

**Data mining approach to predict student's selection of
program majors**

نهج استخراج البيانات للتنبؤ اختيار الطالب للتخصصات البرنامج.

by

SHARMILA SIDDARTHA

**Dissertation submitted in fulfilment
of the requirements for the degree of
MSc INFORMATICS**

at

The British University in Dubai

June 2019

DECLARATION

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that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.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

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

A handwritten signature in blue ink, reading "Siddaethe", with a horizontal line underneath it.

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

Abstract

Students in higher education do not have access to sufficient information when selecting their program major. Program administrators cannot easily predict majors that will be undersubscribed early enough to take corrective actions. At the same time, institutional databases have large volumes of data relating to student demographic profiles, course grades and academic performance. There is an opportunity to apply data mining to arrive at a model to predict student selection of a major. The nature of academic data relating to student majors is multi class and imbalanced – there is always a niche major with few students enrolled. Hence this needs special considerations within the area of data mining.

The purpose of this study is to develop a data mining approach for predicting student's selection of program majors. The approach includes a methodology to manage data mining projects, sampling techniques to handle imbalanced data and multiclass data, a set of classification algorithms to predict and measures to evaluate performance of models.

The methodology used in this study is the systematic literature review to source, evaluate and synthesize current information in this domain and the CRISP-DM to deploy data mining activities. Several data mining techniques such as data exploration, visualization, sampling and evaluation are presented and applied to the academic data. Datamining experiments are deployed in RapidMiner using Decision Trees, Naïve Bayes, Random Forest, Support Vector Machines, Artificial Neural Networks and Gradient Boosted Trees. Balanced sampling, SMOTE – oversampling of minority classes is used to compare results using the confusion matrix, F1-score and the balanced accuracy. Cross validation is applied to train and test performance of

models. Naïve Bayes, Decision Trees offered the best predictions across the different sampling techniques.

This study presents an approach to design and deploy a data mining project that can be used as a basis for developing systems to enable the selection of student majors.

نبذة مختصرة

لا يحصل طلاب التعليم العالي على معلومات كافية عند اختيار تخصصهم. لا يمكن لمسؤولي البرنامج التنبؤ بسهولة بالتخصصات التي لن يتم تسجيل أعداد كافية فيها، من أجل اتخاذ الإجراءات التصحيحية. في الوقت نفسه ، تحتوي قواعد البيانات المؤسسية على كميات كبيرة من البيانات الديموغرافية للطلاب ، ودرجات المواد الدراسية والأداء الأكاديمي. هناك فرصة لتطبيق استخراج البيانات للوصول إلى نموذج للتنبؤ باختيار الطالب للتخصص. طبيعة البيانات الأكاديمية المتعلقة بتخصصات الطلاب متعددة الطبقات وغير متوازنة - هناك دائماً تخصص يضم عدداً قليلاً من الطلاب المسجلين. وبالتالي هذا يحتاج إلى اعتبارات خاصة في مجال استخراج البيانات.

للتنبؤ باختيار الطلاب لتخصصات البرامج. ويشمل (Data mining) الغرض من هذه الدراسة هو تطوير نهج لاستخراج البيانات هذا النهج تطبيق إدارة المشاريع لاستخراج البيانات ، وتقنيات أخذ العينات للتعامل مع البيانات غير المتوازنة والبيانات متعددة الطبقات ، ومجموعة من خوارزميات التصنيف للتنبؤ، وإجراءات تقييم أداء النماذج.

المنهجية المستخدمة في هذه الدراسة هي مراجعة الأدبيات المنهجية للحصول على المعلومات الحديثة وتقييمها وتوليها في هذا المجال لنشر أنشطة استخراج البيانات. تم أيضاً تقديم العديد من تقنيات استخراج البيانات ، مثل استكشاف البيانات CRISP-DM ، وأيضاً وأخذ العينات والتقييم وتطبيقها على البيانات الأكاديمية. تم تطبيق تجارب (visualization) والتصور (Data mining) و Naïve Bayes و Decision Trees باستخدام خوارزميات أشجار القرار RapidMiner استخراج المعلومات في برنامج (Artificial Neural Networks) والشبكات العصبية الاصطناعية (SVM) وآلات المتجهات الداعمة Random Forest SMOTE-oversampling تم استخدام أخذ العينات المتوازن ، و (Gradient Boosted Trees) والأشجار المعززة للتدرج F1-score ، ودرجة (confusion matrix) زيادة في اخذ العينات) لفئات الأقليات لمقارنة النتائج باستخدام مصفوفة الارتباك) التحقق من الصحة) لتدريب واختبار أداء النماذج. (cross validation تم تطبيق (balanced accuracy) والدقة المتوازنة أفضل التنبؤات عبر مختلف تقنيات أخذ العينات (Decision Trees) ، و أشجار القرار Naïve Bayes قدمت خوارزميات

تقدم هذه الدراسة مقارنة لتصميم ونشر مشروع استخراج البيانات الذي يمكن استخدامه كأساس لتطوير النظم لتمكين الطلاب من اختيار تخصصاتهم.

Dedication

“If I have seen further, it is by standing on the shoulders of giants”

Attributed to Sir Isaac Newton.

In all humility, I would like to dedicate this study to researchers and life-long learners who
inspire and educate a new generation of learners.

Acknowledgments

In writing this dissertation, I have received much support and guidance. Primarily, I would like to thank Professor Khaled Shaalan for his guidance in developing this idea to fruition and for encouraging me to research and publish.

My appreciation to Professor Sherief Abdalla for his advice and for igniting a deep interest in data mining through his courses.

I wish to acknowledge my alma mater, The British University in Dubai for the support received, the use of learning facilities and library resources, and the Doctoral Training Center for helpful workshops that assisted in this study.

To my employers, for providing several opportunities to learn and for constantly raising the bar so that I continue to push the educational envelope.

I wish to thank my colleagues and fellow students for supporting and inspiring my academic journey. We have discussed research questions, methods to use, data to collect and analyze, and measures to take or not take! When the going was rough, we have nudged each other along. My appreciation to each of you. May we continue to learn forever and foster lifelong learning amongst students!

Finally, to my family for their encouragement, support, and infinite understanding of how important learning is to me. Thank you for your support in this journey of mine.

Contents Listings

Table of Contents

| | |
|---|-----------|
| DECLARATION | |
| ABSTRACT OR EXECUTIVE SUMMARY | |
| DEDICATION | |
| ACKNOWLEDGMENTS | |
| TABLE OF CONTENTS | I |
| LIST OF ILLUSTRATIONS | IV |
| LIST OF TABLES..... | VI |
| CHAPTER 1 INTRODUCTION | 1 |
| OVERVIEW..... | 1 |
| CONTEXT AND PROBLEM DEFINITION | 3 |
| OPPORTUNITY STATEMENT | 4 |
| AIM OF RESEARCH | 5 |
| RESEARCH QUESTIONS..... | 6 |
| METHODOLOGY..... | 7 |
| <i>Systematic Literature Review</i> | 7 |
| <i>Datamining Methodology – CRISP-DM</i> | 8 |
| RESEARCH GOALS | 8 |
| SUMMARY | 9 |
| CHAPTER 2 LITERATURE REVIEW | 10 |
| INTRODUCTION..... | 10 |
| SUMMARIES BY LITERATURE REVIEW CONTENT AREAS..... | 11 |
| <i>Section 1 -Introduction to Data Mining</i> | 11 |

| | |
|--|-----------|
| <i>Section 2 - Educational Data Mining - EDM</i> | 17 |
| <i>Section 3 - Datamining Methodology</i> | 24 |
| <i>Section 4 - Datamining Classification Algorithms</i> | 38 |
| <i>Section 5 - Evaluation Measures for Multi-class Classification Algorithms</i> | 51 |
| SUMMARY | 57 |
| CHAPTER 3 RESEARCH METHODOLOGY | 59 |
| METHODOLOGY – SYSTEMATIC LITERATURE REVIEW..... | 59 |
| RESEARCH QUESTIONS | 64 |
| DATAMINING PROCESSES – CRISP-DM METHODOLOGY | 64 |
| RESOURCES USED | 65 |
| <i>Mendeley Desktop®</i> | 65 |
| <i>Microsoft Office®</i> | 65 |
| <i>RapidMiner®</i> | 65 |
| SUMMARY | 66 |
| CHAPTER 4 RESEARCH IMPLEMENTATION | 67 |
| CRISP-DM METHODOLOGY | 67 |
| <i>Business Understanding</i> | 67 |
| <i>Data Understanding – Extraction of Data</i> | 68 |
| <i>Data Preparation</i> | 69 |
| SUMMARY | 82 |
| CHAPTER 5 DEPLOYMENT OF EXPERIMENTS AND ANALYSIS OF RESULTS | 84 |
| EXPERIMENTS..... | 84 |
| <i>Auto Modelling with Rapid Miner</i> | 84 |
| <i>Systematic Experimentation</i> | 86 |
| <i>Classifiers Used</i> | 86 |

| | |
|---|------------|
| <i>Evaluation Measures Used</i> | 87 |
| STATISTICAL ANALYSIS | 87 |
| <i>Comparative analysis of Sampling</i> | 87 |
| <i>Analysis of Classifiers across types of sampling</i> | 95 |
| SUMMARY | 98 |
| CHAPTER 6 DISCUSSIONS | 100 |
| RESEARCH Q1 | 100 |
| RESEARCH Q2 | 102 |
| RESEARCH Q3 | 103 |
| RESEARCH Q4 | 105 |
| CHAPTER 7 CONCLUSIONS AND FUTURE WORK | 107 |
| CHAPTER 8 BIBLIOGRAPHY | 109 |
| CHAPTER 9 APPENDICES | 117 |

List of Illustrations

| | |
|---|----|
| Figure 1 Top majors in Arts, Education and Information Technology. Source: https://www.moe.gov.ae/Ar/MediaCenter/News/PublishingImages/majors3.png | 2 |
| Figure 2 Implementation of Systematic Literature Review. Author's Own..... | 8 |
| Figure 3 Data mining activities- Predictive and Descriptive | 11 |
| Figure 4 From organizational data to data mined outputs | 13 |
| Figure 5 Machine Learning Processes - Train, Model, Test and Validate | 15 |
| Figure 6 Steps in Cross-Validation. Source: How to Correctly Validate Machine Learning Models (<i>Rapidminer</i> , 2017) | 16 |
| Figure 7 Source (Wanjau, Okeyo and Rimiru, 2016) | 21 |
| Figure 8 Main Data Mining methodologies. Source:(Piatetsky, 2014) Retrieved from https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm | 25 |
| Figure 9 Approaches to managing Datamining Projects (Saltz et al., 2018)..... | 25 |
| Figure 10 CRISP-DM Model. Author's own work based on (Chapman et al., 2000)..... | 27 |
| Figure 11 ASUM data mining methodology Source: IBM Corporation (IBM Analytics et al., 2016) | 28 |
| Figure 12 Author's representation of the Knowledge Discovery Process Model based on (Fayyad, Piatetsky-Shapiro and Smyth, 1996) | 29 |
| Figure 13 SEMMA based on information in (SAS Institue Inc., 2017) | 30 |
| Figure 14 Source: Rogalewicz and Sika, 2016 | 34 |
| Figure 15 Source: Sharma and Osei-Bryson, 2009..... | 37 |
| Figure 16 Distribution of classes | 38 |
| Figure 17 Sample Decision Tree..... | 40 |

| | |
|---|----|
| Figure 18 ROC chart indicating performance of algorithms for binary classification | 54 |
| Figure 19 Steps for conducting a Systematic Literature Review based on (Okoli and Schabram, 2010) | 59 |
| Figure 20 Keyword Search Map for Literature Review | 61 |
| Figure 21 Word cloud of paper titles used in this study | 63 |
| Figure 22 CRISP-DM Processes..... | 67 |
| Figure 23 Statistics demographic attributes in data set..... | 72 |
| Figure 24 Students age across campuses | 72 |
| Figure 25 Enrollment into major across gender..... | 73 |
| Figure 26 High School average across campus | 74 |
| Figure 27 High School Average across student major..... | 74 |
| Figure 28 IELTS band by student major..... | 75 |
| Figure 29 IELTS bands across gender | 75 |
| Figure 30 Statistics of academic attributes | 76 |
| Figure 31 Academic performance across campuses | 77 |
| Figure 32 Statistics of course grades of successful students..... | 78 |
| Figure 33 Student grades across Gender..... | 78 |
| Figure 34 Correlation analysis of all variables | 80 |
| Figure 35 Correlation analysis of demographic attributes | 80 |
| Figure 36 Pairwise correlation of course grades..... | 81 |
| Figure 37 Auto model overview of classifier performance | 84 |
| Figure 38 Auto model - Naive Bayes Confusion Matrix..... | 85 |
| Figure 39 Auto model - Feature weights by correlation | 85 |

| | |
|---|----|
| Figure 40 Comparison of class instances with sampling applied | 86 |
| Figure 41 Confusion Matrix of Naïve Bayes model with Oversampled Minority Class - SMOTE | 91 |
| Figure 42 Comparative chart of Error Rates | 95 |
| Figure 43 Class instances with Sampling | 96 |
| Figure 44 Model Performance with and without Sampling..... | 97 |

List of Tables

| | |
|---|-----------|
| Table 1 Research questions mapped to questions and content code..... | 10 |
| Table 2 Supervised learning and types of target variables | 14 |
| <i>Table 3 EDM themes and applications for various stakeholders in education. Based on (Romero and Ventura, 2010)</i> | <i>19</i> |
| Table 4 Summary of Related work in Educational Data Mining (Peña-Ayala, 2014) | 19 |
| Table 5 Binary classification with three classes | 48 |
| Table 6 Illustration of the Confusion Matrix for a binary label..... | 52 |
| Table 7 Illustration of the Confusion Matrix for a multi-class label | 53 |
| Table 8 AUC Measures for multiple classification algorithms | 54 |
| Table 9 Classification performance measures - Compiled from various sources..... | 56 |
| Table 10 List of attributes in source data..... | 70 |
| Table 11 Multi-class data distribution and expected recall values | 79 |
| Table 12 Parameters used with classification models..... | 87 |
| Table 13 Decision Tree performance vector with unsampled data..... | 88 |

| | |
|---|-----|
| Table 14 Comparative model performance without sampling..... | 89 |
| Table 15 Comparative model performance with undersampling of majority class | 90 |
| Table 16 Comparative model performance with over sampling with SMOTE | 92 |
| Table 17 Overall Model Performance Rankings | 93 |
| Table 18 Classification Algorithms - Pros and Cons..... | 105 |

Chapter 1 Introduction

Overview

In recent times, students embarking on higher or undergraduate education in the UAE have a plethora of options. The educational sector of the UAE has seen a growth in the higher education sector, offering a wide range of programs in science, engineering, arts and, business. In a report by (Colliers International, 2018) this growth is attributed to a growing population, professionals seeking part-time education, and also the growth of Dubai as an Educational tourist hub.

Additionally, several higher education universities in the UAE also offer a choice of “majors” within their programs. These majors enable students to develop “specializations” in an area relating to their program. Selecting the right major enables students to develop skills and depth of knowledge in an area which also fits their passion and capabilities. Programs with the appropriate majors also ensure high employability of students. Currently, there is a strong drive towards improving the employability of new graduates. Hence it is important that students select programs and majors that help them secure employment as soon as they graduate.

In the UAE, universities offer programs such as Engineering or Information Technology with several majors or areas of specialization. These majors are offered based on market needs, with inputs provided by industry professionals and the expertise of teachers employed. In a study conducted by the UAE’s Ministry of Education (MOE), the most popular majors of Information Technology were Information Security, Networking, and Information Systems (MOE, 2018).

Figure 1 is from a study conducted by the Ministry of Education in the UAE, highlighting some of the popular majors in higher education.

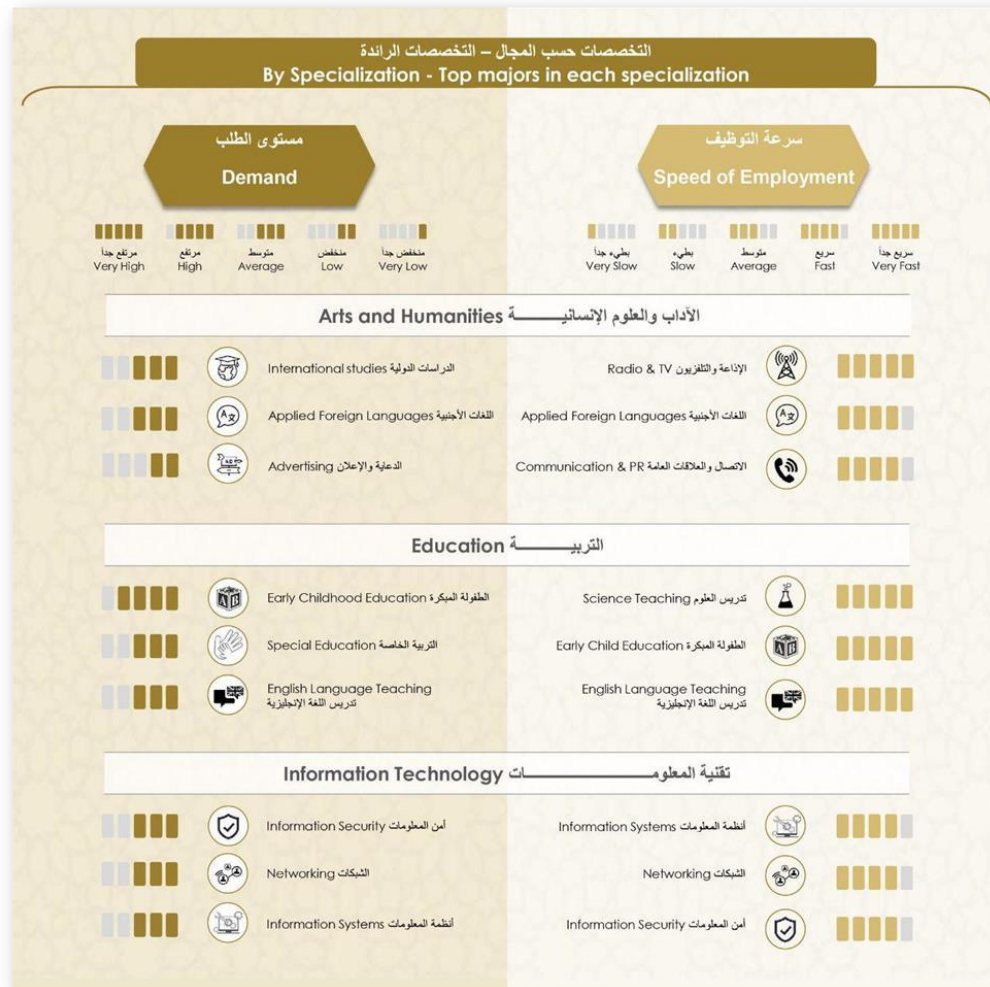


Figure 1 Top majors in Arts, Education and Information Technology. Source: <https://www.moe.gov.ae/Ar/MediaCenter/News/PublishingImages/majors3.png>

Usually, students enroll in a common first year and then select a major. The ideal situation is when universities offer majors that are demanded by the industry, and students subscribe to these enthusiastically. However, this is not always the case. Students are usually undecided or go into a major that is popular with their peers. A small number of students decide on their major based

on mandates from their employer. There is limited awareness of career prospects, employment opportunities, or academic prerequisites that trigger the selection of a major. When the choice of a major is not the right fit with a student's interests and abilities, students are demotivated and are at risk of dropping out. Occasionally, program administrators may have to close majors due to lack of interest from students or demand for graduates from the industry. In both cases, there is a considerable loss of time, effort, and wasted opportunities. In this scenario, there is a need for universities to be able to tap into their data along with inputs from external stakeholders and create majors that are both attractive to students, adaptable to external environmental demands and enable resource optimization in terms of available faculty expertise.

Context and Problem Definition

This study is related to the enrollment of students into the majors offered by the Computer Science Division of a large university in the UAE. The duration of the program is four years and offered to students interested in pursuing a career in the realm of Computer Science and Information Technology. Students enroll in a common year and take various general and technology courses in the first year. By the end of their third semester, they are required to choose their major that are currently offered by the program. A small segment of sponsored students must choose their major as mandated by their employers, but most students make their own decisions. To help students choose wisely, the division offers information sessions regarding the majors on offer. Students are also required to meet with their advisors to discuss their choices. However, students indicate that very often they select their major based on peer influence or advice from family and friends. There is very little evidence that students reflect on their interests, capabilities, employment, and growth opportunities before they choose their major. This phenomenon leads to several problems, such as:

- When students do not enroll in planned courses in specific majors, program managers need to cancel courses due to poor enrollment. Cancellation of courses triggers dissatisfaction amongst teachers and students. Faculty are frustrated when they are asked to teach courses outside their area of specialization, and students are dissatisfied when courses of interest are canceled due to poor enrollment.
- Program managers cannot predict or promote programs as they do not have the data indicating the factors that determine how students select their majors.
- There is also the risk of students graduating with a major that is not in demand in the local industry when they select majors based on peer influence. Students not gaining employment soon after graduation or outside their major is contrary to the vision and mission of the University
- When students choose a major that does not fit their capabilities or passion, the attrition and failure rates are high. High attrition or failure rates have a direct impact on the sustainability of the Program. High failure rates also lead to student's repeating courses, which costs them time and resources and creates poor publicity for the program and major.
- A lack of graduates in the majors demanded in the industry triggers employers to look for employees from outside the country which is not favorable.

Opportunity Statement

With the emergence of data science and the availability of organized transactional data with universities, there is an opportunity to investigate how data mining can be employed to learn the factors that affect the selection of majors by students. However, there is a paucity of published research in educational data mining in this region. Current studies are mostly about the performance of students in academia (Saa, 2016), evaluation of coursework assessments and

their impact on the final exam grade (Anzer, Tabaza and Ali, 2018), student satisfaction with learning experiences in overseas campuses (Wilkins and Balakrishnan, 2013), measuring student satisfaction with blended learning (Naaj, Nachouki and Ankit, 2012) and mining assessment data for alignment with learning outcomes (Hussain *et al.*, 2017). There is very limited published research about the use of data mining in the selection of programs or majors, particularly in the UAE. Hence there is an opportunity to address this gap and apply data mining tasks to the process of students choosing their major by looking for patterns in enrollment and grades attained in the first year of the program.

There are residual opportunities to investigate a methodology for implementing a data mining study and for conducting a systematic study to uncover the most suited algorithms to predict the selection of majors. The opportunities are summarized as:

- Apply data mining to predict selection of student majors
- Identify a suitable methodology to implement data mining projects,
- Explore academic data to develop profiles of students in majors
- Identify models and measures to build prediction models
- Tackle multi-class and imbalanced data which are an inherent aspect of academic data

Aim of Research

This study means to develop a data mining approach to predict a student's selection of program majors by using academic data available in the institutional database.

This study proposes to use institutional data relating to student demographics, grades in courses taken before students chose their major and academic scores from high school, and college placement exams to predict the selection of major.

This study also seeks to identify a suitable framework for data mining projects and identify algorithms that are best suited to the task of prediction using the available data.

This study is important because it offers educational institutions a framework to apply data mining to the process of students selecting their major. Program administrators can become aware of employing data science and classification techniques to understand patterns in student enrollments and the profile of students that seek specific majors. This insight would allow for better marketing of majors and courses to students. The results of this research would guide advisors design a system to help students make informed decisions regarding their selection of major.

Research Questions

To develop a data mining approach, this study aims to answer the following questions based on literature reviews and data mining experiments. The responses to these questions will constitute the approach to predicting the selection of student majors in any university

[R1] Can data mining be applied to the predict student's selection of program majors in higher education?

[R2] What is an appropriate framework to be used for data mining projects?

[R3] What are the most suitable algorithms that can be used to classify and predict multi-class labels with imbalanced data?

[R4] What is the most appropriate evaluation measure to be used with multiclass datasets?

Methodology

This study used the Systematic Literature Review methodology to plan, develop, and implement the research in conjunction with the CRISP-DM methodology. A sound literature review is of utmost importance to develop a good understanding of the relevant literature while a data mining methodology is necessary to implement the solution. Hence a systematic literature review was conducted to obtain the necessary knowledge, and the CRISP-DM methodology applied for the data mining task.

Systematic Literature Review

The purpose of a systematic literature review is to study the underlying theory and context relating to the topic of the research and arrive at the hypothesis or the research questions. It aims to extract and aggregate the most current and succinct information relating to the topic being studied (Rowley and Slack, 2004; Keele and others, 2007; Kitchenham *et al.*, 2010; Okoli and Schabram, 2010). Structured literature reviews are considered secondary research and follow a methodology to ensure an effective study of the topic and enable the successful deployment of the primary study. Such a literature review is also recommended to synthesize information and evaluate results derived from primary research (Keele and others, 2007).

The process recommended for a systematic literature review includes planning the purpose of research, setting up protocols and training, searching, screening, quality appraisal of extracted literature, analysis of findings and writing of reviews (Okoli and Schabram, 2010). This study implements a simpler version of the systematic literature review, as shown in Figure 2.

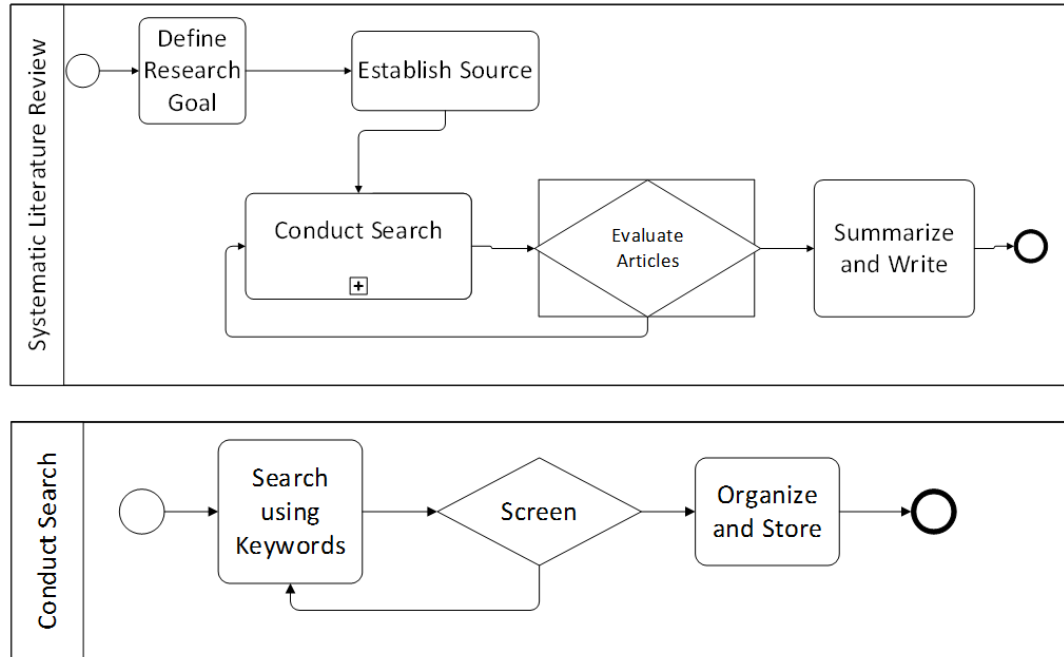


Figure 2 Implementation of Systematic Literature Review. Author's Own

Datamining Methodology – CRISP-DM

This study applied the popular CRISP-Data mining methodology as part of the implementation.

