filled, studymode=Part-time -> Status = Registered	4	94%
6	Residcountry = United Arab Emirates, Streamname=Not filled, studymode=Part-time -> Status = Registered	3	94%
7	Residcountry = United Arab Emirates, Final_Perce__Grade__binarized=1, Streamname=Not filled, studymode=Part-time -> Status = Registered	4	94%
8	Residcountry = United Arab Emirates, Degre_Name=Bachelor, Final_Perce__Grade__binarized=1, studymode=Part-time -> Status = Registered	4	93%
9	Residcountry = United Arab Emirates, Degre_Name=Bachelor, studymode=Part-time -> Status = Registered	3	93%
10	Degre_Name=Bachelor, studymode=Part-	2	93%

	time -> Status = Registered		
11	Residcountry = United Arab Emirates, Final_Perce__Grade__binarized=1, studymode=Part-time -> Status = Registered	3	92%
12	Final_Perce__Grade__binarized=1, studymode=Part-time, SemName = Jan -> Status = Registered	3	92%
13	EmirateName = Dubai -> Status = Registered	1	92%
14	EmirateName = Dubai, Final_Perce__Grade__binarized=1 -> Status = Registered	2	92%
15	Residcountry = United Arab Emirates, studymode=Part-time -> Status = Registered	2	92%
16	studymode=Part-time, SemName = Jan -> Status = Registered	2	92%
17	Residcountry = United Arab Emirates, Degre_Name=Bachelor, Streamname=Not filled -> Status = Registered	3	92%
18	Residcountry = United Arab Emirates, Degre_Name=Bachelor, Final_Perce__Grade__binarized=1, Streamname=Not filled -> Status = Registered	4	92%
19	Final_Perce__Grade__binarized=1, Streamname=Not filled, studymode=Part-time -> Status = Registered	3	92%
20	Streamname=Not filled, studymode=Part-	2	92%

	time -> Status = Registered		
21	Final_Perce__Grade__binarized=1, studymode=Part-time -> Status = Registered	2	91%
22	Residcountry = United Arab Emirates, Streamname=Not filled -> Status = Registered	2	91%
23	Residcountry = United Arab Emirates, Final_Perce__Grade__binarized=1, Streamname=Not filled -> Status = Registered	3	91%

Table 15: Apriori Associator Rules Combinations & Confidence Rate for students who will register and continue the course enrolled

6. Recommendations & Conclusion:

Thus several classification techniques as well as Apriori Association rule were used on the university data file to obtain the desired results. The best classification techniques that could derive the best prediction accuracy and the major attributes contributing towards decision making were identified. The results also prove that there is a symbiotic relationship between various attributes in the data file provided by the university. The positive results derived from this study on Early Detection System for Incomplete Application were discussed in the above sections and have also contributed in addressing our research questions as follows.

6.1 What attributes can best contribute for designing the Incomplete Application Prediction Model (IAPM):

The following attributes have contributed best to formulate our strategic plan.

- IELTS Taken
- TOEFL Taken
- Study Mode
- Final Percentage
- Residence Country
- Emirate Name
- Degree Name

Where IELTS Taken and TOEFL Taken are major contributors for formulating the strategic plan as shown in the visual tree structure from J48 classification technique as shown in figure 18.

The two attributes Study Mode and Final Percentage were identified by both Apriori Association rule and Decision Tree J48 classification technique.

Residence Country, Emirate Name, Degree Name are other attributes identified by Apriori Association rule.

6.2 Which classification techniques are suitable for building IAPM:

Naïve Bayes Tree classifier with Cross Validation technique is of course superior analysis method over Decision Tree J48 since in this case, we are more concerned with the prediction accuracy rather than the time taken that is much negligible.

6.3 Strategy Plan for Student Retention (SPSR):

Here the emphasis is not on the teaching and the learning methodology in the University since mostly students are independent learners at the Master's level. Moreover the incomplete application rate is high for students who register and decide to leave before attending any classes.

Decision Tree J48 focus on the TOEFL & IELTS Taken attributes for deciding whether a student has a Registered or Incomplete status as discussed in the analysis section of this paper.

