PREDICTING STUDENTS' ACADEMIC ACHIEVEMENT USING METHODS OF EDUCATIONAL DATA MINING

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Frau Sarah Alturki
aus Riad, Saudi Arabien

Mannheim, 2022

Dekan:	Dr. Bernd Lübcke, Universität Mannheim
Referent:	Prof. Dr. Heiner Stuckenschmidt, Universität Mannheim
Korreferent:	Prof. Dr. Dirk Ifenthaler, Universität Mannheim

Tag der mündlichen Prüfung: 19.12.2022

ACKNOWLEDGMENTS

This journey has taken a long time and could not have been done without the support of many people. First and foremost, I would like to express my gratitude to my country for offering me the lifetime opportunity to pursue my studies abroad. I am honored and forever grateful to be one of the recipients of such generous support.

I cannot begin to express my special appreciation to my advisor, Prof. Dr. Heiner Stuckenschmidt, for his mentoring and guidance during my Ph.D. study.

To my daughter Lina, this journey started when you were born. Thanks for being alongside me every step of the way. Thank you for cheering me up at the end of long working days. Your love is what got me through when I wanted to give up.

I also want to acknowledge my sincere gratitude to my husband, Dr. Yousef Alzaben, for believing in me even when I doubted myself. Thanks for always being there for me during my ups and downs. Thanks for listing to all the boring details about my studies. Thanks for wiping my tears during my toughest moments.

A big thank you goes to my parents for their love, encouragement, and all the long flights they have taken to be with me. I cannot find the appropriate words to properly describe my appreciation for their devotion, support, and faith in my abilities. To my father, thank you for calling me a doctor ever since I was a child. This title encouraged me to keep up the hard work till I actually earn it. A big thank you goes to my mother for her endless prayers. Words can't begin to explain how much I appreciate everything you have done for me.

To my sister Dr. Nazik, thank you for influencing me to take this journey and paving the way for me. Thanks for being such a great role model. I look forward to starting working with you.

Finally, thanks to all of the students who participated in the study. This research would not have existed without you.

ABSTRACT

The tremendous growth in educational data forms the need to have meaningful information produced from it. Educational Data Mining (EDM) has become an exciting research area that can reveal valuable knowledge from educational databases. This knowledge can be used for many purposes, including identifying dropouts or weak students who need special attention and discovering extraordinary students who can be presented with lifetime opportunities.

This thesis allows the reader to grasp the field of EDM from all its angles, with more details on academic prediction tasks. It provides a comprehensive background for understanding EDM and discusses the different methods and applications of data mining in education. It also provides a rich literature review on predicting students' academic achievement and covers related works from 2007 to 2022. Furthermore, it examines the application of machine learning algorithms to predict students' academic achievement on two diverse datasets.

The first dataset has been obtained from the Computer and Information Science College at Princess Norah University (PNU) in Riyadh, Saudi Arabia. In this work, 300 undergraduate students' records have been used to predict their final academic achievement. We used the Weka software to compare the performance of eight data mining algorithms in predicting students' academic achievement. Those algorithms are C4.5, Simple CART, LADTree, Support Vector Machine, Naïve Bayes, K-nearest-Neighbor, Artificial Neural Networks, and Random Forest and validated the models using 10-folds cross-validation. The empirical results show that: (i) In the College of Computer and Information Science, the following features are the most essential to predict student academic achievement: the student GPA in each semester, the number of failed courses during the first four semesters, and the grades of three core courses; (ii) Naïve Base performs the best in predicting students' achievement followed by Random Forest; (iii) A student's proficiency in English does not play a major role in their success at the college of Computer and Information Sciences, and (iv) Students who attend an orientation year do not have a greater chance of success at that college.

The second dataset represents the records of the Business Informatics master's students at the University of Mannheim in Germany. In this work, more than 700 undergraduate students' data have been used to predict their final academic achievement using different machine learning libraries in python. We compared the performance of nine data mining algorithms in predicting students' academic achievement. Those algorithms are Logistic Regression, Naïve Bayes, K-nearest neighbor, Artificial Neural Networks, Support Vector Machine, Random Forest, Gradient Boosting, Light Gradient Boosting, and Extreme Gradient Boosting and validated the models using 10-folds cross-validation. The empirical results show the following: (i) Bagging and Boosting algorithms produce a better predictive performance as compared to individual classifiers, and (ii) the semesters' grades are the most significant features for the predictive model, followed by students' culture and distance from students' accommodation to university campus.

The outcomes of the two studies can be used to design a recommender system that enables timely interventions for the undergraduate students of the College of Information and Computer Science and the postgraduate students of the Business Informatics program.

Keywords: Student Performance, Machine Learning, Educational Data Mining, Students' dropout, Predictions, Imbalanced Dataset, Oversampling methods.

ZUSAMMENFASSUNG

Das enorme Wachstum an Bildungsdaten macht es erforderlich, aussagekräftige Informationen daraus zu extrahierenEducational Data Mining (EDM) ist zu einem spannenden Forschungsgebiet geworden, das wertvolles Wissen aus Bildungsdatenbanken offenlegen kann. Dieses Wissen kann für viele Zwecke genutzt werden, einschließlich der Identifizierung von Schul- und Studiumsabbrechern und -abbrecherinnen und schwachen Schülern, Schülerinnen und Studierenden, die besondere Aufmerksamkeit benötigen, und der Entdeckung außergewöhnlicher Schüler, Schülerinnen und Studiernder denen lebenslange Chancen geboten werden können.

Diese Arbeit ermöglicht es dem Lesenden, das Gebiet der EDM aus all seinen Blickwinkeln zu erfassen, mit Fokus auf akademische „Prediction tasks". Es bietet einen umfassenden Hintergrund zum Verständnis von EDM und diskutiert die verschiedenen Methoden und Anwendungen von Data Mining in der Bildung. Es bietet auch eine umfassende Literaturübersicht zur Vorhersage der akademischen Leistung von Schüler und deckt verwandte Arbeiten von 2007 bis 2022 ab. Darüber hinaus untersucht es die Anwendung von Algorithmen für maschinelles Lernen, um die akademischen Leistungen von Schülern auf zwei verschiedenen Datensätzen vorherzusagen.

Der erste Datensatz wurde vom Computer and Information Science College der Princess Norah University (PNU) in Riad, Saudi-Arabien, bezogen. In dieser Arbeit wurden die Aufzeichnungen von 300 Studierenden im Grundstudium verwendet, um ihre endgültigen akademischen Leistungen vorherzusagen. Wir haben die Weka-Software verwendet, um die Leistung von acht Data-Mining-Algorithmen bei der Vorhersage der akademischen Leistung von Schülern zu vergleichen. Diese Algorithmen sind C4.5, Simple CART, LADTree, Support Vector Machine, Naive Bayes, K-nearest-Neighbor, Artificial Neural Networks und Random Forest und validierten die Modelle mit 10-facher Kreuzvalidierung. Die empirischen Ergebnisse zeigen, dass: (i) Im College of Computer and Information Science die folgenden Merkmale am wichtigsten sind, um die akademischen Leistungen der Studierenden vorherzusagen: der studentische GPA in jedem Semester, die Anzahl der nicht bestandenen Kurse in den ersten vier Semestern und die Noten von drei Kernfächern; (ii) Naive Base schneidet am besten bei der Vorhersage der Studierendenleistungen ab, gefolgt von Random Forest; (iii) Englischkenntnisse spielen keine große Rolle für den Studienerfolg an der Hochschule für Informatik und Informationswissenschaften, und (iv) Studierende, die ein Orientierungsjahr besuchen, haben an dieser Hochschule keine größeren Erfolgschancen.

Der zweite Datensatz repräsentiert die Aufzeichnungen von Masterstudierenden der Wirtschaftsinformatik an der Universität Mannheim in Deutschland. In dieser Arbeit wurden die Daten von 700 Studenten im Grundstudium verwendet, um ihre endgültigen akademischen Leistungen mithilfe verschiedener Bibliotheken für maschinelles Lernen in Python vorherzusagen. Wir haben die Leistung von neun Data-Mining-Algorithmen bei der Vorhersage der akademischen Leistungen von Schülern verglichen. Diese Algorithmen sind

Logistische Regression, Naive Bayes, K-nearest Nabour, Artificial Neural Networks, Support Vector Machine, Random Forest, Gradient Boosting, Light Gradient Boosting und Extreme Gradient Boosting und wurden mit zehnfacher Kreuzvalidierung validiert. Die empirischen Ergebnisse zeigen Folgendes: (i) Bagging- und Boosting-Algorithmen erzeugen eine bessere Vorhersageleistung im Vergleich zu individuellen Klassifikatoren, und (ii) die Semesternoten sind die wichtigsten Merkmale für das Vorhersagemodell, gefolgt von der Kultur und der Entfernung der Studierenden von ihrer Wohnung bis zum Universitätscampus.

Die Ergebnisse der beiden Studien können genutzt werden, um ein Empfehlungssystem zu entwerfen, das zeitnahe Interventionen für die Bachelor-Studenten des College of Information and Computer Science und die Masterstudierenden des Studiengangs Wirtschaftsinformatik ermöglicht.

Schlüsselwörter: Schülerleistung, maschinelles Lernen, Data Mining im Bildungsbereich, Studienabbruch, Vorhersagen, unausgeglichener Datensatz, Oversampling-Methoden.

PUBLICATIONS

Alturki, S., Cohausz, L, & Stuckenschmidt, H. (2022). Predicting Master's Students' Academic Performance: An Empirical Study in Germany [Manuscript submitted for publication]. Business Informatics and mathematics, University of Mannheim

Alturki, S., Alija, S, & Stuckenschmidt, H. (2022). Online Delivery in Higher Education during Pandamics: Students' Perspective. *Technology Education Management Informatics Journal,* 11 (2), 882–892. DOI: 10.18421/TEM112-49

Alturki, S., & Stuckenschmidt, H. (2021). Assessing Students' Self-Assessment Ability in An Interdisciplinary Domain. *Journal of Applied Research in Higher Education.* https://doi.org/10.1108/JARHE-01-2021-0034

Alturki, S., & Alturki, N., (2021). Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions. *Journal of Information Technology Education: Innovations in Practice*, 20, 121–137. https://doi.org/10.28945/4835

Alturki, S., Hulpus, I., & Stuckenschmidt, H. (2020). Predicting Academic Outcomes: A Survey from 2007 Till 2018. *Technology, Knowledge and Learning*, 27, 275–307. https://doi.org/10.1007/s10758-020-09476-0

LIST OF TABLES

Table 1: Pros and cons of the different DM algorithms used in performing academic predictions...... 43
Table 2: Pros and cons of the different data mining tools used in performing academic predictions .. 45
Table 3: Prediction of students' academic performance or GPA at a degree level 49
Table 4: Prediction of students' failure or drop out of a degree 54
Table 5: Prediction of students' results on particular courses 59
Table 6: Discription of the features that have been used to build the predictive models...................... 64
Table 7: Classification of academic grading 66
Table 8: Feature importance on the prediction models 67
Table 9: Accuracy of predicting the final academic grade after the 2nd academic semester 71
Table 10: Accuracy of predicting the final academic grade after the 3rd academic semester 71
Table 11: Accuracy of predicting the final academic grade after the 4th academic semester 72
Table 12: Description of the collected data that is used to predict the academic achievement 77
Table 13: Features level of importance of the predictions performed after the 1st semester................ 82
Table 14: Features level of importance on the predictions performed after the 2nd semester............. 82
Table 15: Performance of the different DM algorithms in predicting the completion of a degree....... 84
Table 16: Performance of the different DM methods in predicting students' academic grade............ 85
Table 17: Performance of the DM methods in predicting the completion of a degree using SMOTE. 86
Table 18: Performance of the DM methods in predicting the academic grade using SMOTE............ 87
Table 19: F1-Scores for historical Study Outcome Prediction for different feature sets 93
Table 20: Relationship between SSA and culture using Cramer's V (Alturki and Stuckenschmidt, 2021) 98
Table 21: Classification of EDM studies based on their method and application 119
Table 22: Master courses of the Business Informatics program and their prerequisites..................... 121
Table 23: Relationship between the accuracy of SSA and gender using Cramer's V 129

LIST OF FIGURES

Figure 1: The cycle of hiring data mining in educational systems (Romero and Ventura 2007) 16
Figure 2: CRISP-DM Process Life Cycle (Kurgan and Musilek, 2006)... 21
Figure 3: Mostly used features in predicting students' academic outcomes....................................... 37
Figure 4: Mostly used DM algorithms in performing academic predictions 41
Figure 5: Analysis of the DM tools that have been used in the viewed literature 44
Figure 6: Academic predictions based on country.. 46
Figure 7: Percentage of the academic predictions performed in different degrees.............................. 47
Figure 8: The final academic grade distribution for the collected data... 64
Figure 9: The relation between the achieved GPA in each semester and the academic success 68
Figure 10: The relation between the number of failed courses each year and academic success......... 69
Figure 11: The relation between academic language skills and academic success.............................. 69
Figure 12: The relation between previous academic achievement in secondary school and academic success 70
Figure 13: The relation between studying an orientation year and academic success.......................... 70
Figure 14: Students' studying duration at the Business Informatics master's program........................ 77
Figure 15: Students' demographical features 78
Figure 16: Students' academic status 79
Figure 17: Students' academic grade 79
Figure 18: A comparison between the achieved F1 scores after the 1st semester for the minority class (Not_completed) using DM methods with and without SMOTE 89
Figure 19: A comparison between the achieved F1 scores after the 2nd semester for the minority class (Not_completed) using DM methods with and without SMOTE 89
Figure 20: A comparison between the accuracy of the academic predictions after the 1st and 2nd semester using binary classification.. 90
Figure 21: A comparison between the accuracy of the academic predictions after the 1st and 2nd semester using multi-class classification 91
Figure 22: Relationship between the accuracy of SSA and culture (Alturki and Stuckenschmidt, 2021) 98
Figure 23: Students' academic achievement at the Business Informatics master's program based on culture 99
Figure 24: Feature importance for predicting the academic status after the 1st semester.................... 122
Figure 25: Feature importance for predicting the academic status after the 2nd semester 123
Figure 26: Feature importance for predicting the academic grade after the 1st semester.................... 124
Figure 27: Feature importance for predicting the academic grade after the 2nd semester 125
Figure 28: Relationship between academic achievement and each semester's grade in the Business Informatics master's program 126
Figure 29: Relationship between academic achievement and the number of failed courses each semester in the Business Informatics master's program 126
Figure 30: Relationship between the accuracy of SSA and gender 129
Figure 31: Relationship between the accuracy of SSA and GPA 130

ACRONYMS

ADT	Alternating Decision Tee
CRISP-DM	Cross-Industry Standard Process for Data Mining
CS	Computer Science
DM	Data Mining
EDM	Educational Data Mining
EFB	Exclusive Feature Bundling
GDPR	General Data Protection Regulation
GOSS	Gradient-Based One-Side Sampling
GPA	Grade Point Average
GRE	Graduate Record Examinations
ID3	Iterative Dichotomiser 3
IS	Information Systems
IT	Information Technology
Jrip	Repeated Incremental Pruning
KDD	Knowledge Discovery in Databases
KNN	K-nearest Neighbor
KSA	Kingdom of Saudi Arabia
LightGB	Light Gradient Boosting
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-layer Perceptron
NB	Naïve Base
NN	Neural Networks
PNU	Princess Nora University
RN	Random Forest
ROC	Receiver Operating Characteristic
RT	Random Tree
repTree	Reduced Error Pruning Tree
SMOTE	Synthetic Minority Over-sampling Technique
SSA	Students' Self Assessment
SVM	Support Victor Machine
TPP	Temporal Point Processes
UAE	United Arab Emirates
USA	United States of America
XGBoost	eXtreme Gradient Boosting machine

TABLE OF CONTENTS

ACKNOWLEDGMENTS 3
ABSTRACT 4
ZUSAMMENFASSUNG 6
PUBLICATIONS 8
LIST OF TABLES 9
LIST OF FIGURES 10
ACRONYMS 11
CHAPTER 1: INTRODUCTION 16
1.1 Research Motivation 17
1.2 Research Questions 17
1.3 Research Objectives 18
1.4 Thesis Structure 19
CHAPTER 2: EDUCATIONAL DATA MINING BACKGROUND 21
2.1 Phases of Performing Educational Data Mining 21
2.2 Educational Data Mining Methods 23
2.2.1 Relationship mining 23
2.2.2 Prediction 23
2.2.3 Clustering 23
2.2.4 Distillation of data for human judgment 24
2.2.5 Discovery with models 24
2.3 Educational Data Mining Applications 24
2.3.1 Student modeling 24
2.3.2 Domain Modeling and Knowledge Structures 25
2.3.3 Pedagogical support 25
2.3.4 Scientific research 26
2.4 Data Mining Algorithms 26
2.4.1 Decision Trees (DT) 26
2.4.2 Naive Bayes (NB) 27
2.4.3 K-nearest neighbor (KNN) 27
2.4.4 Support Vector Machine (SVM) 27
2.4.6 Logistic Regression (LR) 28
2.4.7 Ensemble learning 28
2.5 Challenges in Educational Data Mining 29

2.5.1 Lack of Knowledge 30

2.5.2 Security issues 30

2.5.3 Personal privacy 31

2.5.4 Ethics 31

2.5.5 Quality of Data 32

2.6 Summary of the Chapter 33

CHAPTER 3: LITERATURE REVIEW ON PREDICTING STUDENTS' ACADEMIC ACHIEVEMENT 35

3.1 Measures of Determining Academic Success 35

3.2 Comprehensive Review of Academic Achievement Prediction Literature 36

3.2.1 Review of the Features Used in Predicting Students' Academic Achievement 36

3.2.2 Review on the used DM algorithms in Predicting Students' Achievement 41

3.2.3 Review on the Used Data Mining Tools in Predicting Students' Achievement 43

3.2.4 Review on the academic prediction studies based on Country 45

3.2.4 Review of the Prediction studies based on the degree 46

3.3 Per Task analysis 47

3.3.1 Prediction of students' academic performance or GPA at a degree level 47

3.3.2 Prediction of students' failure or drop out of a degree 52

3.3.3 Prediction of students' results on particular courses 57

3.4 Summary of the Chapter 60

CHAPTER 4: PREDICTING THE ACADEMIC ACHIEVEMENT OF BACHELOR'S DEGREE STUDENTS 61

4.1 Introduction 61

4.2 Overview of the Bachelor's program 62

4.3 Research Methodology 63

4.3.1 Students' dataset 63

4.3.2 Grading classification 65

4.2.3 DM algorithms 66

4.2.4 DM tools 66

4.2.5 Evaluation method 66

4.2.5 Research challenges 66

4.4 Feature Importance on the Overall Predictions 67

4.5 Results of the academic Predection Models 70

4.5.1 Predicting the graduation grade after the 2nd semester 70

4.5.2 Predicting the graduation grade after the 3rd semester 71

4.5.3 Predicting the graduation grade after the 4th semester 72

4.6 Disscusion and Conclusion 72

4.7 Study Limitations and Future Work 74

4.7.1 Investigating the use of other predictive features 74

4.7.2 Investigating the predictive models on other bachelor’s programs 74

4.7.3 Investigating the use of other DM algorithms 74

CHAPTER 5: PREDICTING THE ACADEMIC ACHIEVEMENT OF MASTER'S DEGREE STUDENTS 75

5.1 Introduction 75

5.2 Overview of the Business Informatics Master’s Program 76

5.3 Research Methodology 77

5.3.1 Students’ dataset 77

5.3.2 Analysis of the students’ dataset 78

5.3.3 DM algorithms 79

5.3.4 Programming language 80

5.3.5 Evaluation method 80

5.3.6 Research challenges 81

5.4 Feature Importac on the Overall Predictions 81

5.4.1 Feature importance on the predictions performed after the first semester 82

5.4.2 Feature importance of the predictions performed after the second semester 82

5.5 Prediction Models 83

5.5.1 Predict students' academic status using binary classification 83

5.5.2 Predict students’ academic grade using multi-class classification 85

5.6 Dealing with Imbalanced Data Using SMOTE 86

5.7 Discussion and Conclusion 88

5.8 Study Limitations and Future Work 92

5.8.1 Predict students’ studying duration in the master’s program 92

5.8.2 Investigating the use of Temporal Point Processes in predicting academic achievement .. 93

5.8.3 Predict students’ academic achievement prior to admission 94

CHAPTER 6: AN IN-DEPTH EVALUATION ON THE IMPACT OF CULTURE ON ACADEMIC PREDICTIONS 96

6.1 The Role of Culture on Students’ Academic Achievement 96

6.2 Using Culture as A Feature for Performing Academic Predictions 98

6.3 Conclusion 99

CHAPTER 7: CONCLUSION 100

7.1 Summary of the Empirical Studies 100

7.2 Answers to the research questions 101

7.3 General Recommendations for Future Work 103
REFERENCES 105
APPENDICES 119
Appendix A: Examples of EDM studies based on their method and application 119
Appendix B: Prerequisite Courses for succeeding the Buiness Informatics master’s program 121
Appendix C: Results of using permutation feature importance function on different algorithms .. 122
Appendix D: Relationship between academic achievement and other features in the master’s program 126
Appendix E: Self-assessment and Self-test Survey 127
Appendix F: Relationship between SSA and different attributes 129

CHAPTER 1: INTRODUCTION

We are drowning in information but starved for knowledge - John Naisbitt, 1982

Daily, a tremendous amount of data is generated from various sources, e.g., social networks, business transactions, medical records, and educational institutions. However, these data are usually stored in databases as raw data. The information overload from the growing data requires introducing new data processing approaches into everyday activities (Kwon and Sim, 2013). Knowledge discovery in databases (KDD) is a way to discover and extract hidden patterns from large data repositories. It is "The non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data" (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). Educational Data Mining (EDM) is a recent trend in the KDD field (Abu Saa, 2016). EDMS refers to techniques, tools, and research designed to extract useful information and patterns from vast quantities of data generated by or related to student learning activities in an educational context (Nithya, Umamaheswari, and Umadevi, 2016). It is based on a variety of literature, including data mining (DM) and machine learning (ML), psychometrics and statistics, information visualization, and computational modeling (Romero and Ventura, 2007). It also concerns social science as it deals with students' behavior from social and cultural aspects. Figure 1 by Romero and Ventura (2007) illustrates the usage of DM in educational environments. It represents an iterative process of hypothesis generation, testing, and refinement in which systems can be shaped to meet the needs of each member of the educational community. The EDM practice transforms raw data from educational organizations into helpful information that can significantly impact educational research and practice (Han, Kamber, and Pei, 2011; Merceron and Yacef, 2010).

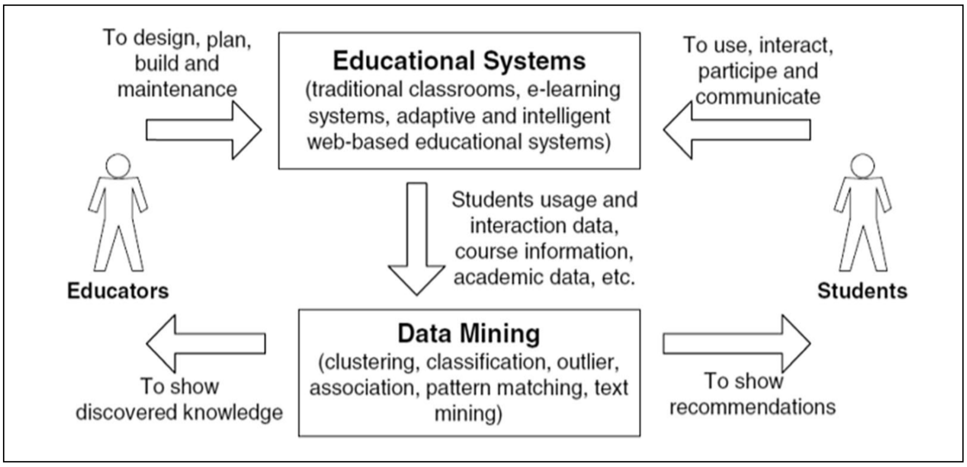

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 1: The cycle of hiring data mining in educational systems (Romero and Ventura 2007)

This chapter of the thesis outlines the research motivation, research questions, and research objectives.

1.1 Research Motivation

The quality of educational institutions relies on providing services that meet students' needs, educational staff, and other participants in the educational environment (Osmanbegović, Suljic, and Suljić, 2012). Therefore, the main goals of the education sector are (i) to improve the learners' experience, (ii) to expand the educators' efficiency, and (iii) to deliver an appropriate, efficient, and effective educational environment (Riffai et al., 2016).

With the continuing growth in higher education enrollments worldwide, students' dropout or failure becomes a significant concern. For example, in the Kingdom of Saudi Arabia (KSA), the average graduation rate for higher education institutions is 65% with students in four-year programs and 35% to 50% among two-year programs (Aljohani, 2016; Riyadh Economic Forum, 2011). In the case of Germany, the number of newly registered students in a degree program has increased by more than 34% in the last decade (Villwock, Appio, and Andreta, 2015). Nevertheless, the graduation rates remain below average as only 35% of the students graduated with a degree (OECD indicators, 2015). Likewise, dropout rates are significantly high in New Zealand, Hungary, Mexico, the UK, Poland, and Norway (Aina et al., 2018). From the previously mentioned examples, it can be concluded that student dropout is a common issue for universities worldwide.

EDM solutions can be used in various approaches to validate and assess an educational scheme, improve teaching and learning process quality, and lay the basis for a better learning experience (Cristóbal Romero, Ventura, and De Bra, 2004). In addition, the EDM applications can assist in studying the students' learning processes by exploring their interactions with the learning environment (Liu, 2014). EDM can help identify possible reasons for students' failure (Algarni, 2016), generate alerts to students in need (Cristóbal Romero et al., 2013), develop strategies to increase the graduation rate, and enable decision-makers to understand students' behavior. This can enhance the learning experience and improve the efficiency of the institution (Dutt, Ismail, and Herawan, 2017).