The CRISP-DM includes six phases for business understanding, data understanding, data preparation, modeling with machine learning algorithms, and deployment phase. All these phases were conducted based on the extensive study reported in Section 3 - Datamining Methodology of this report. Chapter 2 - The Cross-Industry Standard Process for Data Mining or CRISP-DM of this study discusses this methodology in detail.

Research Goals

The main research goal of this study is the application of data mining in the education sector.

The subsidiary goals are linked to the research questions and are listed below

- [R1] – Explore and study literature in educational data mining relating to the selection of program majors

- [R2] – Investigate and identify data mining project methodologies
- [R3] – Study the management of multi-label and imbalanced data in classification tasks
- [R4] – Study and implement classification algorithms and evaluation metrics used in predictive tasks

The research goals mentioned above are rather broad as they include a study of the sector, data mining methodology, as well as the experimentation. This study intends to address these cornerstones and pave the way for future work in these areas.

Summary

This study intends to achieve the research goals and answer questions set out in this study. To this effect, this chapter introduces the educational sector in the UAE and the context for student's selection of a program major. This chapter demonstrates a business understanding of problems and opportunities relating to the selection of majors and reviews it from the perspective of students, administrators, faculty, and employers. The first chapter also introduces the dual methodology to be applied in this study and presents the research goals and questions.

The next chapter introduces the literature review conducted to support the development of this study.

Chapter 2 Literature Review

Introduction

As introduced in Chapter 1, this section of the report will present a detailed literature review arranged in the order of research questions. The table below shows the mapping of each research question to a content code and a section in this chapter.

| Q# | Title | Section in Chapter | Content Code |
|------|---|------------------------|--------------|
| [R1] | Can data mining be applied to the predict student's selection of program majors in higher education? | Section 1 Section 2 | DM, EDM |
| [R2] | What is an appropriate methodology to be used for data mining projects? | Section 3 | DMF |
| [R3] | What are the most suitable algorithms that can be used to classify and predict multi-class labels with imbalanced data? | Section 4 | DMCA |
| [R4] | What is the most appropriate evaluation measure to be used with multiclass datasets? | Section 5 | DMEV |

Table 1 Research questions mapped to questions and content code

The literature review included a study of material sourced from books, journals, and online resources. These were sourced, screened, reviewed, and summarized using the systematic literature review methodology. The following sections provide a summary of the literature in each content area. Section 1 introduces data mining and all the related activities and provides a basis for understanding literature reviewed in sections 2, 3, 4, and 5.

Summaries by literature review content areas

Section 1 -Introduction to Data Mining

The field of data mining strives to find interesting patterns in data (Kotu and Deshpande, 2014).

With the maturity of organizational capabilities in creating and managing information systems to capture and process data in large volumes, there has been a growth in activities relating to data mining and knowledge discovery. Hence data mining is mandated to be an automated elicitation of patterns and analysis of large volumes of data stored electronically (Witten et al., 2016).

Sophisticated software tools are employed to discover patterns and relationships amongst data sets stored in large data stores (Archana and Elangovan, 2014). The outcome of data mining activities is knowledge discovery.

Data mining Activities

Data mining activities include extracting meaningful patterns, building abstracted models from data sets to understand relationships and perform predictions (Kotu and Deshpande, 2014).

Figure 3 indicates the data mining activities categorized as predictive and descriptive tasks (Fadzilah Siraj and Mansour Ali Abdoulha, 2011).

| | |
|-------------|----------------------|
| Predictive | Classification |
| | Prediction |
| | Regression |
| | Time-series analysis |
| Descriptive | Clustering |
| | Summarizations |
| | Association rules |
| | Sequence analysis |

Figure 3 Data mining activities- Predictive and Descriptive

Predictive Tasks

Predictive tasks use historical data to predict the outcome of a variable based on one or more input variables. Classification techniques work with non-numeric data by dividing data into two or more classes based on the values in the independent or predictor variables. Regression works with numeric data. Time series analysis works with predicting trends based on historical data and has several applications in sales analysis.

Descriptive Tasks

Descriptive tasks use historical datasets to find patterns and relationships and present them in a comprehensible format. Clustering techniques are useful to find natural groupings of data that are not easily seen by the human mind. This technique can also be used to group data to manage prediction better. Associations are useful to understand relationships, particularly in marketing. In education, it can be used to group students based on their academic progress. For example, students that achieved a high grade in course Introduction to Programming also chose Introduction to Logic.

Machine Learning

Machine learning is the application of algorithms or models in predictive and descriptive data mining tasks and is an area of data mining. Machine learning is the ability of machines to learn, think, and solve a problem in the way that humans do without being explicitly programmed to do so (Kashyap, 2018). Machine learning is driven by algorithms or programs that help computers learn iteratively from experience and generate knowledge. They work on input variables that are either labeled or unlabeled to produce outputs that are either predictive or descriptive. As illustrated in Figure 4, machine learning converts large sources of data into predictions or

descriptions using algorithms. One of the challenges is to identify the most appropriate modeling algorithm for this task.

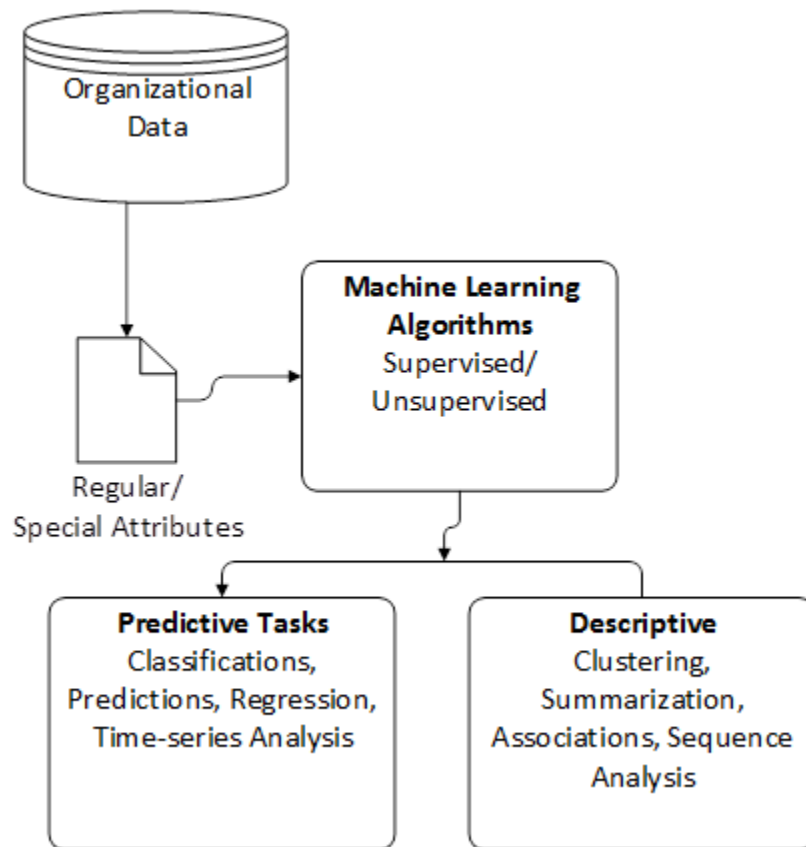


Figure 4 From organizational data to data mined outputs

Supervised Learning

Supervised machine learning is performed on labeled training data and the algorithm tries to learn the relationships between attributes to predict the value of a special variable. In this type of learning, algorithms use various input variables to classify labeled training data into predefined classes that are binary or categorical. The model that is developed during this training is applied to unseen data and the output is derived. Decision tree, Bayesian statistics, neural networks, the support vector machine (SVM), regression, nearest neighbors are algorithms used in supervised learning.

In an educational context, Table 2 provides examples of binary and nominal or categorical target variables in bold:

| Binary Target Variables or variables | Nominal or Categorical Target Variables |
|---|---|
| Will the student graduate ? Yes or No | The program a student is enrolled. Eg.: Engineering, Science, Arts, or Business. |
| Is the student at risk of failing ? Yes or No? | The performance of a student. Eg.: Good, Excellent, or Bad. |
| Will the student achieve a distinction grade? Yes or No | The sector where a student is employed. Eg. Private, Public, or Self. |

Table 2 Supervised learning and types of target variables

Unsupervised Learning

The goal of unsupervised learning is to find patterns in data based on relationships between unlabeled data points or values (Kotu and Deshpande, 2014). The output of unsupervised learning is not a target variable but a description in the form of charts, heat maps, cluster tables, and scatter plots. An example of unsupervised learning is clustering, where algorithms see natural groupings of data. Association rules are formulated by using unsupervised learning to define relationships between items.

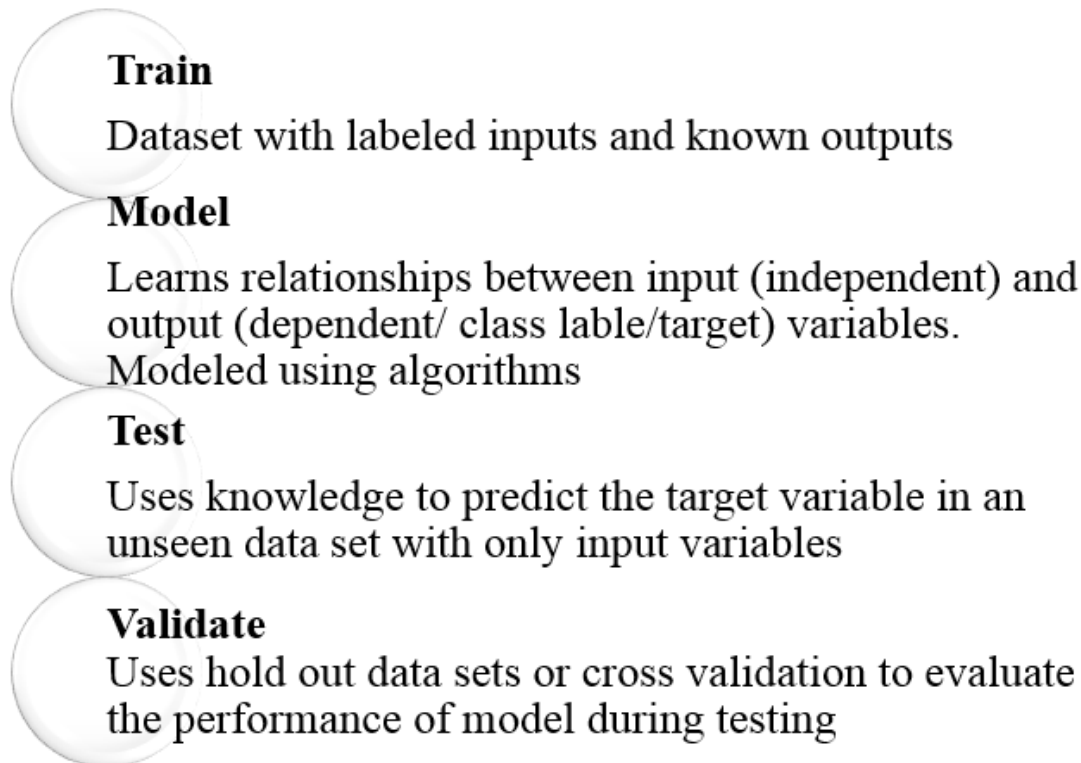


Figure 5 Machine Learning Processes - Train, Model, Test and Validate

The learning process in machine learning includes segmenting data into training and testing samples shown in Figure 5 . The training data is used by the machine learning algorithm to learn patterns in data and develop a model to predict. The testing data is used to check if the model works well on new unseen data. Measures such as accuracy and error rates are used to evaluate the performance of models during testing. A comparative analysis of how well a model performs with the training data set as against the testing data set is used to confirm if the classification algorithm is appropriate or not. This process is termed validation and can be achieved either by splitting the training data to create a holdout set used for testing. The usual practice is to split the

available data into training, and testing sets using a ratio of 70:30 or 80:20 (Kotu and Deshpande, 2014). The data set not used for learning, but testing is known as the hold-out set. Creating hold-out sets becomes challenging when the size of the training data set is small or in the case of imbalanced data. The testing set may not be a good representation of the samples found in the training set, causing inaccurate predictions. A possible solution is using the cross-validation technique with k-folds (Rapidminer, 2017) where k is a numerical value such as 5 or 10. The cross-validation operator ensures that the complete data set is divided into equal-sized k parts. One fold is used for testing, and the other k-1 folds used for learning. This step repeats for all folds. While this is time-consuming, cross-validation ensures a better performance.

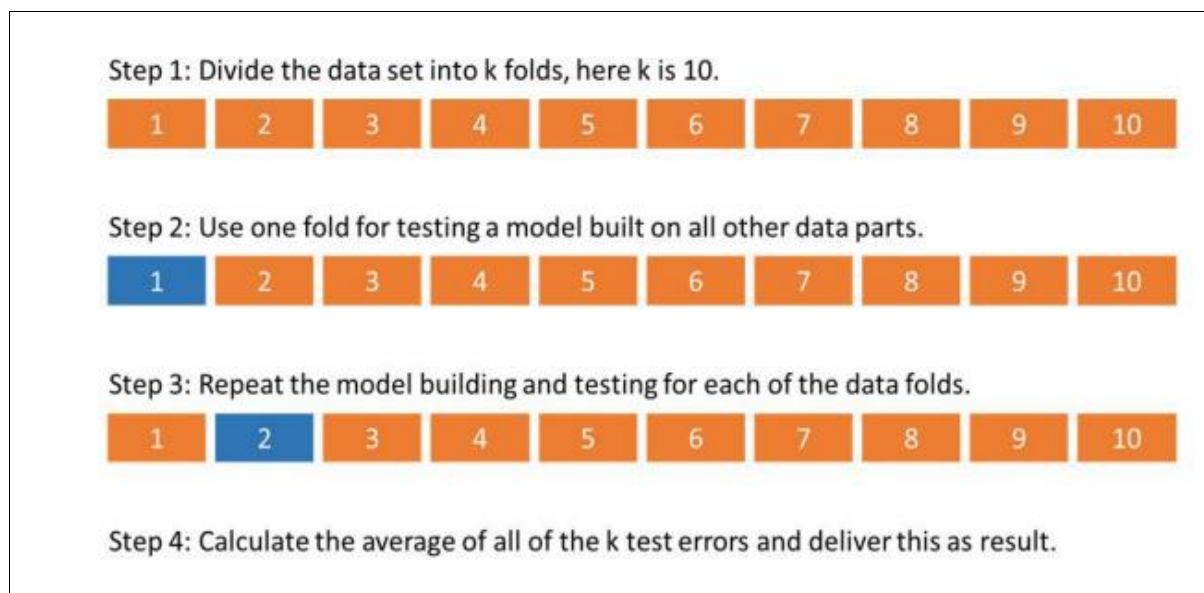


Figure 6 Steps in Cross-Validation. Source: How to Correctly Validate Machine Learning Models (Rapidminer, 2017)

Section 2 - Educational Data Mining - EDM

Several studies describe Educational Data Mining or EDM as the application of data mining techniques to educational data characterized by high volumes of transactions. In the educational sector, there is an increase in the usage of data mining tools to generate visual statistics, variety of useful information and help to discover new knowledge (Romero and Ventura, 2010; Peña-Ayala, 2014; Sgouropoulou et al., 2014b; Kaur, Singh and Josan, 2015; Fernandes et al., 2019).

Data mining applications in EDM

Educational institutions also have distinct stakeholders such as students, teachers, course developers, student advisors, academic administrators, educational technology providers, and facilities administrators (Romero and Ventura, 2010; Peña-Ayala, 2014). The expectation and information requirement of each of these stakeholder groupings have created several opportunities for data mining applications. Table 3 presents the applications of data mining tasks per stakeholder group and themes in the educational sector. Author citations suffixed with a * indicates that the study from the UAE.

| Stakeholder Group | EDM-Academic Theme | Application | Data mining tasks | Authors of Papers |
|---------------------------|--------------------------|---|---|--|
| Students | Learning | Analyze performance in assessments and online learning Develop adaptive learning experiences implemented in Recommender Systems or Tutorial systems | Classification, Clustering, Predicting | (Abdous, He and Yen, 2012), (Anzer, Tabaza and Ali, 2018)* |
| Teachers/ Administrators | Performance | Analyze student's preferences for learning, style of instruction and usage of instructional material Analyze grades from assessments to reveal patterns that help identify students at risk or predict outcomes. Analyze the relationship between student performance and social factors and other indicators | Data visualization, classification, clustering, prediction, regression, text mining | (Saa, 2016)* (Costa <i>et al.</i> , 2017) (Kotsiantis, 2012) (Đurđević Babić, 2018) (Mueen, Zafar and Manzoor, 2016) (Ahadi <i>et al.</i> , 2015) |
| Course/ Curricula Writers | Assessment | Analyze student performance in course assessments to reveal levels of difficulty Analyze assessment items for relevance to cognitive levels or learning outcomes defined in the curricula | Data visualization, Classification and text mining | |
| Advisors | Advising | To study patterns in student performance in various courses and provide suitable recommendations for course enrollment, study plans, career choices, selection of major, etc. | Data visualization, classification, and prediction | (Chanamarn and Tamee, 2017) (Vialardi <i>et al.</i> , 2011) |
| Administrators | Enrollment and Attrition | Analyze student academic performance data and enrollment data and alumni employment data to create relevant programs and course offerings. Analyze student assessment to predict progress, attrition, success/failure rates. | Classification, Prediction, Regression | (Fadzilah Siraj and Mansour Ali Abdoulha, 2011) (Aher and Lobo, 2013) (KumarYadav and Pal, 2012) (J. Kovacic, 2017) (Nandeshwar and Chaudhari, 2009) |

| Stakeholder Group | EDM-Academic Theme | Application | Data mining tasks | Authors of Papers |
|------------------------------------|--------------------|--|--|---|
| Governing Councils/Senior managers | Governance | To analyze student engagement and satisfaction with courses, teachers, learning, and other facilities. To build optimized operational models for the usage of facilities and registration into courses and programs. | Correlation Analysis, prediction, classification, association analysis | (Naaaj, Nachouki and Ankit, 2012)* (Wilkins and Balakrishnan, 2013)* |

Table 3 EDM themes and applications for various stakeholders in education. Based on (Romero and Ventura, 2010)

An extensive literature survey of 240 items of EDM research conducted between 2010 and the first quarter of 2013, conducted by (Peña-Ayala, 2014) summarizes EDM functionalities, as shown in Table 4

Table 10

Counting of EDM approaches organized according to six functionalities that are presented in Sections 3.1 to 3.6.

| Functionalities that represent the EDM approaches introduced in Sections 3.1 to 3.6 | Counting | Percentage (%) |
|---|----------|----------------|
| 1. Student behavior modeling | 48 | 21.62 |
| 2. Student performance modeling | 46 | 20.72 |
| 3. Assessment | 45 | 20.27 |
| 4. Student modeling | 43 | 19.37 |
| 5. Student support and feedback | 21 | 9.46 |
| 6. Curriculum, domain knowledge, sequencing, teacher support | 19 | 8.56 |
| Total | 222 | 100.00 |

Table 4 Summary of Related work in Educational Data Mining (Peña-Ayala, 2014)

Related Work – Educational Data Mining

J. Kovacic, 2017.

The stated objective of this study by (J. Kovacic, 2017) is to build models to predict the pass rate in a specific course, evaluate models, and present findings. To this purpose, institutional course and enrollment data are explored using pivot tables and descriptive statistics. Their project uses the CRISP-DM methodology. Data preparation activities include attribute creation, feature selection data cleansing, data merging. This study uses 483 rows of data and variables relating to student demographics and academic environment to predict the outcome of the student's performance in a course.

The implementation uses four types of decision trees i.e., CART, CHAID, exhaustive CHAID, and QUEST. Models are evaluated using cross-validation, prediction accuracy, and gain charts. The results indicate that accuracy levels of all trees were less than 61% with CART, which is a regression model performed best. This research concludes that profiling students who are likely to fail by creating a set of rules can be used to identify students that are likely to fail and can be used to apply intervention, thereby preventing failure. However, this model needs to be tested with other classification models and with more features such as high school results, work experience that students may have had, etc. to make the study more reliable.

Wanjau, Okeyo and Rimiru, 2016.

The objective of this research by (Wanjau, Okeyo and Rimiru, 2016)

was to find the impact of personal, socio-demographic and school level factors to classify students who could enroll or not into the STEM courses in higher education.

This study used the CRISP-DM methodology for data mining. This study analyzed 232 responses to a survey administered to students enrolled in the university. Data

preprocessing includes attribute removal, replacing missing values, feature selection, and dimensionality reduction. Data transformation and analysis is performed using WEKA. This study includes the attribute ranking and selection as part of the preprocessing activity. Models are built using six classification algorithms like the CART - Decision Tree, K-NN, Artificial Neural Networks, Support Vector Machine, and logistic regression with training and testing datasets. Evaluation of performance is done using error and kappa statistics, confusion matrix, and prediction matrix.

Results of this study indicate that feature selection can indicate the most impactful predictors.

The decision tree classifier achieved the best prediction accuracy, as seen in Figure 7. The K-

Table 6: Performance Measures of the Classifiers using 10-fold Cross Validation

| | <i>Accur acy</i> | <i>Miscalcul ation Rate</i> | <i>ROC Area (AUC)</i> | <i>Precisi on</i> | <i>Speed</i> |
|--|----------------------|-------------------------------------|-------------------------------|-----------------------|--------------|
| <i>Decision Tree (CART)</i> | 85.18 | 14.81 | 0.884 | 0.835 | 0.05 Sec |
| <i>Naïve Bayes</i> | 75.30 | 24.69 | 0.571 | 0.701 | 0.02 Sec |
| <i>Artificial Neural Network (MLP)</i> | 70.37 | 29.62 | 0.576 | 0.664 | 0.13 Sec |
| <i>k-Nearest Neighbor</i> | 77.77 | 22.22 | 0.6 | 0.733 | 0 Sec |
| <i>Support Vector Machine</i> | 76.54 | 23.45 | 0.512 | 0.682 | 0.02 Sec |
| <i>Logistic Regression</i> | 75.30 | 24.69 | 0.544 | 0.701 | 0.02 Sec |

Figure 7 Source (Wanjau, Okeyo and Rimiru, 2016)

NN took the shortest time to produce the result while A-NN took the longest. This paper presents a comprehensive study of an application in educational data mining.

Fadzilah Siraj and Mansour Ali Abdoulha, 2011.

This paper by (Fadzilah Siraj and Mansour Ali Abdoulha, 2011) relates to the evaluation of factors relating to student enrollment. The purpose of the study is to evaluate the impact of student gender and program/faculty on a student's enrollment status, which is a multi-class label.

This study follows the CRISP-DM methodology and uses 6830 records with 38 features from the university. Data mining techniques such as preprocessing, modeling are included. The study conducts an iterative analysis of data to arrive at the best predictors for the target attribute, which is the student enrollment status without disclosing techniques used. Models are built using descriptive and predictive algorithms including

- Descriptive: Frequency Tables, Correlation Analysis, Cross tabulation analysis, Clustering with Kohonen networks.
- Predictive: Logistic Regression, Decision Trees and Neural Networks

Evaluation of models is done using classification accuracy, confusion matrix, and analysis of clusters. The narrative includes description experiments conducted and the results. Neural networks performed best in terms of accuracy. This study successfully employs several data mining techniques and demonstrates the usefulness of these in student registration. However, it does not include plans for implementing necessary changes based on the results of the study.

Nandeshwar and Chaudhari, 2009.

This research by (Nandeshwar and Chaudhari, 2009) relates to the prediction of student enrollment using admissions data and applies the CRISP-DM methodology. Data were extracted and preprocessed to reduce rows, combine features, and convert data type in MS Access. This study includes data visualized using charts developed in Weka. Preprocessing includes attribute creation, data filtering, and data subset selection. Feature selection is implemented using Wrapper and Info Gain. Data were modeled using Decision Trees, Naïve Bayes, and RIDOR, which is also a rule-based classifier. Performance of the algorithm was evaluated using a ten-fold Cross-Validation. This paper compares the accuracy measures for each learning algorithm and uses t-test used to measure confidence. Deployment of experiments and analysis of results indicate that J.48 decision tree performed best. The study concludes that financial aid was an important factor in ensuring student enrollment.

Vialardi et al., 2010.

The goal of the study by (Vialardi *et al.*, 2010) is to provide student recommendations for course enrollment based on similarities with profiles of successful students. This paper uses institutional data relating to courses and student performance to classify students. Preprocessing activities such as data normalization, aggregation, creation of synthetic attribute, cleaning, and filtering are applied. Models are built using basic classification and ensemble algorithms. Evaluation is conducted using error rates and paired t-tests to reject the null hypothesis. Deployment phase included specific objectives for empirical verification of data. Data experiments are evaluated to prove/disprove the hypothesis. Recommender system was built based on the best performing model and deployed with the student enrollment module. Results from this implementation were measured and evaluated. This paper is good example of how data

mining projects should be taken from a model to an operational system. The paper does not discuss any post implementation reviews hence challenges or success of the project is not available,

Section 3 - Datamining Methodology

As part of the literature review, a detailed study was conducted to obtain insights into data mining methodologies. The goal is to identify the most appropriate methodology for data mining projects. This section summarizes the salient features of the most commonly used methodologies, CRISP-DM, KDD, SEMMA, and ASUM.

A methodology is a set of procedures, processes, tools, and techniques necessary to complete a certain task is necessary to manage data mining projects. In the opinion of (Saltz *et al.*, 2018), not following a methodology can lead to several problems such as inefficient collaboration and communication, scope creep due to poorly defined requirements leading to project overruns and poor engagement of stakeholders.

An initial survey of papers relating to such projects indicated that the most common methodologies for conducting data mining projects were CRISP-DM, SEMMA, and Knowledge-Discovery KDD. Of these methodologies, CRISP-DM is the most popular, as indicated by a KDNugget's poll conducted in 2014 (Piatetsky, 2014). Figure 8 below shows the results of KDNugget's poll in 2014 and shows that CRISP-DM is the most used methodology.