In order to get the other major attributes for classification, both TOEFL and IELTS taken field was removed and applied classification technique and Apriori Association. Now the results of both Apriori Association and Decision Tree J48 focus on two major prediction attributes, which are

- Study Mode
- Final Percentage

The result from Apriori Association also focus on combination of other attributes as given below for acquiring high confidence rate for registered students:

- Residence Country
- Emirate Name
- Degree Name

Study mode: There is a very less possibility of an incomplete application in part-time student compared to full time student intake and incomplete students. Statistically also it is proved that the rate of incomplete applications in part time students is only 9% compared to that of full time student, which is over 39%. Improving the facilities for full-time students such as hostel, transport may help to improve the confidence of students to continue the course.

Residence Country: According to (María Cubillo et al., 2006) higher education is described by a superior measure of complexity, divergence, and customization than any other types of businesses. Many features that are quality oriented in higher education cannot be observed, sensed, or tried ahead of time. This brings complications to the assessment of a program, especially for an international candidate. Out of 195 students registered as full-time, 104 students reside in UAE and the remaining 91 students data indicate residence is not UAE. Over 55 UAE non-residents students are with incomplete status leading to an alarming rate of over 60% students not joining the course. According to (Gronroos, 1994), one of the features of educational services is that, throughout the initial step of the internationalization procedure, the facility must be offered at the host country. Through this method, potential candidates will get the essential services encompassing the main education facility and some supplementary facilities, related to education activities at the host organization and ancillary services, in this case, associated to their stay at the host country and the host city. Also the choice to study abroad is one of the most important and expensive initiatives taken by any student. Moreover, the excessive costs of studying abroad make it even more a difficult decision to take. Most complex and expensive decisions are prone to involve customer negotiation to a great extent (María Cubillo et al., 2006). A prospective international candidate will take in to consideration various facets related to staying and living in the host country that involving cost of living, tuition fee, safety, security, culture, international background, university environment, quality of living, and visa and entry requirements etc. Therefore ensuring and improving the facilities such as cultural activities, university environment, hostel, food and

transport amenities at the university level at the same time maintaining a nominal tuition fee may help to retain international students.

Emirate Name: Student's selection for Emirate Name as Dubai also gives a high confidence rate when combined with the other rules as specified in table 15. The focus here can be on the transport facility from different emirates so the students can reach the University on time. Also the college and class timings should be designed comfortable for students travelling from different Emirate as most of them will be working and will not prefer to go back late due to several reasons like going back to work next day if it gets late.

Degree Name: When the previous degree is a Bachelor's, then the possibility of student continuing to Master's is probably more than for a student who wants to pursue a second Master's degree. For the students who's previous degree is already a Master's.

Final Percentage: High scores in the student's previous Bachelor's or Master's degree in table 15 Apriori Results, rule 5 affirms the statement by (college student success) that students' high school class rank and high academic performance (GPA) are important factors in students' persistence, students' precollege academic performance and academic integration greatly influenced their persistence. (Adelman, 1999) Also indicates that a high school curriculum of high academic intensity and quality is a strong factor in a student's likelihood of completing a college degree; the factors such as class rank, AP course, SAT scores and high school GPAs contribute a lot towards college completion.

In our study, the result from decision tree classification also affirms this fact that if a student has taken IELTS and TOEFL, then the student is most likely to be registered and continue the course. So care should be taken to retain the students who have not taken either of these courses. An alternate option can be given such considering Language marks in school or Bachelor degree certificate or time extension to

complete these certifications can be provided. This might boost the confidence of the student to continue the course.

It is recommended in this study that the specific information entered by the students in the application form can be used as a major tool by the admissions office to effectively predict incomplete applications before hand in order to apply custom-oriented principles as mentioned above for retaining these retain students.

Further, since we are focusing on Master's program, and expect mostly working candidates, important information such as whether the student is funded from his work place or any other source for his higher education, whether or not if the student requires any financial support, is the student working, what is the office working hours, what is his current income if he is working, etc if included in the application form could have helped to make a better prediction of whether or not if the student is able to join the course successfully. Moreover this is also supported in our literature review that the financial aid is one of the most important decision criteria for students to pursue higher education.