Educational institutions maintain and store students' data, from academic data to personal records (Dutt, Ismail, and Herawan, 2017). The availability of such data serves as an inspiring motivation for exploring DM abilities in educational settings. With the great opportunities offered by DM to educational societies, the EDM field is expected to continue to expand (Johnson, Adams, and Cummins, 2012). Being part of this expansion serves as a great motivation to carry out this research.

1.2 Research Questions

Following are the primary and sub-research questions addressed for this study:

RQ 1. Is it possible to predict the students ' final academic achievement in bachelor's and master's programs at an early stage?

- Is it likely to have reasonably accurate predictions after the first and second semesters of the students' enrollment?

- Is it likely to have fairly accurate predictions using multi-class classifications (i.e., using 3 or more classes)?

RQ 2. What measurable aspects predict student academic achievement for bachelor's and master's programs in computer science majors?

- To what extent, if any, can student demographics predict students' academic achievement?
- To what extent, if any, can previous knowledge and GPA predict students' academic achievement?
- To what extent can students' behavior after enrollment (e.g., academic load, number of failed courses) predict students' academic achievement?
- To what extent does the distance from students' accommodation to university influence students' academic performance?
- To what extent, if any, do academic language skills influence students' academic performance?

1.3 Research Objectives

The present thesis aims to create new scientific insights and concepts to improve our understanding of EDM by investigating students data of two different programs. The first is the College of Computer and Information Science bachelor's program at PNU (discussed in chapter 4), and the second is the Business Informatics master's program at the University of Mannheim (discussed in chapter 5). Through those two studies, we try to provide sound evidence of the power of EDM. The expected outcomes of this Ph.D. research are as follows:

- Analyze the bachelor's program students' data for the College of Computer and Information Science at Princess Nora University.

- Identify and select appropriate DM algorithms for developing bachelor's degree predictive models for the College of Computer and Information Science.

- Identify and choose proper predictor features that can be employed as the inputs of predictive models for performing predations on the bachelor's students.

- Validate the developed models to identify honorary or at-risk of failure students at the earliest stage possible for the College of Computer and Information Science.

- Analyze the master's program students' data for the Business Informatics master's program at the University of Mannheim.

- Identify and select appropriate DM algorithms for developing predictive models for the Business Informatics master's program.

- Identify and select proper predictor features that can be employed as the inputs of predictive models for performing predations on the master's degree students.

- Find proper techniques to deal with imbalanced datasets in the Business Informatics Master’s program.

1.4 Thesis Structure

This section outlines the contents of each chapter of the thesis.

Chapter 1: Introduction. This chapter gives a brief introduction to the research at hand. It assists in realizing the wide-ranging benefits that can be gained from employing DM solutions in educational settings. This chapter also contains the research motivation, research questions, and the objectives of the study.

Chapter 2: Educational Data Mining Background. This chapter provides a comprehensive background for understanding EDM. It discusses the six phases of the CRISP-DM. It also discusses the different methods and applications of DM in education. Moreover, this chapter briefly explains some of the commonly used algorithms in EDM. It also outlines the challenges that the EDM community faces and gives suggestions for overcoming each challenge.

Chapter 3: Literature Review on Predicting Academic Success. This chapter outlines the measures for determining academic success. It summarizes the pros and cons of the different DM algorithms used in performing academic predictions and the pros and cons of the different DM tools. It contains a rich literature review on predicting academic achievement and related works from 2007 to 2022. The viewed studies are represented in tables based on the country where each study has been performed, the academic degree it covers (i.e., undergraduate or postgraduate), the type of academic prediction, and the type of features that have been used for performing the predictions.

Chapter 4: Predicting the Academic Achievement of Bachelor's Degree Students. This chapter provides an empirical study on predicting the academic achievement of bachelor’s students in the College of Information and Computer science at PNU University in Saudi Arabia. Moreover, this study tries to discover honorary and at-risk of failing students at an early stage of the program. It shows the results of using multi-class predictions. This chapter also shows which features are most significant to performing academic predictions for this college and which features are not. The limitations of this study are also outlined here, and recommendations for future research are provided.

Chapter 5: Predicting the Academic Achievement of Master's Degree Students. Here, an empirical study on predicting the academic achievement of the Business Informatics master’s students at the University of Mannheim in Germany is provided. This chapter shows the results of performing different types of predictions, after the first and second semesters of the studying program. It outlines the benefits of using oversampling techniques to deal with imbalanced datasets. This chapter also shows which features are most significant to performing academic predictions for the Business Informatics master's program. The limitations are also outlined, and recommendations for future research are provided.

Chapter 6: An In-Depth Evaluation on The Impact of Culture on Academic Predictions. This chapter studies the impact of culture on academic success. It explains why it is vital to consider culture as a predictive feature when performing academic predictions, especially if the predictions are to be made in institutes that admit international students.

Chapter 7: Discussions and Conclusion. This chapter concludes the work presented in this thesis by summarizing the results of the two empirical studies, comparing them, and answering each research question. Moreover, it outlines the main limitations both studies have in common and suggests future lines of research.

CHAPTER 2: EDUCATIONAL DATA MINING BACKGROUND

You can have data without information, but you cannot have information without data. - Daniel Keys Moran

After explaining EDM and its profits to the teaching and learning system in chapter 1, this chapter provides a comprehensive background for understanding the phases of EDM, its methods, the commonly used algorithms, and its applications. This chapter also describes the most common challenges associated with EDM and gives suggestions for overcoming those challenges.

2.1 Phases of Performing Educational Data Mining

In order to perform EDM tasks, it is essential to comprehend its deployment phases. Figure 2 demonstrates the six phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM) by (Kurgan and Musilek, 2006).

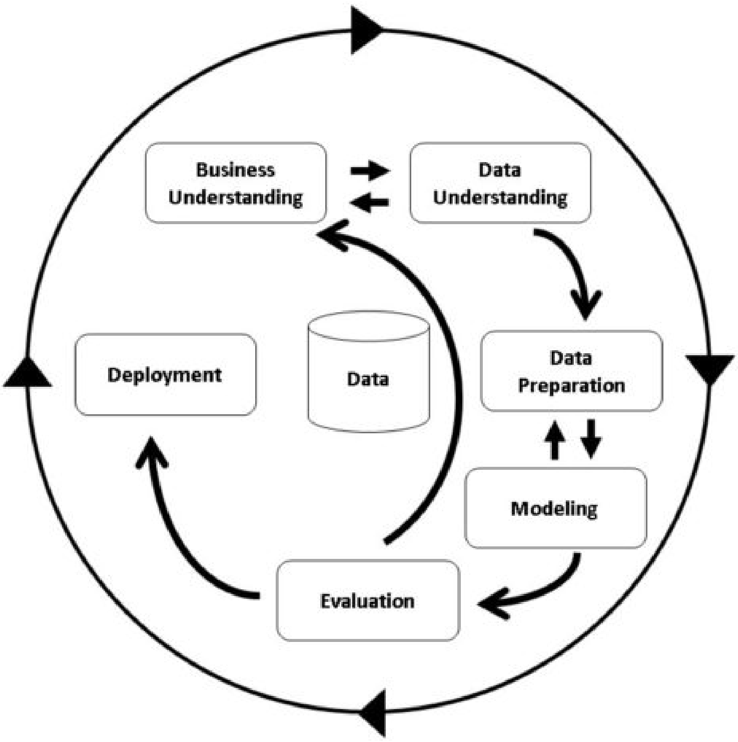


Figure 2: CRISP-DM Process Life Cycle (Kurgan and Musilek, 2006)

The cycle starts with business understanding, in which the key stakeholders in the study are recognized, and any valuable information is understood. At this point of the cycle, the objectives of the EDM study are defined. When performing academic predictions, the overall business objectives are usually related to assisting higher education institutes' management in addressing students' challenges and helping them achieve their long-term success. The detailed

DM objectives to which the two empirical studies relate are discussed indivisualy in Chapter 4 and chapter 5.

In the second phase, researchers and decision-makers try to understand the data at hand, and data is verified for completeness, redundancy, and missing information. At this stage, data significance in terms of achieving the aimed goal is approved. As illustrated through the arrows between the first and the second phase in Figure 2, the objectives might need to be redefined. Cases that could cause such a need are insufficient data or various missing or unreliable values in the dataset.

Once rational objectives are formed, and the required data resources are accessible to meet the analysis goals, the third phase of the CRISP-DM (Data Preparation) starts. In this phase, the data is prepared by cleaning and selecting the relevant feature subset. The goal of this step is to attain a dataset that matches selected methods of data mining. This phase also includes detecting and removing outliers, which could risk the analysis results. Moreover, features are transformed as required to fit the DM algorithm. For instance, some algorithms, such as Neural Networks, need a particular data format since they can only be performed with numerically coded features with a value range between 0 and 1.

In the fourth phase, the DM methods (discussed in section 2.2) and algorithms (discussed in section 2.4) are selected to be used for knowledge generation and testing. In order to find the most robust results, a number of models are generated by using various DM algorithms that are appropriate for resolving the problem. If it is impossible to create a model with reasonable performance outcomes, it may be necessary to consider whether preprocessing of the data can enhance the model outcomes or if more data resources are accessible. Data preprocessing could include resampling techniques, for instance.

A model or models that are most likely to solve the predefined objectives are identified in the fifth phase of the CRISP-DM (Evaluation). Thus, in addition to evaluating the performance measures, it is also necessary to measure the extent to which the tasks defined in the Business Understanding phase have been achieved. In the event of not being able to create a model with good performance metrics and insights that are conducive to achieving the predefined task, the CRISP-DM must start over. Furthermore, it is possible that the project plan will need to be modified or that the results will reveal additional DM analysis opportunities that will need to be addressed (Cleve and Lämmel, 2020). Having produced a model or models that perform well and assist in attaining the pre-defined goals, the process continues with the final step.

In the sixth phase (Deployment), the results are formulated and submitted to the decision-makers. If the pre-defined tasks are met, and the goals are reached, the models can be integrated into the institutional processes. In this phase, Students, instructors, and academic institutions can start benefiting from the built DM models.

2.2 Educational Data Mining Methods

In the educational field, many DM methods can be used. However, while some are used often, others are rarely used. Baker, Isotani, and Carvalho (2011) present an EDM taxonomy that is divided into five sub-areas as follows:

2.2.1 Relationship mining

Relationship mining detects relationships between features and encodes them in rules for future use. Detecting a relationship could be by trying to discover which features are primarily linked with a single feature of sprcial interest or may embrace the practice of trying to determine which relationships between any two features are most robust (Peterson and Baker, 2010). Relationship mining is possibly the most used method in EDM (Daniel, 2014). There are four different approaches for performing relationship mining as follows: (1) Association rule mining, which uses if-then rules, e.g., if a {learner's final grade is less than 3.00, and the student's attendance is low} → {the student will not complete the program}. (2) Correlation mining attempts to discover (positive or negative) linear correlations between features. (3) Sequential pattern mining involves discovering statistically significant patterns between data instances where the values are distributed sequentially, e.g., mining sequential patterns of students' logs in an online learning platform (Poon et al., 2017). (4) Causal DM finds causal relationships in a dataset, e.g., what features of students' behavior cause dropout. All these types of relationships can then be used to form suggestions for content that is expected to motivate students or even assist in modifying the teaching approaches (Merceron and Yacef, 2010).

2.2.2 Prediction

Prediction is used to infer a target feature from other features to forecast a future event. It is performed using supervised algorithms, which aim to infer a function from labeled training data. To perform the prediction, one can use methods of (1) Classification, which uses previous knowledge to form a learning model and then employs that model as a binary or categorical variable for the new dataset, (2) Regression, which predicts continuous variables, or (3) Density estimation, which is built on various kernel functions, e.g., Gaussian functions. In the educational field, prediction is frequently used to forecast an event based on available features such as gender, age, academic performance, students' behavior, and class attendance.

2.2.3 Clustering

Clustering is another DM method that is used to uncover remarkable patterns in data and segment each data point into a particular group that was not previously defined. Ideally, data points in a single group should have features in common, while data points in different groups should have different features. It is performed using unsupervised algorithms, e.g., K-means clustering and Gaussian Mixture Model algorithm. This means that discovering inherent patterns in the data is made automatically without training the data. For example, clustering can group similar course materials or learners based on their knowledge and communication links in the educational field.

2.2.4 Distillation of data for human judgment

Data distillation for human judgment is used to improve student models (Baker, 2010). It aims to illustrate data to support researchers to quickly recognize structures in the data (Daniel, 2017) and highlight helpful information that helps with decision-making, e.g., help instructors quickly find features of student learning activities or identify patterns in students' behavior (Mazza and Milani, 2004). This is performed by presenting data using summaries, visualizations, and interactive interfaces. Recognizing learning patterns and students' variances from visualizations is primary for investigating educational datasets (Baker, 2010). In addition, the distillation of data for human judgment can have a significant role in labeling data for usage in the future employment of prediction models. The primary usage of identification with distilled data is inferences from students' learning curves.

2.2.5 Discovery with models

Discovery with models uses a preexisting validated model (such as clustering or prediction) as a component to be applied to a different dataset with the aim of further investigation (Algarni, 2016; Daniel, 2017). It is perhaps the most rarely used DM method (Algarni, 2016). In the educational field, discovery with models can be adapted to determine relationships between learners' behavior and characteristics, explore research questions through different settings, and integrate psychometric modeling frameworks into machine-learning prototypes (Merceron and Yacef, 2010). For example, determining which educational material subcategories deliver the most benefits to students (Beck and Mostow, 2008) and how a lesson design could affect students' understanding (Jeong and Biswas, 2008).

2.3 Educational Data Mining Applications

Various EDM applications are described by Romero and Ventura (2013), e.g., predicting students' academic achievement, scientific inquiry, providing instructors with feedback, providing students with recommendations, creating alerts, and much more. However, according to Costa et al. (2012), applications of EDM can be classified into four critical categories as follows:

2.3.1 Student modeling

Student modeling is a critical theme in educational software research (Peterson and Baker, 2010) and one of the emerging research disciplines in this field (Baker and Yacef, 2009). Modeling in EDM categorizes students based on their characteristics and individual differences. This categorization enables the software to estimate each student's present knowledge state and respond to each one differently based on his/her needs and requirements. Therefore, student models are considered primary elements of adaptive intelligent educational systems (Tacoma et al., 2018). Student modeling could be performed using prediction, clustering, or methods to distill data for the judgment of humans (Baker, 2010). For instance, clustering can be performed to create student models for numerous types of educational platforms.

2.3.2 Domain modeling and knowledge structures

Another crucial area of EDM applications is determining or enhancing the domain's knowledge structure models. The phrase "knowledge" includes understanding theories and facts about a particular subject, its relations, and mechanisms for solving problems in that subject (Gaevic, Djuric, and Devedic, 2006). The knowledge components of domain models can be different in numerous ways, thus fitting the model design to explicit features of the domain and the educational setting (Tacoma et al., 2018). Knowledge components can signify elements of (i) procedural knowledge ('how'- knowledge), which defines procedures in the domain, or (ii) declarative knowledge ('what'-knowledge), which defines essential theories and facts (Brusilovsky and Millán, 2007). From the perspective of students: Declarative knowledge refers to students' knowing or understanding (e.g., the student knows that python is a programming language), and Procedural knowledge is that the student can perform something (e.g., a student can build a program using python). Various methods have been designed to detect accurate domain models directly from educational data. These methods typically integrate psychometric modeling frameworks with advanced space-searching algorithms and pose prediction challenges for model discovery (Costa et al., 2012). A designer of a model could decide to breakdown the knowledge in the domain into minimal components, thus improving the possible precision of the model or express knowledge components at the level of broader categories and topics, thus facilitating more direct content modeling and linking learning tasks to knowledge elements (Tacoma et al., 2018) Concept models of the learning materials and models clarify the correlations of knowledge in a domain (Barnes, 2005). Trying to predict whether individual activities will be accurate or not by employing various domain models is typical for producing such models (Peterson & Baker, 2010).

2.3.3 Pedagogical support

Studying and improving the pedagogical support delivered by learning software is critical in EDM applications. Pedagogical support means providing services to students to support their instructional programs. Modern education offers various educational assistance to students, such as helping students with the needed learning material, online tutorials, and encouraging them with different learning activities. To assist students in their educational experience, one can use any of the previously discussed data mining methods or a blend of numerous methods; this depends on the task to be performed. Identifying at-risk students to perform early intervention or provide additional and more tailored support is a typical example of pedagogical support.

2.3.4 Scientific research

The main goal of scientific research about learning and learners is to establish experiential evidence. Such evidence approves or articulates scientific concepts, frameworks, and educational phenomenas. Another primary goal of scientific research in EDM is to formulate a new hypothesis (Calvet Liñán and Juan Pérez, 2015). This can assist in determining the vital core components of learning and, as a result, design better learning systems. Scientific research is applied to answer questions regarding student models, domain models, or pedagogical support. Relationship mining promotes discovery for scientific researchers (Baker, 2010). Studies exploring whether state or trait attributes are better predictors of how much a student would game the system (Baker, 2007) is one of the approaches of EDM research. Experiencing the factors that primarily affect students' academic achievement is another common approach of EDM studies.

For further understanding the methods and applications of DM that have been explained in sections 2.2 and 2.3, Appendix A gives more examples of DM studies performed in educational settings.

2.4 Data Mining Algorithms

There are two types of DM algorithms: supervised and unsupervised algorithms. By using unsupervised algorithms, we can reveal hidden patterns in unlabeled data, which will allow us to find patterns within a dataset, but there are no output variables to be predicted. In contrast, supervised algorithms (also referred to as predictive or directive algorithms) predict the value of output variables according to the input variables. In this thesis, we discuss techniques for performing prediction tasks using supervised DM algorithms. The model is developed from training data in which inputs and outputs have previously been labeled. By generalizing the relationship between inputs and outputs, the model can be used to predict other datasets in which only inputs are known (Witten et al., 2017). In EDM, a variety of DM algorithms are used. In general, the literature review suggests that no single DM method is effective in all situations. This section presents the algorithms which have been used in this study.

2.4.1 Decision Trees (DT)

A classification algorithm in which each core node presents a "test" on a feature, each branch represents the test's outcome, and each leaf node corresponds to a class label (decision taken after computing all features). There are many types of DT. In the research presented, the following DT algorithms are used:

> ***SimpleCart*** uses a learning sample, a historical dataset with preassigned classes for all observations for constructing decision trees. It is a learning technique that offers the results as either classification or regression trees, based on the categorical or numeric data set. It uses cross-validation or a large independent test sample to select the best tree from the sequence of trees considered in the pruning process. During the

implementation phase of CART, the dataset is split into two subgroups that are the most different concerning the outcome. This process is continued on each subgroup up to some minimum subgroup size is attained (Kalmegh, 2015).

LadTree is a classification method based on learning a logical expression. Since LAD is a binary classifier, it can differentiate between positive and negative samples (Amudha et al., 2011). For a dataset processed by LAD, a large set of patterns is produced, and a subset of them is chosen to satisfy the above assumption such that each pattern in the model fulfills specific conditions in terms of prevalence and homogeneity (Buhmann, 2003).

C45 (J48 in Weka), which was developed by Quinlan in 1992, is an expansion of the Interactive Dichotomize 3 (ID3) algorithm. This algorithm is used to create a tree-shaped structure that symbolizes sets of decisions. At each tree node, C4.5 selects the data attribute that most efficiently splits its set of samples into subsets enriched in one class or the other. The splitting measure is the normalized information gain. The attribute with the highest normalized information gain is chosen to make the decision.

2.4.2 Naive Bayes (NB)

A supervised classifier based on Thomas Bayes's work, the Naive Bayes, simplifies learning by supposing that features are independent of classes. Based on the hypothesis that the data belong to a class, Naive Bayes calculates the probability that the hypothesis is correct. Thus, only one scan is required of the data. The probability of a hypothesis being true can be incrementally increased or decreased with every training example. Thus, Naïve Bayes is a perfect fit for domains that contain uncertainty (Nielsen & Jensen, 2007). It is based on the Bayes theorem, which states that if event B has happened, then we can find the probability of event A and represented as follows: $P(A|B) = (P(B|A)* P(A)) /P(B)$.

2.4.3 K-nearest neighbor (KNN)

A supervised DM method for estimating the likelihood that a data point will become part of one group using the distance between a classified instance and its closest training examples (Clark, 2013). The classification decision is made based on a majority vote among k empirically observed instances that are most similar to the instance under consideration.

2.4.4 Support Vector Machine (SVM)

A supervised DM method that seeks to find the hyperplane that best separates the data points in high dimensional space by maximizing the margin (Clark, 2013). Maximizing the margin distance enhances the confidence with which future data points can be classified. In order to improve data reparability, SVM can transform the data into a higher dimensional space through several kernel functions.

2.4.5 Artificial Neural Networks (ANN)

A collection of algorithms designed to identify underlying relationships in a set of data by mimicking the information-processing processes of the human brain. In general, it consists of two phases. As a first step, the network is trained on paired data to locate the input-output mapping. Afterward, the weights associated with the connections between neurons are fixed, and the network is used to establish the classifications of a new data set. To encode this dataset, it must adhere to a standardized format with values ranging from 0 to 1(Larose and Larose, 2015). The majority of datasets must therefore be preprocessed before they can be analyzed.

2.4.6 Logistic Regression (LR)

A statistical model that is typically applied to a binary dependent features. A logistic model is one in which the log odds of the probability of an event are derived from a linear mixture of independent or predictor variables. In the presented thesis, the Binary logistic regression is used in the cases where the dependent feature has only two possible outcomes, and the Multinomial logistic regression is used where the dependent feature has three.

2.4.7 Ensemble learning

In ensemble learning, the same or several types of independently trained models are combined to produce a prediction using a meta-algorithm (Karalar, Kapucu, and Gürüler, 2021), i.e.; it is a family of algorithms that seek to create a “strong” classifier based on a group of “weak” classifiers (Zhao et al., 2020). As a result, achieve better predictive tasks than could be acquired from any of the basic learning algorithms alone. Depending on the range of learning algorithms incorporated into the model, ensemble learning can either be homogeneous or heterogeneous. Heterogeneous models apply the same training data to several learning algorithms or to the same algorithms with various parameter settings. Homogeneous models divide training data into subsets and apply the same learning algorithm to each subset by the number of subsets (Wang et al., 2018). In this study, we use two of the commonly used homogeneous methods and are described as follows:

Bagging algorithms

Ensemble learning is also known as bootstrap aggregation and is commonly used for reducing variance in data sets. The bagging technique selects a random sample of data from a training set with replacement, meaning that the individual data points may be chosen more than once. One of the most powerful bagging algorithms is Random Forest (RF), which has been used in this research.

Random Forest, As the name implies, it is a tree-based supervised classifier that acts as an ensemble according to a group of random variables (Cutler and Stevens, 2012). In the training stage, RF applies the general technique referred to as “bagging” to individual trees in the ensemble (Caie, Dimitriou, and Arandjelović, 2021). With bagging, a random sample is chosen from the training set and trees are fit to these samples without pruning. It uses voting mechanisms from multiple decision trees to improve the shortcomings of a single DT and get more accurate predictions. Trees in the random forest provide class predictions, and the class with the greatest votes is the model's prediction.

Boosting algorithms

The Conceptual basis of Boosting is to merge simple “rules” to create an ensemble such that the operation of the single ensemble member is enhanced, i.e., “boosted” (Meir and Rätsch, 2003). In this research, three boosting algorithms have been used as follows:

Gradient Boosting (GBM) is a ML algorithm that gives a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

Extreme Gradient Boosting (XGboost) is an implementation of a stochastic gradient boosting machine. In this research, we use the decision tree learner as implemented in Xgboost. Each decision tree is trained from a randomly sampled set of rows and columns. Each tree is grown to a maximum depth using a leaf growing algorithm that estimates whether an additional leaf will produce a better or worse tree (Chen and Guestrin, 2016). Xgboost models have hyperparameters that define how the model is to be constructed prior to the model being trained. These hyperparameters govern two classes of model design: (i) how the boosting functions and (ii) how the trees are grown and structured.

Light Gradient Boosting (LightGBM) has many of XGBoost's advantages, such as sparse optimization, parallel training, multiple loss functions, and early stopping. Yet, a significant difference between them is in the tree structure. LightGBM does not develop a tree level-wise (row by row). Instead, it develops trees leaf-wise (Joseph, 2020). Moreover, LightGBM does not use the widely used sorted-based DT learning algorithm, which searches for the best split point on sorted feature values, as XGBoost does. Instead, LightGBM executes a highly optimized histogram-based DT learning algorithm, which produces many benefits in both proficiency and memory consumption. The LightGBM algorithm utilizes two novel procedures called Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which allow the algorithm to operate quicker while preserving a high level of accuracy (Ke et al., 2017).

2.5 Challenges in Educational Data Mining

Despite the relatively extensive literature on EDM, it is still a recent field of research. The development of DM in the area of education was relatively late compared to other areas (Silva and Fonseca, 2017). As a consequence, some challenges still need to be addressed. This section identifies the main challenges associated with employing EDM solutions and gives suggestions to overcome those challenges.

2.5.1 Lack of knowledge

A significant obstacle to the employment of EDM solutions is the lack of knowledge (Wolf et al., 2014) regarding (i) the employment of the required tools, (ii) understanding the outputs, (iii) figuring the appropriate conclusions, and (iv) determining which actions to be taken. In addition, educators might not know how to conduct EDM solutions in their practice and might also not understand their importance (Selwyn, 2011). Due to that, it is challenging to find skilled specialists with the essential knowledge of how to fetch, analyze and produce effective use of data and recognize the constructs of the learners' cognition to promote in technology-enhanced education settings (Daniel, 2014).