The article also suggests that in comparison to the 2004 poll, CRISP-DM has remained the same while the application of SEMMA has increased. Almost 38% of those who responded used methodologies specific to their personal preference or those of their organizations', indicating a lack of standardization in this area.

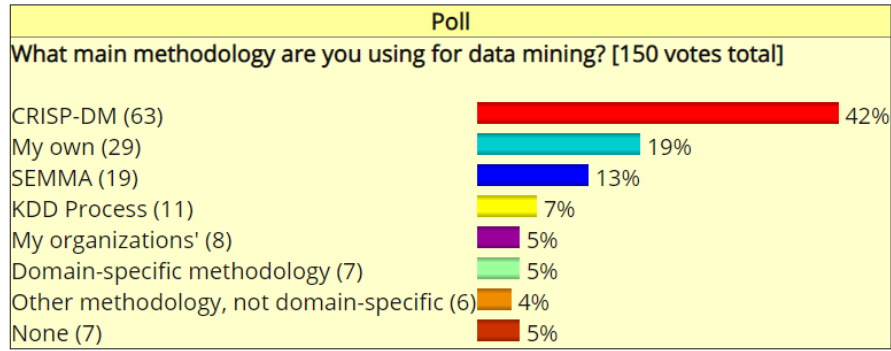


Figure 8 Main Data Mining methodologies. Source: (Piatetsky, 2014) Retrieved from https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

In a more recent study (Saltz *et al.*, 2018) indicate that more data mining projects are beginning to use agile methodologies. However, more than 80% of the respondents were either not sure of the methodology being used or apply processes as they go along in the project.

The study by (Saltz *et al.*, 2018) also categorizes the approaches to data mining projects, as shown in Figure 9:

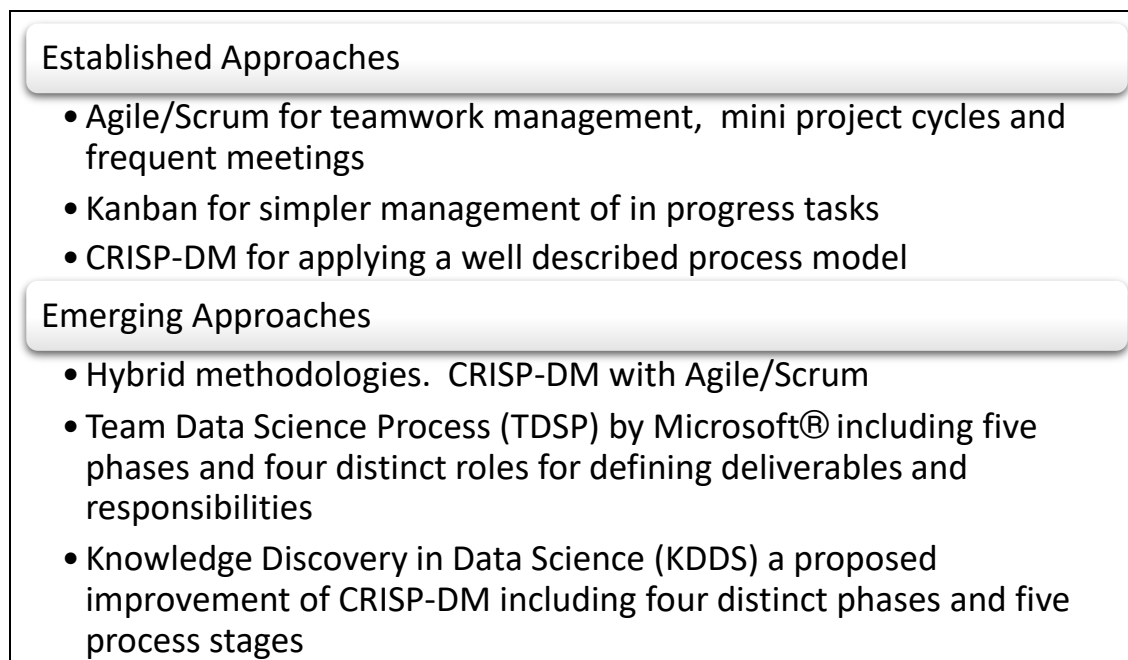


Figure 9 Approaches to managing Datamining Projects (Saltz *et al.*, 2018)

A survey of data mining methodologies conducted by (Mariscal, Marbán and Fernández, 2010) discusses fourteen data mining methodologies indicating that Knowledge Discovery (KDD) and CRISP-DM as being the most popular. In their study (Daderman, Antonia; Rosander, 2018) discuss that KDD, CRISP-DM, and SEMMA as being the most popular methodologies for managing data mining projects.

Data mining project implemented without a proper methodology can lead to several issues, including inefficient management of teamwork, slow information dissemination, poor requirement gathering, and stakeholder engagement, poor coordination of project activities, repetition of tasks due to lack of reproducible objects, poor documentation, scope creep leading extended project life cycles and projects that cannot be deployed to production.

The following section describes the most popular methodologies:

The Cross-Industry Standard Process for Data Mining or CRISP-DM

This de facto industry standard was introduced in 1997 by a consortium of five organizations as a non-proprietary and freely available tool to standardize data mining projects. (Chapman et al., 2000). This framework adopted by IBM® and incorporated into SPSS® is a popular tool for data scientists. (IBM, 2011). While the CRISP-DM framework has not progressed into newer versions, it is still the most popular methodology for data scientists as indicated by the KDnuggets poll conducted in 2014. (Piatetsky, 2014).

The CRISP-DM process model consists of 6 interrelated phases, as shown in Figure 10.

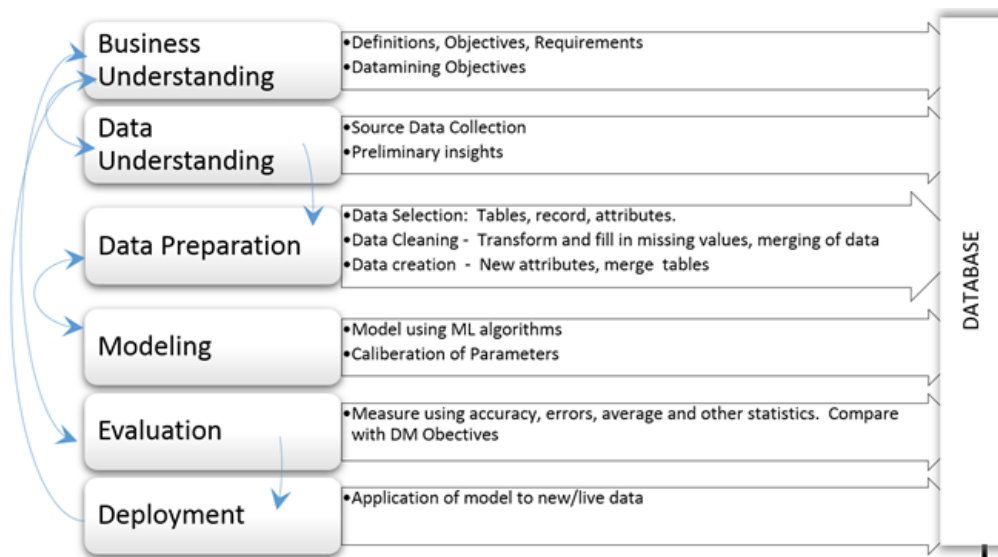


Figure 10 CRISP-DM Model. Author's own work based on (Chapman et al., 2000)

This model is more of an agile model than a linear model and allows for interactions between phases noted by (Chapman et al., 2000; Nandeshwar and Chaudhari, 2009).

Business understanding of drivers for data mining, assessment of the situation, and development of goals and objectives for a data mining project is considered important. Data understanding is akin to collecting and analyzing requirements and provides an insight into the quality of data and its readiness for modeling. Data preparation, modeling, and evaluation are typical data mining tasks and are iterative. During this stage, it may be necessary to revisit project objectives and goals. Deployment, which is the final phase of the CRISP-DM model is not prescriptive enabling adopters to define the depth and breadth. Deployment could be the experiments

conducted with test data or the implementation of new systems with approved models for real users.

As with any life cycle model, the CRISP-DM model is a continuum of processes. After deployment, the system is in use until the objective/requirements change. In this case, the focus is back to the first phase of business understanding.

Analytics Solutions Unified Methodology or ASUM

ASUM is an acronym for Analytics Solutions Unified Method and is a methodology advocated by IBM Corporation to manage data mining projects (IBM Analytics *et al.*, 2016). ASUM is a hybrid methodology incorporating principles of agile methodology into the CRISP-DM model. In their study (Beyadar and Gardali, 2011) present the ASUM methodology as a refined CRISP-DM model as it includes better business understanding and project management activities to communicate project information, monitor project quality and deploy the solution. There are six iterative phases engaged in producing knowledge and project management processes applied to monitor implementation, as shown in Figure 11.

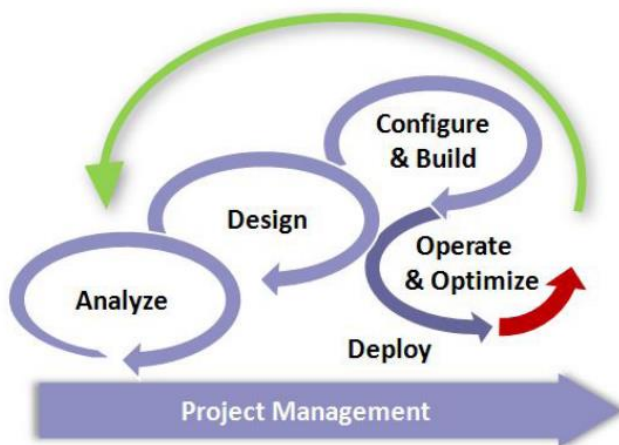


Figure 11 ASUM data mining methodology Source: IBM Corporation (IBM Analytics *et al.*, 2016)

This methodology proposed by (Fayyad, Piatetsky-Shapiro and Smyth, 1996) to unify data mining and knowledge discovery consists of five phases necessary to generate knowledge from large sources of data. These phases are Selection of Data, Pre-Processing, Transformation, Data Mining, and finally, Interpretation/Evaluation. The KDD model embeds the data mining processes, which is considered integral to knowledge discovery. The KDD is a model which is interactive and iterative and is generally user-driven (Mariscal, Marbán and Fernández, 2010).

Figure 12 shows the flow of processes through the five phases of knowledge discovery.

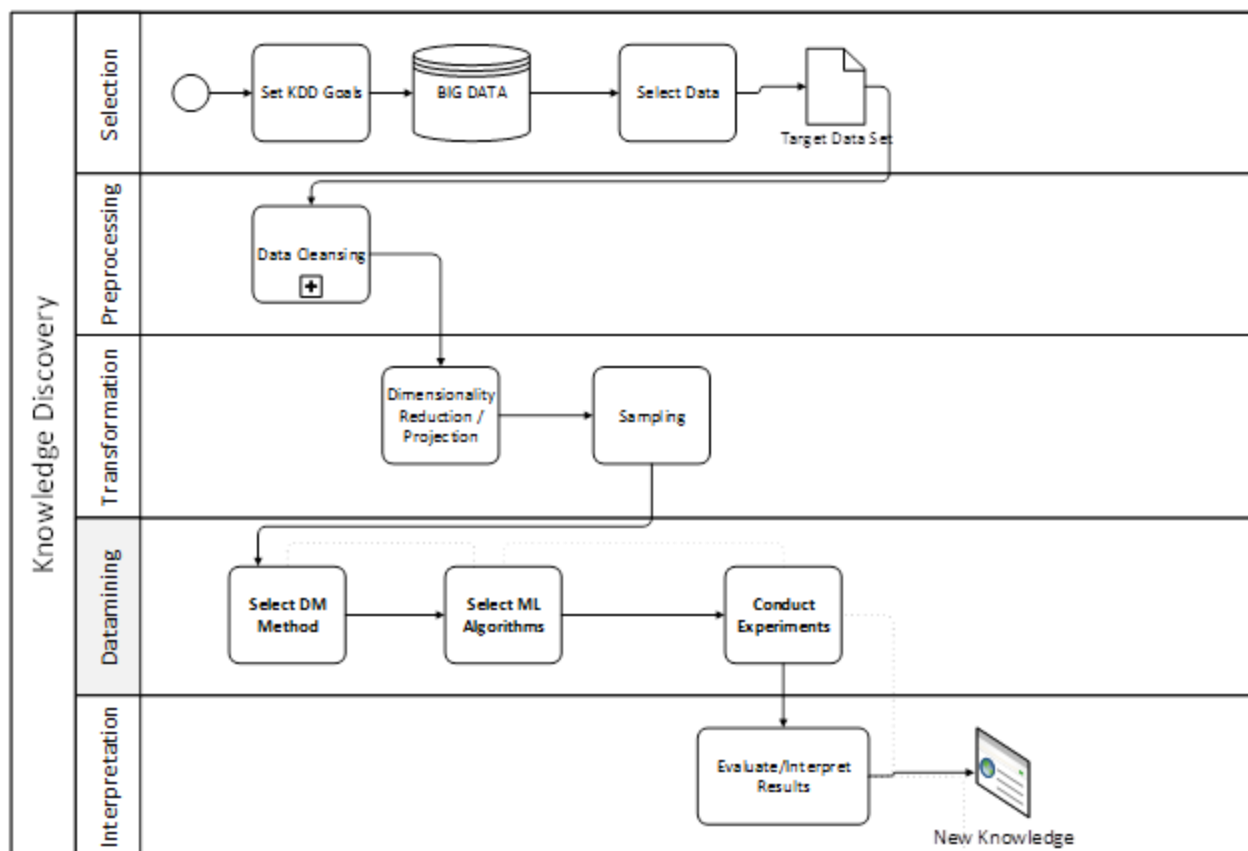


Figure 12 Author's model of the Knowledge Discovery Process based on (Fayyad, Piatetsky-Shapiro and Smyth, 1996)

In the KDD methodology, the data mining processes refer to selecting data mining methods and relevant algorithms. Data mining processes are also iterative and requires going back and forth until the required results are achieved, discovering new patterns or knowledge. The basic disadvantage of this model is that it does not consider the business need to discover new knowledge, or the computerization of knowledge discovery, or the feasibility of these projects.

Sample, Explore, Modify, Model and Assess or SEMMA

SEMMA is an acronym for Sample, Explore, Modify, Model, Assess, and is a data mining methodology developed by SAS Corporation. This methodology is implemented within the Enterprise Miner software suite developed by SAS. (SAS Institute Inc., 2017). The SEMMA methodology includes the following processes

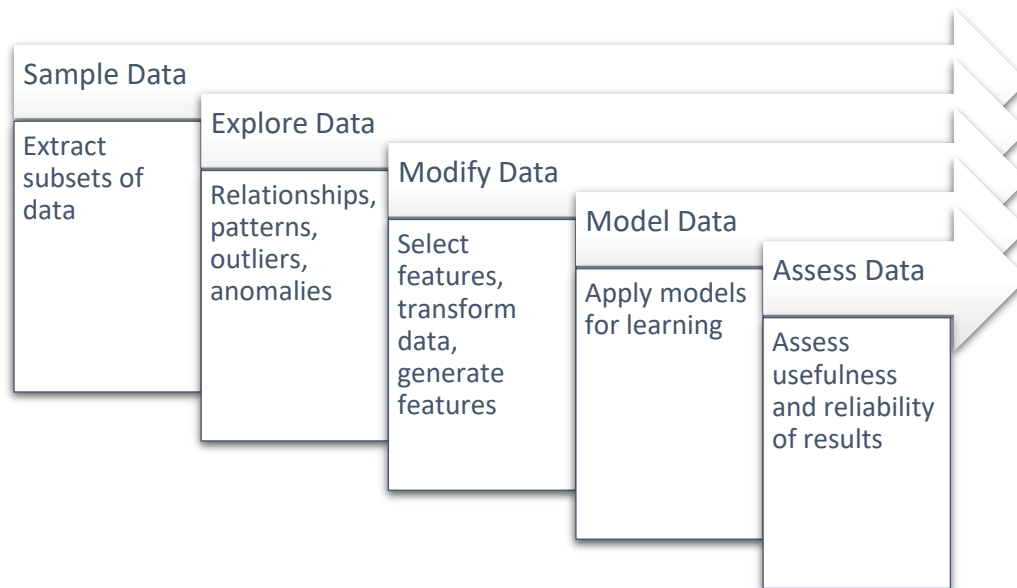


Figure 13 SEMMA based on information in (SAS Institute Inc., 2017)

The main drawback of this methodology as compared to KDD and CRISP is that SEMMA is associated with proprietary software. The SEMMA methodology is critiqued for not including phases to understand the business domain or deploy knowledge once it is learned (Mariscal,

Marbán and Fernández, 2010). Hence it is better used in conjunction with other methodologies that advocate these phases.

Related work – Data Mining Methodologies

Saltz *et al.*, 2018

In their study (Saltz *et al.*, 2018), present a list of problems that arise by applying ad hoc methodologies to manage data mining problems. These problems include The authors also suggest there is very little research available to evaluate the methodologies used and improvements necessary to achieve better data mining projects. To address this gap, the author's survey 78 professionals involved in various roles associated with data mining projects. The goal of this survey was to understand the current methodologies used and whether the incorporation project management processes are necessary for data mining projects. The findings of their survey are that less than 20% of the surveyed participants used specific methodologies such as CRISP-DM or agile methodologies. Most respondents - 85% also agreed that the outcome of their projects would improve by applying a systematic process methodology.

Daderman, Antonia; Rosander, 2018

The study conducted by (Daderman, Antonia; Rosander, 2018) conducts a comparative analysis of CRISP-DM, SEMMA, and KDD methodologies for a data mining project with Saab. The motivation for this study is to identify the best data mining methodology of the three most commonly used ones.

A detailed study of the three methodologies and a summary of cases is presented. The paper presents the advantages and limitations of each methodology. The study indicates that the advantages of CRISP-DM are its well-defined processes, availability of documentation, non-

dependency on specific software tools, and support for data mining techniques. The advantages of this methodology outweigh its disadvantages. The KDD is disadvantaged by limited documentation and SEMMA for its dependence on the SAS software and for not including an understanding of the business case. The study also developed criteria for evaluating these methodologies by interviewing users and studying the organizational culture. These criteria included whether the methodology

- Considers business perspective
- Is distinct in usage
- Is Suitable for Saab's project development

The authors conclude that the CRISP-DM is the most appropriate methodology for the data mining project at Saab.

Angée *et al.*, 2018

The research by (Angée et al., 2018) presents a use case for the application of ASUM methodology for managing a big data project. The author's reason for adopting ASUM is owing to the absence of project management and knowledge management practices within CRISP-DM and SEMMA. The reason for using the ASUM methodology is because it incorporates project management and supports a systems life cycle making. The paper also emphasizes the need for a methodology to manage data mining projects implemented across disciplines and multiple data locations.

The authors present a use case of a large bank requiring a complete data analysis of their corporate customers. This project used a modified ASUM methodology incorporating best practices from project management and other systems life cycle methodologies. This study does

not include details of the data mining activities conducted. However, it is indicated that teams working were geographically dispersed requiring the use of collaborative methods to ensure successful completion of a project.

The salient features of the use case include the following additional processes to supplement the ASUM methodology:

- Analysis Phase: Implementation of Big Data's 5 V's model in by assessing the volume and variety of data, the velocity of data produced, veracity or reliability of data and value realization of data mining.
- Build Phase: Build and evaluate a prototype to ensure the engagement of stakeholder requirements. Define a workflow amongst multiple organizational units using the BPMN model to designate ownership and responsibilities of processes.
- Project management activities: Create a project plan documenting risks, constraints, and project management methodologies. Conduct weekly meetings and virtual meetings to track project status. Lessons learned shared amongst stakeholders. Implement Configuration management to ensure consistency and standardization of deliverables. Create document management systems to manage the traceability of decisions and store all technical documentation for further reference.

This use case has demonstrated that in the present day context where data mining projects are conceptualized and managed at organizational levels, strong project management and better business understanding techniques are essential to ensure project success.

Rogalewicz and Sika, 2016

In their study (Rogalewicz and Sika, 2016) discuss a generic framework with three stages for data mining, namely, Preprocessing, Main Processing, and Post-processing with activities and resources recommended for each stage. They also discuss the key factors of the most frequently used frameworks: CRISP-DM, KDD, SEMMA, and VC-DM. They highlight the importance of business understanding required in CRISP-DM and to some extent, in the VC-DM frameworks. The section on data mining methods shows the grouping of methods under four categories: Description, classification, regression, clustering, and association. The application of data mining in mechanical engineering is categorized and tabulated with concise descriptions. This paper concludes that the CRISP-DM is a better choice for implementing data mining projects as it considers all the phases in detail.

Table 1
Comparison of selected Data Mining methodologies in three main aspects of knowledge discovery process [own work].

| DM Methodology | Pre-Processing | | Main-Processing | Post-Processing |
|--|--|---|-----------------|-------------------------------|
| CRISP-DM <i>Cross-Industry Standard Process for Data Mining</i> | Business Understanding Data Understanding Data Preparation | | Model | Evaluation, Deployment |
| KDD <i>Knowledge Discovery in Database</i> | Selection, Pre-Processing Transformation | | Data Mining | Interpretation and Evaluation |
| SEMMA <i>Sampling, Exploration, Modification, Model, Verification</i> | Sample, Explore, Modify | | Model | Assess |
| VC-DM <i>Virtuous Cycle of Data Mining</i> | Identify | Transform (Pre-Processing and Main-Processing) | | Act, Measure |

Figure 14 Source: Rogalewicz and Sika, 2016

Mariscal, Marbán and Fernández, 2010

In this paper, the authors (Mariscal, Marbán and Fernández, 2010) present a detailed overview of various process models that help data scientists work on data mining and knowledge discovery projects. This paper is a complete study of features, methodologies, and evolution of fourteen data mining models and includes an in-depth analysis of KDD and CRISP-DM. Includes features of all other process models mapped to the CRISP DM and the KDD models. This study recommends a new process model named "Refined Data Mining Process" to address the gaps in the CRISP-DM.

The study refers to the development of the CRISP-DM process model that was independent of tools and techniques as a milestone since it listed all activities in each of the six phases. Inputs and outputs for each task were detailed to ensure reusability, standardization, and improve the quality of data mining projects. The authors indicate that the major drawback was that the CRISP-DM provided a list of "What to do?" and not "How to do?"

The authors also present the notion that data mining is a subprocess of the knowledge discovery process and that data mining is used to produce knowledge. Researchers and industry experts use KDD and DM interchangeably.

This paper proposes a new methodology to incorporate project management techniques into the CRISP-DM. The Refined Data Mining Process consists of three main process groups and 17 subprocesses. The figure below maps the Refined Data Mining processes to the CRISP-DM phases.

| | Processes | Subprocesses | CRISP-DM Phases |
|--|--------------------|------------------------------------|--|
| Refined DM Processes Marban et. al 2010 | <i>Analysis</i> | Life cycle selection | Business Understanding Data Understanding |
| | | Domain Knowledge elicitation | |
| | | HR Identification | |
| | | Problem specification | |
| | | Data prospecting | |
| | | Data cleaning | |
| | Development | Preprocessing | Data Preparation Modeling Evaluation Deployment |
| | | Data reduction and projection | |
| | | Choosing the data mining function | |
| | | Choosing the data mining algorithm | |
| | | Build model | |
| | | Improve model | |
| | | Evaluation | |
| | | Interpretation | |
| | | Deployment | |
| | | Automate | |
| | <i>Maintenance</i> | Establish On-going Support | |

The paper does not include a case relating to the implementation of the proposed model.

Sharma and Osei-Bryson, 2009

In their study (Sharma and Osei-Bryson, 2009) discuss the CRISP-DM framework in detail.

However, the focus of the paper is the development of a framework for understanding the business drivers for data mining, which they identify as poorly defined in the CRISP-DM I. The authors suggest four mandatory tasks to understand business perspective as determination of

business objectives, situation assessment, determination of data mining goals and a project plan for implementation. Each of these tasks has specific outputs that feed into all phases of the CRISP-DM framework. Interactions of the business understanding in CRISP-DM is limited to the data understanding and evaluation phases. The suggested improvements include performing risk identification and analysis, assessment of resources, cost estimations, cost-benefit analysis, and establishment of success criteria.

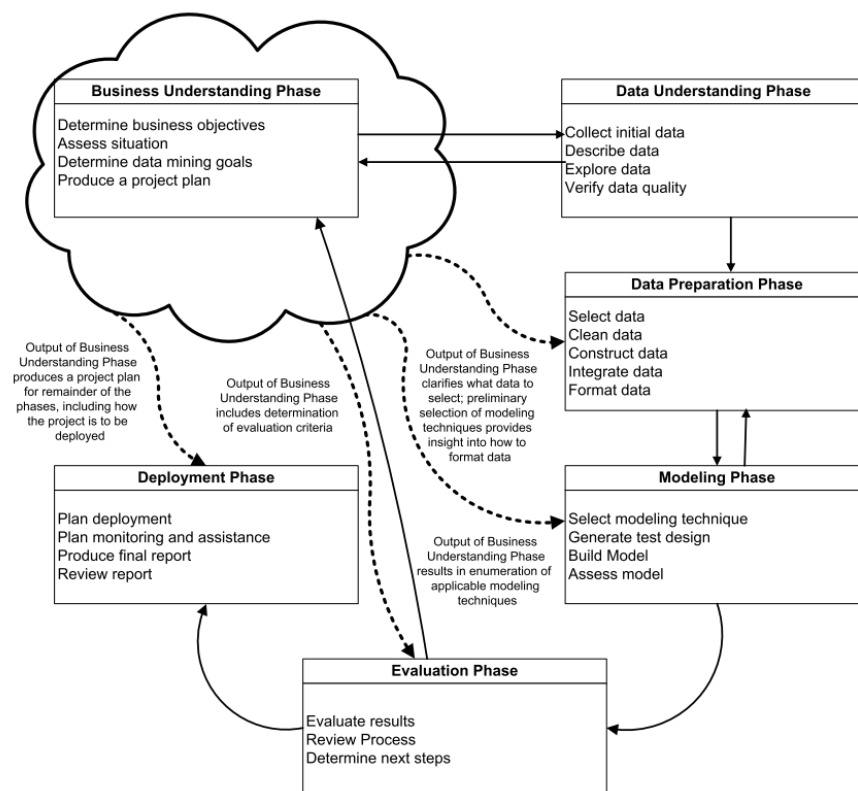


Fig. 2. BU phase pervades all phases of the DM project.

Figure 15 Source: Sharma and Osei-Bryson, 2009

This modification for business understanding phase is applied to a financial services company for assessing credit risk, and a case study is presented. This research conducted in 2009 has adopted best practices from project management, analyzing the feasibility of data mining

projects and developing goals and objectives applying a strategic management framework, which is relevant in the current times as well.