This study is only a part of ongoing effort to model the strategic plan for student retention at the university for which the research is undertaken. Moreover the file size received form the university was quite small with no indication of the year of enrollment. Results could have been more accurate if we had more records form at least four to five years. Further studies would be generalized to address the issues for universities located in a specific geographical location or an entire emirate in future.

7. Bibliography

[1] DeShields Jr, O.W., Kara, A. and Kaynak, E., 2005. Determinants of business student satisfaction and retention in higher education: applying Herzberg's two-factor theory. International journal of educational management, 19(2), pp.128-139.

[2] Sabnani, S.V., 2008. Computer security: A machine learning approach. Master's thesis, Royal Holloway, University of London, 2007.

[3] Rojas-Méndez, J.I., Vasquez-Parraga, A.Z., Kara, A.L.I. and Cerda-Urrutia, A., 2009. Determinants of student loyalty in higher education: A tested relationship approach in Latin America. Latin American Business Review,10(1), pp.21-39.

[4] Reisberg, L., 1999. Colleges Struggle To Keep Would-Be Dropouts Enrolled. Chronicle of Higher Education, 46(7).

[5] Dunham, M.H., 2006. Data mining: Introductory and advanced topics. Pearson Education India.

[6] Kemerer, F.R., Baldridge, J.V. and Green, K.C., 1982. Strategies for effective enrollment management. Amer Assn of State Colleges and Universities.

[7] Kotler, P. and Fox, K., 1995. Strategic marketing for educational organizations.

[8] University of Minnesota Duluth (n.d.). Classification methods. [Accessed 9 December 2016]. Available at: http://www.d.umn.edu/~padhy005/Chapter5.html

[9] Anderson, R.P., Lew, D. and Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological modelling, 162(3), pp.211-232.

[10] Grönroos, C., 1994. Quo vadis, marketing? Toward a relationship marketing paradigm. Journal of marketing management, 10(5), pp.347-360.

[11] Greenhaus, J.H. and Parasuraman, S., 1986. Vocational and organizational behavior, 1985: A review. Journal of vocational Behavior, 29(2), pp.115-176.

[12] María Cubillo, J., Sánchez, J. and Cerviño, J., 2006. International students' decision-making process. International Journal of Educational Management,20(2), pp.101-115.

[13] Adelman, C., 1999. Answers in the Tool Box. Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment. U.S. Department of Education. Washington, DC: Office of Educational Research and Improvement.

[14] Kavšek, B. and Lavrač, N., 2006. APRIORI-SD: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence, 20(7), pp.543-583.

[15] Crosling, G., Heagney, M. and Thomas, L., 2009. Improving student retention in higher education: Improving teaching and learning. Australian Universities' Review, 51(2), pp.9-18.

[16] Aarinen, A., 2012. University websites as facilitators of international student decision-making. Master Thesis, Aalto University, School of Business, Altoo, Finland.

[17] Zemke, R., 2000. The best customer to have is the one you've already got.Journal for Quality and Participation, pp.33-35.

[18] Anderson, E.W. and Sullivan, M.W., 1993. The antecedents and consequences of customer satisfaction for firms. Marketing science, 12(2), pp.125-143.

[19] Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G. and Loumos, V., 2009. Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers & Education, 53(3), pp.950-965.

[20] Kovacic, Z., 2010. Early Prediction of Student Success: Mining Students Enrolment Data. Proceedings of Informing Science & IT Education Conference (InSITE). Pp 647-649.

[21] Tharp, J. (1998). Predicting persistence of urban commuter campus students utilizing student background characteristics from enrollment data. Community College Journal of Research and Practice, 22, 279- 294.

[22] Nandeshwar, A. & Chaudhari, S. (2009). Enrollment prediction models using data mining [online]. [Accessed 12 December 2016]. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.469.1540&rep=rep1&type=pdf

[23] Vergolini, L. and Zanini, N., 2015. Away, but not too far from home. The effects of financial aid on university enrolment decisions. Economics of Education Review, 49, pp.91-109.

[24] Fielding, A.H. and Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental conservation, 24(01), pp.38-49.

[25] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp.10-18.