It is vital to increase knowledge in the community to enhance acceptance and grow a data-driven society in educational surroundings (Cristobal Romero & Ventura, 2013). It is also crucial to improve technical training in educational societies. Researchers should further assist in overcoming this issue by publishing their results in international conferences and journals. They should also evaluate their proposals by collaborating with instructors and students in different educational institutes and universities.

2.5.2 Security issues

Security is one of the foremost challenges when dealing with students' data. Since 2014, according to (Bissell, Lasalle, and Cin, 2019), cybercrime has increased by 64%, and since 2018, it has increased by 11%. These statistics show the seriousness of the issue of data security. Cyber-attacks are possibly a more significant threat for educational institutes with fewer resources to create and manage appropriate data security methods (Greller & Drachsler, 2012). The tremendous amount of information belonging to educational institutes can cause a tragedy if misused on personal, industrial, governmental, or country levels (Bamiah, Brohi, and Rad, 2018).

Unfortunately, many education institutes lack adequate guidelines for managing intellectual property and who can have access to it (Daniel, 2014). Moreover, traditional security techniques (e.g., strong password policies, disaster recovery plans, firewalls, antivirus software) are insufficient to secure data. Institutions with massive databases need to take a holistic security vision (Tankard, 2012). This means that data must be secured with all the necessary methods and technologies to uphold the integrity and validity of users' data. It is vital to consider identifying the various data sources, the origin and creators of data, and who can access it (Moura and Serrão, 2015). It is also essential to correctly identify significant data and associate it with the institution's information security plan to enforce access control and data handling procedures (Kindervag et al., 2012). Moreover, the security of the IT infrastructure itself must be taken into account, such as placing security controls at the edge of the network (Moura and Serrão, 2015). It is highly recommended to execute security audits frequently on the existing dataset to identify the attempted threats to the security system (Leonard, 2015). Also, in order to deliver security at the source of the data, a variety of security structures should be placed closer to the data and its sources (Kindervag et al., 2012).

According to IBM (Madia, 2012), following the three-step Forrester's data-centric security framework is crucial when applying or enhancing security measures. The framework requires (i) defining, (ii) dissecting, and (iii) defending the data environment from security attacks. First, the definition includes detecting and classifying data as either structured or unstructured. Second, dissection consists of implementing intelligence and analytics tools to the data at hand. Finally, defending the data includes implementing data access, threat identification, and termination mechanisms.

2.5.3 Personal privacy

Personal privacy is considered one of the significant barriers in EDM (Greller & Drachsler, 2012). Information privacy is the capacity of a person or more to prevent their data from being visible to others than to those who provide the information (Jain, Gyanchandani, and Khare, 2016). Based on Warren and Brandeis (1890), "The right to privacy", is the "right to be let alone". This perception was additionally developed by Westin (1968), who clarified that recent technologies had changed the stability of power between privacy and social techniques.

Educational institutes necessitate transparency that exposes the identity of the learner to track their learning. However, a malicious insider can track students (e.g., education records, performance, as well as how, when, and students' log data). Additionally, data mining tools and methods to predict learners' future outcomes may violate their privacy if no privacy regulations have been applied.

Indeed, student privacy and data ownership differ from one country to another. For example, while Germany has a highly comprehensive and firm legislative platform regarding data protection, the USA lacks comprehensive data protection legislation (Hoel and Chen, 2018). In Germany and other European countries, the European Union endorsed various principles in the Data Protection Directive under which learners' privacy matters are primarily included. These rules necessitate unambiguous consent of individuals prior to gathering or studying personal data and prohibit gathering sensitive information with few exceptions (Yee, 2007). The restrictions required to comply with this law significantly impact EDM adoption or any data-based solution.

To avoid privacy issues when employing EDM solutions, it is vital to understand and follow the rules and guidelines regarding data privacy where the study is conducted. It is also essential that institutional policies are transparent. Additionally, students should be informed how their data will be used and how their findings will be acted upon by the educational institution.

2.5.4 Ethics

One of the significant barriers in EDM is related to ethics (Greller & Drachsler, 2012). With the rise of data velocity, volume, variety, resolution, flexibility, scalability, and indexical properties (Kitchin, 2014), so does the degree of ethical issues and challenges (Prinsloo and Slade, 2013). Furthermore, realizing what is considered ethical behavior varies and changes actively over time and in nations (Greller and Drachsler, 2012). This makes ethical applications even more challenging.

Although various theoretical and conceptual types of research are available regarding the ethical collection, analysis, and use of students' data, there are very few examples of how institutes react to ethical challenges and issues (Prinsloo and Slade, 2013, 2015). This points to the need for further reporting regarding institutes' responses to such issues.

To avoid ethical issues when employing EDM applications, it is vital to consider ethics at all stages of the EDM process, from data collection to the interpretation of outputs and, finally, the decision-making. It is essential to better understand transparency regarding the aim for which data is being gathered and how to deal with sensitive data. Also, Petersen (2012) discusses the importance of hiding the actual identity of the individuals by anonymizing the data before it is made available. It is also essential to avoid declarations that might lead to discriminatory behaviors, e.g., religious beliefs.

2.5.5 Quality of data

As data is a critical resource in any organization, its quality is essential in identifying performance issues (Batini, Di Milano, and Maurino, 2009; Tee et al., 2007). The use of accurate and high-quality data is essential for producing value from big datasets. Research on data quality began in the 1990s (Cai & Zhu, 2015), resulting in many scientists proposing different explanations of data quality in different fields and periods. The Total Data Quality Management group of MIT University defined "data quality" as "fitness for use" (Wang and Strong, 1996).

Based on Cai & Zhu (2015), the data quality standard comprises five measurements: (1) Availability: The concept of data availability is a core issue in determining the quality of EDM solutions. The restrictions on how data is gathered, how it is stored, and how it is used often make accessing and using educational data more challenging. The increasing constraints could minimize the funding available to EDM research. (2) Usability: Improving data usability (or trustworthiness) is one of the critical challenges of EDM applications. Data usability is the ability to derive useful information from data. While many educational institutes have a wealth of data, this data is usually disorganized or cannot be evaluated effectively to produce useful information that supports target decision-making. (3) Reliability: Data reliability is another primary challenge when implying EDM solutions. It refers to whether we can trust the data. Based on Cai & Zhu (2015), reliability consists of five elements: (i) accuracy, which refers to whether an entity's data values are correct. (ii) consistency, which refers to the usability of data Contents. (iii) completeness represents the degree to which all required data are available in the dataset (iv) adequacy, which is the state of being sufficient for the purpose concerned. (v) auditability elements examine key metrics in order to draw conclusions about a data set's properties. Inaccuracies or missing information can arise in collecting, maintaining, processing, and reporting data. Based on Rothman (2007), low reliability creates several problems as it generates issues when trying to find relationships between variables caused by the high error rates and creates problems for establishing the validity of the data measure. (4) Relevance: One of the main aspects of data quality is the relevance of data. It can be defined as the consistency between the data content and the area of the user's interest. Based on Cai & Zhu (2015), the requirements are divided into two levels: (i) the extent to which accessed data is used by users, and (ii) how closely the data created is aligned with the requirements of users, including the definitions of indicators, elements, classifications, etc. (5) Presentation quality: Presentation quality means that the data classification, description, and coding content satisfy the specification and are simple to comprehend. Based on Cai and Zhu (2015), the extents of presentation quality are (1) readability; which is the capability of data content to be properly described based on known or well-defined phrases, features, abbreviations, units, or other data, and (2) structure; which is the level of complexity in converting unstructured or semi-structured data to structured data using technical tools.

There are two types of approaches for enhancing data quality, namely (i) data-driven, which is an approach for enhancing the quality of data by modifying the data value directly, and (ii) process-driven redesigns the process which is produced or edited data to increase its quality. Each strategy employs various techniques (Batini, Di Milano, and Maurino, 2009). However, both methods aim to enhance data quality (Sidi et al., 2012). Examples of data-driven techniques are acquiring new datasets, standardization, error detection and alteration, record linkage, data integration, source trustworthiness, and cost optimization (Batini, Di Milano, and Maurino, 2009). On the other hand, a process-driven scheme contains two fundamental practices: process control and process redesign. In the first practice, data is checked and managed during the manufacturing process, while in the second, the reasons for low quality will be removed, and new methods will be generated to increase quality. Employing an action that can control the data format prior to storage is another fact in the redesign process (Batini, Di Milano, and Maurino, 2009)

2.6 Summary of the Chapter

In this chapter, the phases of performing data mining in educational settings have been explained. We have discussed and given examples of the different DM methods that are used in educational settings, which are (1) Relationship mining, (2) prediction, (3) clustering, (4) distillation of data for human judgment, and (5) discovery with models. Moreover, the various aims behind performing DM approaches have been clarified, which are (1) student modeling, (2) domain modeling and knowledge structures, (3) pedagogical support, and (4) scientific research. We have also briefly explained the most common DM algorithms.

We also analyzed the challenges associated with implementing DM in educational environments. Based on our research, there are five main challenges associated with implementing EDM applications. The first challenge is related to the lack of knowledge regarding the importance of EDM applications and how to apply them. The second major challenge is regarding security and the severity of cyber-attacks. The third challenge is related to personal privacy and ownership of the data. The fourth challenge is regarding ethical issues. Finally, the fifth challenge concerns data quality and its five dimensions: availability, usability, reliability, relevance, and presentation quality. For every challenge, we outlined some essential measures to avoid or deal with it. However, as EDM is still a developing research field, we expect that its additional development will result in a sufficient understanding of its challenges and how to overcome them.

CHAPTER 3: LITERATURE REVIEW ON PREDICTING STUDENTS' ACADEMIC ACHIEVEMENT1

"An investment in knowledge pays the best interest." – Benjamin Franklin

Due to the developments in the automatic analysis of educational data, various efforts to enhance educational success have been carried out (Chatti et al., 2012). However, the EDM literature is vastly expanding and requires to be updated regularly to consider new studies. The more recent the literature is, the more we can expand our understanding of EDM, its role in the academic community, and the most recent trends regarding the used methods, applications, and algorithms that aim to increase students' success and reduce failure or dropout.

Romero and Ventura (2007) reviewed published studies on EDM from 1995 to 2005. In their review, two forms of educational systems were examined (traditional classes and distance education). The authors also discussed how DM has been applied to education systems. Another EDM survey was performed by Peña-Ayala (2014) that analyzed EDM works published between 2010 and 2013. In the survey, the weaknesses, strengths, risks, and opportunities of EDM are discussed. A systematic literature review of clustering methods was presented by Dutt, Ismail, and Herawan (2017). As part of their research, they examined the applicability and usability of clustering methods in the context of EDM. In their study, they observed that the key advantage of clustering algorithms was that they provided a relatively explicit schema of students' learning styles by clustering several attributes (e.g., how long it took to carry out tasks, how well students performed in class, and how motivated they were). Kumar, Singh, and Handa (2017) carried out a literature review on students' achievement prediction from 2007 to 2016. The authors reported the accuracy and effectiveness of the DM methods used. In spite of this, no information was provided regarding the DM tools used. According to their research, GPA and internal marks are significant indicators of academic performance.

The focus of this chapter is on prediction tasks in EDM based on different and more recent published work (2007-2022). In this chapter, we start by outlining the measures of determining academic achievement. Then, we analyze the frequently used features, tools, and algorithms in predicting students' academic performance. Furthermore, we highlight the types of features that previous scholars found to be meaningful for academic prediction, as well as the advantages and shortcomings of the DM algorithms and tools.

3.1 Measures of Determining Academic Success

Since higher education serves as a fundamental role in the progress of a society (Pinheiro et al., 2015), improving student success is a long-term target for higher education institutions. In order to maximize students' success rate, it is essential to identify and understand academic

[1] This chapter is a modified version of an article published in Technology, Knowledge, and Learning and has been reproduced with the permission from Springer Nature.

success. Unfortunately, the definition of “academic success” is complex and wide-ranging, depending on the type of institution, its nature, and its mission; thus, it is commonly misrepresented within educational research. However, York, Gibson, and Rankin (2015) provide a theoretically based definition of academic success that consists of six key elements. Those elements are (1) academic achievement, primarily measured by course grades and grade point averages (GPA), (2) satisfaction, which is mainly determined by course evaluations or institutional surveys, (3) persistence, which can be measured by the rate of retention between particular years of college and the rate of degree attainment, (4) acquisition of skills and competencies, which can be assessed through assignments and course evaluations, (5) attainment of learning objectives, which can also be assessed through assignments and course evaluations, and finally, (6) career success, based on factors such as job attainment rates, promotion histories, career satisfaction, and professional achievement. It is important to note that all the academic predictions that have been viewed in this chapter try to predict the first component of academic success which is academic achievement.

Another vital requirement for maximizing students’ success is identifying the features that affect academic performance. Knowing students’ success features may help in reaching the highest level of quality education (Yassein et al., 2017). It can potentially assist in delivering a clear and robust description of the types of knowledge and behavior that are linked with adequate performance. Such awareness can be obtained by using data mining methods over educational records.

3.2 Comprehensive Review of Academic Achievement Prediction Literature

The purpose of this section is to review the different types of predictions performed in higher education institutions. A summary of thirty-two studies concerning the prediction of academic accomplishments in higher education is provided. The studies were performed in various countries around the world between 2007 and 2022. In addition, we present some of the most significant findings from the literature review by reviewing some of the primary outcomes of various previous studies.

3.2.1 Review of the features used in predicting students’ academic achievement

Researchers have been able to expand student modeling towards determining what factors predict student failure in higher education courses or a higher education degree (Dekker, Pechenizkiy, and Vleeshouwers, 2009). Based on the literature reviewed, Figure 3 illustrates the most commonly used features for predicting academic outcomes in higher education, despite their impact. As can be seen, Gender is used in most of the viewed studies (62.5%) followed by GPA, which has been used in more than half of the studies (53.13 %). Their frequencies are followed by those of course grades (53.13 %), age (46.88 %), and language proficiency (31.25%). Other features including nationality, employment status, income, marital status, and attendance are included in fewer than 30% of publications. The following section provides a more detailed analysis of these features.

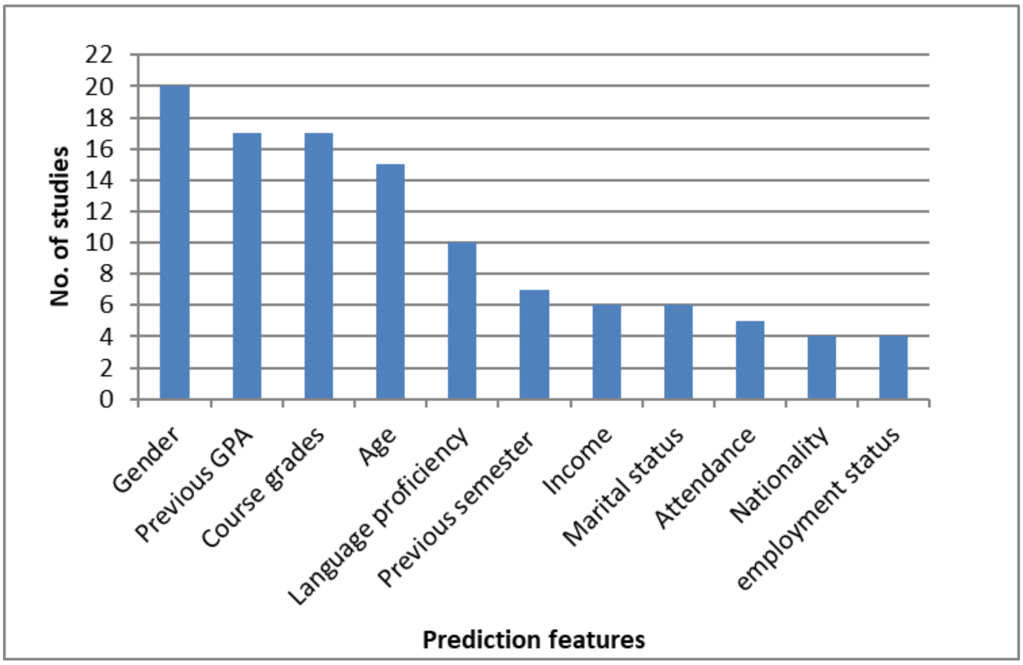


Figure 3: Mostly used features in predicting students' academic outcomes

Based on the viewed literature, the features used to predict academic achievement can be grouped into three classifications. They are (i) pre-enrollment features, (ii) demographics, and (iii) post-enrollment features.

Pre-enrollment features

Pre-enrollment features are the features that are associated to students' achievements before their enrollment. The most used pre-enrollment features by researchers to perform academic predictions are as follows:

GPA: Concerning using students' previous qualifications for predicting academic achievement, the most undertaken feature is GPA (Abu Saa, 2016; Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Osmanbegović, Suljic, and Suljić, 2012; S. Huang and Fang, 2013; Pal and Pal, 2013; Kabakchieva, 2013; Kovačić, 2010; Ibrahim and Rusli, 2007; Zimmermann et al., 2011; Rotem, Yair, and Shustak, 2020; Raj and Manivannan, 2020; Abu Zohair, 2019; Zhao et al., 2020). As compared to demographics and pre-enrollment factors, Ibrahim and Rusli (2007) found that GPA is most significant in predicting students' success (with an 87% correlation).

Academic Language Skills: Also, academic language skills have been considered often as a feature to predict student achievement (Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Abu Saa, 2016; Badr et al., 2016; Asif et al., 2017; Rotem, Yair, and Shustak, 2020; Raj and Manivannan, 2020; Zhao et al., 2020). The language used in textbooks, in classrooms, and on tests and exams is referred to as academic language. There has been some research (Arsad, Buniyamin, and Manan, 2014) that indicates that academic language skills do not influence a student's accomplishment in knowledge courses or non-linguistic courses. However, other research (Wait and Gressel, 2009) has demonstrated that academic language skills do impact student success. According to their study, the ability of students to acquire new knowledge through listening and reading is greatly enhanced by proficiency in the teaching language. In addition, students with language skills are better equipped to express their thoughts through oral discussions and examinations. In terms of assessing the significance of predicting academic achievement based on language proficiency, most of the research did not report any significance. Nevertheless, Badr et al. (2016) reported that their prediction model was more accurate (67.33%) when it was based solely on language skills and no other features.

Other pre-enrollment features: some other pre-enrollment features that have not been used often by researchers include scores earned in GREs (Zhao et al., 2020), scores earned in SAT exams (Aulck et al., 2016), enrollment examination (Asif et al., 2017; Alemu Yehuala, 2015; Osmanbegović, Suljic, and Suljić 2012), and previous academic institute (Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Kabakchieva, 2013; Daud et al., 2017).

To sum up, we believe that it is important to forecast students' academic success using pre-enrollment information. Unfortunately, there is little research assessing the effects of the individual predictor features, and the few studies that have done so have found that academic language proficiency and prior GPA have a beneficial influence on academic predictions. Although pre-enrollment features can help in making academic predictions at an early stage, they are difficult to collect.

Demographical features

As the name suggests, Demographical features represent the characteristics of the students. The following are the demographic characteristics that researchers most frequently use to make academic predictions:

Gender: Figure 3 shows that, when compared to other demographic factors, gender has been the factor most frequently employed to predict academic achievement. According to the literature targeted by this study, some researchers found that students of different genders did not perform significantly better (Goni et al., 2015), while others found that, depending on the subject, male students (Chang, 2008) or female students (Simsek and Balaban, 2010) performed better. Unfortunately, the prediction task was not undertaken in any of the aforementioned experiments. Regarding the research publications that used gender for predicting academic outcomes, we identified 20 publications. However, only six studies (Kovačić, 2010; Osmanbegović, Suljic, and Suljić, 2012; Rotem, Yair, and Shustak, 2020; Zhao et al., 2020; Nasrullah et al., 2021; Karalar, Kapucu, and Gürüler, 2021) reported its impact on the overall prediction. They all came to the conclusion that gender has no influence on the prediction.

Age: In terms of predicting academic success, age is the second most common demographic factor. Past research has found a positive relationship between age and performance (Sturman 2003; Watkins and Hattie, 1985), which may explain its prevalence. Previous studies explained the positive correlation between age and academic achievement by suggesting that older students are more highly motivated, experienced, and possess efficient study habits. Unfortunately, most studies that we target did not report this feature's individual impact.

Exceptions include the studies of Kovačić, 2010; Rotem, Yair, and Shustak, 2020; Nasrullah et al. 2021; Zhao et al., 2020. They all found that age does not affect academic success significantly.

Marital Status: There is also substantial literature on the relationship between Status of marriage and academic achievement. Therefore,18.75% of the studies we surveyed used Marital status. Yess (2009) explored the impact of marital status on the scholastic achievement of 240 Community College students in the USA. The outcomes showed that it was a substantial predictor of achievement. There is also a study by Ma, Wooster, and A. (2009) that investigated the effects of marital status on the academic performance of 374 college students. Their research revealed that married students had better grades than unmarried ones. It was reported, however, that marital status had no significant impact on predictions by Nasrullah et al. (2021) and Zhao et al. (2020).

Other Demographic Features: In addition, other demographic factors, such as income, have been used as predictors (Daud et al., 2017; Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Pal and Pal, 2013; Ali et al., 2013; Villwock, Appio, and Andreta, 2015; Yadav and Pal, 2012). Among these studies, Ali et al. (2013) examined the factors that affect graduate students' academic performance, including their socioeconomic status. The authors concluded that income contributes significantly to the success of students based on a sample of 100 randomly selected students. In addition, employment status has been shown to predict academic achievement in several studies (Daud et al., 2017; Kovačić, 2010; Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Mohamadian et al., 2015). Among these studies, Mohamadian et al. (2015) explored the relationship between employment status and academic achievement using data collected from 235 students. According to their findings, unemployed students had significantly higher academic achievement than employed students. In their view, working students are less likely to devote adequate time to their studies, which results in less success.

Despite the heavy use of demographic features, the extent to which they contribute to academic achievement prediction is not yet clear due to multiple studies either not reporting the individual contributions of these features or reaching opposing conclusions concerning their effectiveness. In light of previous research claiming that gender, age, marital status, and other areas of demographics affect students' academic success, the latest EDM research utilizes these factors as features to predict academic success, yet with questionable effectiveness.

The results of our analysis show that the choice of demographic features and their usage is likely to be strongly influenced by the cultural background of the countries where the study takes place. For instance, when the study is performed in a collectivistic country (e.g., India, Indonesia, and Malaysia), we observe features that are relevant to the family of the student, such as family support (Sembiring et al., 2011; Nasrullah et al., 2021), family income (Yadav and Pal, 2012; Pal and Pal, 2013; Villwock, Appio, and Andreta, 2015; Abu Saa, 2016; Daud et al., 2017), family size (Yadav and Pal, 2012; Raj and Manivannan, 2020), and parent's qualifications (Abu Saa, 2016; Nasrullah et al., 2021; Raj and Manivannan, 2020). This is not the case in studies performed in individualistic countries. (e.g., the United States and Europe).

A culture that is more individualistic tends to place more emphasis on achieving one's own goals, whereas a culture that is more collective places more emphasis on achieving goals as a family and team (Kim, 1995). It is possible that students from individualistic cultures may be more competitive than those from collectivistic cultures in this regard. Therefore, we believe that further research is needed to better understand the impact of culture on academic performance.

Post-enrollment features

Post-enrollment features are related to students' achievements after their enrollment. Following are the most commonly used post-enrollment features for performing academic predictions, regardless of their importance to the prediction:

Grades: Among the most commonly used features to predict the academic achievement of students after enrollment are grades earned in quizzes and examinations. (Zimmermann et al., 2011; Al luhaybi, Tucker, and Yousefi, 2018; Badr et al., 2016; Huang and Fang, 2013; Pradeep and Thomas, 2015; Villwock, Appio, and Andreta, 2015; Yadav, Bharadwaj, and Pal, 2011; Yassein et al., 2017; Aulck et al., 2017; Rotem, Yair, and Shustak, 2020a; Nasrullah et al., 2021; Smirani et al., 2022; Abu Zohair, 2019; Karakose et al., 2021; Arun et al., 2021; Kemper, Vorhoff, and Wigger, 2020). Based on Huang and Fang's (2013) study, the achieved grade on a mid-term exam is the most relevant feature influencing prediction accuracy.

Results in Previous Semester: The success rate of the previous semester, which is typically measured by GPA, has also been used often (Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Kabakchieva, 2013; Alemu Yehuala, 2015; Abu Saa, 2016; Asif et al., 2017; Al luhaybi, Tucker, and Yousefi, 2018; Kemper, Vorhoff, and Wigger, 2020) in the studies we have reviewed. That is since students' success depends on previously acquired knowledge or skills. For example, Asif et al. (2017) found that the marks of a four-year program's first and second-year courses play a role in predicting the graduation performance in a program. Likewise, Al luhaybi et al. (2018) found that the results of the core modules of the first year of the academic program have a high impact on the prediction of the high risk of failure students.

Attendance: A number of studies have used attendance as a predictor of students' academic success (Al luhaybi, Tucker, and Yousefi, 2018; Pradeep and Thomas, 2015; Yadav, Bharadwaj, and Pal, 2011; Yassein et al., 2017; Nasrullah et al., 2021; Karalar, Kapucu, and Gürüler, 2021) since higher attendance is considered an indicator of success among students. Using data from a course at a university in which attendance to classes was not mandatory, Lukkarinen, Koivukangas, and Seppälä (2016) examined the relationship between students' class attendance and learning performance. The researchers found that attendance is positively and significantly connected to the academic performance of students. Also, in a study conducted by Alija (2013), binary logistic regression was applied to examine the relationship between attendance and student achievement. They found that there is a greater likelihood of students receiving passing grades if they attend lectures regularly.

Other Post-enrollment Features: A balanced academic load is essential to academic success. This is determined by the number of credit hours and the difficulty of the course (Szafran and Austin, 2002). The choice of courses taken (Alemu Yehuala, 2015; Aulck et al., 2017) and the total number of credit hours taken (Alemu Yehuala, 2015; Abu Saa, 2016; Rotem, Yair, and Shustak, 2020) have therefore been used as indicators of academic success. It has been found by Alemu Yehuala (2015) that credit hours are one of the most significant variables associated with academic success.