Section 4 - Datamining Classification Algorithms

This section presents a literature review of classification algorithms used in predictive tasks. The data mining requirement of my study is to predict the major that students are likely to select based on several input variables. The target variable is multi-class polynomial value being one of the following: SECURITY, APPSDEV, BUS-SOL, NETWORK, MULTI-MEDIA. The imbalance ratio of the highest class to the lowest class (694 to 18) is 38%, indicating class imbalance shown in Figure 16.

Predictive tasks with imbalance are a challenge because algorithms are biased towards the majority class while the minority class may be more significant from a domain perspective. For example, in the case of universities, it may be more interesting to know the profile of students that are choosing majors such as NETWORK and MULTI-MEDIA to encourage other students to enroll in them.

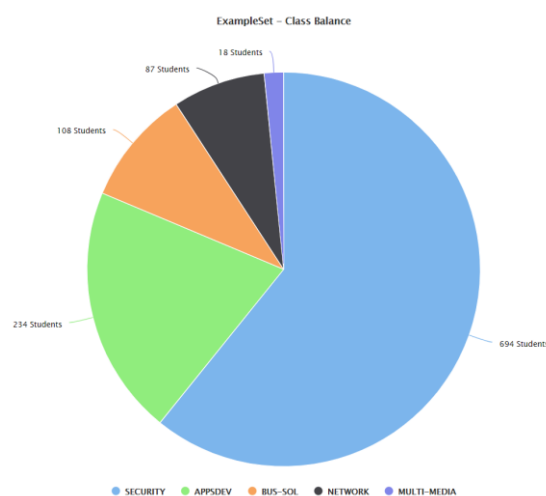


Figure 16 Distribution of classes

There are several ways to manage the modeling imbalanced dataset to ensure efficient processing. Data level methods include oversampling and undersampling techniques to form balance data; modify existing algorithms to reduce bias towards the majority class by assigning costs and hybrid methods (Krawczyk, 2016).

Hence my study focuses on those algorithms that are suitable to multi-class, imbalanced datasets.

Classification Algorithms

Classification or the prediction of a class is an activity that involves using data in the input variables to segregate data into the predictand classes. The most common machine learning or data mining tasks for classifications are explained in the following sections.

Decision trees

In machine learning, decision trees are rule-based algorithms used to split datasets on homogenous values until a decision or a classification is reached. The initial data is progressively divided into smaller subsets based on “attribute-selection” criterion. (Miguéis *et al.*, 2018). This process of division or branching is repeated until a target class, or a leaf node is reached.

Decision trees can work on categorical or nominal data for classification and numerical data for regression analysis. Several studies relating to multi-class labels use decision trees as they are easy to interpret and work well with a few relevant attributes (Ashari, Paryudi and Min, 2013). Decision trees also help to understand the relationship between the target variable and the predictors. (Đurđević Babić, 2018). However, they are also prone to overfitting, especially with multi-class labels. Studies also recommend reducing the depth of the tree or pruning to obtain better performance and reduce overfitting. (KumarYadav and Pal, 2012)

As can be seen in Figure 17, the best predictor attribute is at the highest level in the tree and is known as the root node. The root node is split into leaf nodes based on rules that are inferred by the algorithms. Splitting ends when a leaf node or the target variable class is reached. In this Figure 17, the best predictors for the various classes are the High School Average and the grades in courses Hardware and Networking, Information Systems and Programming.

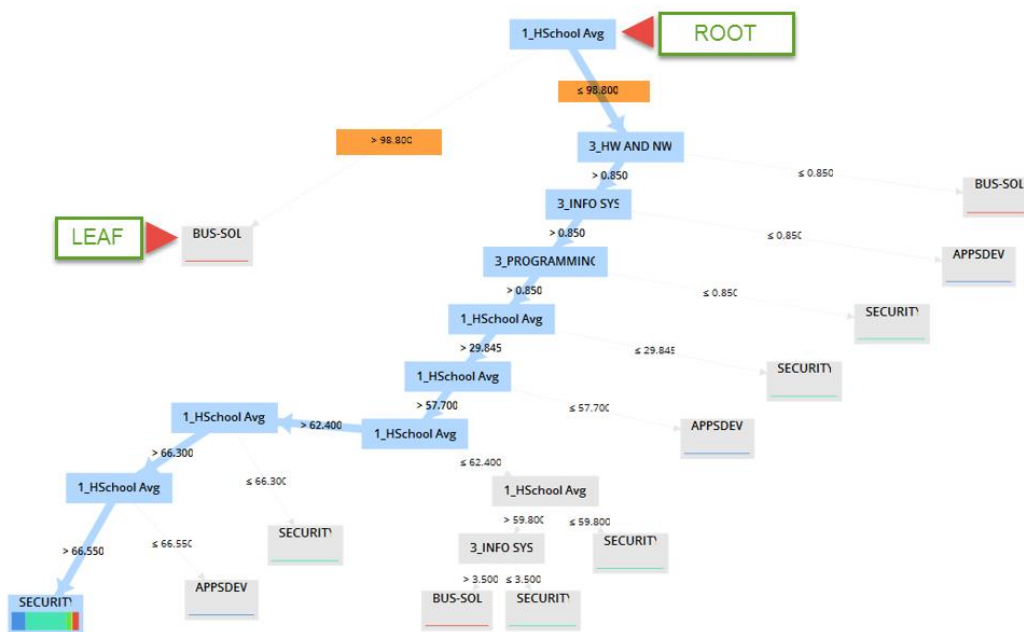


Figure 17 Sample Decision Tree

Random Forest

The Random Forest classification algorithm was created by Leo Breiman and Adele Cutler and is an ensemble of decision trees. (Salford Systems, 2019)

In this algorithm, sampling is random, and a voting system ensures that each tree votes on the classification, with the highest vote considered (Ben Youssef et al., 2018). The most predicted class or the class with the highest voted is the prediction for the forest. This algorithm is known

to provide high levels of accuracy, particularly where there is a high number of input variables and large data sets. It has also been used with imbalanced class data sets and is known to balance errors. Apart from classification, Random Forests can also be used rank features or variables for prediction and determine the statistical significance of intervention. (Spoon et al., 2016). In a study relating to student enrollment, random forests performed best both in terms of accuracy as well as handling imbalanced data (Chau and Phung, 2013).

Naïve Bayes

The Naïve Bayes is a classification algorithm derived from Naïve Bayes theory that believes attributes used for classification are mutually independent (Jishan *et al.*, 2015). It is a probabilistic model that calculates the probability of the target class variable for the given values of the predictor variables. (Kotu and Deshpande, 2019). Naïve Bayes is preferred since it is faster to compute than the other algorithms. It is also versatile, used for both binary and multi-class prediction tasks.

Support Vector Machines – SVM

SVM is a supervised learning classification algorithm based on the Vapnik-Chervonenkis theory which works by separating data into two areas using a hyperplane or a boundary. During training, a boundary is fitted for a data belonging to one particular class. During testing, each sample is evaluated for its fit inside the boundary or not.

SVM is appropriate to be used with text mining, imbalanced data sets or with image processing and hence can be applied to many real-world problems. However, the computational cost of this algorithm is high, making it difficult to use with datasets with a high number of attributes (Kotu, Deshpande, 2015).

Artificial Neural Networks – ANN

Artificial Neural Networks or ANN is a supervised learning algorithm based on the workings of the biological neural networks. The building blocks layers consisting of the input neurons, learning as a brain would the output layer. Popular models are the Multi-layer Perceptron and deep learning.

The ANN is used to model non-linear relationships between attributes using hidden layers of previously learned information about the relationships between inputs and outputs. However, the cost of this is high because it uses every row in the data set to estimate errors as the difference between the actual output and the predicted output. (Kotu, Deshpande, 2015)

The ANN model works only with binary/numeric data and does not handle missing data well. Hence data preprocessing is critical. This model is apt for predictive analytics with large volumes of numeric data such as test scores, the sale price of goods, etc.

Gradient Boosted Trees - GBT

GBT is an ensemble learning algorithm based on decision trees using a boosting technique. Decision trees are added one at a time to minimize the loss function. This classifier takes longer to execute, but can deal with categorical attributes and handle multi-class labeled data. In their study (Fernandes *et al.*, 2019) use this classifier for its ability to process large amounts of data at higher levels of accuracy and ranking variables based on how they influence the target variable.

Related work – Classification Algorithms

Pristyanto, Pratama and Nugraha, 2018

This research by (Pristyanto, Pratama and Nugraha, 2018) is about establishing the importance of data preprocessing to handle imbalanced data in the context of Educational Data Mining. The

secondary goal is to improve the performance of the SVM algorithm in handling multi-class imbalanced data.

The sampling techniques used are SMOTE and OSS or One-sided selection. OSS is about undersampling the majority class with careful selection of data to ensure proper representation while retaining all samples of the minority class. SMOTE is about oversampling the minority by creating synthetic samples. The study also conducts experiments using 403 rows of data obtained from the university. The nature of data is not specified, although class imbalance in target data is presented. The classification algorithm used is the Support Vector Machine – SVM algorithm with data split into training and test. The study does not mention the use of cross-validation. The evaluation measures used are accuracy, sensitivity, specificity, and geometric mean. The analysis includes a comparison of the performance of SVM without sampling, sampling with OSS, sampling with SMOTE, and sampling with OSS + SMOTE. The proposed mode of OSS+SMOTE performs the best across all evaluation measures, with specificity being the highest. While the study demonstrates that using SMOTE and OSS provides good results with SVM, there is no evidence of parameters used with SVM or a comparison with other algorithms to be certain.

Hartono et al., 2018

In their study ((Hartono *et al.*, 2018)) discuss the problems relating to class imbalance and how they impact prediction tasks. This paper discusses misclassification errors, the prediction accuracy of minority classes, and poor performance of the SVM classification algorithm. The proposed solution of implementing the Biased SVM and Weighted-SMOTE is presented. These models are applied to the Iris dataset and evaluated using the confusion matrix, accuracy, and sensitivity measures. The accuracy levels of 0.86 are proof of successful handling of imbalanced

data by the combination of Biased SVM classifier with weighted-SMOTE sampling techniques. The study, however, does not provide any comparative analysis of accuracy levels without the application of the specified methods.

Costa et al., 2017

In their paper (Costa et al., 2017) discuss the effectiveness of data mining techniques in identifying student progress in courses and the impact of fine-tuning algorithms on EDM techniques. Hence their study bridges both the business understanding and the technical application of data mining techniques. The context is the failure rates in an introductory programming course.

For comparison, two sets of data from sources relating to distance learning and on-campus course were used in to predict failures early in the course to implement interventions. Data preprocessing included data cleansing and formatting, attribute selection, and discretization. Features were based on weighted ranking using information gain and applying SMOTE sampling techniques.

The classification algorithms used were the Decision Tree, SVM, Neural Networks, and Naïve Bayes and evaluated using F-Measure, Precision, and Recall. The highlights of this study are the comparative analysis of model performance with and without preprocessing, with and without fine-tuning of classifier parameters. T-tests were used to evaluate the effectiveness of pre-processing and fine-tuning of algorithms. Overall the Decision tree classifier was able to predict with high levels of accuracy on both sets of data at early in the course to introduce intervention. The Decision tree was able to achieve high prediction for Distance learning at 50% of course

completion and 25% for the on-campus course. The study has not investigated the reason for this deviation since it was not within the scope of the study.

Krawczyk, 2016

In this paper, (Krawczyk, 2016) discuss issues and challenges that need to be considered to enhance knowledge relating to the mining of imbalanced data. The mode of research is literature review without practical implementation. This study includes a complete discussion on how imbalanced data should be tackled using three approaches: Data level preprocessing, algorithms for classification and hybrid models with ensemble classifiers with sampling techniques for both two-class and multiclass data. It also presents a list of applications of multi-class imbalanced data with the top three being image processing, sentiment analysis, and software code analysis for defect prediction. Interestingly educational data mining does not feature in this list of applications. The paper presents challenges to managing imbalanced data, particularly where there is extreme class imbalance and lack of studies in this area is stressed.

More, 2016.

This paper by (More, 2016) is a survey of resampling techniques with imbalanced data. The study uses a synthetic data set with two classes 10,000 sample rows of data. For analysis, the dataset is split into 70% training and 30% testing and validated using the five-fold cross-validation on the training set. The focus of this paper is to compare the undersampling of majority class and oversampling of minority class using a variety of sampling methods such as One side selection, ADASYN, SVM-SMOTE, SMOTE-Boost, cluster-based oversampling, Kernel-based methods and active learning. The model applied was logistic regression. Precision and recall measures were used to compare the results for the different types of sampling.

Eventually, the study found that the best performing sampling method using the linear regression as the classification method was SMOTE+ ENN (Edited Nearest Neighbor).

Sa, 2016

In this research, (Saa, 2016) aims to study the relationship between a student's personal and social factors and their academic performance for a university in the UAE. The data is gathered using a survey administered to university students, and 270 responses evaluated. The study includes twenty-five attributes of which twenty-three are personal and social factors, and the others relate to academic performance in high school and university. The predictand attribute is the student's GPA, which is multi-class with four distinct values. This data was analyzed using different types of Decision Tree classification algorithms: C4.5, ID3, CART, and CHAID C4.5. The confusion matrix and the accuracy and precision measures evaluate the performance of the classification algorithms. The study reports that CART performed best. However, the applied discretization technique did not help classifier to learn well. The analysis also included using the Naïve Bayes model to generate the probability distribution matrix to analyze the relationship between attributes and the predictand value. This paper provides an interesting application of using rule-based classifiers for classification and Naïve Bayes for understanding the relationships between attributes. The study can be extended to see if classification results would be different or better by including only those attributes that were found to be interesting. Another improvement would be the inclusion of some of the rules generated by the various tree algorithms for comparative analysis.

Shang et al., 2016

In their (Shang et al., 2016) have compiled a literature review of 162 papers from 2006 to 2016 relating to the management of imbalanced data and indicate that there is a growing trend in published research in this area. This paper discusses approaches to conducting literature surveys, pre-processing techniques such as sampling and feature selection techniques relating to imbalanced data, algorithms, and evaluations of imbalanced data. A survey of papers that used cost-sensitive learning is detailed. The lack of literature in this area happens due to the challenges in modifying algorithms necessary for cost-sensitive learning. Their study also reveals that most studies on imbalanced data relate to binary classes, and hence, they have constructed their experiments on multi-class learning of imbalanced data.

Research relating to classification algorithms focuses on ensemble classifiers, subgrouping them as iterative based applying Gradient Boosting and Decision Trees and Adaboost and parallel based classifiers including various bagging algorithms. The paper also summarizes published research involving modification of classifications relating to SVM, KNN, and weighted KNN, Naïve Bayes, Neural Network – Extreme Learning Machine.

This paper also records descriptions and application of evaluating metrics used for classifiers. Included is a count of papers using AUC/ROC, Accuracy/Error, G-Mean, F-Measure, Sensitivity, Specificity and Precision for binary classifications and Micro and Macro averages for use with multi-class predictions.

The final segment of this paper is a study of the application domains featuring imbalanced data. As per the survey, the high ranked sectors with studies in imbalanced learning are chemical and biomedical engineering, financial management, anomaly detection in information technology security sector. Interestingly, imbalanced learning relating to online-learning in the educational sector is indicated as an emerging trend.

Agrawal, Viktor and Paquet, 2015

This paper by (Agrawal, Viktor and Paquet, 2015) is specifically about the redressing the gap that exists in handling imbalanced datasets with multiple classes. Their study also discusses the issues arising from the use of class decomposition methods viz. One-vs-One(OVO) and One-vs-All(OVA). With the OVA, the multiclass is converted into multiple binary classes where each class is labeled positive while all other classes collectively become negative. In the case of three classes, the results of three binary classifications are voted or combined into one result. With OVO, each class results in multiple pairs of binary classifications. In the case of three classes, there would be six binary classifications which result in the learner either overfitting or underfitting depending on the degree of skewness between classes.

| Binary classifications in case of three classes – C1, C2, and C3 | |
|--|--|
| One-Vs-All (OVA) | One Vs. One (OVO) |
| Number of binary classifications = Number of classes | Number of classifications = Number of Classes x (Number of Classes – 1)/2 |
| C1 Vs. C2+C3 | C1 Vs. C2, C1 Vs. C3 |
| C2 Vs. C3+C1 | C2 Vs. C3, C2 Vs. C1 |
| C3 VS C1+C2 | C3 Vs. C2, C3 Vs. C1 |

Table 5 Binary classification with three classes

To address the problem, the authors propose a novel hybrid sampling algorithm called SCUT (SMOTE and Clustered Undersampling Technique). This technique attempts to balance the sampling within imbalanced datasets using clustering techniques for undersampling and oversampling, reducing the problems of samples being ignored or unlearned otherwise. The paper presents the algorithm which is trained and tested on seven multi-class datasets from the KEEL

repository. The classification algorithm applied were Decision Tree, K-NN, SVM, and Naïve Bayes with ten-fold cross-validation. The performance of models was evaluated using G-Mean, F-Measure, and AUC. The other comparative analysis was the performance with different sampling techniques such as SMOTE, Random Undersampling – RU and CUT with the proposed SCUT. To evaluate the statistical significance of results, the study used the Friedman statistical test, obtaining a result of < 0.05 . The major finding of this paper is that SCUT worked best with Naïve Bayes and Support Vector Machine in correctly classifying minority classes with undersampling. With the Decision Tree and K-NN classifiers, SMOTE worked better indicating that undersampling did not have an impact on the performance of these algorithms. The SCUT sampling algorithm also performed best with a data set with the highest number of classes – ten, indicating SCUT's sensitivity to minority classes. The study does not conclusively demonstrate that SCUT is better than the other methods but is successful in establishing its suitability to handle an imbalance in multi-class datasets.

Jishan et al. 2015.

In their study, ((Jishan *et al.*, 2015)) use data from a university to predict student performance using various attributes such as GPA and performance in course assessments. Classification algorithms Naïve Bayes, Decision Tree, and Neural Network were used to predict the Student Grades based on grades in quizzes and the GPA. The predictand attribute is the course grade which can have one of 5 classes, i.e. grades A, B, C, D and F. With a total of 180 rows of continuous data the highest count is for grade 'C' at 48 and the lowest count for grade 'F' at 10 students, making the dataset imbalanced. To deal with this imbalance, the authors employ the equal width binning and the SMOTE oversampling technique and noted a significant improved

of the instances of the minority classes. In the implementation, the balanced data set was split into training and testing using the 80:20 ration. Data were modeled using Naïve Bayes, Decision Tree, and Neural Network classifier algorithms on four sets of data: Base data, discretized data, balanced data with oversampling and discretized with oversampled data. Models were evaluated using accuracy, confusion matrix, F-Measure, and AUC for each of the classes - one against all.

The analysis of measures included the comparison of results across models and the Pearson correlation coefficient evaluation of accuracy measures across models. Naive Bayes and Neural Network classifiers achieved an accuracy of 75% with discretized and oversampled data significantly higher as compared to results on the base data. The Pearson correlation coefficient used indicates a high correlation between accuracy and the F-measure and a moderate correlation between accuracy and AUC measures.

Fernández et al., 2013.

This paper by (Fernández et al., 2013) reviews methodologies available for managing imbalanced data, conducting experiments, and presenting results. The methodology includes a review of data preprocessing, sampling, cost-sensitive learning, and use of ensemble-based classifiers. Their study used sixty-six data sets with different attributes. While it is not clear how many of these were multi-class, it is indicated that some of the datasets are highly imbalanced. The algorithms used were Decision trees, Support Vector Machines, and KNN with Euclidean distance metric. The dataset was split into training and testing with five-fold cross-validation. The evaluation measure used the AUC to compare the performance of models. Sampling techniques used were SMOTE and variants, ADASYN, and SPIDER. Their study also applied cost-sensitive classifiers such as CostSensitive, MetaCost, and CS-Weighted Classifiers.

Algorithms used were boosting and bagging based ensemble classifiers such as AdaB-M1, AdaC2, RUSB, and SBAG and EasyEnsemble.

The research indicates that SMOTE+ENN and SMOTE emerged the best sampling techniques across all classifiers. In terms of cost-sensitive learning, the CS-Weighted approach was ranked higher than others, as demonstrated by the Shaffer Post hoc test for detecting differences. In terms of ensemble classifiers, the best performance was noted with SBAG or SMOTE-BAG while RUSB performed well on an average across all classifiers. This paper is an excellent reference since it incorporates an extensive survey of literature and robust comparative analysis of results.

Section 5 - Evaluation Measures for Multi-class Classification Algorithms

This section provides a study of measures used to evaluate the effectiveness of classification algorithms. There are three ways to measure the performance of a classification algorithm which is confusion matrix, receiver operating curves (ROC) and the Area under the Curve or (AUC), lift or gain charts (Kotu, Deshpande, 2015). In their study (Sokolova and Lapalme, 2009), discuss 24 measures used with binary, multi-class, multi-topic/multi-label, and hierarchical classification.

Confusion Matrix

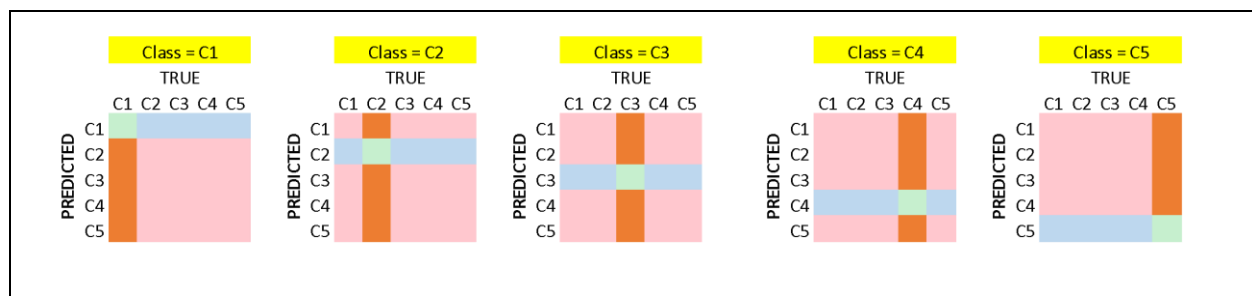
Confusion Matrix is a tool used to represent data classes, their actual values, and the predicted values of classifier models. The rows are representing the predicted class labels and columns indicating the true class labels. In binary classes there are only two classes positive or negative. In this case the confusion matrix is a 2 x 2 matrix with four combinations.

As seen in Table 6 Illustration of the Confusion Matrix the intersecting cell contains a count for predictions representing true positives- TP, false positives – FP, false negatives – FN or true negatives – FN.

| | | Confusion Matrix: Binary Class X and Y | |
|---------------------|---|--|-------------|
| | | ACTUAL CLASS | |
| | | X | Y |
| PREDICTED | X | TP Error | FP Error |
| | Y | FN Error | TN |
| TP - True positive | TP is a count of positives predicted as positive | | |
| FP - False positive | FP is a count of positives predicted as negative. This is an error. | | |
| FN - False negative | FN is the count of negative predicted as positive. This is also denoted as an Error | | |
| TN - True negative | TN is the count of negatives predicted as negative | | |

Table 6 Illustration of the Confusion Matrix for a binary label

However, in the case of multi-class labeled data, the matrix is determined by the number of items in the class. In the case of multi-class labels, the majority class is negative while the minority class is considered positive. In this study as seen in Table 7 Illustration of the Confusion Matrix for a multi-class label, it is shown as 5x5 matrix of 25 values.



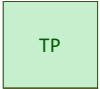

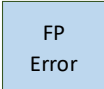
| | | |
|---------------------|---|---|
| TP – True Positive |  | Is class that is predicted correctly and is a diagonal on the matrix |
| FN- False Negative |  | False negative for each class column is calculated as total of errors in that column |
| FP – False positive |  | False positive for each class row is sum of errors in that row |
| TN – True Negative | | True negative for the matrix is the Total Population – (TP+FN+FP) |

Table 7 Illustration of the Confusion Matrix for a multi-class label

The confusion matrix helps to determine the accuracy, precision and recall values for each class.

The metrics for the model are determined by computing averages of each measure

Receiver Operating Curve/Area Under Curve

The ROC curve plots the true positive rate or sensitivity/recall against the false positive rate, which is computed as $1 - \text{specificity}$. A ROC curve is created by plotting the fraction of true positives (TP rate) versus the fraction of false positives (FP rate). The ROC chart shows connection or tradeoff between the TP Rate and FP Rate. Using this ROC chart, the AUC- the area under the ROC curve can be used to measure the classifier's ability to discriminate between true positive and true negatives. The AUC and ROC are used to compare the performance of different classification algorithms used on the same dataset (Kotu, Deshpande, 2015). As shown in Table 8, classification using Deep Learning is the best performer.

| Model | AUC |
|-------------|-----|
| Naive Bayes | 0.6 |

| | |
|------------------------|------------|
| Deep Learning | 0.7 |
| Decision Tree | 0.5 |
| Random Forest | 0.5 |
| Gradient Boosted Trees | 0.6 |
| Support Vector Machine | 0.5 |

Table 8 AUC Measures for multiple classification algorithms

A classifier can have very high accuracy on a data set, particularly when there are too many instances of one class and have very poor class recall and precision rates. Hence, in this case, accuracy may not be the best measure. Several authors (Kotu, Deshpande, 2015, Manning, 2009, Sokolova and Lapalme, 2009) recommend using the ROC and AUC to measure the performance of classifiers.

The best performance for a classifier is the AUC value of 1.0, and anything lower than 0.5 is considered a poor performance. (Kotu, Deshpande, 2015) suggest selecting those classifiers that not only have a ROC curve that is closest to the best and have an AUC value above 0.8. Figure 18 indicates that classifiers are performing poorly as they are all less than 70% and none of the curves reach the top left corner of the graph.

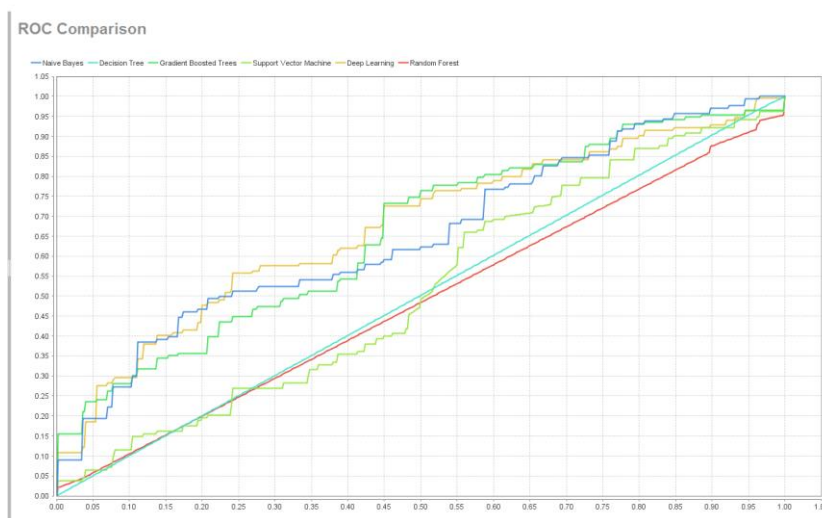


Figure 18 ROC chart indicating performance of algorithms for binary classification

Performance Measures and their formulas

Performance measures are obtained during the validation phase of the data mining task and helps to evaluate the classification algorithm's ability to evaluate the number of correctly recognized class samples or true positives and errors in a confusion matrix (Sokolova and Lapalme, 2009; Tharwat, 2018). Measures such as accuracy, recall, error rates, specificity and balanced accuracy are computed to understand the performance of a classifier. When more than one classifier is used, then these measures can be used to compare performances and identify the better algorithm.