As a conclusion, employing post-enrollment features for the prediction of students' academic outcomes have a role in maximizing the accuracy of the prediction. This is because such features correspond to students' current condition in the program instead of depending on their previous condition only.

3.2.2 Review on the used DM algorithms in Predicting Students' Achievement

Mostly used DM algorithms in predicting students' achievement

In EDM, many prediction algorithms have been investigated. Since there is no answer to the question of which is the best DM algorithm, we notice that most researchers explore several algorithms to predict students' success and do not rely on the results of just one algorithm. They frequently compare the results of each algorithm to determine the best-fit technique for the specific dataset and thus ensure the greatest accuracy rates when employing the model. Figure 4 shows the frequencies of the used DM algorithms.

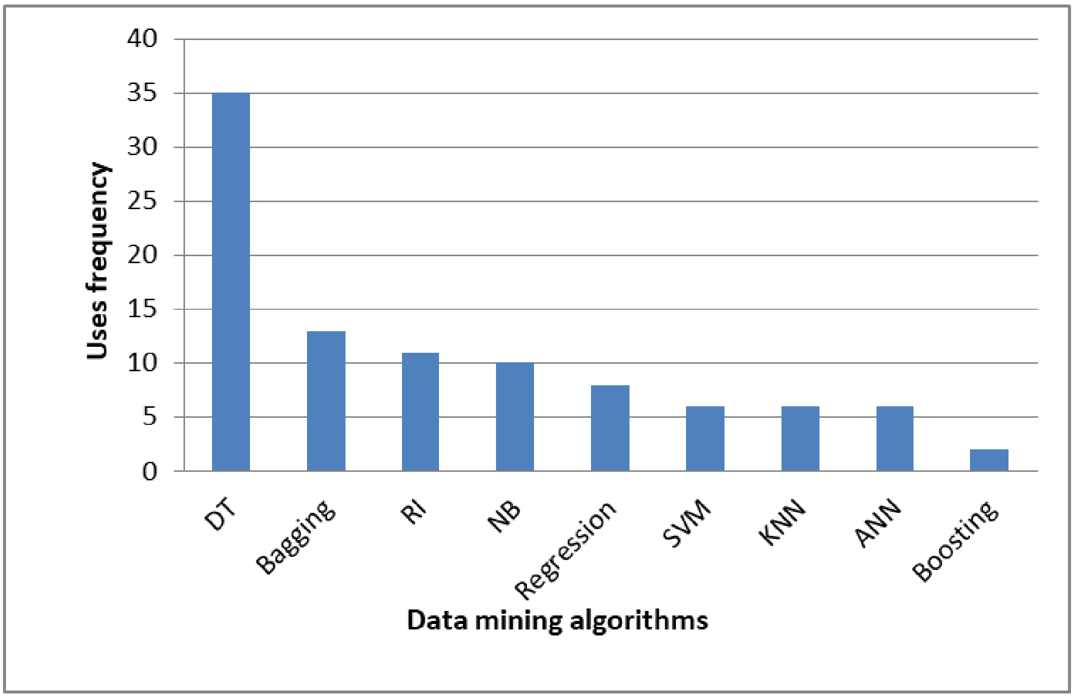


Figure 4: Mostly used DM algorithms in performing academic predictions

As can be seen, Decision trees are the most commonly used DM methods in the studies covered. Their ease of use and efficiency has made them one of the most popular and influential methods in machine learning since they were introduced in the 1960s (Song and Lu, 2015). Under this algorthim, knowledge models can be directly transformed into IF-THEN rules. The CHAID, CART, C4.5, and ID3 algorithms (Jain et al., 2017) are all decision tree algorithms. There is, however, greater popularity for C4.5 than for the other algorithms for decision trees. Fourteen studies have used this method, resulting in an accuracy range of 0.364 to 0.945. There were five studies that indicated that it was the best scoring method (Alemu Yehuala, 2015; Kabakchieva, 2013; Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Yadav and Pal, 2012) and two studies that indicated that it was the second-best scoring method (Osmanbegović, Suljic, and Suljić, 2012; Abu Saa, 2016). CART has also been used in 5 of the reviewed studies leading to a range of accuracies between 0.40 and 0.622. It was the best scoring method in three cases (Kovačić, 2010; Yadav, Bharadwaj, and Pal, 2011; Abu Saa, 2016). In four studies, ID3 was applied and was rated as the best method in the study of (Pal and Pal, 2013) with 0.78 accuracy, and the worst method in the study of (Abu Saa, 2016) with 0.333 accuracy. ADT was used in 2 studies only. In the first study by Pal and Pal (2013), it produced 0.6950 accuracy, while in the second by Pradeep and Thomas (2015), it obtained an accuracy of 0.995 and was assessed as the best scoring method. CHAID was also used in two studies only. It achieved an accuracy of 0.594 in the first study (Kovačić, 2010) and an accuracy of 0.341 in the second (Abu Saa, 2016). Rule-based classifiers such as JRip, NNge, OneR, and Ridor (Lakshmi, 2012) have also been used several times by researchers. The results were often satisfactory, with an accuracy of 0.545 (Kabakchieva, 2013) in its worst cases and 0.982 (Pradeep and Thomas, 2015) in its best cases.

Although black-box algorithms can be complicated for people to comprehend, they may outperform logistic regression and decision trees regarding prediction accuracy. For instance, Naïve Bayes produced outstanding results, above 0.75 in most cases. In fact, it was found to be the best performing algorithm in four of the viewed studies (Asif et al., 2017; Al luhaybi, Tucker, and Yousefi, 2018; Kovačić, 2010; Shakeel and Anwer Butt, 2015). Moreover, SVM was found to be the best performing algorithm in the case of Daud et al. (2017) with a 0.867 accuracy. ANN was used in six studies by Zhao et al., 2020; Arun et al., 2021; Nasrullah et al., 2021; Huang and Fang, 2013; Abu Zohair, 2019; Karalar, Kapucu, and Gürüler, 2021. Unlike other researchers, Nasrullah et al. (2021) depended on only one algorithm to perform their predictions. The algorithm they relied on was ANN, which produced a high accuracy (0.924).

Researchers have also investigated ensemble methods. For instance, Asif et al. (2017); Zhao et al. (2020); Arun et al. (2021); Shakeel and Anwer Butt (2015); Aulck et al. (2017); and Smirani et al. (2022)used random forest, which is a bagging algorithm. In the case of Zhao et al. (2020); and Arun et al. (2021), the Random Forest gave the highest prediction accuracy. On the other hand, only one of the thirty-two viewed studies used Boosting algorithms (Smirani et al., 2022), in which they investigated two boosting algorithms, namely LightGBM and XGBoost. Both boosting algorithms gave similar accuracies, 0.96 and 0.965, respectively.

Pros and cons of the different DM algorithms used in performing academic predictions

Table 1 below outlines the main advantages and disadvantages of the commonly used DM algorithms for performing academic predictions.

Table 1: Pros and cons of the different DM algorithms used in performing academic predictions

DM Algorithm	Pros.	Cons.
Rule Induction (Domingos, 1995)	Cost-effective computational space Statistical measures can be used efficiently to reduce noise	The training process is slow Has difficulty recognizing exceptions or small, low-frequency sections of the space
Decision Trees (Clark 2013; Kaushal and Shukla 2014; Yu-Wei 2015)	Simple to understand Can handle missing values	Could suffer from overfitting Less accuracy with continuous variables
Logistic Regression (Geng 2006; Yu-Wei 2015)	Easy to comprehend Provides probability outcome	Difficult in handling missing values May suffer from over-fitting
K-Nearest Neighbor (Clark 2013; Yu-Wei 2015)	Nonparametric Simple to understand the output Robustness to noisy training data	Black box Mixed data types are challenging to handle Considers all features equally significant Outlier-sensitive
Support Vector Machine (Clark 2013; Harrington 2011; Tomar and Agarwal 2013)	High accuracy Can handle different data types Effective in high dimensional space	Black box Training processes may consume time High algorithmic complexity
Naïve Bayes (Harrington 2011; Yu-Wei 2015)	Simple to use Can deal with missing and noisy data	Black box Assumes that all features are independent and equally important
Neural Networks (Clark 2013; Kaushal and Shukla 2014; Tomar and Agarwal 2013)	High accuracy Can manage missing and noisy data	Black box Big data is difficult to manage Complicated
Bagging (Kotsiantis, Tsekouras, and Pintelas, 2005)	A group of weak learners can be more effective than a single strong learner Can produce high accuracy Avoids overfitting	Computationally expensive May result in high bias if it is not appropriately modeled and thus may result in under-fitting
Boosting (Karalar, Kapucu, and Gürüler, 2021)	Reduce the bias and variance in a supervised learning technique Works well with two-class classification problems Can deal well with missing data	Hard to implement in real-time due to the increased complexity of the algorithm

3.2.3 Review on the used data mining tools in predicting students' achievement

In this section, we outline the commonly used data mining tools and reveal the advantages and disadvantages of each tool.

Mostly used DM tools in performing academic predictions

Based on the studies we have viewed in this chapter, most researchers rely on DM tools rather than programming languages. The open-source Weka tool appears to be the most widely used DM tool for predicting academic results, as 63% of the researchers used it to perform their predictions (figure 5). It is intended for ML and DM and was developed at the University of Waikato in New Zealand. Weka supports several standard DM tasks like data clustering, classification, regression, pre-processing, visualization, and feature selection. Weka has become popular with academic researchers recently due to its highly active community. SPSS has also been used by EDM researchers quite often (17%) compared to the rest of the DM tools. A major benefit of the IBM SPSS tool is its ability to offer the user extensive control and allow the development of predictive models quickly with business expertise (Brahmeswara Kadaru and Umamaheswararao, 2017). Similarly, RapidMiner, formerly referred to as Yale, offers a number of advantages, including multiple deployment options. However, it is less popular in the EDM community as it was used in 8% of the studies only. As per pure programming, only a few of the viewed researchers (8%) performed predictions using python. Python has many data-oriented feature packages that can speed up and simplify the processing of data allowing it to be a good choice for performing academic predictions on large datasets.

There are more DM tools that are suitable for performing DM tasks. However, according to our knowledge, researchers have not yet investigated them in the educational field. Examples of such tools are KNIME, Orange, Spark, and KEEL

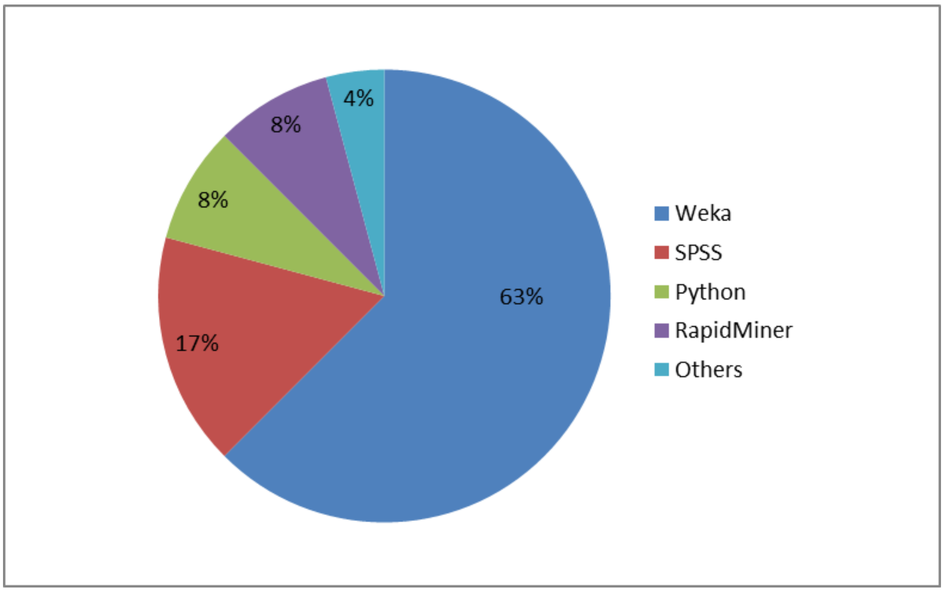


Figure 5: Analysis of the DM tools that have been used in the viewed literature

Pros and cons of the different DM tools used in performing academic predictions

Table 2 summarizes the seven tools that offer algorithms that are capable of modeling and predicting educational data (Slater et al., 2017). All these applications are well-documented and can be run on Microsoft Windows, Linux, and Mac OS platforms. A comparison of the main advantages and disadvantages of each tool is also presented in the table.

Table 2: Pros and cons of the different data mining tools used in performing academic predictions

DM tool and source	Programming language	Pros.	Cons.
Rapid Miner[2] (commercial)	Java	Provides both a command line interface and a graphical user interface (GUI) Provides visualizations performs Multi-level cross-validation Assesses models using a variety of metrics	Limited functionality for engineering new features out of existing features
SPSS[3] (Commercial)	Java	Both command line and GUI are supported Visualize the process easily New features can be created from existing features	Minimum support for modeling Less flexible than other tools Difficult to customize Slow in handling large data sets
WEKA[4] (Open source)	Java	Supports command line and GUI Displays visualizations	Does not support the creation of new features
KNIME[5] (Open source)	Java	Supports GUI Provides visualizations Capable of integrating data from various sources Provides extensions for R, Python, Java, and SQL	Does not support interactive execution Not all nodes can be streamed
Orange[6] (Open source)	Python Cython C ++ C	Supports command line and GUI Customizable visualizations Easy to understand interface	Limited in the scale of data that it can work with, comparable to Excel Less suitable for big projects
Spark MLLib[7] (Open source)	Scale SQL Java R Python	Displays visualizations Can connect with several programming languages through API	Purely programmatic tool (less usability for non-programmers)
KEEL[8] (Open source)	Java	Supports command line and GUI Visualizes data Algorithms for discretization are supported Support many feature selection algorithms support missing data	Limited functionality for engineering new features out of existing ones Limited support for clustering and factor analysis Limited support for association rule mining

3.2.4 Review on the academic prediction studies based on Country

Based on the literature we have viewed, most prediction studies happened to be performed in south Asia (figure 6), more specifically, six studies in India (Yadav, Bharadwaj, and Pal, 2011;

[2] rapid-i.com/content/view/181/190/

[3] https ://www.ibm.com/analy tics/dk/da/techn ology /spss/

[4] https ://www.cs.waika to.ac.nz/ml/weka/

[5] https ://www.knime .com

[6] https ://www.orang e.biola b.si

[7] https ://www.spark .apach e.org/mllib /

[8] https ://www.sci2s .ugr.es/keel/

Yadav and Pal, 2012; Pal and Pal, 2013; Pradeep and Thomas, 2015; Arun et al., 2021; Raj and Manivannan, 2020) and three studies in Pakistan (Daud et al., 2017; Shakeel and Anwer Butt, 2015; Asif et al., 2017). Southwest Asia comes in the second place, mainly two studies in the United Arab Emirates (UAE) (Abu Zohair, 2019; Abu Saa, 2016) and three studies KSA (Yassein et al., 2017; Badr et al., 2016; Smirani et al., 2022). The third most academic prediction studies are performed in Southeast Asia, including Indonesia (Nasrullah et al., 2021), Thailand (Nguyen Thai Nghe, Janecek, and Haddawy, 2007), Vietnam (Nguyen Thai Nghe, Janecek, and Haddawy, 2007), and Malaysia (Sembiring et al., 2011). With the same number of studies in West Europe, precisely one study in Germany (Kemper, Vorhoff, and Wigger, 2020), one study in Switzerland (Zimmermann et al., 2011), one study in Norway (Jeno, Danielsen, and Raaheim, 2018), and one study in the United Kingdom (UK) (Al luhaybi, Tucker, and Yousefi, 2018). Next is North America (USA), with three performed studies (Zhao et al., 2020; Huang and Fang, 2013; Aulck et al., 2017). Then, west Asia with two studies, namely in Israel (Rotem, Yair, and Shustak, 2020) and Turkeya (Karalar, Kapucu, and Gürüler, 2021). On the other hand, locations as South America (Villwock, Appio, and Andreta, 2015), East Europe (Kabakchieva, 2013), southeast Europe (Osmanbegović, Suljic, and Suljić, 2012), Australia (Kovačić 2010), and Africa (Alemu Yehuala, 2015), performed one study each.

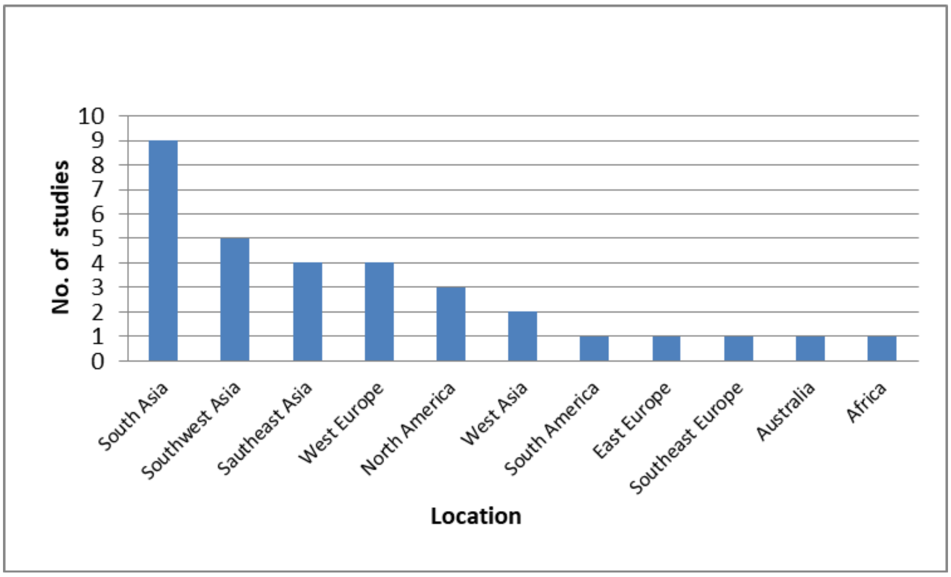


Figure 6: Academic predictions based on country

3.2.5 Review of the prediction studies based on the degree

According to Rotem, Yair, and Shustak (2020), the research on students' dropout and postponement at the undergraduate level are more than at the postgraduate level, and no solid predictive models are to be found for postgraduates. In our reviewed literature, most researchers performed academic predictions at a bachelor's degree level (figure 7). Although postgraduate students also face challenges leading to dropout or delay in the program, only 27% of the reviewed studies performed predictions in a master's degree level (Nguyen Thai Nghe, Janecek, and Haddawy, 2007; Yadav and Pal, 2012; Zimmermann et al., 2011; Zhao et al., 2020; Jeno, Danielsen, and Raaheim, 2018; Rotem, Yair, and Shustak, 2020; Badr et al., 2016; Abu Zohair, 2019).

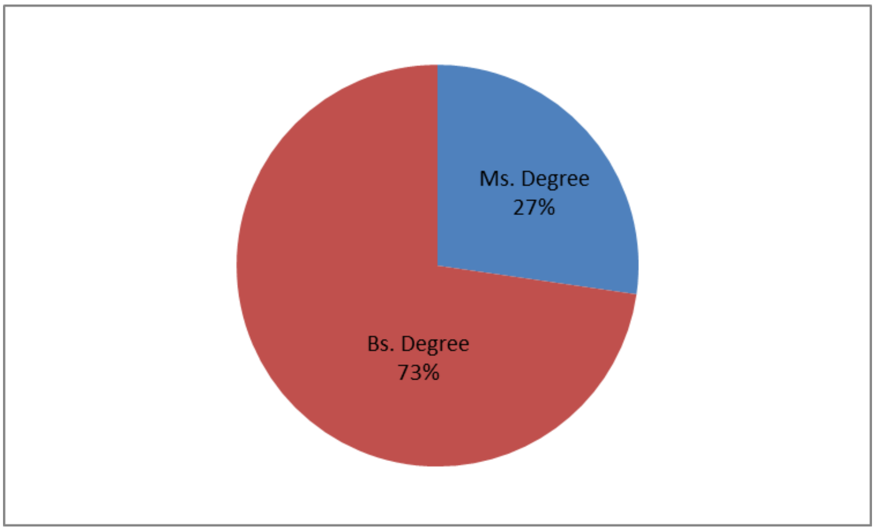


Figure 7: Percentage of the academic predictions performed in different degrees

3.3 Per Task analysis

Having reviewed the features, algorithms, and tools used in the literature and categorizing the studies based on the targeted academic degree and where the study has been performed, we now analyze in detail the same literature by shifting the focus to the sub-tasks that comprise the academic achievement task.

Based on the viewed literature, there are three main types of predictions of students' performance in higher education: (1) academic performance or GPA at a degree level, (2) failure or dropout, and (3) academic performance at a course level. In this section, the literature reviewed is presented using bullet points and tables in order to facilitate random access and simplify comparisons. There are only a few studies that have demonstrated the importance of certain features in prediction. However, the tables provide a comprehensive view of all studies viewed. This includes the prediction task, the location of the study, the features used for the prediction, the DM tool, the DM algorithm, and the accuracy of the prediction.

3.3.1 Prediction of students' academic performance or GPA at a degree level

One of the most widely used metrics for assessing the quality of universities is their students' academic performance. Among the primary applications of EDM is predicting students' GPAs or overall academic performance, e.g., above average, average, and below average. Such predictions are useful in a variety of contexts at universities, such as identifying outstanding students for scholarship allocation. The following studies have investigated the impact of success factors on the prediction of students' academic performance at an undergraduate or graduate level:

- Sembiring et al. (2011) sampled 300 students from the faculty of computer systems and software engineering in order to predict their final grades. Unlike the rest of the studies, they used innovative features. Each of the features was tested for significance using multivariate analysis. Family support was found to have the greatest impact (52.6%) on the prediction, followed by engaging time, followed by study behavior, and finally by study interest. Meanwhile, students' own beliefs had no impact on the study.

- Kabakchieva (2013) used a total of 10,330 student records to predict their achievement based on 5 classes (Bad, average, good, very good, and excellent). According to their findings, the classifiers perform differently for each of the five classes. Another finding is that the post-enrollment characteristics of students, such as their university admissions score and their number of failures at first year exams are the most influential factors for the prediction.

- Abu Saa (2016) conducted a survey of 270 students in order to predict students' performance in an IT Department by collecting data from daily classes and online. In their study, they found that students' performance is not entirely determined by post-enrollment factors, e.g., their academic efforts, but, on the contrary, many other factors are equally significant, if not more so. This includes demographical features, such as gender, mother occupation, pre-enrolment features, high school grade, and University fees discount.

- Asif et al. (2017) used a sample of 210 undergraduates to predict students' performance. Only marks were used to perform the prediction. Using only student pre-university marks and student first and second-year marks, their study demonstrated that graduation performance in a four-year university program can be reasonably predicted.

- Zhao et al. (2020) used only 132 records to explore the problem of identifying a good admissions strategy for a Master of Science program in Data Science. They used features such as gender, age, previous GPA, previous major, language proficiency, and GRE. Their findings suggest that students with an undergraduate major in Business, Economics, International Studies, Humanities, and Communications are poor candidates for the data science program, while applicants with Computer Science, Electrical Engineering, (Applied) Mathematics, or Statistics backgrounds are more likely to succeed in the program. moreover, they found that high GRE scores, undergraduate GPA, and School Ranking are positive indicators of success.

Table 3 summarizes the studies we analyzed in lines of the country, degree, features, algorithms, tools, and accuracy. While nine of the following studies are performed on bachelor’s students, only four of the studies are performed on master’s students.

Table 3: Prediction of students' academic performance or GPA at a degree level

Authors	Prediction	Country	Deg.	Features	Algorithms	Tool	Results
Nghe et al. (2007)	Predict students' GPA at the Asian Institute of Technology	Thailand	Ms.	Demographics Pre-enrolment	DT- C4.5 DT- BT	Weka	C4.5 produced better accuracy (91.98%) for 2 classes (pass/fail), (67.74%) for 3 classes (Fail/Good/Very Good) and (63.25%) for 4 classes (Fail/Fair/Good/Very Good)
Nghe et al. (2007)	Predict students' GPA at the end of the third year in Can Tho University	Vietnam	Bs.	Demographics Pre-enrolment Post-enrollment	DT- C4.5 DT- BT	Weka	C4.5 produced better accuracy than BT.(92.86%) for 2 classes (pass/fail), (84.18%) for 3 classes (Fail/Good/Very Good) and (66.69%) for 4 classes (Fail/Fair/Good/Very Good)
Yadav, Bharadwaj, and Pal (2011)	Predict computer science students' performance at VBS Purvanchal University	India	Ms.	Post-enrolment	DT- CART DT- ID3 DT- C4.5	Weka	CART produced the best accuracy (56.25%) followed by ID3 (52.08%), then C4.5 (45.83%)
Sembiring et al. (2011)	Predict final grade of students from the faculty of computer systems and software engineering at the University of Malaysia Pahang	Malaysia	Bs.	Demographics Post-enrollment	SVM	Rapid-Miner	SVM produced high accuracy (83%)
Zimmermann et al. (2011)	predict the achievement of students at a Computer Science program at ETH Zurich	Switzer-land	Ms.	Pre-enrolment	Bagging- RF	n.a	Third year bachelor's achievements are more predictive than the first-year grades in predicting the Ms. Students' GPA
Yadav and Pal (2012)	Predict Engineering student academic performance at VBS Purvanchal University in Jaunpur	India	Bs.	Demographics Pre-enrolment	DT- ID3 DT- CART DT- C4.5	Weka	C4.5 produced the best accuracy (67.77%) followed by ID3 and CART with the same accuracy (62.22%)

Pal and Pal (2013)	Predict student performance at VBS Purvanchal University in Jaunpur	India	Bs.	Demographics Pre-enrolment	DT- ID3 DT- ADT Bagging	Weka	ID3 produced the best accuracy (78%) followed by bagging (73%) then ADT (69.50%)
Kabakchieva (2013)	Predict students' performance (Bad, average, good, very good and excellent) at the University of National and World Economy	Bulgaria	Bs.	Demographics Pre-enrolment Post-enrollment	DT- C4.5 KNN RI- JRip RI- OneR Bayesian	Weka	C4.5 produced the best accuracy (66.5%) followed by JRip (63%) then KNN (60%) and Bayesian (≈ 60%) then finally OneR (54.5%)
Abu Saa (2016)	Predict students' outcomes in the IT Department at Ajman University of Science and Technology	UAE	Bs.	Demographics Pre-enrolment Post-enrollment	DT- CART DT- C4.5 DT- ID3 DT- CHAID NB	Rapid-Miner and Weka	CART produced the best accuracy (40%) followed by C4.5 (36.40%) then NB (35.19%) then CHAID (34.07%) then finally ID3 (33.33%)
Asif et al. (2017)	Predict students' achievements using 2 classes (low/high) at the end of the third year of their IT degree	Pakistan	Bs.	Pre-enrolment Post-enrollment	DT RI NB NW KNN Bagging- RF	Rapid-Miner	NB produced best accuracy (83.65%) followed by 1-nearest neighbor (74.04%) then RF (71.15%) then DT (69.23%) then neural NW (62.50%) then finally rule induction (55.77%)
Yassein et al. (2017)	Predict students' academic outcomes at Najran University	KSA	Bs.	Post-enrollment	DT- C4.5 Clustering	SPSS & clementine	n.a
Zhao et al. (2020)	Predict student success in a data science degree program at Fordham University	USA	Ms.	Demographics Pre-enrolment	DT SVM ANN NB KNN Ensemble Learner L Bagging- RF	n.a	RF and the ensemble learner L achieved the two best overall predictive accuracy

					Regression-LR		
Arun et al. (2021)	predict the final year GPA of Computer Science and Engineering students at B.M.S College of Engineering	India	Bs.	Post-enrollment	SVM KNN ANN Bagging- RF Bagging- RT Regression-LR Etc…	Weka	RF performed the best

3.3.2 Prediction of students' failure or drop out of a degree

It is well known that student failure or dropout is a significant concern in the education and policymaking communities (Demetriou and Schmitz-Sciborski, 2011). An educational institution's reputation is adversely affected by high dropout rates and poor academic performance among students. Both individuals and educational institutions are harmed by students' drop out of the educational system. The prevention of educational dropouts therefore poses a significant challenge to higher education institutions. By identifying at-risk students at an early stage, we are able to prevent these incidents from occurring. A number of studies have been conducted in order to predict students' likelihood of failing or dropping out of college:

- Pradeep and Thomas (2015) predicted the dropout of undergraduate student using the transcripts of 150 students who have been enrolled in a Technology program. Using the Attribute Selection Algorithms provided in the WEKA tool, the number of features used was reduced from 67 to the 13 best features. Most of the features selected were post-enrollment features including attendance, taking notes during lessons, and some course scores. A number of factors, including age, gender, and religion, were not taken into account, since these factors did not have an impact on the overall prediction.

- Alemu Yehuala (2015) examined 11,873 student records to predict university students who are at risk of failure. According to their research, six major factors determine whether a student will succeed or fail: the number of students in a class, the number of courses offered in a semester, higher education, the student's entrance certificate, the result of the examination, and the gender of the student.

- Villwock, Appio, and Andreta (2015) examined the features which may impact students' decision to drop out of a Mathematics major. There was evidence that the courses that contributed to dropouts in the Major varied between years. Considering only the subjects studied in the first year, the course that most contributed to dropouts was “Differential and Integral Calculus I”, and considering the first 2 years, it was “Finite Mathematics”. Furthermore, it was concluded that the work factor contributed most to the dropout rate. It is believed that this is due to the limited time available to working students for doing extracurricular activities. Additionally, they found that marital status and age were associated with dropping out.

- Daud et al. (2017) examined 776 instances of students across multiple universities in Pakistan in order to predict whether they would complete their studies or drop out. A total of 23 features (selected through the feature extraction process) were used in the experiment. According to their findings, natural gas expenditures, electricity expenditures, self-employment, and location are the factors most influential in predicting students' performance.

- Aulck et al. (2016) studied the dropout rate in a department of Electrical Engineering using a dataset of more than 32,500 students. The analysis of individual features revealed that the most significant predictors of dropout are a student's GPA in math, English, chemistry, and psychology courses, year of enrollment, and age.

• Rotem et al. (2020) combined anonymized yearly data files regarding students from 2007 to 2017. The dataset contained detailed information on 22,761 master's students. The total dropout rate was 12.33%. They found that background variables (Students' socioeconomic background and gender) have practically no value in predicting dropout outcomes from master's programs. They also found that pre-academic achievements have no contribution to the dropout prediction at the master's level. The three variables are the main predictors: academic load, achieved grades, and failed examinations.

• Nasrullah et al. (2021) used 19 attributes to predict student dropouts. Those factors include gender, age, marital status, family support, and many more. They found that the attributes that primarily affect the prediction are students' grades, followed by the failed courses, then attendance records.

Table 4 summarizes the studies we analyzed regarding predicting students' dropout in the country where the study was held, the used features, tools, algorithms, and accuracy of the prediction. While ten of the following studies are performed on bachelor's students, only two of the studies are performed on master's students.

Table 4: Prediction of students' failure or drop out of a degree

Authors	Prediction	Country	Deg.	Features	Algorithms	Tool	Results
Pradeep and Thomas (2015)	Predict student dropout in Technology program at Mahatma Gandhi University	India	Bs.	Pre-enrolment Post-enrollment	DT- ADT DT- C4.5 DT- REP tree Bagging- RT RI -JRip RI- NNge, RI- OneR RI- Ridor	Weka	ADT obtained the best accuracy (99.5%) followed by JRip (98.02%), then NNge and RT with same accuracy (97.02%), then Ridor (96.53%) then REP Tree (95.05%), then C4.5 (94.55%) then finally OneR (89.60%)
Alemu Yehuala (2015)	Predict university students at risk of failure at Debre Markos University	Ethiopia	Bs.	Demographics Pre-enrolment Post-enrollment	DT- C4.5 NB	Weka	C4.5 produced better accuracy (91.62%-92.33%) than NB (86.3%-87.4%)
Shakeel and Anwer Butt (2015)	Predict students who are likely to drop out and students needing further help in the University of Gujrat	Pakistan	Bs.	Demographics Pre-enrolment Post-enrollment	DT- C4.5 NB Bagging-RF Regression- LR	Weka	NB produced the best accuracy (91.93%), followed by RF (88.71%), then C4.5 (87.09%), then finally LR (66.13%)
Villwock et al. (2015)	Predecting student's drop out of the Mathematics Major at Universidade Estadual do Oeste do Paraná	Brazil	Bs.	Demographics Pre-enrolment Post-enrollment	DT- C4.5	Weka	C4.5 produced a high predictive accuracy (91.84%)
Daud et al. (2017)	Predict the completion or dropout of students from different universities	Pakistan	Bs.	Demographics Pre-enrolment	SVM Bayes network NB DT- C4.5 DT- CART	Weka	SVM produced the best accuracy (86.7%), followed by Bayes network & NB with the same accuracy (84.8%), then C4.5 (76.6%), then finally CART (71%)
Aulck et al. (2017)	Predict student dropout using the first semester's grades in the	USA	Bs.	Demographics Pre-enrolment Post-enrollment	Regression- LR Bagging- RF KNN	n.a	LR produced the best accuracy (66.59%), then KNN (64.60%), then RF (62.24%)

	Electrical Engineering department at the Eindhoven University of Technology						
Kemper, Vorhoff, and Wigger (2020)	Predict student dropout at Karlsruhe Institute of Technology	Germany	Bs.	Demographics Post-enrollment	LR DT	n.a	DT produced slightly better accuracy (91.3%) than LR (90.08%)
Jeno, Danielsen, and Raaheim (2018)	predict dropout intentions among biology students using Self-Determination Theory	Norway	Bs. & Ms.	Post-enrollment	Regression standard errors	SPSS & AMOS	The motivational dynamics for bachelor's and master's students differ from each other
Rotem, Yair, and Shustak (2020)	Predict students' dropout at the Hebrew University of Jerusalem	Israel	Ms.	Demographics Post-enrollment	Regression- LR	n.a	Using LR can accurately predict academic failure
Raj and Manivannan (2020)	predicts the students who have the likelihood of failing in a bachelor of business administration degree in a private university	India	Bs.	Demographics Pre-enrolment Post-enrollment	DT- repTree RI- Jrip, Bagging- RF Bagging- RT NB	n.a	repTree performed the best (72.44% accuracy), followed by RF (66.14%), then NB (62.20%)
Nasrullah et al. (2021)	predicting student dropouts at the Health Study Program in one of the tertiary institutions	Indonesia	Bs.	Demographics Post-enrollment	ANN	n.a	ANN produced a high accuracy (92.4%)
Smirani et al. (2022)	Predict student failure using a	KSA	Bs.	Post-enrollment	The SGFP model mixes ensemble	n.a	MLP produced the highest accuracy (97.3), followed by RF (97.1), then XGBoost (96.5), then LightGBM

	stacked generalization-based algorithm (SGFP) using data from LMS and grade containers at Umm al-Qura University.				learning classifiers: Boosting- Light GBM Boosting- XGBoost Bagging- RF Multilayer Perceptron (MLP)		(96). However, using an SGFP approach performs better than one classifier.

3.3.3 Prediction of students' results on particular courses

By predicting a student's achievement at a course level, teachers are able to gain insight into how their classes perform, and, as a result, take practical measures to enhance students' learning. For example, if the prediction indicates that some of the students in the class are likely to fail the course, educators may take proactive measures to help these students succeed. Various active and cooperative learning strategies can be used to accomplish this. An overview of some studies that have been conducted in order to predict students' results in specific courses is provided below:

- Kovačić (2010) conducted a study on 453 students in order to predict the performance of these students in an "Information Systems" course. Specifically, they investigated whether successful students could be distinguished from unsuccessful students based on demographic characteristics (such as gender, age, ethnicity, and disability) or by study environment (such as course program, faculty, or course block). In their study, the researchers found that the information collected during the registration process (demographics, secondary school, employment status, and early enrollment) is insufficient to distinguish between successful and unsuccessful students.

- Osmanbegović, Suljic, and Suljić (2012) examined 257 student records to predict their level of achievement in a "Business Informatics" course. They performed an analysis to determine the importance of each feature individually. Based on their analysis, they found that GPA has the greatest impact on the prediction, followed by entrance exams, study material, and average weekly study hours. Meanwhile, the number of household members, the distance from the faculty, and gender had the least impact on the prediction.

- Huang and Fang (2013) analyzed the performance data of 323 undergraduate students to predict the performance of these students in a Dynamic course. Among their interesting findings is that the grades students earn in pre-requisite courses may not accurately reflect their knowledge of those topics. The reason for this is that they may have taken pre-requisite courses in the past, and by the time they take the dependent course, their knowledge of the pre-requisite courses may have improved.

- Badr et al. (2016) used students' records in order to predict how they would perform in a "Programming" course. Upon analyzing the relationship between programming and the other courses, only the English courses directly impacted their predictions.

- Al luhaybi, Tucker, and Yousefi (2018) gathered data from 129 students to predict the students at high risk of failure in four computer science core modules. The predicted class feature is the “Overall Grade”, which is the final grade obtained by the student in the targeted module. There are five possible grades for the overall grade: A: Excellent, B: Very Good, C: Good, D: Acceptable, and F: Fail, which have been classified into Low risk, Medium risk, and High risk of failure to improve the classification results. In their study, they found that student qualifications for program entry significantly impacted their academic performance. In addition, some of the final grades received in previous modules have an effect on the students' academic performance in the current module.

- Karalar, Kapucu, and Gürüler (2021) Performed a prediction study at a Turkish state university. The data set consists of the data of a 15-week compulsory course entitled "Information Technologies" for all students in the first semester and obtained at the end of the 2020-fall semester. This study incorporates activity data from the university LMS (Moodle) and Conference Management Software (Adobe Connect) spanning a course semester. According to the study, quiz scores, degrees, number of lecture notes downloaded, number of other course materials downloaded, and amount of time spent watching recorded course videos all play a significant role in predicting at-risk students.

Table 5 offers a summary of the published research regarding predicting students’ results on particular courses, with respect to the country where the study was held, the used features, tools, algorithms, and accuracy of the prediction. While five of the following studies are performed on bachelor’s students, only two are performed on master’s students.

Table 5: Prediction of students' results on particular courses

Authors	Prediction	Country	Deg	Features	Algorithms	Tool	Results
Kovačić (2010)	Predict successful and unsuccessful student in Information Systems course	New Zealand	Bs.	Demographics Pre-enrolment	DT- CHAID DT- CART	SPSS	CART produced better accuracy (60.5%) than CHAID (59.4%)
Kovačić (2010)	Predict students' success in business informatics course in Faculty of Economics	Bosnia and Herzegovina	Bs.	Demographics Pre-enrolment Post-enrollment	DT- C4.5 NB Multilayer	WEKA	NB produced the best accuracy (76.65%), followed by C4.5 (73.93%), then finally multilayer prediction (71.2%)
Huang and Fang (2013)	Predict student academic performance in Engineering Dynamics at Utah University	USA	Bs.	Pre-enrolment Post-enrollment	Regression ANN SVM	SPSS	The developed predictive models have an average prediction accuracy of 86.8–90.7%
Badr et al. (2016)	Predicting students' grades in programming for the KSU mathematics department	KSA	Ms.	Post-enrollment	RI- CBA rule-generation	LUCS-KDD	CBA rule-generation produced an accuracy between 62.75% to 67.33%
Al luhaybi et al. (2018)	Predict 2nd-year computer science student academic performance in 4 computer science core courses at Brunel University	UK	Bs.	Demographics Pre-enrolment Post-enrollment	DT- C4.5 NB	Weka and Java API	NB produced slightly better accuracy (78.79%) than C4.5 (77.3%)
Abu Zohair (2019)	Predicting student dissertation project grade using four classes at Dubai's British University	UAE	Ms.	Demographics Pre-enrolment Post-enrollment	ANN NB SVM KNN LDA	Python and R	SVM and LDA perform the best using a small dataset
Karalar, Kapucu, and Gürüler (2021)	predicting students at risk of failure in online course "Information Technologies"	Turkey	Bs.	Demographics Post-enrollment	ET DT ANN Bagging- RF Regression-LR	Python	The proposed ensemble model made a good prediction with a specificity of 90.34%. It consists of combinations of ET, RF, and LR classification algorithms.

3.4 Summary of the Chapter

Researchers and practitioners worldwide are experiencing exciting opportunities in the area of EDM. The purpose of this chapter was to provide a literature review on predicting academic achievement in higher education over the past 16 years (between 2007 and 2022). It was revealed in the chapter that considerable work had been conducted regarding the analysis and prediction of academic performance. This study demonstrated that classification and regression algorithms can be used successfully to predict students' academic achievement both on a course and degree level. In our review, we discovered that most of the EDM research conducted in the past decade has been carried out using the open-source machine learning software Weka.

Additionally, we found that decision tree algorithms are the most commonly used algorithms for predicting academic achievement, with C4.5 being a popular choice among them. This is most likely due to the fact that such white box classification algorithms obtain models that can be explained by IF-THEN rules. Teachers, for example, who are not experts in the use of DM, can interpret the rules in a simple manner and use them in direct decision-making. As compared to the other methods of DM, neural networks, support vector machines, and K nearest neighbors did not appear to be frequently used. Since these methods rely on black-box mechanisms, they may not be preferred by researchers. Moreover, most researchers still rely on traditional machine learning. For instance, boosting algorithms are not common, as only one study in 2022 investigated them. More studies should explore their impact on academic predictions. In spite of the increasing popularity of deep neural networks in the machine learning community, in particular, in the context of applications to natural language processing, they have not yet been incorporated into the EDM literature. The reason for this may be related to the fact that they require extensive training data, whose acquisition is problematic in educational settings. Furthermore, we found that the features used widely vary according to the specific settings of each institute, culture, and country. There are, however, certain characteristics that researchers agree on when predicting students' academic achievement in higher education regardless of their environment, i.e., where they come from and what they believe in, such as gender, age, prior GPA, and proficiency in the academic language.

The surveyed literature has a significant limitation in that only a few studies examine and report the significance of each predictor. Instead, most studies report only the final results, making it difficult to evaluate the value of each feature, even for those that are widely used. This confirms that the prediction of students' performance is still a very actively researched problem, whose current solutions can still be improved. In addition, the factors that primarily influence academic outcomes and hence can be used to predict future performances are still not widely understood. In conclusion, more research is required first to understand the contributions of each employed feature and, second, to explore different sets of features and methodologies. This could further improve the current prediction accuracy.

CHAPTER 4: PREDICTING THE ACADEMIC ACHIEVEMENT OF BACHELOR'S DEGREE STUDENTS[9]

I am a big believer in early intervention. -Temple Grandin

4.1 Introduction

Students' graduation rates are seen as one of the foundations of measuring higher education institutions' quality and accountability. As seen in the previous chapters, the strengths of the DM practices can deliver valuable perceptions for predicting the final academic performance of students. This type of prediction can help both teachers and students in many ways. It enables decision-makers to take appropriate interventions at the earliest stage possible. Additionally, instructors can be aware of each student's capabilities and thus can customize the teaching tasks based on students' needs, e.g., recommending extracurricular learning material to students facing obstacles, using different teaching tactics, and providing online tutoring videos for students needing it. Although previous studies mainly focused on underachieving students, EDM can also have a high potential for extraordinary students. EDM permits discovering honorary students early in the program. We believe that this group of students is also of particular interest as universities increasingly value talented students and rely on them to present the perfect image of the university they belong to. Therefore, discovering such students at an early stage can serve universities in endless ways. Moreover, Opportunities can be offered to well-deserving individuals, e.g., scholarships, internships, and workshops. State of the art in chapter 3 shows that most of the studies that produced high accuracies used features that are quite challenging to gather, e.g., marital status, employment status, and students' attendance. Moreover, we can see that only three academic prediction studies have been performed in the Kingdom of Saudi Arabia. Nevertheless, none of those studies predicted the specific graduation grade and none of them explored predicting honorary students.

As mentioned in chapter 2, section 2.3, there are four main EDM applications. In this study, we address two of them. First, student modeling by (i) using methods of relationship mining to understand the relationship between students' success and the different factors, (ii) predicting students' academic achievement using various DM algorithms, and (iii) using methods to distill data for the judgment of humans, i.e., visualize the students' data and make inferences about it. Second, scientific research to establish experiential evidence of the EDM potential. As we try to present an approach that may be put into practice with quite ease at other academic institutions, we perform the academic predictions using features that can be obtained easily from any university's database. In our study, we do not only predict students at risk of failure but also honorary students. Moreover, we investigate predicting students' specific grades using not only three classes but five classes as well (which has never been investigated in the viewed literature).

[9] This chapter is a modified version of an article published in the Journal of Information Technology Education: Innovations in Practice and has been reproduced here with the permission of the Informing Science Institute

The main objectives of this chapter are to (i) test whether removing the orientation year from the College of Computer and Information Systems at PNU is a significant change in the programs offered by that college, (ii) find the significance of correlation attributes on the prediction of students' academic performance, and (iii) predict students' final graduation grade at an early stage of their studying journey. For those objectives, we hypothesize the following:

- Studying an orientation year can help students in increasing their success at the college of Computer and Information Science.
- Previous knowledge and GPA can assist in predicting the academic achievement of the students of Computer and Information Science college.
- English skills do not influence students' academic performance at the Computer and Information Science college.
- The number of failed courses per year can influence the prediction of students' academic achievement.
- Students' academic load per semester can affect the prediction of their academic achievement.
- It is not likely to have reasonably accurate predictions after the first and second semesters of the bachelor's program.
- It is likely to have accurate predictions using multi-class classification.

The following sections are organized as follows. The next section presents an overview of the bachelor's program. After that, we outline the study's methodology. Then we present details of the achieved results and discussion. Finally, a summary of the study's main findings is provided, as well as a description of our limitations and suggestions for future research.

4.2 Overview of the Bachelor's program

In the College of Computer and Information Sciences at PNU, students' achievement is weighed using a 5.00-grade point average (GPA). Using the total number of credit hours for which grades were given, the sum of all quality points earned is divided by the total number of credit hours. The college contains three majors: Computer Science (CS), Information Technology (IT), and Information Systems (IS). These majors are within the same realm of study. Each of the majors, however, offers courses on specialized aspects of computer science. The CS major focuses on the theory of computational applications. Moreover, it draws special attention to algorithms and mathematics. Students of this major learn the basics of programming languages, linear and discrete mathematics, and software design and implementation. They study the machine itself and understand how and why various computer processes run in the sense they do.

The IT major pays more attention to network models, their protocols, the types of traffic generated, and their quality-of-service requirements. Students of that major learn how to fix performance issues in networks and how to use different techniques for optimize performance. They also focus on internet design principles, internet routing design, internet application protocols, cryptography, and security.

Students enrolled in the IS major focus primarily on meeting the requirements of users in an organizational context. This is done through the selection, development, implementation, integration, and administration of computing technologies. IS students also gain knowledge of how to take advantage of current technical concepts and practices. They also come to know how to analyze, and define the requirements that must be satisfied to address IT problems or opportunities faced by organizations or individuals. Moreover, they learn the fundamentals of effectively designing IT-based solutions and integrating them into the user environment, identifying and evaluating current and emerging technologies, and discussing their applicability to solve the users' needs.

Formerly, all three majors (i.e., CS, IT, and IS) could be completed within five years, i.e., ten semesters (two semesters as part of the orientation (Preparatory) year, eight semesters as part of each major). Orientation is a program designed to help students prepare for higher education. As part of their orientation year, students take English language courses and foundational undergraduate math and physics courses. The aim of this year is to allow students to adjust themselves to the upcoming academic environment and teaching system. In recent years, the programs have been changed to eliminate the orientation year, allowing students to graduate in four years instead of five.

A total of twelve courses are shared among the three majors, which all are taught in English. Programming Language (1), Programming Language (2), Database Fundamentals, and Computer Networks Fundamentals are among the introductory courses taught during the first two years of all three majors.

4.3 Research Methodology

In this section, we start first by describing the selected students' dataset. Then, we describe the data collection phase. After that, the used DM algorithms and tools and the evaluation method are presented.

4.3.1 Students' dataset

A random sample of 300 female students aged 20-22 from the College of Computer and Information Science at PNU was used in this study. Among the records collected, one hundred records from the IS major, one hundred from the IT major, and one hundred from the CS major. In the collected sample, 117 students have studied an orientation year, whereas 183 students have not. Those who study an orientation year do not receive credit for the GPA earned during that year; therefore, the GPA for the orientation does not affect the final GPA. Based on the collected data, most students have accepted GPAs and a minority have poor GPAs. Figure 8 illustrates the distribution of students' final academic grades.

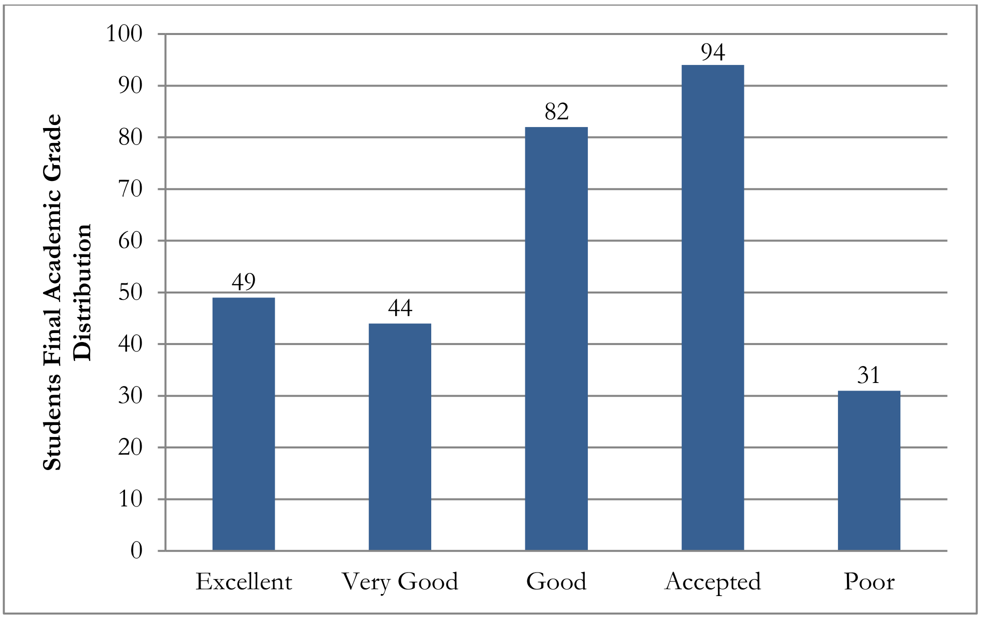


Figure 8: The final academic grade distribution for the collected data

The prediction of academic achievement can be based on three types of features (Alturki, Hulpus, and Stuckenschmidt, 2020): (i) demographics, e.g., gender and age, (ii) pre-enrollment features, e.g., previous GPA, and (iii) post-enrolment features, e.g., course grades. In our study, we applied one pre-enrollment feature, namely the secondary school graduation grades, and the remaining features are pre-enrollment features, namely: two English course grades, the GPA of the first four semesters, the number of academic credits earned in the first four semesters, the number of courses that have been failed in the first four semesters, and the grades of the college's core courses. While there are several common courses across all three majors, we considered only the courses taken in the first four semesters to undertake the prediction before the third academic year. The characteristics of the dataset are shown in Table 6.