In their study (Tharwat, 2018) provide a detailed description of various classification measures recommended for both balanced and imbalanced data sets. The important aspect of their study is the graphical representation of measures and addressing errors and not just accuracy.

A more exhaustive presentation of twenty-four measures of performance for all classification tasks such as binary, multi-class, multi-labelled, and hierarchical is presented by (Sokolova and Lapalme, 2009).

In their paper, (Maratea, Petrosino and Manzo, 2014) discuss the performance measures and suggest a new formula for adjusting the F-Measure to be used with imbalanced data as more effective measure.

A collection of classification performance measures, their description and formulas are listed in Table 9. These measures are applied to the performance results of experiments and are documented in Chapter 5 Evaluation Measures Used.

| Measure | Description | Formula |
|---|--|---|
| Accuracy | Accuracy is the total number of correct predictions divided by the total population | $\frac{TP + TN}{total\ population}$ |
| Error | The total number of incorrect predictions of class labels. High error rates represent a poor performance | 1-Accuracy |
| Precision | Precision is the percentage of results that are relevant and is obtained by dividing the correct positive prediction divided by the total number of positive predictions. Precision measures the accuracy of positive prediction | $\frac{TP}{(TP + FP)}$ |
| Recall or Sensitivity or True Positive Rate | Recall or sensitivity is the percentage of total correct positive results divided by the total number of positives. High recall values indicate a good performance | $\frac{TP}{(TP + FN)}$ |
| Specificity or the True Negative Rate | Specificity is the number of times correct negatives are correctly predicted. Correct negative results divided by the total number of negatives. High specificity values indicate good performance. | $\frac{TN}{(TN + FP)}$ |
| False Positive Rate | It is the number of incorrect positive predictions divided by the total number of negatives. Low FP rates indicate a good performance | 1-Specificity |
| F1-Score | It is the harmonic mean, which is an appropriate measure to evaluate imbalanced datasets given that it punishes extreme values more. | $2 \times \frac{precision \times recall}{precision + recall}$ |
| Balanced Accuracy | Balanced Accuracy sometimes referred to as the AUC helps to evaluate misclassification as it considers both specificity and sensitivity | $\frac{1}{2}(TPR + TNR)$ |

Table 9 Classification performance measures - Compiled from various sources

Summary

This chapter presented a detailed literature review relevant to the research goals and questions of this study.

A description of data mining activities provided the context for the research goals. A summary of applications and their related tasks in educational data mining was presented for various stakeholder groups. A review of relevant contemporary studies in EDM provided information on preprocessing data, classification algorithms used and the evaluation measures to be applied. This review also helped to identify a lack of literature in selecting program majors, providing an opportunity to be addressed by this study.

A review of literature about data mining methodologies helped to identify four methodologies and understand their common elements, strengths and weaknesses. This review also provided an insight into the missing elements of the methodologies and how they could be addressed. The most important part of this review was the study of classification for multi-class imbalanced data. This review helped to identify the data sampling methodology and the most appropriate algorithms that could be used. The insights from this review helped to formulate a strategy for the datamining experiments conducted in this study. The key insights from this chapter are:

- There is a shortage of studies conducted in EDM relating to the selection of student major, but it is an area to be considered
- Data mining projects need to be implemented using a methodology. However, the selected methodology needs to be supported by a project management framework
- Imbalanced and multi-class data sets need to be managed with proper data sampling techniques and or by tweaking parameters of the algorithm. Accuracy is not the best

measure in the case of imbalanced multi-class data and other measures need to be considered.

- Data exploration is critical in understanding data and visualization and correlation techniques can reveal significant relationships and patterns.

The next chapter details the research methodology used to conduct the literature review and the data mining tasks.

Chapter 3 Research Methodology

Methodology – Systematic Literature Review

A well-defined and organized methodology is necessary to review current literature while undertaking studies. My study applied a systematic literature review to ensure a successful outcome. A systematic literature review is considered to be a method used to identify, appraise, and construe current research that is relevant to the research topic (Rowley and Slack, 2004). A literature review is a “secondary study,” as it reviews studies conducted by other researchers (Garousi and Mäntylä, 2016). In their study (Okoli and Schabram, 2010) indicate that there are eight steps in a systematic literature review:

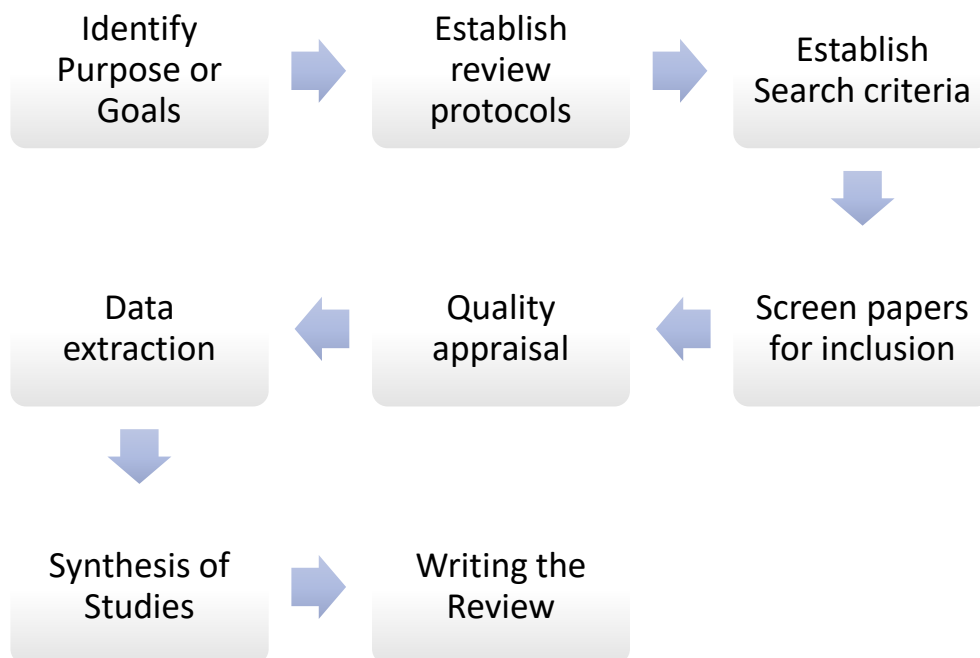


Figure 19 Steps for conducting a Systematic Literature Review based on (Okoli and Schabram, 2010)

The application of these steps recommended by (Okoli and Schabram, 2010) is described in the following section. The data mining project was developed using the CRISP-DM methodology described in detail in Chapter 2 Section 3.

Identify Research Goals

The main research goal of my study is the application of data mining in the education sectors.

The subsidiary goals are:

- Investigating data mining project methodologies
- Selection of program majors relating the undergraduate education in the UAE
- Classification algorithms and evaluation metrics used in the predictive task
- Management of multi-label and imbalanced data

Establish Sources

During the initial stages of this study, a list of current resources was established to review a variety of contemporary studies. These included:

- Online library available at the University with access to e-textbooks and journals
- Google Scholar
- Research Intelligence Portals such as Elsevier, IEEE, ProQuest
- Scopus for evaluating journals and
- Mendeley for managing references and bibliography
- Textbooks that were available from reputed publishers such as O'Reilly.

Conduct Keyword Search

Keyword search is the process of identifying search keywords to find papers based on the research goals. The preliminary search was narrowed by including keywords relating to the sub-concept such as Educational Data Mining in UAE or Multiclass Imbalanced Data Sampling.

Figure 20 shows a map of keywords based on the central concepts of data mining, literature review, and RapidMiner software used in my study.

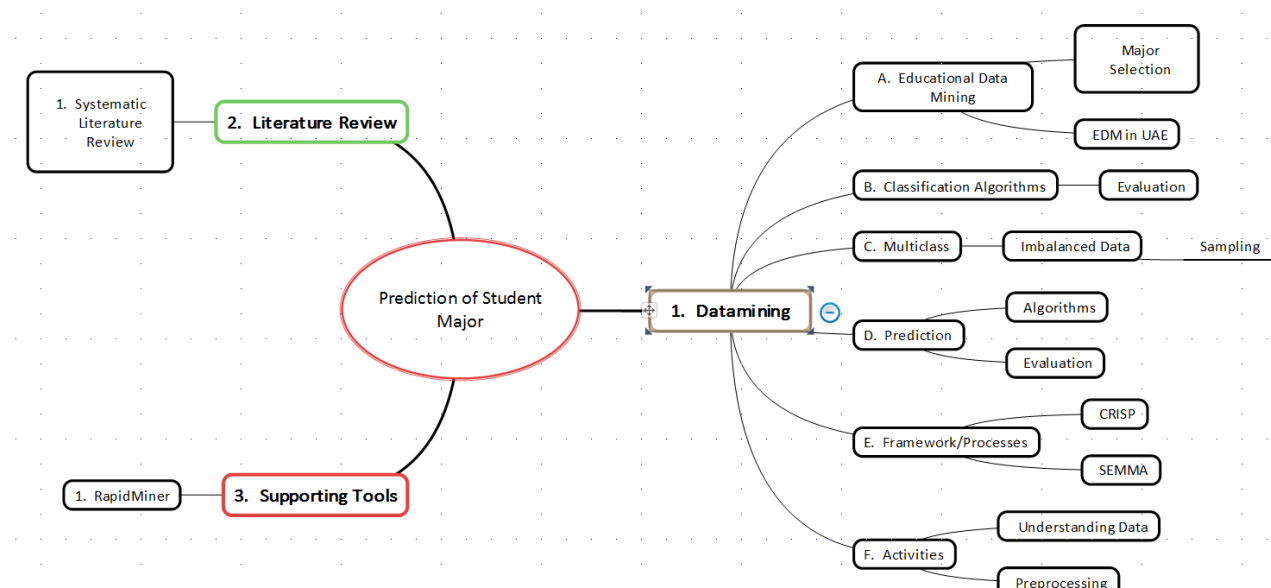


Figure 20 Keyword Search Map for Literature Review

There was a gradual growth in this list of keywords as papers were extracted, screened, and filtered for consideration or not. During this process, it appeared that including North American terminologies such as “Data Science” or “Data Analytics” or “Big Data Projects” provided better results.

Further reading of the extracted papers provided more keywords for search. The data mining experiments conducted in parallel also provided inputs to refine keywords for the search process.

Screen Papers

The year of publication and reputation of the journal was the first criteria for initial criteria for considering papers for review. Papers published in the last six years, i.e. published between 2012 and 2018 were considered. Where there was a shortage of current papers, older papers were reviewed based on the number of citations and reputation of journal.

Initial review included scanning the abstracts, keyword, and conclusions for fit with research goals.

Scanning bibliographies of papers for keywords provided links to useful resources as well.

Organization and Storage of Papers

A digital filing system on google drive ensured ease of access and retrieval of the selected papers. The naming convention used for each stored paper was the year of publication and name of the paper. The desktop version of Mendeley was used to store papers for review, annotation and referencing in Microsoft Word®

Evaluate Articles

During the screening process, these filters helped to sift papers:

- Relevance to research goals
- Relevance to the literature review and research methods

- Relevance to educational processes such as student enrollment, selection of majors and attributing factors, etc.
- Relevance to data mining from non-educational sectors to provide an insight into opportunities and challenges

Figure 21 Word cloud of paper titles used in this study

This study used a summarization technique to elicit the most pertinent information. Tabular summaries were used to distill information relating to:

- Algorithms used
- Results achieved

A tally sheet was also constructed to keep track of the algorithms used in each paper and the evaluation measures used. During the writing process, referring to these notes and summaries helped to focus on the main points easily.

Research Questions

The process of conducting a systematic review help to formulate these research questions as introduced in Chapter 1.

[R1] Can data mining be applied to the predict student's selection of program majors in higher education?

[R2] What is an appropriate framework to be used for data mining projects?

[R3] What are the most suitable algorithms that can be used to classify and predict multi-class labels with imbalanced data?

[R4] What is the most appropriate evaluation measure to be used with multiclass datasets?

Datamining Processes – CRISP-DM Methodology

This study conducted an academic research project in data mining of educational data. Applying the CRISP-DM methodology provided a structure to the data mining tasks of this project and complimented the systematic literature review. From the literature review on data mining methodologies, it appeared that CRISP-DM is most suitable for a project of this size and scope and hence was considered for use. Chapter 1 of this study presents the business understanding

phase. Chapter 4 presents the data understanding and preparation phase. Chapter 5 presents the modeling, evaluation, and deployment of experiments.

Resources Used

Mendeley Desktop®

Reference Management for this paper was done using Mendeley Desktop®. This software allows users to organize papers from journals, books, and URLs of webpages into a personal library. Mendeley provides users the ability to annotate, write notes, generate citations, manage reference documents, and provides a plugin that works with Microsoft Word® to insert citations and bibliographies. (Mendeley, 2018)

Microsoft Office®

In producing this paper, Microsoft Excel® was used to extract and analyse data and Microsoft Visio® was used to produce illustrations. Microsoft Word® was used to create and format the dissertation document.

RapidMiner®

All experiments in this study were deployed using RapidMiner Studio V9.2®, an open source software for data analytics. This software combines data mining functions such as data preparation, machine learning, and modeling for classification and clustering. RapidMiner also allows incorporating additional functionality by providing community-developed extensions. The interface is easy to use and provides drag and drop features to develop process models which can be validated using several operators. The help features and tutorials provide documentation to build comprehensive models. The free educational license used in this study enables the processing of 10,000 rows of data and offers tools such as Turbo Prep that helps with data

cleansing and visualization. The Auto Model functionality helps to identify the best models for the supplied data (RapidMiner, 2018)

(Adhatrao et al., 2013; Jishan et al., 2015; Kotu and Deshpande, 2014; Miguéis et al., 2018; Saa, 2016; Sgouropoulou et al., 2014;) have used RapidMiner® in their studies.

Summary

Adopting a research methodology is critical to the successful planning and implantation of a research paper that includes a literature review and its application. Using only a data mining methodology would limit a proper study of literature. Hence two methodologies i.e systematic literature review and CRISP- DM were incorporated into this study. This has resulted in providing a strong basis for conducting research and applying knowledge to a real life situation. This chapter also introduced the use of essential resources necessary for a researcher to manage the various activities.

The next chapter discusses the implementation of this study using the CRISP data mining methodology.

Chapter 4 Research Implementation

This chapter presents the processes and results of the data mining task of this research. The scope of this task is to use historical educational data to predict a student's choice of major using several input attributes. This is considered a predictive task using classification algorithms applied to a multi-class target variable with an imbalanced data set.

CRISP-DM Methodology

This academic study follows the CRISP-DM methodology, described in detail in Chapter 2 The Cross-Industry Standard Process for Data Mining or CRISP-DM.

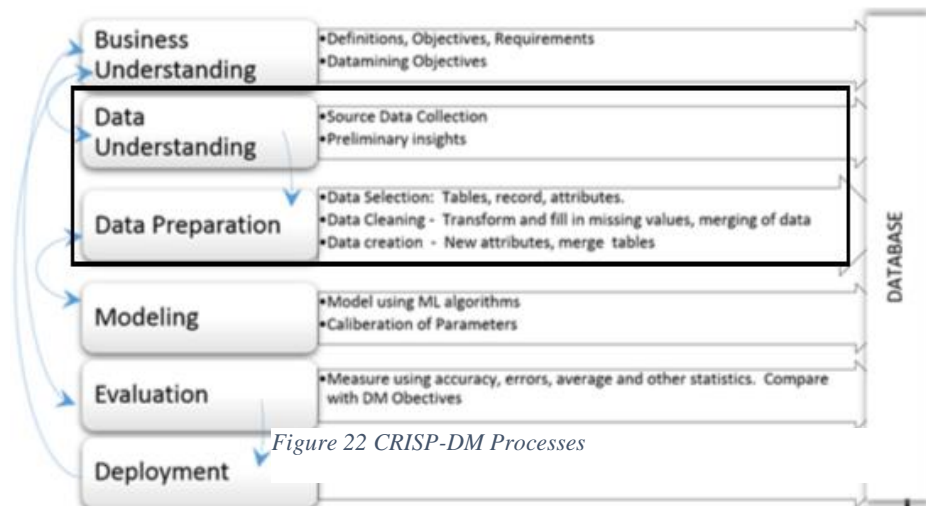


Figure 22 CRISP-DM Processes

The processes undertaken in this study is described in the following sections of this chapter

Business Understanding

To understand the business value of this data mining task, discussions were conducted with colleagues and students. Internal and external strategy documents were reviewed and analyzed using the author's expertise in this domain. Business understanding is presented in Chapter 1 - Context and Problem Definition. The stakeholders who would benefit from this task are the students who can make informed choices, program managers and teachers who would benefit

from knowing a profile of students that chose a major and obtain an insight into low enrollments which are the minority class. The outcome of this study can also help a new generation of students determine the predictors that are important to consider when enrolling for a major enabling them to be better prepared.

This study considers the prospect of using historical data and data mining tools to explore data and build a predictive model to address these issues.

Data Understanding – Extraction of Data

The first step in this process was to extract data from the main institutional database for two academic years for all campuses that offered Computer Studies program. Each academic year has three semesters: Fall, Spring, and Summer. However new students are not enrolled in the summer term; hence data downloaded was limited to Fall and Spring terms of 3 academic years 2015, 2016 and 2017. The institutional database has separate repositories for student master data and transactional data relating to course registration and grades. This required data to be downloaded and merged from separate sources. Hence, student master data and course grades for specific courses for three academic years were downloaded as excel sheets and merged to get 12,000 records. From this merged data the following records were deleted:

- Records of students who were either withdrawn or dismissed for academic reasons
- Records of students who were not from the catalog terms 201510, 201520, 201610 and 201620

This resulted in approximately 3000 rows of data

Data Preparation

Joining and Cross-Tabulating Data

This study also required to evaluate the impact of course grades on the selection of major.

Students take five core courses in their first three semesters of the program: Information systems, Programming, Hardware and networking, Database design and Web technologies. Teachers of these courses discuss that poor grades in these courses often impact the choice of a major. For example, students who get low grades in the course hardware and networking difficult choosing the Networking major and those who find courses in programming difficult refrain from selecting the Applications Development major. Hence it was considered necessary to include course grades with student's master data.

Grades for each course were merged into one master sheet in Excel. Rows for students with a fail grade was removed since they would either repeat the course or be asked to withdraw based on the GPA. These course grades were then joined with the master list of student's demographic information. This data was cross-tabulated to obtain course grades with student demographic data in one row. This resulted in about 1,200 records

Errant records with missing data in the major column were eliminated giving a total of 1,141 examples of usable rows.

Feature Reduction

The merged dataset included 82 features or attributes that described a student. For this study only those attributes relevant to personal demographics, performance data in university placement exams and average grades in high school were retained. The id column was retained

for further processing and all other private information was deleted. To complete this task, the author's judgment and domain expertise were applied.

The retained attributes were organized and prefixed with a number indicating their grouping.

Table 10 shows the organization, grouping and naming of data that made it easier to read results.

| Attribute | Description | Grouping | Prefix |
|---------------------|--|-----------------|---------------|
| Gender | Nominal with two classes | Demographic | 1_ |
| Age | Computed using Date of Birth | Demographic | 1_ |
| High School Average | Numerical Value | Demographic | 1_ |
| CEPA | Math and English Placement Test Grade | Demographic | 1_ |
| IELTS Band | English placement test for entry | Demographic | 1_ |
| Campus | Nominal with several classes indicates the student's campus | Academic | 2_ |
| Major | Multi class label indicates the chosen major in the computer studies program | Academic | 2_ |
| Cgpa | Numerical real value indicates the academic performance at the end of the first academic year | Academic | 2_ |
| Academic Standing | Multi class with 3 values | Academic | 2_ |
| INFO SYS | Multi class with final Letter grade for a core course on the impact of information systems on businesses | Course Grade | 3_ |
| HW and NW | Multi class with final Letter grade for a core course introducing concepts of hardware and networking infrastructure | Course Grade | 3_ |
| WEBTECH | Multi class with final Letter grade for a core course in developing web applications | Course Grade | 3_ |
| DATABASE DES | Multi class with final Letter grade for a core course introducing concepts of database design and implementation | Course Grade | 3_ |
| PROGRAMMING | Multi class with final Letter grade for a core course introducing programming structures | Course Grade | 3_ |
| ID | Retained to be able to join data from grades table. Excluded in the final data set used for experiments | None | None |

Table 10 List of attributes in source data

Imputing of Missing Values

The cleansed dataset was almost complete with very few missing values and these were imputed using K-NN algorithm in rapid miner.

Data Understanding – Exploration of Data

Exploring or understanding of data is achieved by using descriptive statistics or visualization and helps to identify the basic structure of data, relationships between attributes or features of the data. Descriptive statistics is analyzing data using simple mean, deviations and correlation while visualization is an analysis of the graphical representation of data (Kotu, Deshpande, 2015).

In this study three techniques were used in this study which are basic statistical analysis, visualization and correlation. The automodel feature was used to develop a basic understanding of classification algorithms on this data.

Data exploration

This section presents exploration of data using and data visualization and statistical analysis conducted on the three groupings of the data set. Exploring data provided a great understanding of patterns in data, impact of attributes on the target variable and indicators of how students are performing in key courses. The initial statistical analysis of each group of attributes is shown in

Figure 23 Statistics demographic attributes in data set

Figure 30 Statistics of academic attributes

Figure 32 Statistics of course grades of successful students

Analysis of Demographic Attributes

The demographic profile indicates that there are more female students than male students and the average age is 22. The high school average of these students is 85%. The average IELTS band is 5.2 and CEPA score is 174.

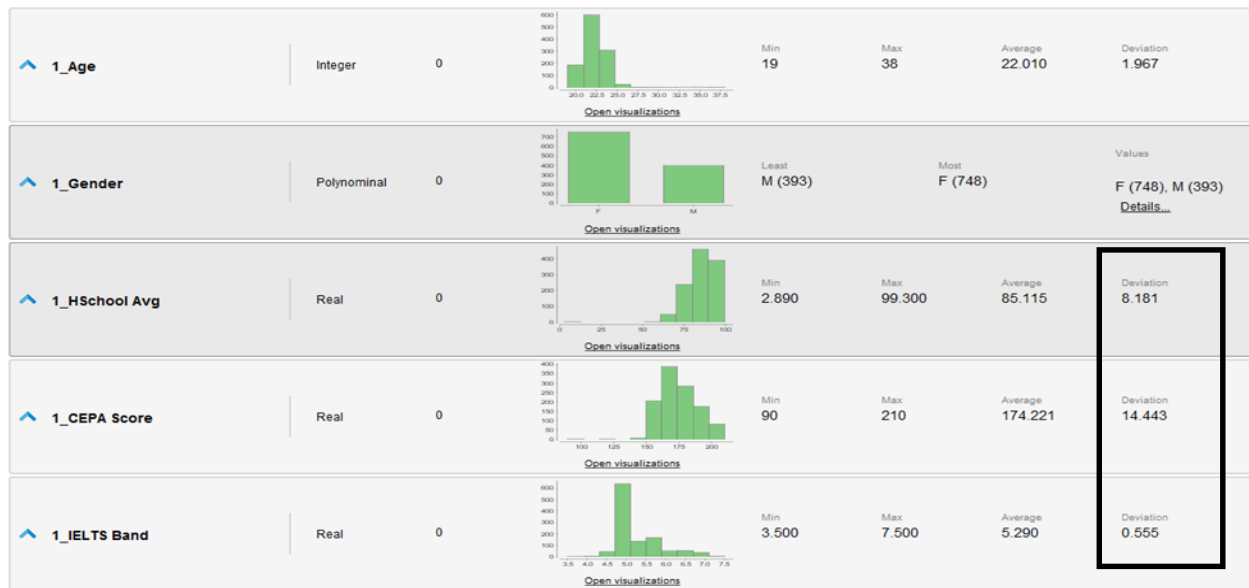


Figure 23 Statistics demographic attributes in data set

- 1_Age: The average age of students is 22 across all campuses indicating that the average student is a young adult and probably not a working student. The spread of age is not very large as indicated by a low standard

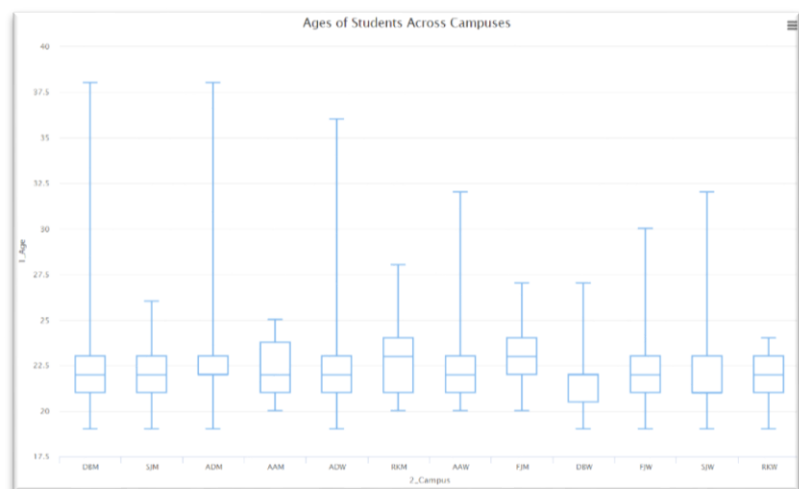


Figure 24 Students age across campuses

deviation. The boxplot chart shown Figure 24 indicates that mature students who are above 30 years are mostly male as indicated by the campus code.

- 1_Gender: The gender is a polynomial or categorical value. It is interesting to note that number of female students are twice as many as male students and could have some relevance to the choice of program major. More female students also take Multimedia, Networking and Business Solutions while there is an equal mix of students taking the other majors



Figure 25 Enrollment into major across gender

- 1_HSchool Avg: The high school average has a high standard deviation indicating outliers.

The box plot shown below further clarifies this.

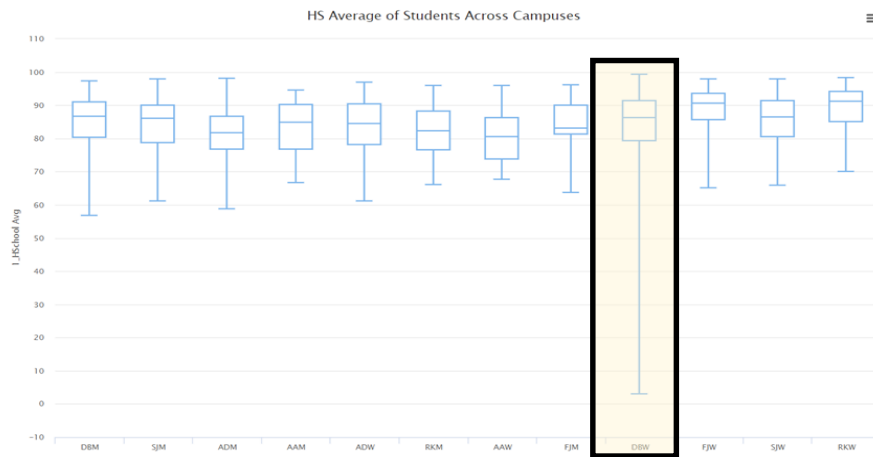


Figure 26 High School average across campus

The scatter plot shows that the high school average for students in all majors are between 75 and 100. Those below 75 are mostly from the Security Major and Business Solutions Major

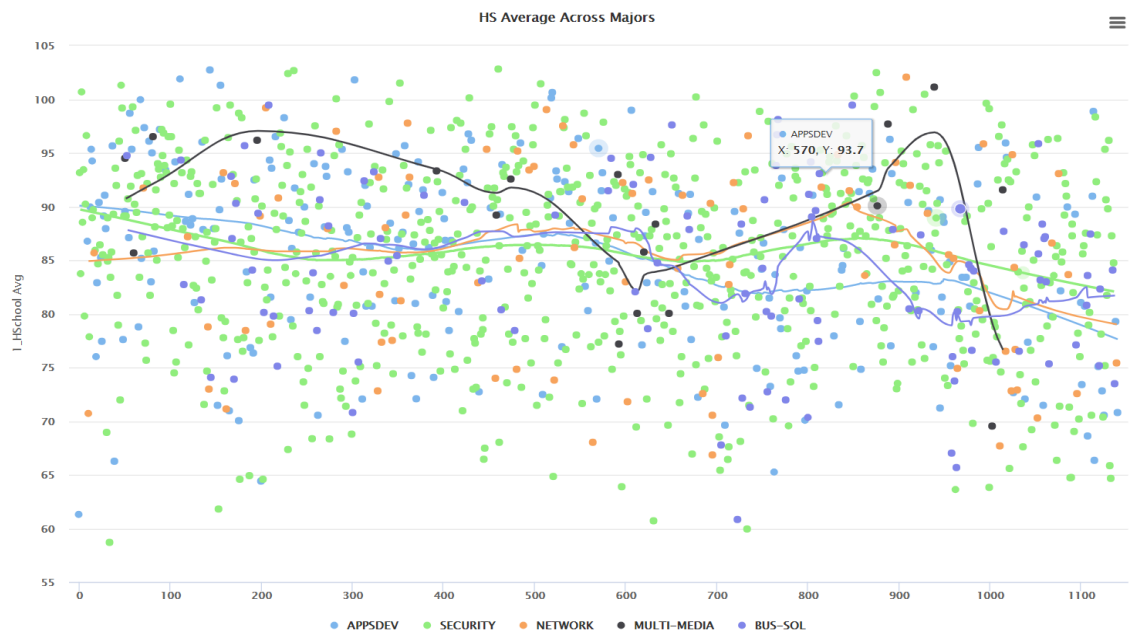


Figure 27 High School Average across student major

1_CEPAScore: The CEPA is a score received by students for a placement exam in English and Math. This attribute has a high standard deviation indicating outliers.

1_IELTS Band: The average of the IELTS band is 5 which is the minimum requirement for acceptance into the program. Students enrolled in the SECURITY and APPSDEV majors have a higher score than those in the other programs.

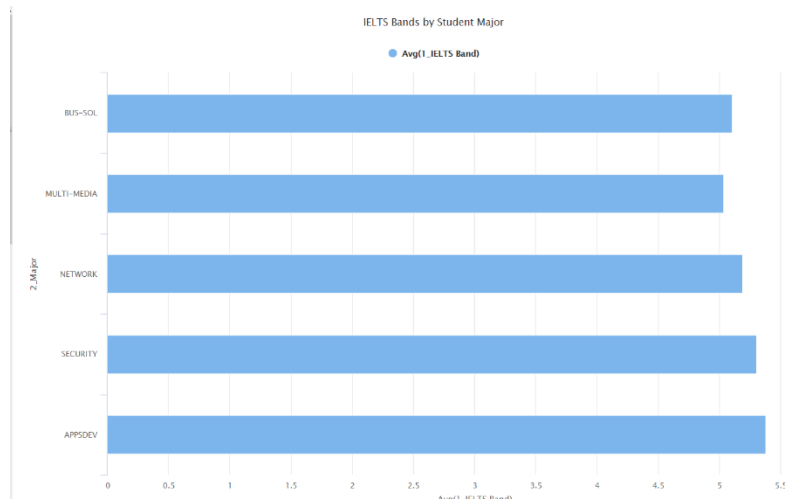


Figure 28 IELTS band by student major

The histogram shown below indicates that more male students have a higher IELTS band than female students.

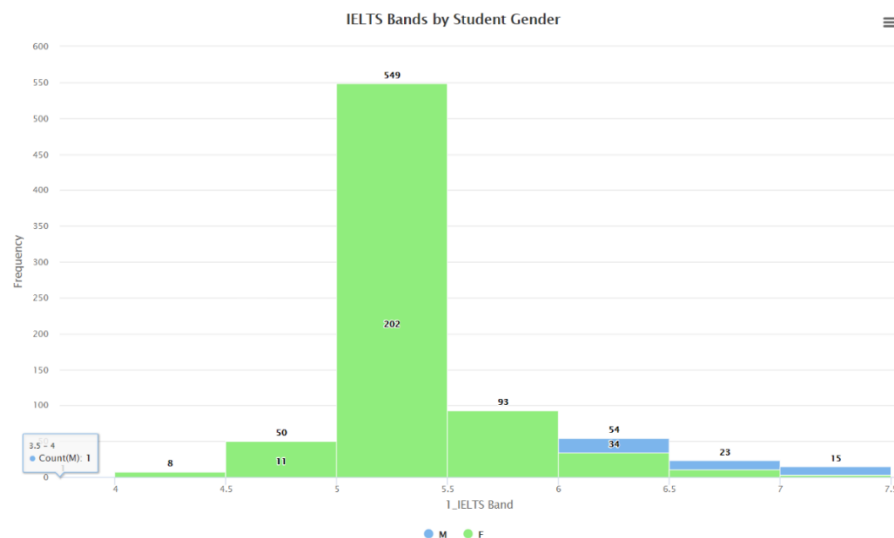


Figure 29 IELTS bands across gender

Analysis of Academic Attributes

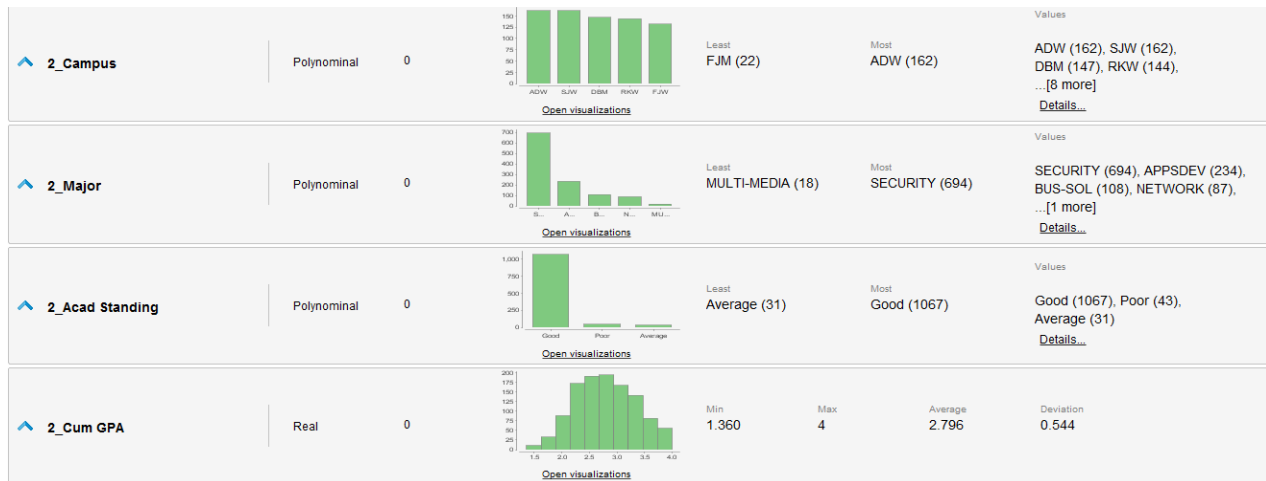


Figure 30 Statistics of academic attributes

The academic attributes are those relating to general academic performance of the student and identifies the campus that they are registered in as well as the program major that they are enrolled in. There is a causal relationship between 2_CUM_GPA which is the cumulative GPA of a student of all courses taken and 2_ACAD Standing. The academic standing good/poor/average is based on the cumulative GPA. A low GPA results in a poor academic standing. The average GPA is 2.796 while two campuses marked A and B in the chart below have a maximum of 3.5 while students in three campuses have achieved the highest possible GPA of 4 by the third semester.

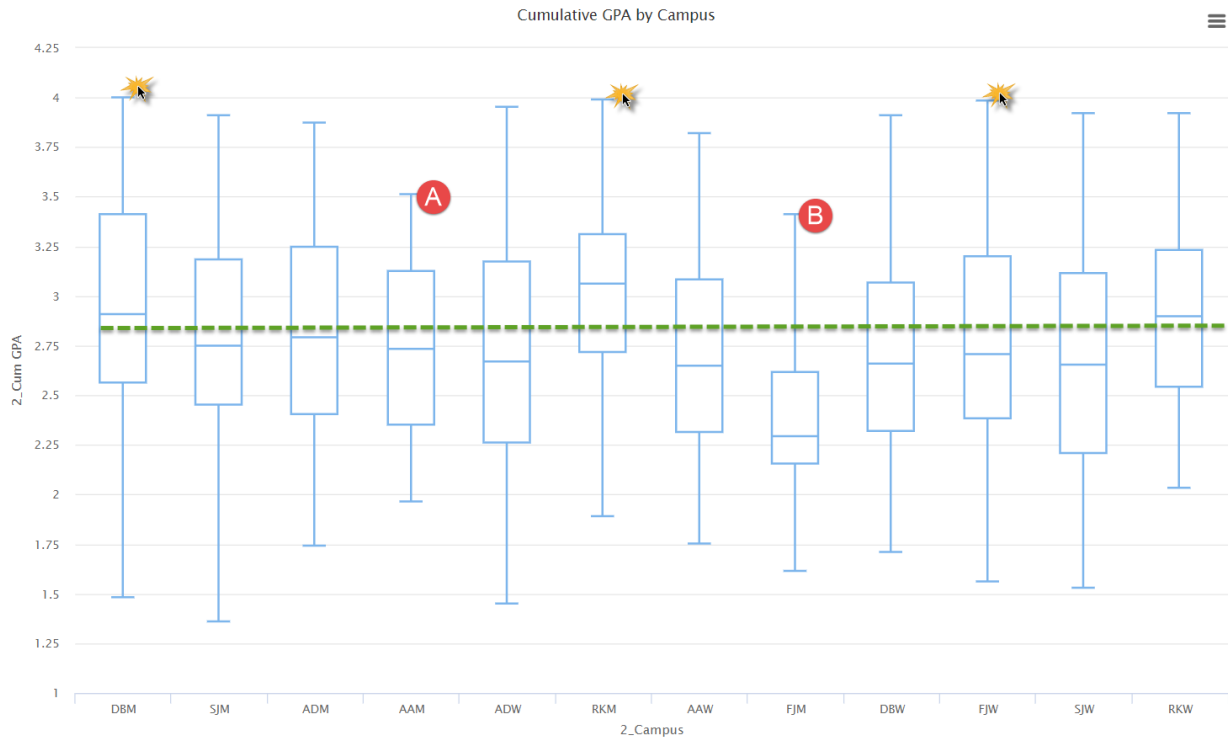


Figure 31 Academic performance across campuses

Analysis of Course Grade Attributes

These are the final grades of IT courses that students take in the first 3 semesters of bachelor's degree program. These courses introduce students to the core concepts relating to information technology theory and skills such as programming, configuring hardware and networks, building websites, designing and building databases and building information systems. The learning outcomes for these courses are a combination of theory and applications. Learning is evaluated in a variety of assessments and students receive a weighted letter grade. For this study, these grades have been converted to equivalent real numeric values based on letter grades



Figure 32 Statistics of course grades of successful students

- It is observed that the average grade for male students is better than their female counterparts in courses taken in the first year.
- Figure 33 shows that female students had higher average grades than the male students only in the database course.

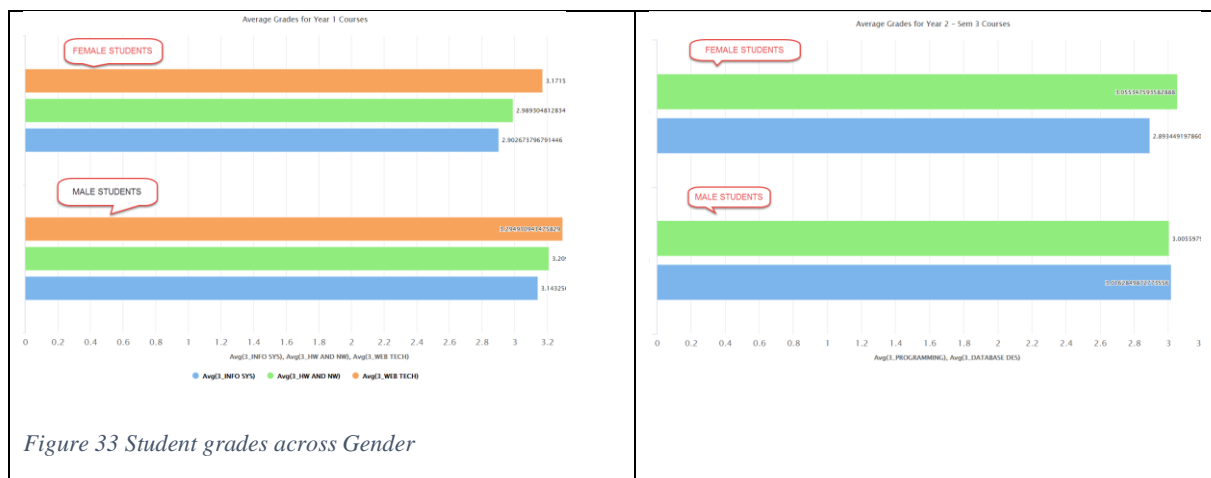


Figure 33 Student grades across Gender

Analysis of Label Attribute and Class balance

The special attribute used in this study is “Major” which is multi-class label with 5 values. It is the label or the independent variable that needs to be predicted. Students in the Computer Studies program can opt to select one of the five majors and hence the data in this column was carefully examined.

The distribution of classes in the attribute “Major” is shown below. The predictand attribute Major is polynomial with one class worth 61% and the other four totaling 39%. This imbalance impacts the type of measures that can be used to evaluate the performance of predicting algorithms. A baseline expected recall was computed based on the percentage of occurrence to offer a point of comparison for evaluating performance of models.

| Multi-class Data Distribution | | | | |
|-------------------------------|-------------|--------------------------|-----------------|-----------------------------|
| Index | Major | Description | Number Enrolled | Percentage/ Expected Recall |
| 1 | SECURITY | Security and Forensics | 694 | 61% |
| 2 | APPSDEV | Applications Development | 234 | 21% |
| 3 | BUS-SOL | Business Solution | 108 | 9% |
| 4 | NETWORK | Network Engineering | 87 | 8% |
| 5 | MULTI-MEDIA | Multi Media Technologies | 18 | 2% |
| | | | 1,141 | |

Table 11 Multi-class data distribution and expected recall values

Correlation Analysis

Correlation analysis is used to measure the statistical relationship between data variables and the correlation matrix shown in Figure 34 is a visual representation highlighting the degree of relationship. Negative correlation exists between age and most academic indicators. High correlation exists between course grades and GPA which can be explored further.

| Attributes | 1_Age | 1_CEPA... | 1_Gend... | 1_HSch... | 1_IELTS... | 2_Cum ... | 2_Major... | 2_Major... | 2_Major... | 2_Major... | 3_DATA... | 3_HWA... | 3_INFO ... | 3_PRO... | 3_WEB ... |
|--------------------|--------|-----------|-----------|-----------|------------|-----------|------------|------------|------------|------------|-----------|----------|------------|----------|-----------|
| 1_Age | 1 | -0.247 | -0.117 | -0.290 | -0.073 | -0.114 | -0.072 | 0.113 | 0.047 | -0.029 | -0.079 | 0.059 | 0.013 | -0.056 | -0.065 |
| 1_CEPA Score | -0.247 | 1 | -0.302 | 0.021 | 0.798 | 0.378 | 0.106 | -0.109 | -0.117 | 0.053 | 0.163 | 0.262 | 0.329 | 0.232 | 0.261 |
| 1_Gender = F | -0.117 | -0.302 | 1 | 0.133 | -0.315 | -0.071 | -0.066 | 0.127 | 0.035 | -0.064 | 0.033 | -0.149 | -0.162 | -0.079 | -0.083 |
| 1_HSchool Avg | -0.290 | 0.021 | 0.133 | 1 | 0.015 | 0.349 | -0.008 | -0.057 | 0.009 | 0.022 | 0.256 | 0.153 | 0.193 | 0.176 | 0.234 |
| 1_IELTS Band | -0.073 | 0.798 | -0.315 | 0.015 | 1 | 0.341 | 0.082 | -0.107 | -0.052 | 0.039 | 0.172 | 0.280 | 0.335 | 0.221 | 0.255 |
| 2_Cum GPA | -0.114 | 0.378 | -0.071 | 0.349 | 0.341 | 1 | 0.096 | -0.094 | -0.062 | 0.015 | 0.652 | 0.651 | 0.672 | 0.593 | 0.634 |
| 2_Major = APPSDEV | -0.072 | 0.106 | -0.066 | -0.008 | 0.082 | 0.096 | 1 | -0.164 | -0.146 | -0.633 | 0.084 | 0.014 | 0.041 | 0.161 | 0.084 |
| 2_Major = BUS-SOL | 0.113 | -0.109 | 0.127 | -0.057 | -0.107 | -0.094 | -0.164 | 1 | -0.093 | -0.403 | -0.112 | -0.068 | -0.101 | -0.094 | -0.123 |
| 2_Major = NETWORK | 0.047 | -0.117 | 0.035 | 0.009 | -0.052 | -0.062 | -0.146 | -0.093 | 1 | -0.358 | -0.049 | 0.019 | -0.032 | -0.089 | 0.008 |
| 2_Major = SECURITY | -0.029 | 0.053 | -0.064 | 0.022 | 0.039 | 0.015 | -0.633 | -0.403 | -0.358 | 1 | 0.042 | 0.023 | 0.054 | -0.006 | -0.005 |
| 3_DATABASE DES | -0.079 | 0.163 | 0.033 | 0.256 | 0.172 | 0.652 | 0.084 | -0.112 | -0.049 | 0.042 | 1 | 0.440 | 0.494 | 0.490 | 0.487 |
| 3_HW AND NW | 0.059 | 0.262 | -0.149 | 0.153 | 0.280 | 0.651 | 0.014 | -0.068 | 0.019 | 0.023 | 0.440 | 1 | 0.590 | 0.351 | 0.552 |
| 3_INFO SYS | 0.013 | 0.329 | -0.162 | 0.193 | 0.335 | 0.672 | 0.041 | -0.101 | -0.032 | 0.054 | 0.494 | 0.590 | 1 | 0.399 | 0.507 |
| 3_PROGRAMMING | -0.056 | 0.232 | -0.079 | 0.176 | 0.221 | 0.593 | 0.161 | -0.094 | -0.089 | -0.006 | 0.490 | 0.351 | 0.399 | 1 | 0.444 |
| 3_WEB TECH | -0.065 | 0.261 | -0.083 | 0.234 | 0.255 | 0.634 | 0.084 | -0.123 | 0.008 | -0.005 | 0.487 | 0.552 | 0.507 | 0.444 | 1 |

Figure 34 Correlation analysis of all variables

| Attributes | 1_Age | 1_HSchool A... | 1_CEPA Score | 1_IELTS Band | 2_Cum GPA |
|---------------|--------|----------------|--------------|--------------|-----------|
| 1_Age | 1 | -0.290 | -0.247 | -0.073 | -0.114 |
| 1_HSchool Avg | -0.290 | 1 | 0.021 | 0.015 | 0.349 |
| 1_CEPA Score | -0.247 | 0.021 | 1 | 0.798 | 0.378 |
| 1_IELTS Band | -0.073 | 0.015 | 0.798 | 1 | 0.341 |
| 2_Cum GPA | -0.114 | 0.349 | 0.378 | 0.341 | 1 |

Figure 35 Correlation analysis of demographic attributes

Figure 35 indicates that there is a strong relationship noticed between the CEPA Score which is a test of Math and English and IELTS which is a test of English. Interestingly a weak correlation between performance in university measured by Cum GPA in the second year of college as compared with the average performance in high school. This information can be used in while selecting attributes for prediction tasks.

Correlation between course grades

As seen in Figure 36 Pairwise correlation of course grades, the strongest correlation is between courses taken in the first semester when students are new and may not have all the study skills necessary to navigate the demands of coursework assessments.

| First Attribute | Second Attribute | Correlation ↓ |
|-----------------|------------------|---------------|
| 3_INFO SYS | 3_HW AND NW | 0.590 |
| 3_HW AND NW | 3_WEB TECH | 0.552 |
| 3_INFO SYS | 3_WEB TECH | 0.507 |
| 3_INFO SYS | 3_DATABASE DES | 0.494 |
| 3_DATABASE DES | 3_PROGRAMMING | 0.490 |
| 3_WEB TECH | 3_DATABASE DES | 0.487 |
| 3_WEB TECH | 3_PROGRAMMING | 0.444 |
| 3_HW AND NW | 3_DATABASE DES | 0.440 |
| 3_INFO SYS | 3_PROGRAMMING | 0.399 |
| 3_HW AND NW | 3_PROGRAMMING | 0.351 |

Figure 36 Pairwise correlation of course grades

Summary

The chapter presents the first phase of the data mining implementation. A business understanding of this project was summarized in continuation of details presented in Chapter 1. Data understanding described the nature of data and steps followed in extracting and consolidation for preparation. Subsequently data was explored in detail using statistical and visualization techniques and interesting results were noted. Using correlation, it was possible to review the relationship between course grades. The information noted here would be useful during the modelling phase where features need to be selected. The key findings of data exploration are

The demographic and academic profile

- ✓ More female students than male students and the average age is 22.
- ✓ More female students also take Multimedia, Networking and Business Solutions
- ✓ The high school average students is 85%. The average IELTS band is 5.2 and CEPA score is 174.
- ✓ A low GPA results in a poor academic standing. – causal relationship.

Course Grades Profile

- ✓ Average grade for male students is better than their female counterparts in courses taken in the first year.
- ✓ Female students had higher average grades than the male students only in the database course.
- ✓ Strong correlation is between courses taken in the first semester

Data set

- ✓ Major is multi class polynomial with 5 values and is imbalanced
- ✓ Data is quite clean with very few missing values

The upcoming chapter 5 will present the results of the experiments implemented in RapidMiner.

Chapter 5 Deployment of Experiments and Analysis of Results

Experiments

This objective of this study was to apply the most suitable algorithms to classify and predict multi-class labels with imbalanced data and evaluate their performance using appropriate performance measures. The literature review helped to formulate a strategy for deployment which included sampling, application of cross-validation, classification algorithms, and the performance evaluation.

Auto Modelling with Rapid Miner

Initial experiments with the final dataset were done using the auto-modeling function available with RapidMiner. Figure 37 Auto model overview of classifier performance provided an indication of the classifiers to use for the systematic experimentations. These results showed the best performing classifiers in terms of accuracy and speed of process execution




| Model | | Accuracy | Standard Deviation | Total Time  |
|--------------------------|---|----------|--------------------|--|
| Naive Bayes |  | 59.6% | $\pm 2.2\%$ | 12 s |
| Decision Tree | | 59.2% | $\pm 1.2\%$ | 30 s |
| Support Vector Machine | | 59.6% | $\pm 2.2\%$ | 3 min 53 s |
| Gradient Boosted Trees | | 59.6% | $\pm 2.2\%$ | 5 min 45 s |
| Generalized Linear Model |  | 59.3% | $\pm 1.7\%$ | 6 min 43 s |
| Logistic Regression | | 59.6% | $\pm 2.2\%$ | 6 min 52 s |
| Fast Large Margin | | 59.6% | $\pm 2.2\%$ | 6 min 56 s |
| Deep Learning | | 59.6% | $\pm 2.2\%$ | 8 min 16 s |
| Random Forest | | 59.6% | $\pm 2.2\%$ | 9 min 20 s |

Figure 37 Auto model overview of classifier performance

As per these results and information gathered from the reading literature relating to the handling of imbalanced data, systematic experiments were conducted. The performance of the fastest classifier, Naïve Bayes seen in Figure 38 Auto model - Naive Bayes Confusion Matrix, indicated an accuracy of 59.64%. However, the recall for the minority classes were 0 showing the algorithm was learning only from the majority class. Hence, it was decided to use sampling techniques to get better performance.

Naive Bayes - Performance

Criterion
accuracy
classification error

Table View Plot View

accuracy: 59.64% +/- 2.22% (micro average: 59.63%)

| | true APPSDEV | true SECURITY | true NETWORK | true MULTI-MEDIA | true BUS-SOL | class precision |
|-------------------|--------------|---------------|--------------|------------------|--------------|-----------------|
| pred. APPSDEV | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. SECURITY | 70 | 195 | 27 | 5 | 30 | 59.63% |
| pred. NETWORK | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. MULTI-MEDIA | 0 | 0 | 0 | 0 | 0 | 0.00% |
| pred. BUS-SOL | 0 | 0 | 0 | 0 | 0 | 0.00% |
| class recall | 0.00% | 100.00% | 0.00% | 0.00% | 0.00% | |

Figure 38 Auto model - Naive Bayes Confusion Matrix

Attribute weighting by correlation helped to identify attributes in the data set that influenced the target or predictand variable “Major”. Course grades in programming, CEPA score, gender are the most significant factors and can be further investigated.