Table 6: Discription of the features that have been used to build the predictive models

Feature	**Description**	**Type**	**Value**
GradGPA	Graduation grade using 5 classes	Nominal	Excellent, Very good, Good, Accepted, and Poor
Academic Success	Graduation grade using 3 classes	Nominal	Above average, average, and below average
SecPer	Secondary school academic achievement	Numeric	0 – 100%
Major	Student enrollment major	Nominal	CS, IS, and IT
PY	Whether the student studied an orientation year or not	Boolean	Yes or no

GPA1	First semester GPA of the student	Numeric	0 - 5
GPA2	Secondsemester GPA of the student	Numeric	0 - 5
GPA3	Third semester GPA of the student	Numeric	0 - 5
GPA4	Fourth semester GPA of the student	Numeric	0 - 5
Hrs/sem1	Academic load of a student during the first semester	Numeric	12 - 24 hours
Hrs/sem2	Academic load of a student during the second semester	Numeric	12 - 24 hours
Hrs/sem3	Academic load of a student during the third semester	Numeric	12 - 24 hours
Hrs/sem4	Academic load of a student during the fourth semester	Numeric	12 - 24 hours
F/year1	Number of courses failed during the first academic year	Numeric	≥ 0
F/year2	Number of courses failed during the second academic year	Numeric	≥ 0
Prog1	Grade achieved in programming (1)	Nominal	A+, A, B+, B, C+, C, D+, D, and F
Prog2	Grade achieved in programming (2)	Nominal	A+, A, B+, B, C+, C, D+, D, and F
DB	Grade achieved in Database's fundamentals	Nominal	A+, A, B+, B, C+, C, D+, D, and F
NW	Grade achieved in Computer Networks fundamentals	Nominal	A+, A, B+, B, C+, C, D+, D, and F
English1	The student's English grade (1)	Nominal	A+, A, B+, B, C+, C, D+, D, and F
English2	The student's English grade (2)	Nominal	A+, A, B+, B, C+, C, D+, D, and F

4.3.2 Grading classification

In the raw dataset, the final GPA is within the range of 0–5.0, where 5.0 is the best possible GPA score. Table 7 shows the grading classification that is used in this study.

Table 7: Classification of academic grading

GPA	Grade	Symbol	Level
4.5 - 5	Excellent	A	Above average
4.00 – 4.5	Very good	B	Average
3.25 – 4.00	Good	C	Average
2.5 – 3.25	Accepted	D	Average
Less than 2.5	Poor	E	Below average

4.3.3 DM algorithms

According to the literature review (Chapter 3), no single classifier is effective in all situations. In this study, we use eight DM algorithms, three of which are decision tree algorithms, namely, C4.5, SimpleCart, and LadTree. We also select random forest as an ensemble algorithm. Moreover, Naïve Base, K-nearest Neighbor and Artificial Neural Networks have also been explored.

4.3.4 DM tools

We have used the WEKA software package to build and validate the predictive models. As for generating the plots that visualize the relation between features and academic success, we have used python with pandas, NumPy, seaborn, and matplotlib libraries.

4.3.5 Evaluation method

Our study evaluates the prediction models using non-exhaustive cross-validation (10-fold cross-validation). Each time, nine of the folds are used for training, one-fold is used for testing the model, and the holdout method is repeated ten times. we compare the performance of eight DM algorithms in terms of the following:

- The accuracy of the classifier, defined as the number of correct predictions divided by the number of predictions made as a function of the dataset.

- Receiver Operating Characteristic (ROC) is used to measure the performance of classifiers by graphing True Positives and False Positives for every classification threshold (the higher the ROC, the better the results).

- F Measure is a measurement that combines both precision and recall into a single measure that captures both properties (a poor F Measure score is 0.0, whereas a best F Measure score is 1.0).

4.3.6 Research challenges

As mentioned in chapter 2, section 2.5, One of the main challenges that EDM researchers face is regarding the lack of knowledge on the benefits of performing EDM studies. Since EDM is a relatively new field, its potentials are still not clear to the students and general society. Therefore, collecting the required data has been challenging for us. To overcome this challenge, we educated the faculty members on the aim of performing this study and its potential benefits in increasing students' success. We also ensured them that their information will be anonymized and will not be used for any other purposes but this study.

4.4 Feature Importance on the Overall Predictions

In order to understand how different features influence the performance of classifiers, we explored different feature selection techniques in Weka. These include Search-Based, Correlation-Based, Information Gain-Based, and Wrapper with Naïve Bayes. The two types of search methods are BestFirst, which searches the space of attribute subsets by greedy hill climbing augmented with backtracking, or Ranker, which ranks attributes based on their evaluations. The results of the selected methods are presented in Table 8.

Table 8: Feature importance on the prediction models

Feature Evaluator	Description	Search Method	Feature Importance
Search Based (CfsSubsetEval)	Examines the impact of a subset of features by considering their predictive capability and the degree of redundancy between them. Preference should be given to subsets of features that are highly associated with the class but have low intercorrelations.	BestFirst	GPA1 (100%) GPA2 (100%) GPA3 (100%) GPA4 (100%) DB (100%) F/year1 (80%) NW (30%)
Correlation Based (CorrelationAttributeEval)	An attribute is evaluated by measuring its correlation (By Pearson formula) with the target attribute.	Ranker	1. GPA3 2. GPA4 3. F/year2 4. GPA2 5. GPA1 6. F/year1 7. SecGPA 8. DB 9. Hrs/sem4
Information Gain Based (InfoGainAttributeEval)	An attribute's influence is evaluated by measuring the amount of information gained about the target attribute.	Ranker	1. GPA3 2. GPA4 3. DB 4. GPA2 5. GPA1 6. prog1 7. prog2 8. F/year2 9. F/year1
Wrapper with Naïve Bayes (WrapperSubsetEval)	The evaluation of attribute sets is conducted using a learning scheme. An assessment of the accuracy of the learning scheme for a set of attributes is made using cross-validation.	BestFirst	GPA3 (100 %) GPA4 (100 %) GPA1 (90%) NW (90%) F/year1 (70%) DB (60%) Prog1 (40%)

In view of the outcomes presented in table 8, we can conclude that the following attributes are the most predictive of students' academic achievement at the College of Information and Computer Science: GPAs for each semester, the number of failed courses in the first and second years, and the grade they achieved in the 'Database fundamentals' and 'Programming 1' core courses. However, 'English skills', 'secondary school grade', 'academic load', and 'programming 2' do not significantly influence the prediction of students' academic achievement.

To further understand the influence of the attributes on the prediction, we present some examples of the relation between the features that have been used for performing the academic predictions and the final academic grade using scatter plots. Figure 9 shows the relation between the students' GPA in their 3rd and 4th semester vs. their final academic achievement. As can be noticed, the GPA achieved each semester significantly relates to academic success. The honorary students', which are considered "above average", tend to have a high GPA from the beginning of the program. On the other hand, the students at risk of failing, which are classified as "below average", tend to have a weak GPA at an early stage.

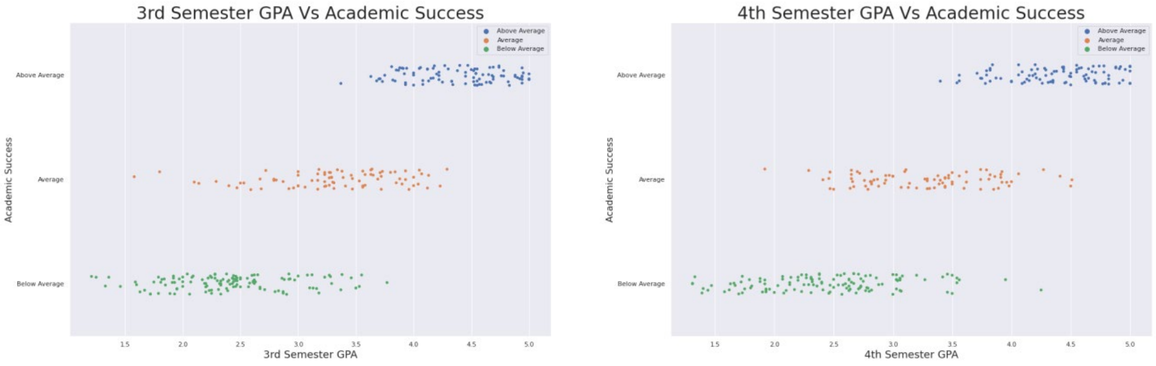


Figure 9: The relation between the achieved GPA in each semester and the academic success

Moreover, figure 10 helps in understanding the influence of the number of failed courses each year on the final academic achievement. The scatter plots show a significant relationship between above-average students and the number of failed courses, as most above-average students have never failed a course. However, although most below-average students have failed at least one course per year, there are a few cases that have not failed. Based on that, we conclude that the number of failed courses is more significant for predicting honorary students than it is for predicting students at risk of failing.

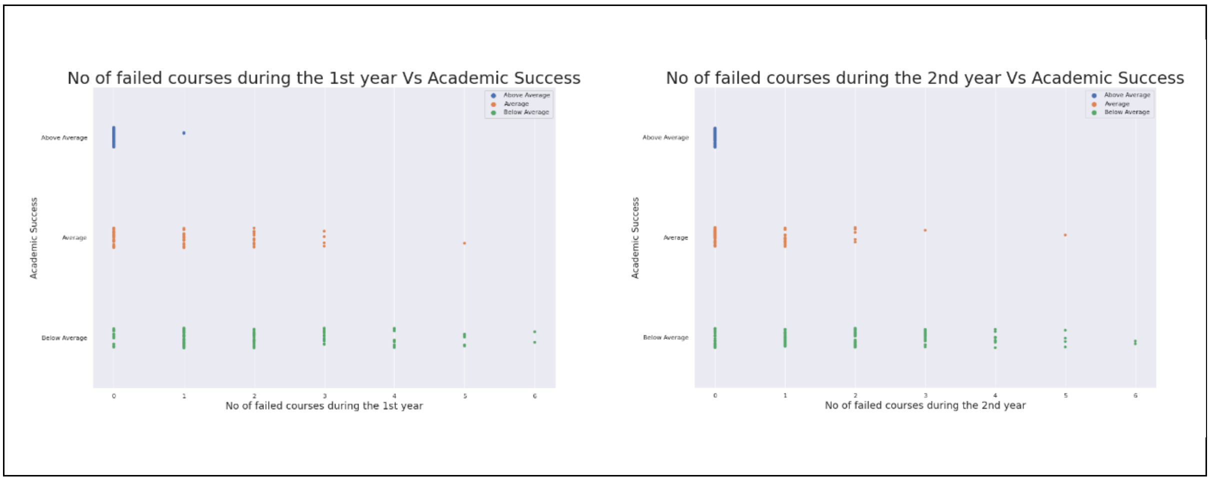


Figure 10: The relation between the number of failed courses each year and academic success

Figures 11, 12, and 13 show the features that were found to have no significant relation to the final academic achievement. For instance, figure 11 represents the relation between the students' language skills and their bachelor's academic achievement. We can notice that there is no clear pattern for each group of students as the grades vary from A+ to C+ for all three groups. Although the number of failures in English courses is zero for the above average students, it is not a good enough predictive feature as the number of failures is significantly low for the rest of the groups.

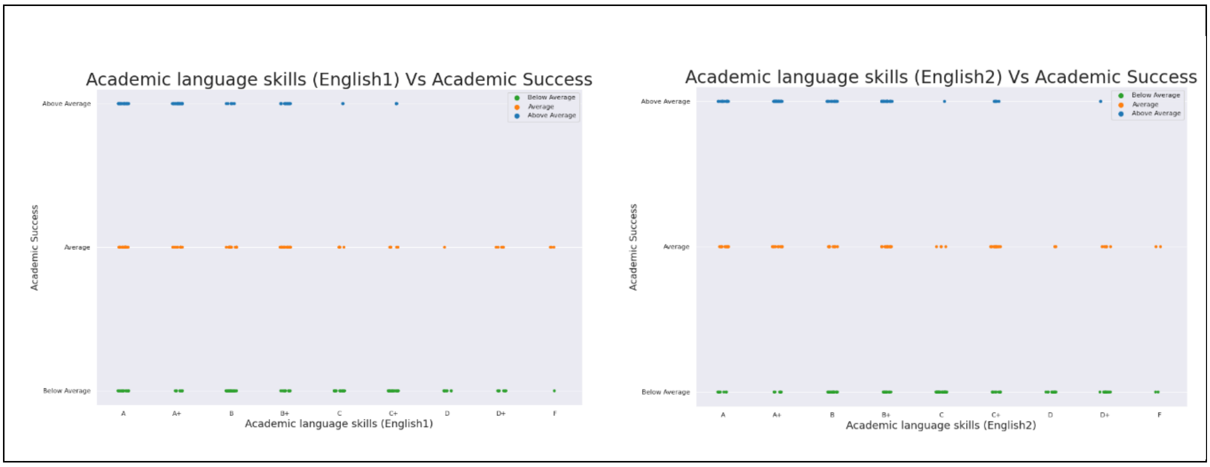


Figure 11: The relation between academic language skills and academic success

Although the previous grade in secondary school has been mostly used by researchers for predicting bachelor's students' academic achievement, we have found it to have no influence on predicting the success of the bachelor's students of the Information and Commuter Science College. Figure 12 shows that all three groups of students had a similar range of grades prior to entering the program. moreover, figure 13 presents the impact of studying an orientation year on academic success. We whiteness that there is no relationship between the two attributes. Therefore, canceling the orientation year from the bachelor's program has no negative impact.

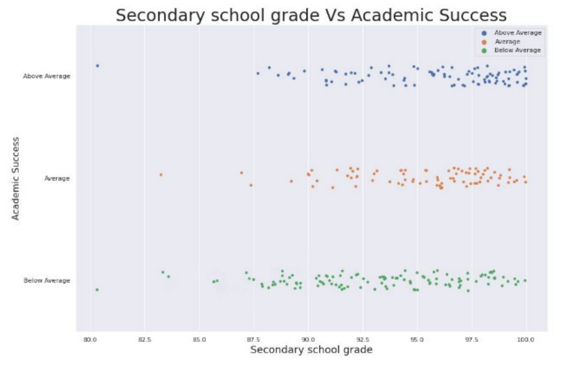


Figure 12: The relation between previous academic achievement in secondary school and academic success

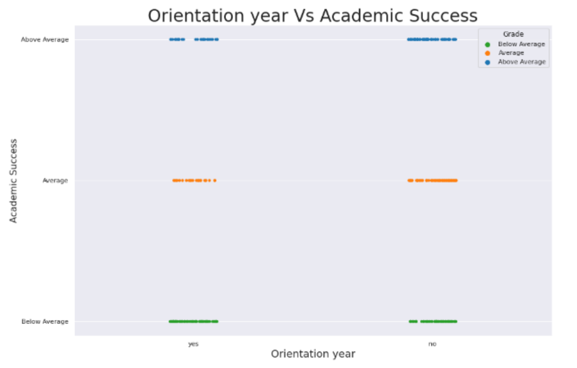


Figure 13: The relation between studying an orientation year and academic success

4.5 Results of the academic Predection Models

This section provides three case studies for predicting students' academic achievement, namely after the second, after the third, and after the fourth semester. In each case, two models are built. The first model uses five classes (Excellent, very good, good, acceptable, and poor), whereas the second uses three classes (Above average, average, and below average). For each model, we compare the outcomes of eight DM algorithms in terms of classification accuracy, ROC, and F1 score for honorary (≥ 4.5 GPA) and at-risk of failing (≤ 2.5 GPA) students.

4.5.1 Predicting the graduation grade after the 2nd semester

In this academic prediction case, four attributes are selected: 1st semester GPA, 2nd semester GPA, the number of failed courses during the 1st year, and the grades of Programming Language (1), which is a core course taken during the first academic year. Table 9 shows the result in predicting students' academic achievement after their 2nd semester.

As a general trend with the performance of all the classifiers, we can see that the accuracy remarkably improves when trying to perform the prediction using three classes only compared to five classes. It is common that more object classes will make the distinction between the classes harder. Although we were able to achieve a high F1 score in case of predicting honorary students, it is not the case for the students at risk of failing. To conclude, performing predictions after the second semester in an 8-10 semester program is challenging at this point of the program, especially when trying to predict a specific grade. For instance, the highest achieved accuracy when performing predictions using five classes is 0.54 by Naïve Bayes.

Table 9: Accuracy of predicting the final academic grade after the 2nd academic semester

DM Algorithm	**Using five classes (Excellent, very good, good, acceptable, and poor)**				**Using three classes (Above average, average, below average)**			
	Accuracy	ROC	F-score (at-risk)	F-score (Honorary)	Accuracy	ROC	F-score (at-risk)	F-score (Honorary)
NB	**0.54**	**0.84**	**0.45**	0.72	**0.74**	**0.89**	**0.80**	**0.83**
KNN	0.48	0.76	0.26	0.73	0.67	0.81	0.74	0.76
SVM	0.49	0.77	0.27	0.74	0.67	0.76	0.77	0.75
MLP	0.51	0.75	0.26	0.70	0.68	0.81	0.77	0.73
J48	0.53	0.77	0.28	0.77	0.71	0.78	0.79	0.82
RF	0.53	0.81	0.26	**0.78**	0.71	0.86	0.78	0.81
SimpleCart	0.51	0.73	0.15	0.78	0.69	0.83	0.75	0.81
LADTree	0.52	0.80	0.24	0.77	0.68	0.85	0.72	0.81

4.5.2 Predicting the graduation grade after the 3rd semester

In this academic prediction case, six attributes are selected: the GPA from the first semester, the GPA from the second semester, the GPA from the third semester, the number of failed courses during the 1st year, and the grades of the two core courses that are taken during the first three semesters, i.e., Programming Language (1), and Database Fundamentals. Table 10 below compares the different classifiers' results in predicting students' academic achievement after their third academic semester. Again, the left side represents performing the predictions using five classes, and the right represents performing the predictions using only three classes.

Just like the previous case, the accuracy remarkably improves when trying to perform the prediction using three classes compared to five classes. Moreover, it can be observed that Naïve Bayes performs the best in general, with accuracy between 0.63 and 0.80 using five classes and three classes, respectively. As we focus on honorary and at-risk of failing students, it is essential to discuss the achieved F1 score for those two classes. Regarding predicting students at risk, Naïve Bayes performs the best in both models. However, when it comes to predicting honorary students, Naïve Base and Random Forest perform similarly. For instance, in the case of predicting using three classes, Random Forest achieved a 0.87 F1 score, and Naïve Base achieved a 0.88 F1 score.

Table 10: Accuracy of predicting the final academic grade after the 3rd academic semester

DM Algorithm	**Using five classes (Excellent, very good, good, acceptable, and poor)**				**Using three classes (Above average, average, below average)**			
	Accuracy	ROC	F-score (at-risk)	F-score (Honorary)	Accuracy	ROC	F-score (at-risk)	F-score (Honorary)
NB	**0.63**	**0.89**	**0.46**	0.84	**0.80**	**0.93**	**0.85**	**0.88**
KNN	0.60	0.81	0.38	0.76	0.73	0.87	0.81	0.78
SVM	0.58	0.83	0.32	0.72	0.77	0.85	0.82	0.84
MLP	0.56	0.80	0.33	0.73	0.73	0.86	0.81	0.79
J48	0.56	0.78	0.18	0.71	0.76	0.85	0.84	0.83
RF	**0.63**	0.87	0.37	**0.85**	0.78	**0.93**	0.83	0.87
SimpleCart	0.60	0.82	0.17	0.76	0.75	0.86	0.83	0.82
LadTree	0.58	0.85	0.39	0.79	0.77	0.89	0.82	0.86

4.5.3 Predicting the graduation grade after the 4th semester

A total of nine attributes are used to predict the graduation grades of students in this academic prediction case: GPAs from semesters 1, 2, 3, and 4, the number of failed courses during the first and second years, the number of failed courses during the second year, and the grade for Programming Language (1), Database Fundamentals, and Computer Network Fundamentals. In Table 11, different DM algorithms are compared for their ability to predict the academic achievement of bachelor's students after their fourth semester.

We witness that Naïve Bayes outperformed all algorithms with an accuracy of 0.70- 0.85using five and three classes, respectively, followed by Random Forest with 0.68- 0.82 accuracy. Like the previous case, Naïve Bayes performs the best in predicting at-risk students in both models. However, Random Forest and naïve base perform quite similarly in predicting honorary students.

Table 11: Accuracy of predicting the final academic grade after the 4th academic semester

DM Algorithm	**Using five classes (Excellent, very good, good, acceptable, and poor)**				**Using three classes (Above average, average, below average)**			
	Accuracy	ROC	F-score (at-risk)	F-score (Honorary)	Accuracy	ROC	F-score (at-risk)	F-score (Honorary)
NB	**0.70**	**0.92**	**0.54**	**0.85**	**0.85**	**0.95**	**0.88**	0.91
KNN	0.63	0.83	0.29	0.76	0.78	0.91	0.85	0.82
SVM	0.64	0.86	0.42	0.77	0.81	0.89	0.86	0.87
MLP	0.56	0.82	0.34	0.66	0.77	0.89	0.81	0.87
J48	0.62	0.82	0.26	0.78	0.79	0.89	0.81	0.91
RF	0.68	0.91	0.38	0.82	0.82	**0.95**	0.84	**0.92**
SimpleCart	0.64	0.83	0.37	0.82	0.80	0.89	0.83	0.91
LADTree	0.63	0.88	0.48	0.78	0.79	0.91	0.78	0.90

4.6 Disscusion and Conclusion

Due to the fact that evaluating bachelor's degrees should be an ongoing process, our first objective was to confirm that eliminating the orientation year from the College of Computer and Information Systems at PNU is a significant change in the college's programs and would not negatively impact the academic success of students. Since previous studies reported that the orientation year can improve students' academic achievement (Davig and Spain, 2003; McMullen, 2014), we hypothesized that the orientation year influence students’ success. However, we found that students' success in the College of Computer and Information Science is not affected by it, as evidenced by the results. Thus, the removal of the orientation year from the study programs was a reasonable decision and did not have adverse consequences for the students.

The second objective of this study was to find the significance of correlation attributes predictors. We used four feature selection methods to reach this objective, and 9 out of 18 proposed features were found to be significant predictors of students' academic achievement. Across all four semesters, earned GPA played a significant role in predicting academic achievement. This is consistent with our hypotheses and the findings of Asif et al. (2017). Nevertheless, the third and fourth semester earned GPAs of students have a greater impact on prediction than the first- and second-semester grades. It is reasonable considering that courses become more challenging as students progress through semesters and their skills begin to vary. The number of failed courses in the first two years of the program has also been found to play a significant role in predicting students' academic success. This is in line with the results of Kabakchieva (2013) and supports our hypotheses. Meanwhile, we have found that academic load does not significantly influence the success of students. This contradicts our hypothesis and the findings of Alemu Yehuala (2015) and Rotem et al. (2020), possibly due to the similar academic workloads that students experience. While prior GPA from secondary school is one of the most widely used factors to predict academic success (Abu Saa, 2016; Aluko et al., 2018; Garg, 2018; Huang & Fang, 2013; Kabakchieva, 2013; Kovačić, 2010; Osmanbegović & Suljic, 2012; Pal & Pal, 2013; Thai-Nghe et al., 2007), we found it to have no significant effect. This contradicts our hypothesis. Additionally, we concluded that English language proficiency does not affect students' success at the College of Computer and Information Science. This supports our hypotheses and in accordance with the findings of Arsad, Buniyamin, and Manan (2014) and Bani-Salameh (2018). We suppose that such results are since the nature of courses in the College of Computer and Information Science is not linguistics and rather more scientific.

The third objective of this study was to predict students’ final grades at a degree level at an early stage of their studying journey, with a particular focus on two groups of students, the at-risk students, and the honorary students. We have built four main models using eight supervised DM algorithms. As a general observation, it is clear that, with the increase of attributes, the models’ accuracy increases as all eight algorithms performed the best (they acquired a higher accuracy, ROC, and F-score) in the third case and performed the worse in the first. This supports the findings of Zimmermann et al. (2011). Moreover, performing predictions using three classes yields much more significant results than when using five classes. This is in line with the results of Nguyen Thai Nghe, Janecek, and Haddawy (2007). As far as which classifier predicted academic achievement better, the results are similar to those of (Asif et al., 2017; Al luhaybi, Tucker, and Yousefi, 2018; Kovačić, 2010; Shakeel and Anwer Butt, 2015) in which Naïve Base performed the best compared to the rest of the used algorithms. Naïve Base produced a general accuracy of 0.54 in its worst case and 0.85 in its best. The basis behind the successful performance of Naïve Bayes is described by Domingos and Pazzani (1996) as follows: "Naïve Bayes is commonly thought to be optimal, in the sense of achieving the best possible accuracy, only when the independence assumption holds, and perhaps close to optimal when the at-tributes are only slightly dependent. However, this very restrictive condition seems to be inconsistent with the Naïve Bayes' good performance in a wide variety of domains, including many where there are clear dependencies between the attributes." The second-best performing algorithm is Random Forest, with accuracy between 0.53 and 0.82. In general, Random Forests produce better results because they are more robust than a single decision tree; by aggregating many decision trees, they are able to reduce errors and overfitting as a result of bias. Moreover, random forests search for the best feature among a random subset of features, while single decision trees search for the most essential feature when splitting a node. A model with such a wide range of diversity is usually more accurate. Nevertheless, binary classification trees, such as CART or J48, allow instances to follow only one path through the tree.

Moving to our focus, which is about the prediction of honorary and at-risk students early in their academic journey, our conclusion is that, by utilizing three class predictions, it is possible to identify those groups of students early in their bachelor's studies at the College of Computer and Information Sciences (80 – 92 F1 score).

4.7 Study Limitations and Future Work

4.7.1 Investigating the use of other predictive features

The present study has several limitations, one of which is that all the participants are females since PNU accepts only female students. To reach more accurate results that can improve students' learning outcomes, this study could be extended to include more distinctive features, including demographics such as gender, and the student behavior following enrollment, such as attendance. Moreover, with the great increase in online classes after the Covid-19 pandemic, it is now possible to collect students’ learning activity information with quite ease.

4.7.2 Investigating the predictive models on other bachelor’s programs

In order to meet time constraints, this study has only included students from a single university and a single college. It may be of interest to examine the role of language skills in social science programs and the role of orientation years in other programs, such as medicine. This will help in better understanding the influence of such feature.

4.7.3 Investigating the use of other DM algorithms

In this research, we explored the use of eight DM algorithms. However, future research could examine other algorithms, such as boosting algorithms. In predictive data analysis, boosting algorithms reduce errors by sequentially training multiple models to enhance the overall accuracy, i.e., multiple weak learners are converted into a single robust learning model.

CHAPTER 5: PREDICTING THE ACADEMIC ACHIEVEMENT OF MASTER'S DEGREE STUDENTS

The privilege of university education is a great one; the more widely it is extended, the better for any country. -Winston Churchill

5.1 Introduction

Pursuing a master's degree is considered a well-established postgraduate qualification in higher education. It supports building students' current abilities and helps them acquire new skills related to a particular profession. With the increasing interest in pursuing a master's degree worldwide, there is a growing number of failures and dropouts. The drop-out rate for master's programs in Germany reached 15% for German students and 28% for international students (Kercher, 2018). In most German universities, students are not closely monitored and are allowed to stay enrolled for extended periods without progressing towards completing their degree program (Berens et al., 2019). Although many support programs aim to reduce student attrition at German universities, those programs are not explicitly targeted at the group of students at risk of not completing their degree but are offered to the general student body (Berens et al., 2019).

The power of EDM can provide tremendous insights, not only for undergraduate students but for postgraduates as well. To minimize wasting financial and human resources caused by failure or dropouts, it is vital to build models that can predict dropouts at the earliest stage possible. Although the attrition rates of master's students are widely reported, no solid predictive models exist (Rotem et al., 2020). The literature review in chapter three shows that the number of EDM studies covering postgraduate degrees is unfortunately limited. Nevertheless, none of those master's studies have been performed in Germany. Since the feature sets used for performing academic predictions differ from one country to another and from one degree to another, it is vital to investigate which feature sets are important for the academic predictions performed at a master's degree level in German universities. Moreover, the literature shows that most researchers still rely on traditional machine learning algorithms for performing academic predictions. For instance, boosting algorithms are still not common, as only one study by Smirani et al. (2022) investigated using XGBoost and LightGBM in predicting student failure at a bachelors' level.

In chapter 2, section 2.3, we mentioned the four main applications of EDM. This chapter mainly falls under the umbrella of two EDM applications, student modeling and scientific research. We present academic predictions that are intended to support the Business Informatics master's program at the University of Mannheim in Germany. Along with exploring the traditional DM algorithms, we investigate three boosting algorithms not only for predicting students at risk of not completing their degree but for predicting students' academic performance using three classes as well. The key objectives of this study are: (i) predicting the students' academic achievement at an early level of their master's program, (ii) finding methods to deal with the imbalanced students' dataset, and (iii) figuring out which features are mostly correlated to the predictions performed on the master students of the Business Informatics program. For those objectives, we hypothesize the following:

- It is likely to have reasonably accurate predictions after the first and second semesters of the students' enrollment.
- It is likely to have accurate predictions using multi-class classification.
- Student demographics could have a role in predicting academic achievement.
- Students' behavior after enrollment (e.g., academic load, number of failed courses) can predict academic achievement.
- Distance from students' accommodation to university influences academic performance.

The rest of this chapter is organized as follows: we start by providing a general idea of the Business Informatics master's program offered by the School of Business Informatics and Mathematics at the University of Mannheim. Following that, we present the research methodology that has been adopted to perform this study. Then, we cover the importance of each feature on the academic predictions. Afterward, a comparison between the different academic prediction models is provided. Finally, we discuss the shortcomings of this study and outline future lines of research.

5.2 Overview of the Business Informatics Master's Program

The Business Informatics program is highly interdisciplinary, combining aspects of Informatics with Business Administration. A number of courses are offered in Computer Science, Data Analytics, Business, and Mathematics. The program is intended to last for four semesters (two years), with approximately 120 European Credit Transfer System. However, it usually takes students up to six semesters to complete the degree (Figure 14). The most important knowledge that the applicant should have before starting the master program are in the area of linear algebra, probability and statistics, databases, algorithms and programming, logics and combinatorics and management of enterprise systems. For viewing the master's courses that rely on those topics, see Appendix B. Unfortunately, students often suffer from insufficient background knowledge (i.e., incredibly diverse background, from pure computer science to management education) to successfully attend specific courses, leading to a lot of friction and dissatisfaction among both students and instructors.

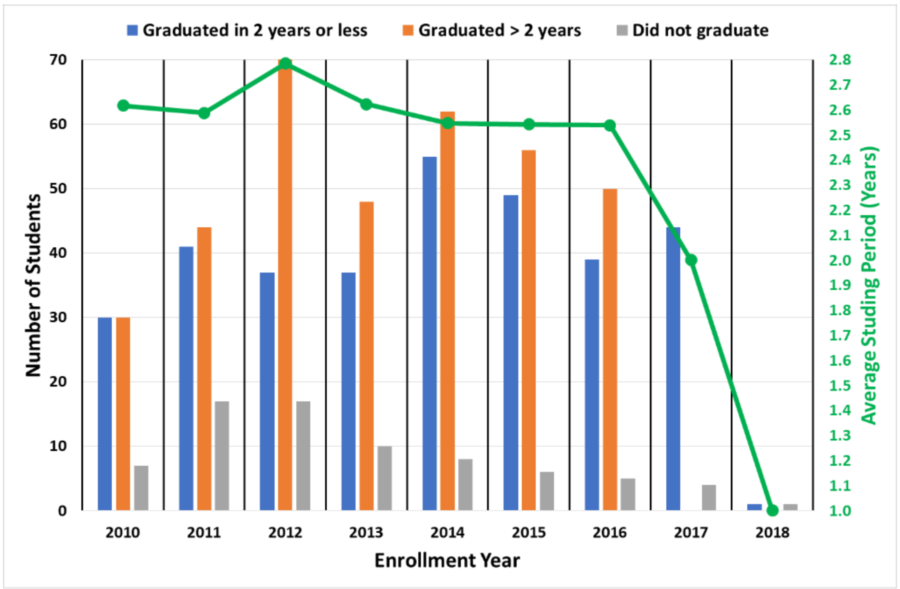


Figure 14: Students' studying duration at the Business Informatics master's program

5.3 Research Methodology

This section presents an overview of the type of collected data, the data analysis, the used DM algorithms, and evaluation methods.

5.3.1 Students' dataset

The data set of 700 students used in this study has been obtained from the Business Informatics and Mathematics faculty at the University of Mannheim from 2010 till 2018. To ensure data efficiency, we excluded the students that did not graduate prior to the Covid-19 pandemic from our study. The exclusion is because the nature of examination and learning style drastically changed, e.g., examinations are performed online. Table 12 describes the type of data collected to perform our prediction study.

Table 12: Description of the collected data that is used to predict the academic achievement

Feature	Description	Type	Value
Academic_status	Whether the student completed the degree or not	Nominal	Completed, and Not_completed
Academic_grade	Student's final achieved grade	Nominal	Above average, Average, and Below average
Gender	Student's gender	Nominal	Male, Female
Enrollment_age	Student's age at the time of enrollment	Numeric	21-38
Culture	Student's culture	Nominal	Collectivistic, and Individualistic
Distance	Distance from accommodation to the university campus	Numeric	≥ 1km

Grade_sem1	Student’s average grade in the 1st academic semester	Numeric	1 – 5
Grade_sem2	Student's average grade in the 2nd academic semester	Numeric	1 – 5
F_sem1	Number of failed courses in the 1st semester	Numeric	≥ 0
F_sem2	Number of failed courses in the 2nd academic year	Numeric	≥ 0
Unregistered_exams1	The number of courses that have been studied in the 1st semester, however, did not take the exam	Numeric	≥ 0
Unregistered_exams2	The number of courses that have been studied in the 2nd semester, however, did not take the exam	Numeric	≥ 0
Registered_exams1	The number of courses that have been examined in the 1st semester	Numeric	≥ 0
Registered_exams2	The number of courses that have been examined in the 2nd semester	Numeric	≥ 0

5.3.2 Analysis of the students’ dataset

Before performing the academic achievement predictions, it is essential to analyze the dataset. As shown in figure 15, the number of male students significantly exceeds the number of females. Furthermore, the number of students coming from individualist cultures slightly exceeds those from collectivistic cultures. We can also notice that most enrolled students are 24 and 23 years old, and only very few are in their thirties. Regarding students 'performance (Figures 16 and 17), we can see that most enrolled students passed the master's program. However, a considerable amount of failure and dropout needs to be given attention. Moreover, the “Above average” students represent the largest number of students, followed by the “Average” students, then finally the “Below average” students.

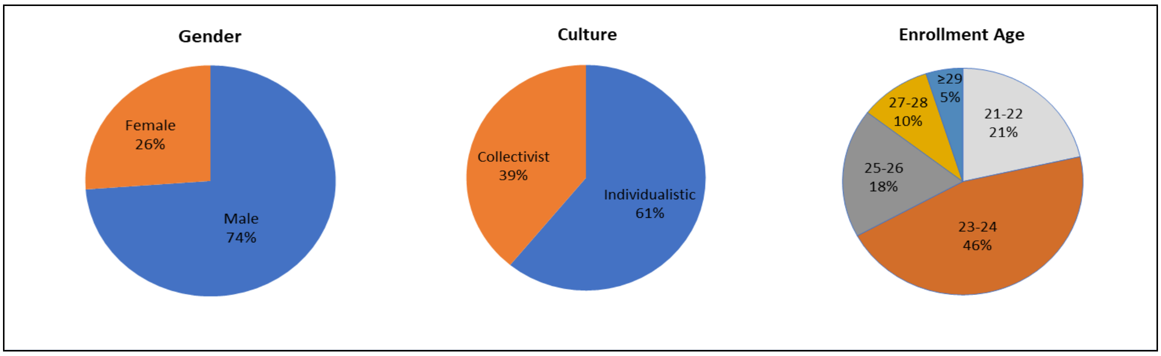


Figure 15: Students' demographical features

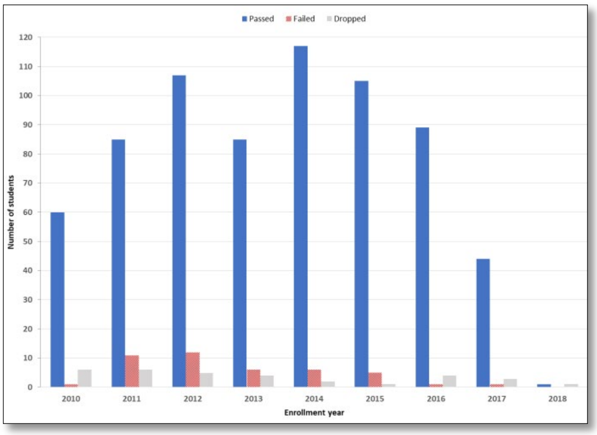


Figure 16: Students' academic status

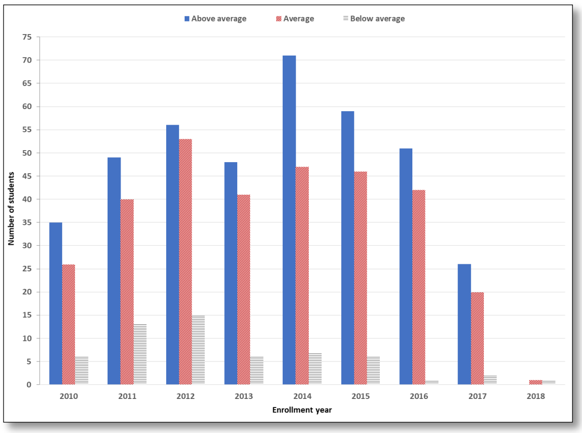


Figure 17: Students’ academic grade

5.3.3 DM algorithms

In this study, we investigate nine DM algorithms as follows:

- Logistic Regression: We use Binary logistic regression in the cases where the dependent feature has only two possible outcomes (completed and not completed the degree) and Multinomial logistic regression, where the dependent feature has three possible outcomes (above average, average, and below average). In both types of predictions, we use the default parameters.

- Random Forest: Our study uses the default parameters of Random Forest (“n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min _samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_node s=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_jobs=Non e, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha =0.0, max_samples=None”)

- K-Nearest Neighbour: Grid-search function has been used to optimize candidate algorithms. This function performs an iterative search to find the optimal hyperparameter values for a particular learning algorithm. We use grid search with (K= 3, 5, 7, and 9) and reported the results of K= 5.

- Artificial Neural Networks: We use the default parameters of multi-layer perceptron classifier (MLP) (1 input layer, 1 hidden layer, 100 units for each hidden layer, 1 output layer, learning_rate= ’constant’, and activation= ’relu’).

- Naive Baise: We use the Gaussian NB, meaning that the likelihood of the features is assumed to have a Gaussian distribution (Normal distribution). We have used the default parameters: priors=None, and var_smoothing=1e-09.

- Support Vector Machine: We tested "Linear", "Radial Basis Function (rbf)", "sigmoid (sgd)", and "Polynomial (poly)" as kernels. However, we report the results of poly SVM where the model tries to maximize the width of the margin between classes using a polynomial class boundary (“C=1.0, kernel='poly', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False, random_state=None”).

- Gradient boosting: we use the default parameters (“loss='log_loss', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0, max_depth=3, min_impurity_decrease=0, init=None, random_state=None, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, validation_fraction=0.1, n_iter_no_change=None, tol=0.0001, ccp_alpha=0”).

- Extreme gradient boosting: we use the default parameters: (“max_depth=3, learning_rate=0.1, n_estimators=100, silent=True, objective='binary:logistic', booster='gbtree', n_jobs=1, nthread=None, gamma=0, min_child_weight=1, max_delta_step=0, subsample=1, colsample_bytree=1, colsample_bylevel=1, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, base_score=0.5, random_state=0, seed=None, missing=None”).

- Light gradient boosting: (“boosting_type='gbdt', num_leaves=31, max_depth=-1, learning_rate=0.1, n_estimators=100, subsample_for_bin=200000, objective=None, class_weight=None, min_split_gain=0.0, min_child_weight=0.001, min_child_samples=20, subsample=1, subsample_freq=0, colsample_bytree=1, reg_alpha=0, reg_lambda=0, random_state=None, n_jobs=None, importance_type='split'”).

5.3.4 Programming language

All phases were performed on the Anaconda 4.13.0 (a free OS-independent platform) distribution with Python version 3.8.8. The libraries that have been used in Python are Scikit-learn (ML algorithms), Pandas (to import and build Data Frames), NumPy (array computing), Matplotlib and Seaborn (data visualization), imblearn (imbalanced data manipulation), xgboost and lightgbm (ML algorithms).

5.3.5 Evaluation method

Our study validates the prediction models using 10-fold cross-validation. For evaluation, we consider four measures which are precision, recall, F1 score, and overall accuracy, explained as follows:

- **Precision:** the ratio of correctly predicted positive observations to the overall predicted positive observations. It is calculated as precision= (TP)/ (TP+FP).

- **Recall:** the ratio of correctly predicted positive observations to the total observations in an actual class. It is calculated as recall= (TP)/ (TP+FN).

- **F1 score:** the weighted average of Precision and Recall. It is calculated as F1 score= (2 * Precision * Recall) / (Precision + Recall).

- **Accuracy:** the correctness of value, i.e., the ratio of correctly predicted observation to the total observations. It is calculated as accuracy= (TP+TN)/ (TP+TN+FP+FN).

Where: TP = True positive; FP = False positive; TN = True negative; FN = False-negative.

5.3.6 Research challenges

As mentioned in chapter 2, there are some common challenges in the field of EDM. The main challenge we faced in performing this research was regarding data collection and students' personal privacy, especially with the new laws and directives enforced in 2018 as part of the European General Data Protection Regulation (GDPR). The new laws impose more restrictions regarding data handling. Although students' data are anonymized, getting authorization to work on such dada is a significant concern. Many measures have been taken to overcome this challenge. Those measures include using an encrypted device with no Internet access. Moreover, students' anonymized information cannot be accessed without a decryption code.

5.4 Feature Importac on the Overall Predictions

Feature Importance refers to the techniques that calculate a score for each input feature for a given model where the scores represent the "importance" of each feature. A higher score means that the specific feature will have a more significant effect on the predictive model. There are various functions for generating feature importance in python in which we have explored some of them. However, since Random Forest provided the best predictive accuracy, it is reasonable to present the impact of each feature on the predictions performed by that classifier.

The Random forest permutation importance measurement, which Breiman (2001) introduced, loops through each column in the dataset, shuffles the particular column, and performs predictions with the shuffled column. The error term should increase if a column is significant to making predictions. In other words, the most important columns are those that result in a maximum error increase (loss function). Sections 5.4.1 and 5.4.2 show the feature importance results after the first and second semesters, respectively, using the Random Forest built-in function. For viewing the results of the permutation feature importance using the rest of the algorithms, see Appendix C.

5.4.1 Feature importance on the predictions performed after the first semester

By viewing table 13, one can notice that the most significant attribute for performing the predictions after the first semester is "Grade_sem1", followed by "Distance", then "Culture". Moreover, “Registered_exams1” and “F_sem1” have a small impact. On the other hand, "Gender", "Enrollment_age", and "Unregistered_exams1" have the most negligible impact on the prediction.

Table 13: Features level of importance of the predictions performed after the 1st semester

Feature	Importance measure	
	Predicting the Academic_status after the 1st semester	**predicting the Academic_grade after the 1st semester**
Grade_sem1	**0.27**	**0.36**
Distance	0.18	0.14
Culture	0.17	0.12
Registered_exams1	0.10	0.10
F_sem1	0.10	0.09
Enrollment_age	0.08	0.07
Unregistered_exams1	0.05	0.06
Gender	0.05	0.04

5.4.2 Feature importance of the predictions performed after the second semester

Table 14 shows that the most significant attributes for performing the predictions after the second semester are "Grade_sem2" followed by "Grade_sem1". While "Culture", "Distance, and F_sem2" affect the prediction, the rest of the features have no significant impact.

Table 14: Features level of importance on the predictions performed after the 2nd semester

Feature	Importance measure	
	Predicting the Academic_status after the 2nd semester	**predicting the Academic_grade after the 2nd semester**
Grade_sem2	**0.30**	**0.33**
Grade_sem1	0.14	0.22
Culture	0.11	0.07
Distance	0.10	0.06
F_sem2	0.09	0.06
Registered_exams2	0.06	0.05

Registered_exams1	0.05	0.05
F_sem1	0.05	0.04
Enrollment_age	0.03	0.04
unregistered_exams2	0.03	0.03
unregistered_exams1	0.02	0.03
Gender	0.02	0.02

5.5 Prediction Models

In this section, we present the results obtained from using the nine DM algorithms that have been described in section 5.3.3. The academic predictions represented in this section are of two types, binary and multi-class classification.

5.5.1 Predict students' academic status using binary classification

Table 15 compares the performances of the different DM algorithms that have been used for predicting students’ academic status, which is a binary classification. Typically, binary classification involves two classes, one of which is in a normal state and the other one which is in an abnormal state. In our case, “*completed*” is the normal state and “*not_completed*” is the abnormal.

We can notice that all the DM algorithms generally provide good accuracy ranging between 88 and 92 in the case of performing the predictions after the first studying semester and between 89 and 94 when performing the predictions after the second semester. However, since we are classifying imbalanced classes, the “accuracy” metric should not be given attention in the Confusion Matrix. Instead, we should consider other matrixes such as the precision, recall and F1 score. When doing so, we can see a significant difference between the results of the majority class (Completed) and the results of the minority class (not_completed). For instance, in the case of Logistic Regression, the precision, recall, and the F1 score reached 0.91, 0.98, and 0.95, respectively, for the “Completed” class. On the other hand, the precision, recall, and F1 score are 0.48, 0.14, and 0.21 for the “Not_completed” class. To explain this typical scenario, most ML algorithms assume that data are equally distributed. As a result, when dealing with imbalanced classes, there is a tendency for ML classifiers to be biased towards the majority class, resulting in poor classification of minorities

Table 15: Performance of the different DM algorithms in predicting the completion of a degree

DM Algorithm	Prediction		After the 1st semester				After the 2nd semester			
	Performance Measure		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
LR	Class	Completed	0.91	0.98	0.95	0.90	0.94	0.95	0.95	0.90
		Not_completed	0.48	0.14	0.21		0.40	0.34	0.37	
RF	Class	Completed	0.93	0.98	0.96	0.92	0.95	0.98	0.97	0.94
		Not_completed	0.68	0.35	0.46		0.70	0.42	0.53	
KNN (K=5)	Class	Completed	0.93	0.98	0.95	0.91	0.94	0.98	0.96	0.93
		Not_completed	0.57	0.28	0.38		0.68	0.37	0.48	
NB	Class	Completed	0.95	0.92	0.93	0.88	0.97	0.91	0.94	0.89
		Not_completed	0.41	0.54	0.45		0.41	0.69	0.51	
SVM (poly)	Class	Completed	0.91	0.98	0.94	0.91	0.94	0.98	0.97	0.94
		Not_completed	0.26	0.10	0.14		0.70	0.35	0.48	
ANN	Class	Completed	0.93	0.97	0.95	0.91	0.95	0.99	0.96	0.93
		Not_completed	0.53	0.34	0.41		0.70	0.37	0.48	
GBM	Class	Completed	0.94	0.98	0.96	0.92	0.95	0.97	0.96	0.93
		Not_completed	0.50	0.26	0.34		0.60	0.47	0.53	
XGBoost	Class	Completed	0.94	0.98	0.96	0.92	0.95	0.98	0.97	0.94
		Not_completed	0.60	0.29	0.39		0.69	0.44	0.53	
LightGBM	Class	Completed	0.94	0.98	0.96	0.92	0.95	0.98	0.97	0.94
		Not_completed	0.55	0.29	0.38		0.69	0.44	0.53	

Although all of the explored classifiers performed poorly in terms of predicting the minority class (“not_completed”), Random Forest dealt better with 46-53 F1 score. A Random Forest classifier is usually better suited to deal with imbalanced data for two main reasons. Since it is capable of including class weights, it is cost-sensitive, thus penalizing misclassifications of minority classes. A second feature of this approach is the combination of sampling and ensemble learning, which entails down sampling the majority class and growing trees based on a more balanced sample of data. Although there is a significant difference between the performance of Random Forest and the rest of the classifiers, it is not reliable enough for practical implementation unless using methods of dealing with imbalanced data.

5.5.2 Predict students' academic grade using multi-class classification

Table 16 compares the performances of the different DM algorithms used to predict the academic achievement after the first and second studying semesters using three classes. Unlike binary classification, multi-class classification does not take into account normal and abnormal results. The examples are classified as belonging to one of several classes that have been identified. Just like the previous cases (5.5.1), the model works best in predicting the majority class, which is the "Above average" students, followed by the second major class ("Average"), then finally the "below average", which is the minority class. With such results, we can conclude that balancing the data is necessary before implementation.

Table 16: Performance of the different DM methods in predicting students' academic grade

DM Algo	prediction		After the 1st semester				After the 2nd semester			
	Performance Measure		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
LR	Class	Above average	0.83	0.85	0.84	**0.79**	0.85	0.85	0.85	0.80
		Average	**0.74**	**0.79**	0.76		0.76	0.79	0.77	
		Below average	**0.62**	0.22	0.33		**0.71**	0.49	0.58	
RF	Class	Above average	0.83	0.84	0.84	0.77	**0.87**	0.85	0.86	**0.81**
		Average	0.71	0.77	0.73		0.75	**0.82**	0.78	
		Below average	0.60	0.32	0.41		0.67	0.36	0.46	
KNN (K= 5)	Class	Above average	0.80	0.82	0.81	0.74	0.85	0.85	0.85	0.79
		Average	0.68	0.72	0.70		0.74	0.79	0.76	
		Below average	0.57	0.28	0.38		0.64	0.36	0.46	
NB	Class	Above average	0.82	0.86	0.84	0.76	0.82	0.86	0.84	0.78
		Average	0.73	0.69	0.71		0.76	0.68	0.72	
		Below average	0.48	**0.44**	**0.46**		0.54	**0.69**	**0.61**	
SVM (poly)	Class	Above average	0.79	**0.90**	0.84	0.76	0.83	**0.90**	0.86	0.80
		Average	0.73	0.69	0.70		**0.78**	0.75	0.76	
		Below average	0.40	0.04	0.08		0.69	0.40	0.51	
ANN	Class	Above average	0.81	0.83	0.82	0.75	0.83	0.83	0.83	0.77
		Average	0.69	0.73	0.71		0.71	0.74	0.73	
		Below average	0.57	0.28	0.38		0.55	0.38	0.45	
GBM	Class	Above average	0.84	0.84	0.84	0.77	0.86	0.84	0.85	0.79
		Average	0.72	0.77	0.74		0.74	0.73	0.76	
		Below average	0.44	0.27	0.33		0.53	0.42	0.49	
	C	Above average	**0.86**	0.84	**0.85**		**0.87**	0.84	0.85	

XGBoost		Average	0.73	0.80	0.79	**0.79**	0.74	0.79	0.76	0.80
		Below average	0.60	0.27	0.37		0.54	0.42	0.48	
LightGBM	Class	Above average	0.83	0.84	0.83	0.77	0.86	0.85	0.86	0.80
		Average	0.71	0.75	0.73		0.75	0.78	0.76	
		Below average	0.50	0.27	0.35		0.59	0.49	0.54	

5.6 Dealing with Imbalanced Data Using SMOTE

By viewing the results in section 5.5.1 and 5.5.2, one can notice that the classifiers generally achieved high accuracy. However, low precession, recall, and F1 score for the minority classes. These misleading results are typical when analyzing imbalanced data. Several techniques have been proposed to solve the problems associated with learning from imbalanced data. Those techniques include (i) resampling (by either oversampling the minority class or under-sampling the majority class), (ii) generating synthetic samples, (iii) cost-sensitive learning, which focuses on assigning different costs to the misclassification errors that can be made, then using specialized methods to take those costs into account, and (iv) collecting more data.