Weights by Correlation

| Attribute | Weight ↓ |
|----------------|----------|
| 3_PROGRAMMING | 0.087 |
| 1_CEP A Score | 0.087 |
| 1_Gender = F | 0.077 |
| 3_DATABASE DES | 0.071 |
| 1_IELTS Band | 0.068 |
| 2_Cum GPA | 0.057 |
| 1_Age | 0.056 |
| 3_INFO SYS | 0.053 |
| 3_WEB TECH | 0.048 |
| 1_HSchool Avg | 0.031 |
| 3_HW AND NW | 0.028 |

Figure 39 Auto model - Feature weights by correlation

Systematic Experimentation

In this study, the results of the three sets of experiments were presented. The same data set was used for each set and processed using six classifiers from different families. Each set of experiments differed in the types of sampling used as seen in Figure 40.

- the use of classifying operators for training
- sampling techniques used and the number of instances or examples in the data set



Figure 40 Comparison of class instances with sampling applied

Classifiers Used

Based on the literature reviewed for the most appropriate classifiers for multi class labels, ensemble, nested and plain supervised classification operators were selected. All experiments applied Cross-Validation with leave one out parameter. The following table lists other parameters used:

| Type | Code | Classifiers | Parameter |
|--------------------------------------|----------|--|--|
| Supervised, classification | DT1 | Decision Tree | Gain ratio With pruning applied |
| | NB1 | Naïve Bayes | With Laplace correction enabled |
| | NN1 | Neural Net – Deep Learning | - |
| Supervised, Ensemble, classification | RF1 | Random Forest | 20 Trees Gain Ratio With/without pruning |
| | GBT1 | Gradient Boosted Tree H20. | 20 Trees Maximal depth of 10 |
| Nested Classification Operator | PbyB-SVM | Polynomial by Binomial Classification with SVM. Used for classification of polynomial labels | - |

Table 12 Parameters used with classification models

Evaluation Measures Used

The classification performance operator was used to obtain a confusion matrix with accuracy, weighted recall and weighted precision measures. The confusion matrix provided the per class accuracy and prediction percentages as shown in the table below. Using these details, the Error Rate, Specificity and False Positive Rates, F1-Score and Balanced Accuracy were computed.

Statistical Analysis

Comparative analysis of Sampling

The following section enumerates results of these experiments in tabular and graphical formats. Analysis and discussions are presented at the end of the section.

Experiment Set 1 – Experiments using the original data set without sampling

In this set, six experiments were conducted, and each model was evaluated using the Performance Classifier. Confusion matrix was recorded for each model to arrive at other performance measures.

Confusion Matrix for Decision Tree Classifier

Table 13 shows the confusion matrix for the data modeled by the Decision Tree Classifier in RapidMiner. The columns are true labels, and rows are the predicted labels. This matrix includes the model accuracy which is an average of accuracy computed for each class which in this case is 57.76%. The recall of the majority class label security is 92%, and the precision is 61.72.

Values from each of the experiments were used to arrive at the model wise performance.

| Confusion Matrix or Performance Vector | | | | | | | |
|---|---|---------------|--------------|------------------|--------------|-------|-----------------|
| Classifier: | Decision Tree | | | | | | |
| Sampling | None applied | | | | | | |
| Model Accuracy | Accuracy: 57.76% +/- 49.42% (micro average: 57.76%) | | | | | | |
| N=1141 | true APPSDEV | true SECURITY | true NETWORK | true MULTI-MEDIA | true BUS-SOL | TOTAL | class precision |
| pred. APPSDEV | 15 | 20 | 1 | 0 | 3 | 39 | 38.46% |
| pred. SECURITY | 206 | 640 | 82 | 17 | 92 | 1037 | 61.72% |
| pred. NETWORK | 10 | 32 | 3 | 1 | 12 | 58 | 5.17% |
| pred. MULTI-MEDIA | 0 | 0 | 1 | 0 | 0 | 1 | 0.00% |
| pred. BUS-SOL | 3 | 2 | 0 | 0 | 1 | 6 | 16.67% |
| TOTAL | 234 | 694 | 87 | 18 | 108 | | |
| class recall | 6.41% | 92.22% | 3.45% | 0.00% | 0.93% | | |

Table 13 Decision Tree performance vector with unsampled data

All classes were weighted equally, and the Average F1-Measure was computed using the per class F1-Score. Table 14 shows the highest measure for each category is printed in bold.

| Comparative model performance without sampling. | | | | | | | |
|---|---------------|---------------|----------------------|---------------|---------------------|-------------------|------------------|
| Description | Accuracy | Precision | Recall - Sensitivity | Specificity | False Positive Rate | Balanced Accuracy | Average F1 Score |
| Decision Tree | 57.76% | 24.40% | 20.60% | 80.55% | 19.45% | 50.58% | 18.17% |
| Gradient Boosted Tree H2O. | 56.44% | 29.59% | 25.63% | 81.81% | 18.19% | 53.72% | 26.14% |
| Naïve Bayes* | 52.32% | 34.78% | 40.43% | 83.23% | 16.77% | 61.83% | 34.96% |
| Neural Net - DL | 60.39% | 35.30% | 26.70% | 81.80% | 18.20% | 54.25% | 25.63% |
| Random Forest | 60.56% | 26.97% | 21.28% | 80.67% | 19.33% | 50.97% | 18.46% |
| SVM. | 60.82% | 12.16% | 20.00% | 80.00% | 20.00% | 50.00% | 15.13% |

Table 14 Comparative model performance without sampling

In these experiments, it was observed that SVM provided the highest accuracy which is marginally higher than Random Forest. However, it is not the best since the recall rates are low and is not predicting minority classes correctly. With higher recall, F1 score and balanced accuracy and low false positive rate, Naïve Bayes is a better classifier. Neural Net is marginally better at precision than Naïve Bayes but falls short on the other measures.

Experiment Set 2 – Experiments with Under Sampling

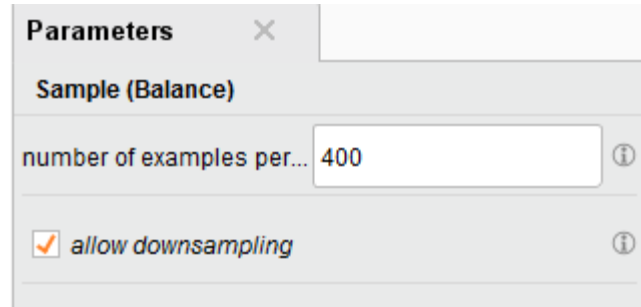
These experiments were devised to use the original data set and the sample balancing operator that allows for up and down sampling. In

In these experiments, the allow down

sampling parameter was enabled and the

number of examples set to 400. With

these parameters, the operator multiplies



data when existing samples are insufficient and reduces sample size per class when the number of samples are higher than the supplied parameter. Table 15 Comparative model performance with undersampling of majority class records results of these experiments.

| Comparative model performance with Undersampling of the Majority class | | | | | | | |
|--|----------|-----------|----------------------|-------------|---------------------|-------------------|------------------|
| Description | Accuracy | Precision | Recall - Sensitivity | Specificity | False Positive Rate | Balanced Accuracy | Average F1 Score |
| Decision Tree* | 83.15% | 82.68% | 83.15% | 67.93% | 32.00% | 75.54% | 62.35% |
| Gradient Boosted Tree H2O. * | 82.55% | 81.68% | 82.55% | 70.11% | 29.89% | 76.33% | 61.48% |
| Naïve Bayes | 59.55% | 57.66% | 59.55% | 56.77% | 43.23% | 58.16% | 56.93% |
| Neural Net - DL | 20.00% | 4.00% | 20.00% | 20.00% | 80.00% | 20.00% | 6.67% |
| Random Forest* | 83.05% | 82.55% | 83.05% | 57.25% | 42.75% | 70.15% | 46.00% |
| SVM. | 20.00% | 0.00% | 20.00% | 20.00% | 80.00% | 20.00% | 0.00% |

Table 15 Comparative model performance with undersampling of majority class

The best performance in this category was that of the Decision tree which ranked the highest with the F1 Score. The accuracy levels were also significantly indicating the positive impact of down sampling the majority class. However, the ensemble GBT fared better in the balanced score and on low False positive rate indicating. The significant learning here is that recall score went up considerably when the majority class was under sampled and other classes were over sampled.

Comparatively the SVM performed poorly when data was down sampled to 400 rows in this experiment.

Experiment Set 3 – Experiments with over sampling using the SMOTE operator

The SMOTE operator synthetically creates data for minority classes using the Synthetic Minority Over-sampling Technique (Chawla *et al.*, 2002). The original dataset had 1141 instances and after oversampling, 1817 rows were produced. As seen in Figure 41, the highest and lowest class had 694 instances indicating that the lowest class was multiplied to match the number in the highest class.

| | Major | TRUE | | | | | TOTAL S |
|------------------|--------------|------------|--------------|-----------|-----------------|-------------|-------------|
| | | APPSDEV | SECURIT Y | NETWORK | MULTI- MEDIA | BUS- SOL | |
| PREDICTED | APPSDEV | 24 | 23 | 3 | 0 | 4 | 54 |
| | SECURITY | 190 | 599 | 57 | 3 | 88 | 937 |
| | NETWORK | 5 | 12 | 13 | 0 | 6 | 36 |
| | MULTI-MEDIA | 11 | 53 | 11 | 690 | 4 | 769 |
| | BUS-SOL | 4 | 7 | 3 | 1 | 6 | 21 |
| | TOTAL | 234 | 694 | 87 | 694 | 108 | 1817 |

Figure 41 Confusion Matrix of Naïve Bayes model with Oversampled Minority Class - SMOTE

Table 16 indicates that Naïve Bayes performed the best using over sampling of minority with the highest balanced accuracy and the F1 Scores. Sampling makes a difference with this classifier.

However, Recall went down when compared to balanced undersampling.

Decision tree had the lowest FP Rate but lower level recall and precision scores resulting in lower balanced accuracy and F1 score.

| | Comparative model performance with over sampling with SMOTE | | | | | | |
|----------------------------|--|------------------|-----------------------------|--------------------|----------------------------|--------------------------|-------------------------|
| Description | Model Accuracy | Precision | Recall - Sensitivity | Specificity | False Positive Rate | Balanced Accuracy | Average F1 Score |
| Decision Tree | 73.03% | 48.03% | 43.75% | 91.85% | 8.15% | 67.80% | 42.53% |
| Gradient Boosted Tree H20. | 70.78% | 48.95% | 45.20% | 91.56% | 8.44% | 68.38% | 45.93% |
| Naïve Bayes* | 69.07% | 50.01% | 49.71% | 91.41% | 8.59% | 70.56% | 49.24% |
| Neural Net - DL | 44.19% | 47.49% | 29.88% | 82.65% | 17.35% | 56.26% | 29.63% |
| Random Forest | 73.31% | 52.56% | 43.30% | 91.75% | 8.25% | 67.53% | 42.98% |
| SVM. | 60.32% | 34.42% | 35.40% | 88.27% | 11.73% | 61.84% | 33.58% |

Table 16 Comparative model performance with over sampling with SMOTE

Overall Model Result Evaluation

Table 17 Overall Model Performance Rankings

| | Description | Decision Tree | Gradient Boosted Tree H20. | Naïve Bayes | Neural Net - DL | Random Forest | SVM. |
|------------------------------|----------------------|---------------|----------------------------|-------------|-----------------|---------------|------|
| SMOTE - OVERSAMPLED MINORITY | Model Accuracy | 2 | 3 | 4 | 6 | 1 | 5 |
| | Error Rate | 5 | 4 | 3 | 1 | 6 | 2 |
| | Precision | 4 | 3 | 2 | 5 | 1 | 6 |
| | Recall - Sensitivity | 3 | 2 | 1 | 6 | 4 | 5 |
| | Specificity | 1 | 3 | 4 | 6 | 2 | 5 |
| | False Positive Rate | 6 | 4 | 3 | 1 | 5 | 2 |
| | Balanced Accuracy | 3 | 2 | 1 | 6 | 4 | 5 |
| | Average F1 Score | 4 | 2 | 1 | 6 | 3 | 5 |
| UNDERSAMPLED MAJORITY CLASS | Model Accuracy | 1 | 3 | 4 | 5 | 2 | 5 |
| | Error Rate | 6 | 4 | 3 | 1 | 5 | 1 |
| | Precision | 1 | 3 | 4 | 5 | 2 | 6 |
| | Recall - Sensitivity | 1 | 3 | 4 | 5 | 2 | 5 |
| | Specificity | 2 | 1 | 4 | 5 | 3 | 5 |
| | False Positive Rate | 5 | 6 | 3 | 1 | 4 | 1 |
| | Balanced Accuracy | 2 | 1 | 4 | 5 | 3 | 5 |
| | Average F1 Score | 1 | 2 | 3 | 5 | 4 | 6 |
| WITHOUT SAMPLING | Model Accuracy | 4 | 5 | 6 | 3 | 2 | 1 |
| | Error Rate | 3 | 2 | 1 | 4 | 5 | 6 |
| | Precision | 5 | 3 | 2 | 1 | 4 | 6 |
| | Recall - Sensitivity | 5 | 3 | 1 | 2 | 4 | 6 |
| | Specificity | 5 | 2 | 1 | 3 | 4 | 6 |
| | False Positive Rate | 2 | 5 | 6 | 4 | 3 | 1 |
| | Balanced Accuracy | 5 | 3 | 1 | 2 | 4 | 6 |
| | Average F1 Score | 5 | 2 | 1 | 3 | 4 | 6 |

In case of SMOTE – **oversampling** of the minority classes, Naïve Bayes was the best in balanced accuracy and F1 Score as it was ranked the highest in recall. Accuracy and Precision were consistently high with Random forest which also had the lowest error rates.

In case of **under sampled data**, the best performer was the ensemble GBT with the highest in Balanced accuracy since it has a very specificity. However, the Decision tree performed best in accuracy, precision as well as recall and hence had the highest F1-Score and the lowest error rate. Hence the overall performance in case of under sampled data was the Decision Tree classifier.

In case of **data without sampling**, Naive Bayes performed well in recalling classes correctly, reflected in the F1 and Balanced Accuracy scores. SVM had the best accuracy but low scores on precision indicating that it favored the majority class. SVM scored highest in false positive confirming that it only classified the majority class.

In the absence of a consistent top-ranking classifier for all data sets, Naïve Bayes did better with oversampled data and data without any sampling and Decision Trees and Gradient Boosted Trees performed best with under sampled balanced data sets.

The lowest error rates were noted with under sampled data, with rule based learners such as Decision Trees, Random Forests and an Ensemble Model – GBT as seen in Figure 42

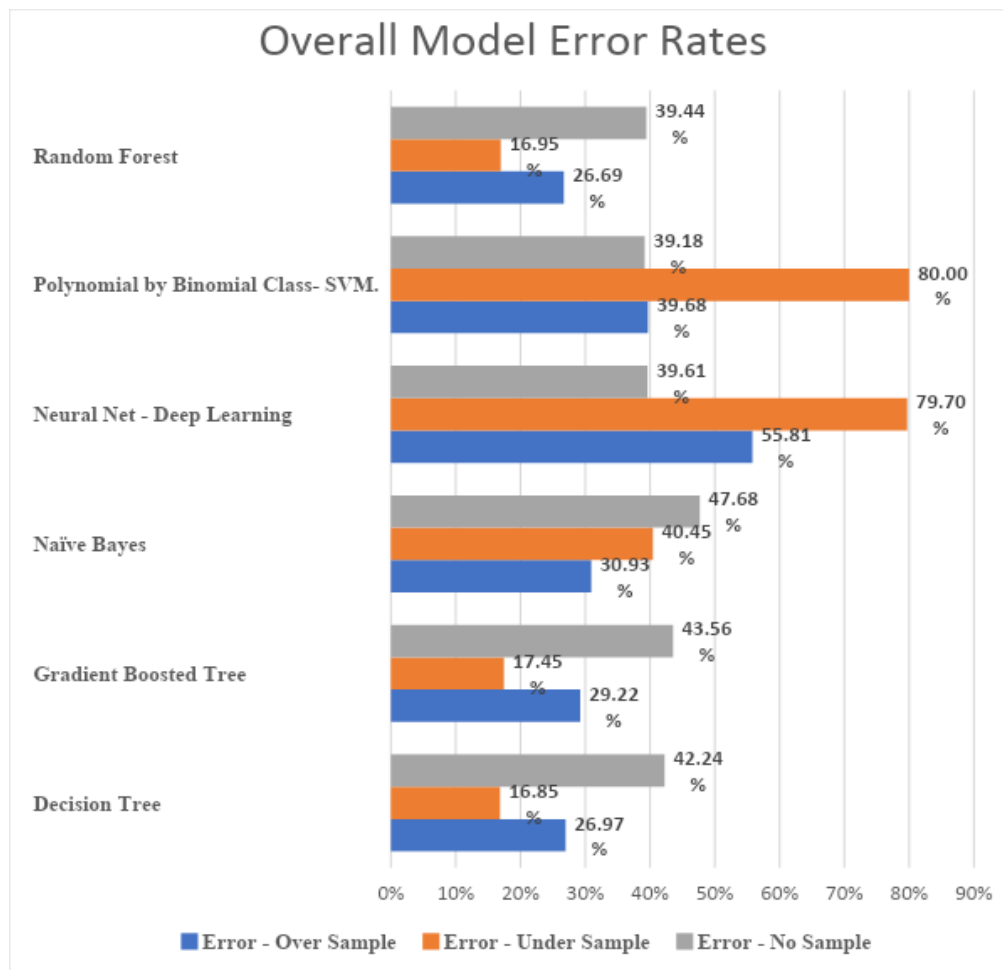


Figure 42 Comparative chart of Error Rates

Analysis of Classifiers across types of sampling

In this section, the performance of each classifier across different types of sampling is noted. A comparison of accuracy, F1score and balanced accuracy was conducted to understand the impact of sampling. Figure 43 shows the per class instances with sampling. As can be seen, over sampling takes the minority class and creates instances to match the number of instances in the majority class. With the down sampled data, each class is set to 400. Instances from the majority class is sampled and instances generated for the other classes to make them all equal.

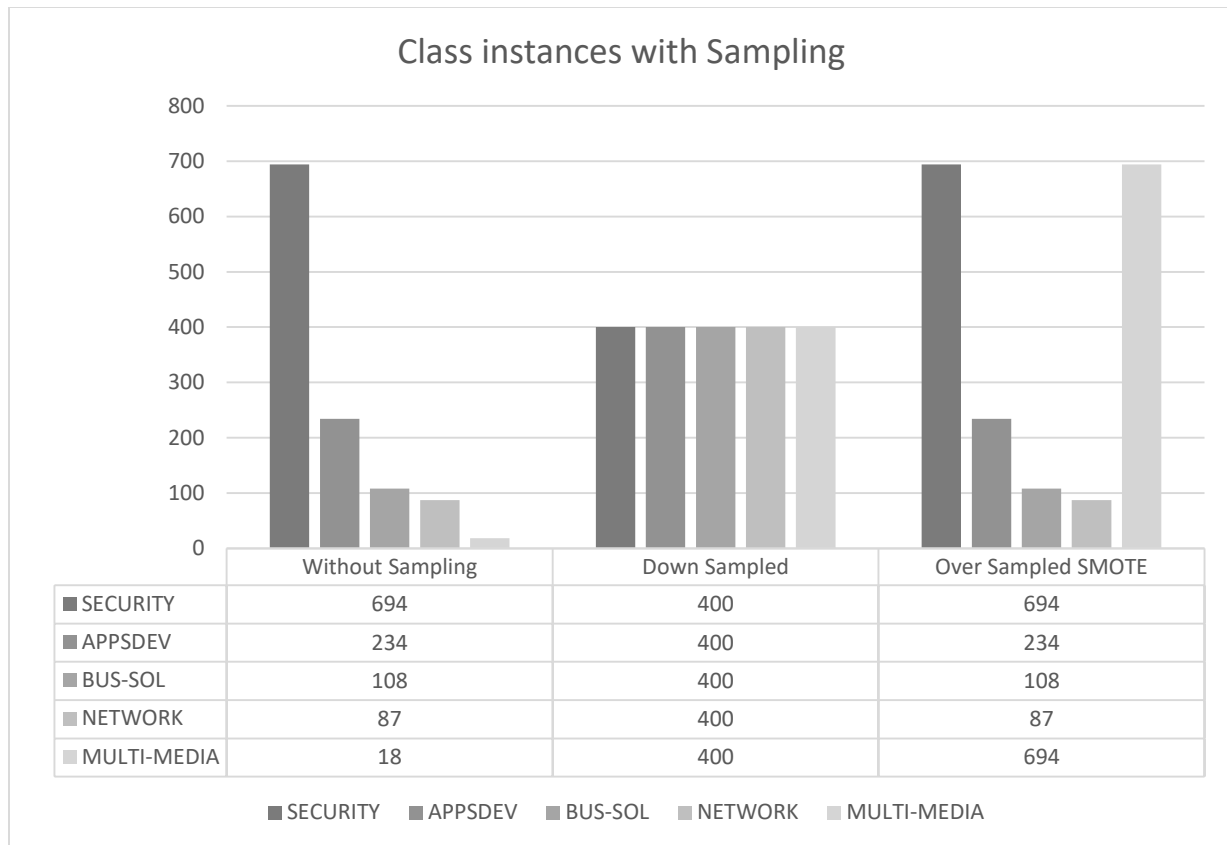


Figure 43 Class instances with Sampling

This study evaluated the performance of each classification algorithm across the different sampling techniques used. Figure 44 shows the performance of models without sampling indicated by NONE, under sampling indicated by UNDER and over sampling with SMOTE indicated by OVER.

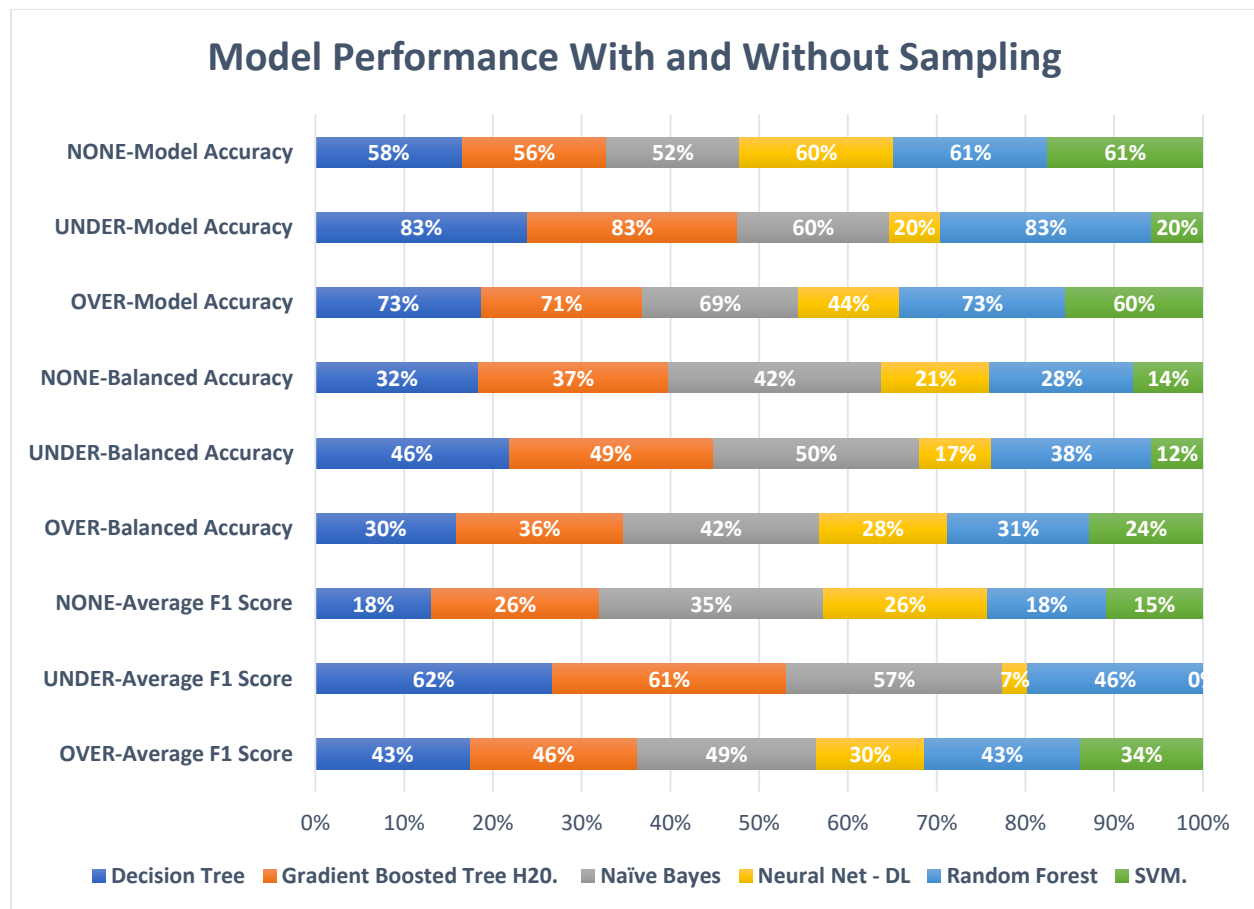


Figure 44 Model Performance with and without Sampling

Decision Trees provided the best accuracy, balanced accuracy and F1 score with under sampled data and performed the poorly when there was no sampling.

Gradient Boosted Tree also had the best accuracy, balanced accuracy and F1 score with under sampled data. Random Forest achieved high indicators with under sampled data and performed poorly when data was not sampled.

Naïve Bayes had the best accuracy with over sampled data. However, the F1 and Balanced Accuracy were the best with under sampled data. Hence the better performance of this classifier was with the under sampled dataset.

SVM had the best accuracy with no sampling due to overfitting. However, with over sampling of the minority class, SVM performed better with the F1 score and the balanced accuracy. SVM needs more instances of the minority class to train.

Neural Net had the best accuracy when there was no sampling because it does not overfit to the majority class. This classifier also performed better with over sampled minority in terms of Balanced accuracy and the F1 score.

Summary

This chapter discussed the deployment of data mining experiments and the analysis. At the outset the Auto Model feature of RapidMiner® was used to determine baseline levels. Initial analysis of results indicated that accuracy levels were around 60% and minority classes were being ignored. This provided an opportunity to explore possibilities for sampling techniques as well as using other measures for evaluating classification algorithms. Three sets of experiments were designed and executed with an identical training data using cross validation for testing. Experiments were without sampling, with balanced sampling/undersampling of majority and oversampling of minority class using six classification algorithms. Results from the confusion matrix were used to compute balanced accuracy and the F1-score using Excel. Results were ranked and used in the analysis. Key findings of this section are:

- Accuracy is not the best measure of prediction when imbalance exists in the data set
- Using measures that include recall-sensitivity and specificity that involve true positives and false negatives are better for assessing performance models
- False Positive Rate measures the performance of the negative samples and hence can provide some indication about classification of the minority samples

- Sampling improved performance of all classifiers. With balanced sampling/undersampling of majority, Decision Tree and Gradient Boosted Tree performed well. In case of Oversampling of the minority with SMOTE, and without any sampling, Naïve Bayes performed best.

This concludes the chapter on deployment. The following chapter includes a discussion on each research question.

Chapter 6 Discussions

This paper discusses the application of data mining and classification to an educational problem, and the methodologies used. The results of the research conducted are discussed in the following section in order of the research question posed.

Research Q1

[R1] Can data mining be applied to the predict student's selection of program majors in higher education?

The literature review conducted at the outset revealed the typical applications of educational data mining, potential for growth and areas that can be considered for further studies.

The reasons for this interest in EDM are the gradual maturity in educational enterprise systems that recording data relating to educational functions of such as enrollment, academic progression, grades, online learning and assessments and student satisfaction with learning. This large collection of data provides a great opportunity for data mining activities.

This study built on the list of stakeholders and EDM applications by (Romero and Ventura, 2010), adding advisors as stakeholders and applications relating to advisement. Applications were grouped into themes as related to student learning, performance in courses, advising, enrollment and attrition, and academic governance. In their study (Peña-Ayala, 2014) categorize papers into similar themes, and indicate that the most popular functionalities relate to student behavior, performance and assessment models.

A count of papers reviewed for this study indicates that the focus is mainly on the prediction of student performance in courses, student enrollment, progression and attrition. However, there is

a shortage of studies relating to factors that affect the selection of programs or majors, a function that impacts several stakeholders i.e. students, managers, advisors and teachers.

Obtaining primary data is time consuming and elaborate and cannot be reused, whereas transactional data in universities can be used efficiently to develop predicting models to help students select a suitable major. Studies by (Saa, 2016) considers social and demographic factors that impact academic performance using data from a survey while all others used data from institutional databases.

(Vialardi *et al.*, 2010) present a case where the registration system evaluates the potential of a student and the difficulty level of the course to predict if students would pass or fail. Prediction is done using a data mining model constructed and integrated into the registration system. The intention is to help students become aware of the challenges in the course that they are trying to register and make an informed decision. Such integration also helps to reduce attrition and ensure successful progression.

There is a shortage of research in the UAE in this sector and none in the area which help students select their major based on their academic data. This data is easily available and can be mined to help students make an informed decision. The challenge is about applying data mining in real time or developing applications that can easily integrate to existing information systems that will provide students, advisors and other stakeholders with real-time information. The other challenge relates to accessibility of student academic data owing to issues relating to confidentiality. Data may not become available soon enough to be used for data mining tasks. These challenges need to be addressed at an organizational level to ensure that data mining is successful.

Despite these challenges, it can be concluded that academic data can be used to predict a student's selection of majors.

Research Q2

[R2] What is an appropriate framework to be used for data mining projects?

This study conducted a literature review of established and emerging methodologies used to manage data mining projects.

The initial findings revealed that CRISP-DM is the most popular, but several studies indicated short comings of this model. The primary short coming of CRISP-DM is the non-consideration of project management tasks (Sharma and Osei-Bryson, 2009; Angée *et al.*, 2018). Hence the SEMMA, KDD and ASUM were reviewed to understand their similarities and improvements over CRISP-DM.

It is clear CRISP-DM is used because it provides an immediate framework to implement a data mining project. The six phases of this methodology allow practitioners to move from business understanding to deployment. Processes in each phase are also well defined and hence it is easy to follow the model. In their study (Mariscal, Marbán and Fernández, 2010) suggest additional subprocesses to make the methodology include life cycle selection, resource management, deployment and maintenance. This recommendation marries a systems development methodology with the proposed three phased model of Analysis-Development-Maintenance.

Whereas the goal of the KDD model is to develop new knowledge with the help of data mining. In this case data mining is a tool to achieve new knowledge. The criticism is the lack of focus on the business understanding phase that may lead to the new knowledge less useful.

SEMMA is associated with a software that enables the implementation of data mining tasks.

ASUM is an improvement on the CRISP - DM and acknowledges the need to incorporate project management into data mining projects.

To conclude, there is no standardized methodology despite the popularity of CRISP-DM. None of the papers reviewed documented the challenges, risks or the feasibility of data mining projects. Hence the answer to this question is there is no best methodology. A good practice suggested by (Daderman, Antonia; Rosander, 2018) is to develop a set of criteria to evaluate a methodology. This study recommends

- Developing criterion for selecting the methodology based on the complexity and scope of the data mining project and business needs of the organization.
- Where the scope is large, and it is required for the project to go into operation, consider adopting iterative or agile development models
- Embed the datamining methodology such as CRISP-DM within a project management framework using PMI framework of process groups and knowledge areas. (*PMBOK® Guide – Sixth Edition (2017)*). This would make the CRISP-DM robust and enable the management of data mining projects using an established framework.

Research Q3

[R3] What are the most suitable algorithms that can be used to classify and predict multi-class labels with imbalanced data?

The most suitable algorithms to be used is totally dependent on the nature of the data, hence it is critical to explore the data in several ways as demonstrated in this study. It is important to understand the data in the target variable whether it is numerical or not, whether it is binary or

multi-class or multi-labelled, and the ratio of imbalance between each class, number of variables and size of dataset.

Understanding data helped to decide on the sampling techniques to be used as suggested by (Krawczyk, 2016). Hence conducting systematic experiments with different sampling techniques helped to identify the best algorithms for the prediction task.

Literature review of algorithms helped to identify the pros and cons and the parameters that need to be used to obtain optimal results. Table 18 shows recommendations by other studies and observations in this study that either contradict or support it.

| Classifier | Recommendation | Pros | Cons | Observed in this study |
|------------------------|--|--|--|---|
| Decision Trees | Reduce depth of pruning (KumarYadav and Pal, 2012) | Work on numerical or polynomial multi-class | Prone to overfitting with multiclass labels | Pruning applied and DT worked well with sampling |
| Random Forest | Use with sampling for better results. Use parameters 2 random features with 300 trees (Chau and Phung, 2013) | Does not overfit imbalanced data. Provides high levels of accuracy with large data sets and smaller feature sets. | Slow to train without pruning | Performed well with 20 trees and gain ratio enabled. Achieved good results with Balanced Sampling |
| Gradient Boosted Trees | | Can be used for feature ranking Few parameters to tune | Slow to train specially with a large feature set. Needs to tune several parameters for optimal performance | GBT performed well with balanced sampling. Performance not observed with large number of records Training took long with 100 trees hence was set at 20 with a depth of 10. |
| Naïve Bayes | | Is fast and versatile and can handle binary and | | Performed the best without sampling and with |

| Classifier | Recommendation | Pros | Cons | Observed in this study |
|----------------------------|--|--|---|--|
| | | multi-class prediction tasks. Does not consider relationship between attributes | | oversampling with SMOTE. |
| SVM | Convert data to numeric and reduce dimensionality before use. | Handles text and images and imbalanced data sets | Slow, high computational costs with high number of attributes | SVM was slow even with few attributes. Does not work well with polynomial attributes and needs data to be converted. |
| Artificial Neural Networks | Using SMOTE oversampling improves prediction of minority class (Jishan <i>et al.</i> , 2015) | Works better with large data sets | Works only with binary or numeric data | The best performance was with without sampling. Oversampling with SMOTE results were better than undersampling. |

Table 18 Classification Algorithms - Pros and Cons

This study recommends the tuning of parameters and progressive experiments to obtain the best results.

Research Q4

[R4] What is the most appropriate evaluation measure to be used with multiclass datasets?

In answering this question, this study reviewed several evaluation measures used to measure the performance of algorithms used in predicting tasks. In multiclass predicting tasks considering only the accuracy can be misleading since the model it considers only the majority class. It may be necessary to correctly identify and predict the significant minority as well. Hence to get a complete picture it is important to consider values that include measurement of negatives as well. The Error rates, F1-Score and Balanced Accuracy incorporate recall and sensitivity and hence

were used to evaluate performance of models. Ranking of results helped to identify the algorithm and the technique.

This study presented the confusion matrix for multiclass data and identified reasons to look beyond accuracy as a measure to verify the results of a classifier.

To summarize, this study has answered all the research questions established at the beginning of the study. The next chapter outlines the conclusion and possibilities of work to be done in the future.

Chapter 7 Conclusions and Future Work

Conclusion

This study provided an approach to applying data mining to a problem in the educational sector.

The approach provides

- A recommendation to use data mining methodologies to manage a data mining project
- Data exploration and visualization techniques to understand data
- Sampling techniques to handle imbalanced data and multiclass data
- A set of classification algorithms to predict and measures to evaluate their performance

This approach was developed by answering a series of research questions relating to these areas.

The study provides tools and techniques for those wanting to embark on a data mining project.

While the context is educational, it can be used for other tasks in other domains that involve multiclass data sets with imbalanced data.

Future Work

This study did not develop a product or an application to deploy this solution in real time and hence the success of prediction is not fully tested. However, it is possible to build an application to take inputs from students and recommend a major. The algorithm for this recommendation can be derived converting the outcome of the best performing decision tree into rules and use these in the application.

This study recommended the use of data available in the institutional database. Obtaining primary data relating to social factors and employer requirements would enrich this study.

This study did not delve into the cost of misclassification in prediction in educational data mining tasks. A further study in this area would relate to evaluate costs of misclassifying and impact on the minority classes.

This study has learnt that the best balanced accuracy and F1-score is when the majority class was down sampled, and all classes were balanced. There is an opportunity to pursue further experiments with sample balancing with larger datasets and to observe performance with samples at different intervals.

Chapter 8 Bibliography

Abdous, M., He, W. and Yen, C. J. (2012) 'Using data mining for predicting relationships between online question theme and final grade', *Educational Technology and Society*, 15(3), pp. 77–88.

Ahadi, A. *et al.* (2015) 'Exploring Machine Learning Methods to Automatically Identify Students in Need of Assistance', pp. 121–130. doi: 10.1145/2787622.2787717.

Aher, S. B. and Lobo, L. M. R. J. (2013) 'Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data', *Knowledge-Based Systems*. Elsevier B.V., 51, pp. 1–14. doi: 10.1016/j.knosys.2013.04.015.

Analytics, D. I. B. M. *et al.* (2016) 'Analytics services Datasheet'.

Angée, S. *et al.* (2018) 'Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects', in *International Conference on Knowledge Management in Organizations*, pp. 613–624.

Anzer, A., Tabaza, H. A. and Ali, J. (2018) 'Predicting Academic Performance of Students in UAE Using Data Mining Techniques', *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, (June), pp. 179–183. doi: 10.1109/icacce.2018.8458053.

Archana, S. and Elangovan, K. (2014) 'Survey of classification techniques in data mining', *International Journal of Computer Science and Mobile Applications*, 2(2), pp. 65–71.

Ashari, A., Paryudi, I. and Min, A. (2013) 'Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation

Tool', *International Journal of Advanced Computer Science and Applications*, 4(11), pp. 33–39.
doi: 10.14569/ijacsa.2013.041105.

Beyadar, H. and Gardali, K. (2011) 'Knowledge management in organizations', *2011 5th International Conference on Application of Information and Communication Technologies, AICT 2011*, (July). doi: 10.1109/ICAICT.2011.6110900.

Chanamarn, N. and Tamee, K. (2017) 'Enhancing Efficient Study Plan for Student with Machine Learning Techniques', *International Journal of Modern Education and Computer Science*, 9(3), pp. 1–9. doi: 10.5815/ijmeecs.2017.03.01.

Chau, V. T. N. and Phung, N. H. (2013) 'Imbalanced educational data classification: An effective approach with resampling and random forest', *Proceedings - 2013 RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for Future, RIVF 2013*, (January), pp. 135–140. doi: 10.1109/RIVF.2013.6719882.

Chawla, N. V *et al.* (2002) 'SMOTE: synthetic minority over-sampling technique', *Journal of artificial intelligence research*, 16, pp. 321–357.

Colliers International (2018) *Higher Education in Dubai*.

Costa, E. B. *et al.* (2017) 'Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses', *Computers in Human Behavior*. Elsevier Ltd, 73, pp. 247–256. doi: 10.1016/j.chb.2017.01.047.

Daderman, Antonia; Rosander, S. (2018) 'Evaluating Frameworks for Implementing Machine Learning in Signal Processing and KDD'.

Đurđević Babić, I. (2018) 'Machine learning methods in predicting the student academic

motivation’, *Croatian Operational Research Review*, 8(2), pp. 443–461. doi: 10.17535/corr.2017.0028.

Fadzilah Siraj and Mansour Ali Abdoulha (2011) ‘Mining enrolment data using predictive and descriptive approaches’, *Malaysia - Knowledge-Oriented Applications in Data Mining*, pp. 53–72.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) ‘Knowledge discovery and data mining: towards a unifying framework’, *Kdd-96*, pp. 82–88.

Fernandes, E. *et al.* (2019) ‘Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil’, *Journal of Business Research*. Elsevier, 94(August 2017), pp. 335–343. doi: 10.1016/j.jbusres.2018.02.012.

Garousi, V. and Mäntylä, M. V. (2016) ‘A systematic literature review of literature reviews in software testing’, *Information and Software Technology*. Elsevier B.V., 80, pp. 195–216. doi: 10.1016/j.infsof.2016.09.002.

Hartono, H. *et al.* (2018) ‘Biased support vector machine and weighted-smote in handling class imbalance problem’, *International Journal of Advances in Intelligent Informatics*, 4(1), p. 21. doi: 10.26555/ijain.v4i1.146.

Hussain, M. *et al.* (2017) ‘Mining Educational Data for Academic Accreditation: Aligning Assessment with Outcomes’, *Global Journal of Flexible Systems Management*, 18(1), pp. 51–60. doi: 10.1007/s40171-016-0143-3.

J. Kovacic, Z. (2017) ‘Early Prediction of Student Success: Mining Students Enrolment Data’, *Proceedings of the 2010 InSITE Conference*, pp. 647–665. doi: 10.28945/1281.

Jishan, S. T. *et al.* (2015) ‘Improving accuracy of students’ final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique’, *Decision Analytics*, 2(1), pp. 1–26. doi: 10.1186/s40165-014-0010-2.

Keele, S. and others (2007) *Guidelines for performing systematic literature reviews in software engineering*.

Kitchenham, B. *et al.* (2010) ‘Systematic literature reviews in software engineering-A tertiary study’, *Information and Software Technology*, 52(8), pp. 792–805. doi: 10.1016/j.infsof.2010.03.006.

Kotsiantis, S. B. (2012) ‘Use of machine learning techniques for educational proposes: A decision support system for forecasting students’ grades’, *Artificial Intelligence Review*, 37(4), pp. 331–344. doi: 10.1007/s10462-011-9234-x.

Kotu, Deshpande, B. (2015) ‘Predictive analytics and data mining. Concepts and practice with rapidMiner [Análisis predictivo y minería de datos. Conceptos y práctica con rapidMiner’.

Kotu, V. and Deshpande, B. (2014) ‘Predictive analytics and data mining : concepts and practice with RapidMiner’.

Kotu, V. and Deshpande, B. (2019) ‘Chapter 1 - Introduction’, in Kotu, V. and Deshpande, B. (eds) *Data Science (Second Edition)*. Second Edi. Morgan Kaufmann, pp. 1–18. doi: <https://doi.org/10.1016/B978-0-12-814761-0.00001-0>.

Krawczyk, B. (2016) ‘Learning from imbalanced data: open challenges and future directions’, *Progress in Artificial Intelligence*. Springer Berlin Heidelberg, 5(4), pp. 221–232. doi: 10.1007/s13748-016-0094-0.

- KumarYadav, S. and Pal, S. (2012) 'Data Mining Application in Enrollment Management: A Case Study', *International Journal of Computer Applications*, 41(5), pp. 1–6. doi: 10.5120/5534-7581.
- Manning, C. D. (2009) 'Intro to Information Retrieval', *Information Retrieval*, (c), pp. 1–18. doi: 10.1109/LPT.2009.2020494.
- Maratea, A., Petrosino, A. and Manzo, M. (2014) 'Adjusted F-measure and kernel scaling for imbalanced data learning', *Information Sciences*, 257, pp. 331–341. doi: 10.1016/j.ins.2013.04.016.
- Mariscal, G., Marbán, Ó. and Fernández, C. (2010) 'A survey of data mining and knowledge discovery process models and methodologies', *Knowledge Engineering Review*, 25(2), pp. 137–166. doi: 10.1017/S0269888910000032.
- Miguéis, V. L. *et al.* (2018) 'Early segmentation of students according to their academic performance: A predictive modelling approach', *Decision Support Systems*. Elsevier, 115(September), pp. 36–51. doi: 10.1016/j.dss.2018.09.001.
- MOE (2018) *Training @ www.moe.gov.ae, Ministry of Education, UAE Study - Majors in Demand*. Available at: <https://www.moe.gov.ae/En/MediaCenter/News/Pages/Training.aspx>.
- Mueen, A., Zafar, B. and Manzoor, U. (2016) 'Modeling and Predicting Students' Academic Performance Using Data Mining Techniques', *International Journal of Modern Education and Computer Science*, 8(11), pp. 36–42. doi: 10.5815/ijmecs.2016.11.05.
- Naaj, M. A., Nachouki, M. and Ankit, A. (2012) 'Evaluating Student Satisfaction with Blended Learning in a Gender-Segregated Environment', *Journal of Information Technology Education*,

11.

Nandeshwar, A. and Chaudhari, S. (2009) 'Enrollment prediction models using data mining', *Retrieved January*, 1(2007), pp. 1–17. doi: 10.1638/2012-0017R2.1.

Okoli, C. and Schabram, K. (2010) '(Okoli, Schabram 2010 Sprouts) systematic literature reviews in IS research', 10(2010).

Peña-Ayala, A. (2014) 'Educational data mining: A survey and a data mining-based analysis of recent works', *Expert Systems with Applications*, 41(4 PART 1), pp. 1432–1462. doi: 10.1016/j.eswa.2013.08.042.

Piatetsky, G. (2014) *What main methodology are you using for your analytics, data mining, or data science projects? Poll, Kdnuggets*. Available at: <https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html> (Accessed: 2 March 2019).

PMBOK® Guide – Sixth Edition (2017). (no date).

Rapidminer (2017) 'How to Correctly Validate Machine Learning Models'.

RapidMiner (2018) *RapidMiner educational license program*. Available at: <https://rapidminer.com/educational-program/> (Accessed: 13 April 2019).

Rogalewicz, M. and Sika, R. (2016) 'Methodologies of knowledge discovery from data and data mining methods in mechanical engineering', *Management and Production Engineering Review*, 7(4), pp. 97–108. doi: 10.1515/mper-2016-0040.

Rowley, J. and Slack, F. (2004) 'Conducting a Literature Review', *Management Research Review*. Emerald Group Publishing, Limited, 27(6), p. 31.

- Saa, A. A. (2016) *Educational Data Mining & Students' Performance Prediction, IJACSA) International Journal of Advanced Computer Science and Applications.*
- Saltz, J. S. *et al.* (2018) 'Exploring Project Management Methodologies Used Within Data Science Teams', *Twenty-fourth Americas Conference on Information Systems, New Orleans, 2018*, pp. 1–5.
- SAS Institue Inc. (2017) 'SAS ® Enterprise Miner TM 14.2: Reference Help', pp. 321–327.
- Sharma, S. and Osei-Bryson, K. M. (2009) 'Framework for formal implementation of the business understanding phase of data mining projects', *Expert Systems with Applications*. Elsevier Ltd, 36(2 PART 2), pp. 4114–4124. doi: 10.1016/j.eswa.2008.03.021.
- Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', *Information Processing and Management*. Elsevier Ltd, 45(4), pp. 427–437. doi: 10.1016/j.ipm.2009.03.002.
- Spoon, K. *et al.* (2016) 'Random Forests for Evaluating Pedagogy and Informing Personalized Learning', *Journal of Educational Data Mining*, 8(2), pp. 20–50.
- Systems, S. (2019) 'Interview with Adele Cutler: Remembering Leo Breiman - Dan Steinberg's Blog', *Salford-systems.com*.
- Tharwat, A. (2018) 'Classification assessment methods', *Applied Computing and Informatics*. The Author. doi: 10.1016/j.aci.2018.08.003.
- Vialardi, C. *et al.* (2010) 'A case study: data mining applied to student enrollment', ... *Educational Data Mining* ..., (June 2014), pp. 333–334.
- Vialardi, C. *et al.* (2011) 'A data mining approach to guide students through the enrollment

process based on academic performance’, *User Modeling and User-Adapted Interaction*, 21(1–2), pp. 217–248. doi: 10.1007/s11257-011-9098-4.

Wanjau, S. K., Okeyo, G. and Rimiru, R. (2016) ‘Data Mining Model for Predicting Student Enrolment in STEM Courses in Higher Education Institutions’, *International Journal of Computer Applications Technology and Research*, 5(11), pp. 698–704. doi: 10.7753/ijcatr0511.1004.

Wilkins, S. and Balakrishnan, M. S. (2013) ‘Assessing student satisfaction in transnational higher education’, *International Journal of Educational Management*, 27(2), pp. 143–156. doi: 10.1108/09513541311297568.

Ben Youssef, Y. *et al.* (2018) ‘A novel online QoE prediction model based on multiclass incremental support vector machine’, *Proceedings - International Conference on Advanced Information Networking and Applications*, AINA. IEEE, 2018-May, pp. 334–341. doi: 10.1109/AINA.2018.00058.

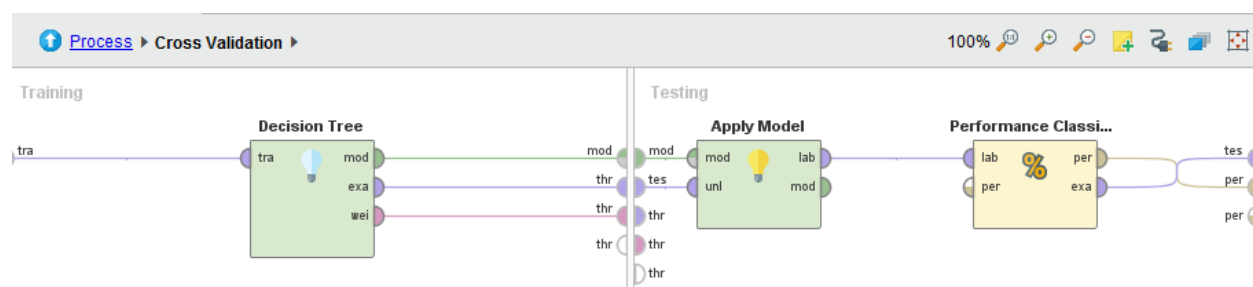
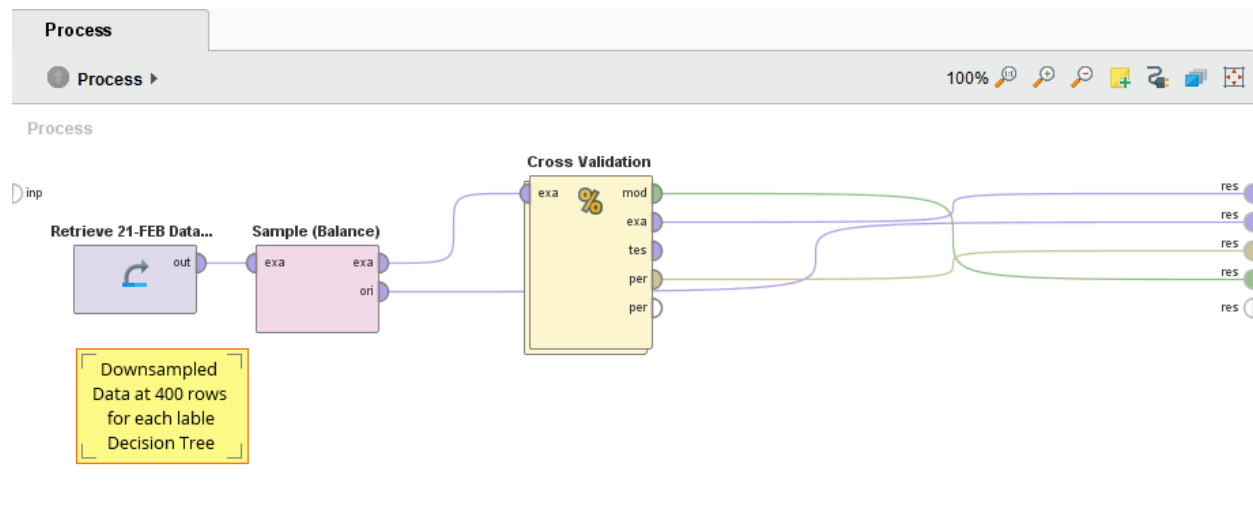
Chapter 9 Appendices

Appendix - 1

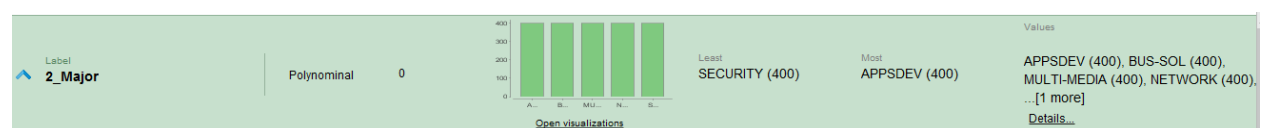
Sample Model from RapidMiner

Experiment: Decision Tree with Down sampled majority with Sample-Balance operator.

Drilled down main-process



Balanced Sample Size



☒ Table View ☐ Plot View

accuracy: 82.65% +/- 37.88% (micro average: 82.65%)

| | true APPSDEV | true SECURITY | true NETWORK | true MULTI-MEDIA | true BUS-SOL | class precision |
|-------------------|--------------|---------------|--------------|------------------|--------------|-----------------|
| pred. APPSDEV | 306 | 104 | 4 | 0 | 9 | 72.34% |
| pred. SECURITY | 50 | 184 | 3 | 0 | 16 | 72.73% |
| pred. NETWORK | 12 | 51 | 393 | 0 | 5 | 85.25% |
| pred. MULTI-MEDIA | 4 | 12 | 0 | 400 | 0 | 96.15% |
| pred. BUS-SOL | 28 | 49 | 0 | 0 | 370 | 82.77% |
| class recall | 76.50% | 46.00% | 98.25% | 100.00% | 92.50% | |