Since we have a limited dataset and it is not possible to collect more data, over-sampling is the optimal approach. Over-sampling simulates data points to enhance balance across the classes. There are several over-sampling techniques. Our study explores using Synthetic Minority Oversampling Technique (SMOTE), which was proposed to improve random oversampling as it overcomes the overfitting problem posed by random oversampling (Chawla et al. 2002). In SMOTE, new minority instances are synthesized between existing (real) minority instances. For each example of the minority class, synthetic training records are generated by selecting one or more of the k-nearest neighbors. Data is then generated by selecting features randomly between those two data points. Once the data have been oversampled, they are reconstructed, and classification models can be applied to the reconstructed data. Tables 17 and 18 show the significant improvements in predicting the minority classes after applying SMOTE.

In general, ensemble algorithms produce better results than single classifiers. Regarding which ensemble algorithm performed the best, Random Forest and LightGBM performed quite similarly as they both produced the same accuracy when performing binary classifications (Table 17). However, Random Forest slightly outperformed LightGBM in terms of predicting the minority classes after the first semester as it gave a higher recall and F1 score, 0.93 and 0.90 respectively.

Table 17: Performance of the DM methods in predicting the completion of a degree using SMOTE

DM Algo	Prediction		After the 1st semester				After the 2nd semester			
	Performance Measure		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
LR	Cl	Completed	0.82	0.79	0.81	0.81	0.84	0.84	0.84	0.84

		Not_completed	0.80	0.83	0.81		0.84	0.84	0.84	
RF	Class	Completed	**0.92**	0.88	**0.90**	**0.90**	**0.93**	**0.91**	**0.92**	**0.92**
		Not_completed	0.88	**0.93**	**0.90**		**0.91**	**0.94**	**0.92**	
KNN (K=5)	Class	Completed	0.77	0.76	0.77	0.77	0.85	0.83	0.84	0.84
		Not_completed	0.76	0.78	0.77		0.83	0.86	0.84	
NB	Class	Completed	0.82	0.78	0.80	0.81	0.84	0.84	0.84	0.84
		Not_completed	0.79	0.83	0.81		0.84	0.84	0.84	
SVM (poly)	Class	Completed	0.76	0.83	0.79	0.78	0.82	0.87	0.84	0.84
		Not_completed	0.81	0.74	0.77		0.86	0.81	0.83	
ANN	Class	Completed	0.84	0.80	0.82	0.82	0.88	0.80	0.84	0.84
		Not_completed	0.81	0.85	0.83		0.81	0.89	0.85	
GBM	Class	Completed	0.90	0.88	0.89	0.89	**0.93**	0.90	0.91	0.91
		Not_completed	0.88	0.91	0.89		0.90	0.93	0.91	
XGBoost	Class	Completed	0.91	0.88	0.89	0.89	**0.93**	0.90	0.91	0.91
		Not_completed	0.88	0.91	0.89		0.90	0.93	**0.92**	
LightGBM	Class	Completed	0.91	**0.91**	0.89	**0.90**	**0.93**	**0.91**	0.91	**0.92**
		Not_completed	**0.90**	0.91	0.88		**0.91**	**0.94**	**0.92**	

As for the multiclass predictions (Table 18), Random Forest produced the highest accuracy after both first and second semester, 0.82 and 0.87 respectively. It also produced the highest precision, recall, and F1 score for all the classes (“above average”, “average”, and “below average”) in most cases. The second best performing classifier is LightGBM, with accuracy between 0.80 and 0.86.

Table 18: Performance of the DM methods in predicting the academic grade using SMOTE

DM Algo	prediction		After the 1st semester				After the 2nd semester			
	Performance Measure		Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
LR	Class	Above average	**0.83**	0.84	0.83	0.75	0.86	0.84	0.85	0.81
		Average	0.65	0.64	0.65		0.71	0.73	0.72	
		Below average	0.76	0.76	0.76		0.86	0.85	0.85	
RF	Class	Above average	**0.83**	0.84	**0.84**	**0.82**	**0.89**	0.85	**0.87**	**0.87**
		Average	**0.75**	**0.72**	**0.74**		**0.80**	**0.81**	**0.81**	
		Below average	**0.88**	**0.91**	**0.89**		**0.92**	**0.94**	0.93	
KNN	Class	Above average	0.76	0.77	0.77	0.70	0.83	0.80	0.81	0.81
		Average	0.63	0.58	0.60		0.73	0.72	0.73	

(K= 5)		Below average	0.71	0.76	0.74		0.87	0.91	0.89	
NB	Class	Above average	0.81	0.85	0.83	0.74	0.81	0.86	0.84	0.80
		Average	0.63	0.58	0.60		0.71	0.68	0.69	
		Below average	0.77	0.78	0.73		0.87	0.86	0.86	
SVM (poly)	Class	Above average	0.74	**0.85**	0.80	0.70	0.81	**0.86**	0.84	0.80
		Average	0.59	0.55	0.57		0.72	0.69	0.70	
		Below average	0.76	0.69	0.72		0.88	0.84	0.86	
ANN	Class	Above average	0.81	0.80	0.80	0.71	0.85	0.80	0.82	0.78
		Average	0.58	0.61	0.59		0.65	0.74	0.69	
		Below average	0.74	0.72	0.73		0.87	0.79	0.83	
GBM	Class	Above average	0.82	0.83	0.83	0.79	0.87	0.84	0.85	0.85
		Average	0.70	0.70	0.70		0.76	0.79	0.78	
		Below average	0.86	0.85	0.86		0.92	0.91	0.91	
XGBoost	Class	Above average	**0.83**	0.83	0.83	0.79	0.87	0.84	0.85	0.85
		Average	0.69	0.69	0.69		0.77	0.80	0.78	
		Below average	0.84	0.84	0.84		0.92	0.91	0.92	
LightGBM	Class	Above average	0.82	0.84	0.83	0.80	0.87	0.85	0.86	0.86
		Average	0.73	0.70	0.72		0.79	0.80	0.80	
		Below average	0.86	0.87	0.87		**0.92**	**0.94**	**0.94**	

5.7 Discussion and Conclusion

Although former studies in EDM used an extensive range of features for predicting students' academic achievement (in terms of (i) achieved grades or (ii) passing and failing), those features are sometimes not obtainable for practical usage, and therefore, the prediction models are not feasible for employment. In this study, we used easy to collect attributes that any institute can obtain. The first objective of performing this study was to accurately predict the academic achievement of master's students at an early stage. We have built four initial models, two are designed to make predictions after the first studying semester, and two are designed to perform the predictions after the second semester.

By going back to Tables 15 and 16, which represent the initial models, we can notice that the results of predicting the largest classes (“Complete” and “Above_average”) are better than the rest of the classes (“Not_completed”, “Average”, and “Below_average”). This finding was also reported by Nguyen Thai Nghe et al. (2007). The question that comes to mind is whether the models are reliable for practical usage. Although we achieved high prediction accuracy in all four models, they are misleading results and unreliable for implementation. That is because other evaluation methods, such as the precision, recall, and F1 score for the minority classes, are not sufficient enough. We worked on that issue by using SMOTE (Tables 17 and 18).

Figures 18 and 19 are examples that compare the F1-score of the minority class (Not_completed) before using SMOTE and after using it when performing the predictions after the first and second studying semester. Although the accuracy of the classifiers slightly decreased, they are more applicable as we were able to have high precision, recall, and F1 score for the minority classes.

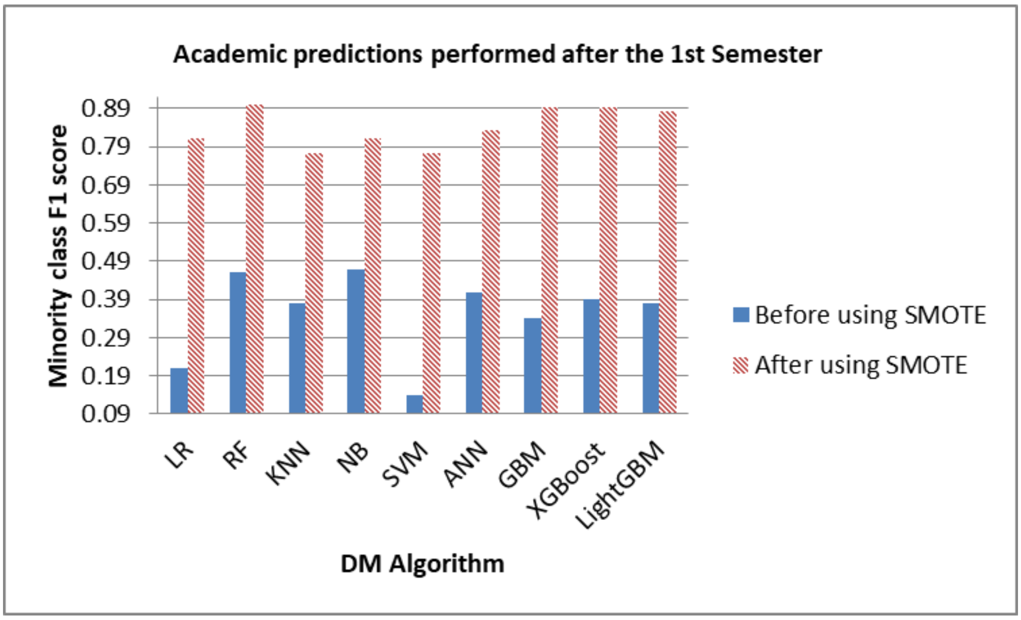


Figure 18: A comparison between the achieved F1 scores after the 1st semester for the minority class (Not_completed) using DM methods with and without SMOTE

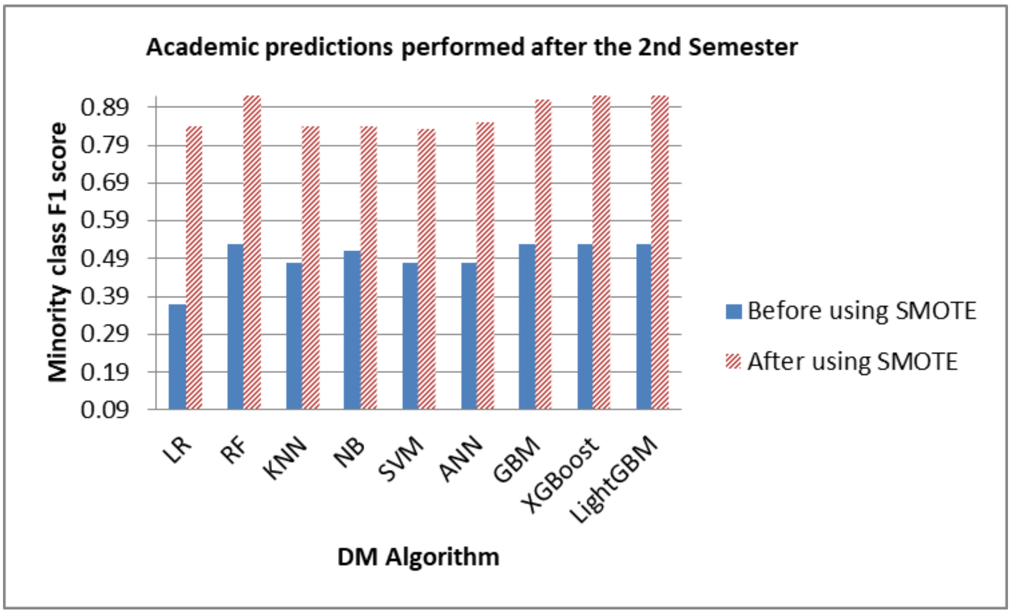


Figure 19: A comparison between the achieved F1 scores after the 2nd semester for the minority class (Not_completed) using DM methods with and without SMOTE

Moreover, as a general trend, the predictions performed after the second semester yield more significant results than those performed after the first semester (Figures 20 and 21). It is reasonable as we have a more realistic vision of the students' performance after the second semester than we do after the first. We can also notice that predicting academic status is more accurate than predicting graduation grade. This supports the findings of Nguyen Thai Nghe et al. (2007), who reported that predicting two-class problems produces more accurate results than predicting three or more class problems (i.e., the more the classes, the more challenging the prediction is). To get into more details regarding the performance of the classifiers, we can see that they gave similar accuracies; however, bagging (Random Forest) and boosting (GBM, XGboost, and LightGBM) ensemble algorithms provided better results. Ensemble algorithms can be more accurate than single models as they tend to repeat the process many times such that the model learns the data and makes proper predictions. For instance, Random Forest produced an accuracy between 0.90 and 0.92 in the case of predicting the academic status and 0.82- 0.87 in the case of predicting the graduation grade. This is not surprising Since Random Forest uses both bagging and decision trees to form the ensemble method. Another reason behind the excellent performance of Random Forest is that it chooses features randomly during the training phase. Hence, it does not depend highly on any specific set of features. This randomized feature selection is a unique trait of Random Forest.

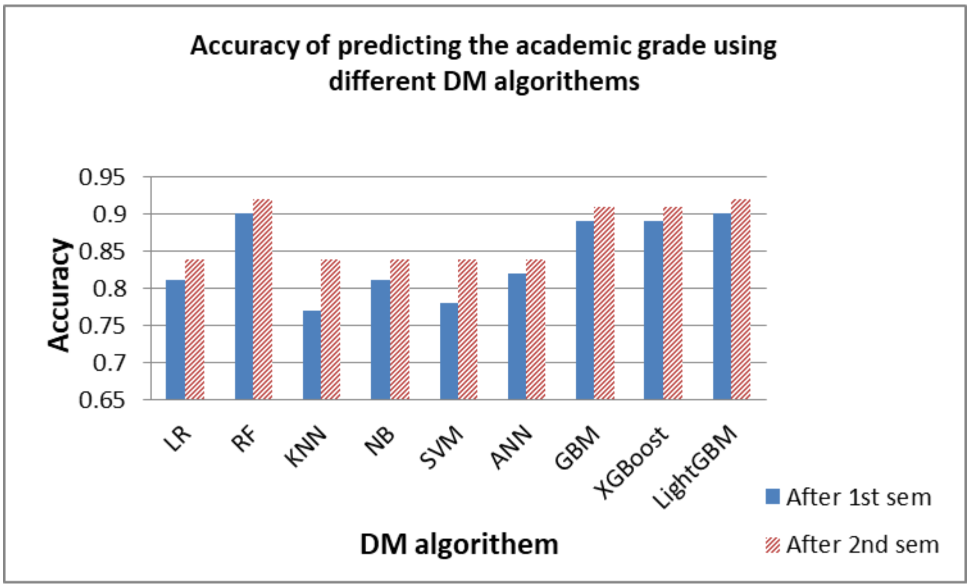


Figure 20: A comparison between the accuracy of the academic predictions after the 1st and 2nd semester using binary classification

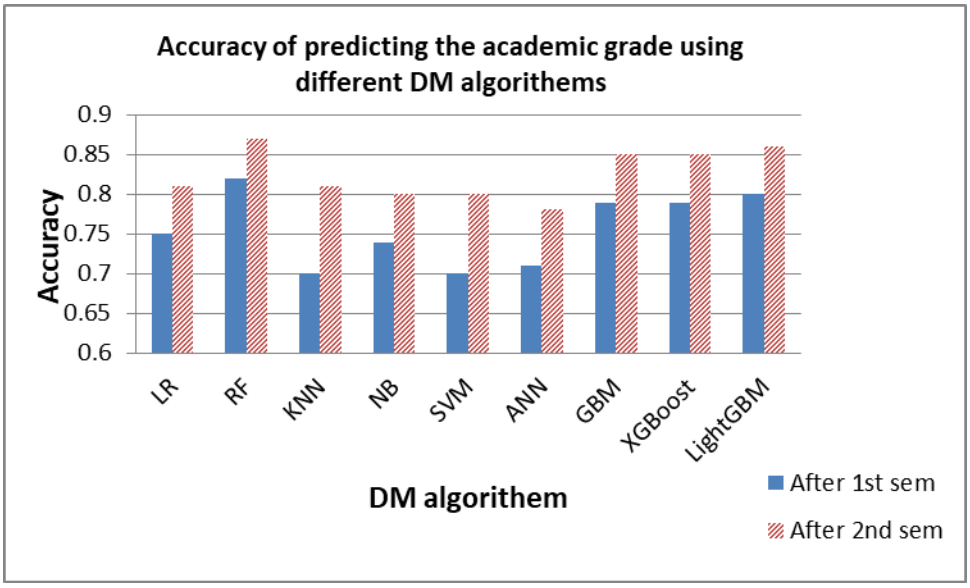


Figure 21: A comparison between the accuracy of the academic predictions after the 1st and 2nd semester using multi-class classification

As for our second objective, regarding finding out which attributes have the most influence on predicting students' academic achievement (represented in table 13 and 14), we have found that the most crucial attributes for performing the predictions after both the first and the second semester are the achieved academic grades in each semester (importance rate: 14-36%). This supports the finding of Rotem et al. (2020) and Asif et al. (2017). To further comprehend the reasons behind those results, Appendix D provides scatterplots that shows the relation of semester grades to the final predictive outcomes.

The second most important features are culture and distance, with an importance of 7-17% and 6-18%, respectively. Distance increases the financial and personal costs associated with attending classes, which limits individual choices and increases the likelihood of low participation. (Vieira et al., 2018). Therefore, students who live far from campus are more likely to fail or drop out. As discussed in the literature, cultures' behavior toward learning may differ. As students from individualistic cultures are more competent, they have higher chances of succeeding the master’s program. However, there are other factors that can influence international students (which are “collectivistic” in our study) to dropout from German educational programs. Those factors include poor linguistic proficiency, financial problems, lack of social and academic integration, and misconceptions about German higher education institutions' teaching and learning culture (Kercher, 2018).

As for the number of failed courses, they have been found to have a minor effect on predicting academic status and graduation grade (with an importance rate of 4% -10%). This contradicts the findings of Alturki & Alturki (2021) and Kabakchieva (2013), who found that the number of failed courses is essential for predicting bachelor's students' achievement. This inconsistency may be because the different groups of students have similar attitude towards failing courses in the Business Informatics master's programs. To get into more details, a large number of those who did not complete their degree and below average students had a zero number of failed courses in their first and second semester. Moreover, many of those who completed their degree had failed some courses in the first and second semester (see appendix D). This means that there is no clear pattern regarding which group of students fails courses and which group does not.

In order for students to achieve academic success, it is vital that they balance their academic load (Alturki et al., 2020). In fact, Alemu Yehuala (2015) found that it is one of the main significant features for predicting students' academic success. We tested this by investigating the impact of the number of registered and unregistered exams per semester. In German universities, the students can take their examinations right after the course has been studied or postpone the exam to one of the following semesters. The number of registered exams represents the amount of studying load, i.e., the more registered exams, the more the load is on the student. We found that the number of registered exams has a minor effect with a 5-10% importance rate. As for the number of unregistered exams, it has even a lower effect (2-6%) compared to the rest of the post-enrollment features.

Moreover, we found that enrollment age has almost no effect (4-8%). This finding contradicts the findings of Aulck et al. (2016). However, it is in line with the findings of Kovačić (2010) and does not surprise us as there is no significant gap between the ages of most applicants. Finally, even though gender has been used more often in the literature than other demographic variables for predicting academic performance (Alturki et al., 2020), it does not necessarily mean that it has a significant effect. In our case, gender does not affect the prediction, as it had an importance rate of only 2-4%. This is also in line with the findings of Kovačić (2010) and Osmanbegović et al. (2012).

5.8 Study Limitations and Future Work

The empirical results presented herein should be viewed in the context of some limitations that can be addressed in future studies.

5.8.1 Predict students' studying duration in the master's program

Higher education students must meet several objectives to obtain degrees, and in several cases, this can extend their period at the educational institution (Yue and Fu, 2017) or, worse case, drop out (Braxton, Milem, and Sullivan, 2000). Extending beyond the intended years of a degree, which is common in the Business Informatics program, can significantly increase the cost of obtaining that degree. In section 5.3.2, we represented the studying duration in the Business Informatics master's program at the University of Mannheim from 2010 to 2018. Although the intended duration for this specific program is two years (four semesters), most students take more time to graduate. This delay could be mitigated by factors such as the student's background, previous and current academic performance, and much more.

Uunderstanding the routes students take toward completing an academic degree can assist faculty and administrators in better-serving students to meet their educational targets (Aiken et al., 2020) and can help decision-makers take action to improve the academic performance of at-risk of failing students. Predicting students' studying duration can serve students and universities with great benefits. Therefore, future studies should investigate performing such predictions.

5.8.2 Investigating the use of Temporal Point Processes in predicting academic achievement

Temporal Point Processes (TPP) are a statistical approach for modeling timed event sequences, i.e., they typically consider histories of events up to a specific time point. Recently, TPP has shown potential for many machine learning and data science applications as they can assist in discovering unexpected trends, finding temporal patterns, and improving prediction outcomes. They are often used in predictions related to earthquake, power outages, criminology, accidents, infectious disease, and behavior-based network analysis. However, their use in educational contexts is still neglected.

Cohausz, Stuckenschmidt, and Alturki (2022) have attempted to validate the approach of integrating temporal information in the analysis of the Business Informatics master's students' performance using TPP. The anonymized dataset contains information about almost 25.000 exam registrations by 1268 Students over the past ten years. The model has been implemented using the PoPPy library (a ML toolbox focusing on point process model) with Self-exciting (Hawkes) Process, exponential decay kernel and maximum likelihood estimation.

To predict the expected outcomes for a possible next semester for each student, they have used the following set of features: (i) number of study semesters so far, (ii) number of passed exams, (iii) number of failed exams, (iv) number of final fails, and (v) numbers of second and third attempts with the aim of. In addition, they used a TPP model to generate more set of parameters (5 Attributes regarding exam status predictions, 11 attributes regarding exam grade predictions, and 10 attributes regarding Exam remark predictions for the next semester) using study terms as time points.

Table 19 shows the classification results on the target classes. We see that the use of the basic feature set produces excellent results for the 'Passed' outcome which is the majority class. However, including temporal knowledge via the TPP predictions produced a better fit of the model, especially with respect to the minority classes (i.e. Failed).

Table 19: F1-Scores for historical Study Outcome Prediction for different feature sets (Cohausz, Stuckenschmidt, and Alturki, 2022)

Features	Passed	Failed	Dropped out
Base	0.93	0.80	0.77
Base + Status TPP	0.93	0.82	0.77
Base + Grade TPP	0.94	0.86	0.78
Base + Remark TPP	**0.94**	0.86	0.78
All	**0.94**	**0.89**	**0.79**

We conclude that including temporal information is a promising approach for improving performance prediction in higher education. With the growing usage of MOOCs and the significant amount of data generating from online courses, collecting the timing and ordering of students' behaviour and interaction with the e-learning system has become more feasible, providing a much better basis for temporal analyses. Therefore, future studies should focus on including temporal features when performing academic predictions.

5.8.3 Predict students' academic achievement prior to admission

Many scholars believe that the students' performance prediction should be received in an early stage of the studying program (Anderson, 2017; de Barba, Kennedy, and Ainley, 2016). However, it is still unclear whether it is possible to predict students' success prior to accepting them into academic programs.

Determining admission to computer science programs is relatively challenging since it is an interdisciplinary field that attracts applicants from diverse backgrounds. Predicting students' success before accepting them in the Business Informatics program will bring massive advantages to the students, instructors, and university. However, its possibility is still an open question that needs to be investigated. We have initiated this investigation by comparing the accuracy of SSA with short tests in the five most critical topics in the Business Informatics master's program (Linear Algebra, Databases, Probability and Statistics, Algorithms & Programming, and Logic and Combinatorics) (Alturki and Stuckenschmidt 2021). The survey tool EvaSys was used to create our self-assessment and test survey. It is automation software that can be used to automate organizational surveys and research projects, course and training evaluations, exams, and assessments ("Survey Automation Software - EvaSys and EvaExam" 2019). The interested reader can view the survey in Appendix E. In addition, information regarding the participant's gender, age, and previous GPA were obtained from the admissions office of the University of Mannheim.

As a result of this study, we were able to identify the most relevant factors affecting SSA accuracy in higher education (Alturki and Stuckenschmidt 2021). SSA's accuracy level was significantly affected by (i) previous GPA; whereas students with higher GPAs tend to be more accurate in their self-assessments, students with lower GPAs tend to overestimate them. (ii) The behavior of SSA differs significantly by gender, with an average association of 0.21; females tend to underestimate their abilities, while males tend to overestimate them. (iii) The behavior of different cultures towards SSA differs significantly with an average association of 0.19, i.e. the percentage of overestimation is higher for collectivist participants, while peers from individualistic cultures tend to conduct self-assessment more accurately. (iv) The SSA accuracy level is significantly influenced by the subject matter, which is to say there is a strong correlation between greater self-assessment accuracy and experience with the topic being assessed. More details regarding the accuracy results can be found in Appendix F.

Having found these results, we highly recommend investigating SSA's significance level for predicting students' academic achievement and determining whether more factors affect their behavior with self-assessment. Researchers should devote more attention to self-assessment studies in higher education. Moreover, instructors should be capable of persuading and educating students about the importance of conducting such assessments and the importance of providing accurate score reports.

CHAPTER 6: AN IN-DEPTH EVALUATION ON THE IMPACT OF CULTURE ON ACADEMIC PREDICTIONS

“Culture is a way of coping with the world by defining it in detail.” – Malcolm Bradbury

It is common to use the terms culture and ethnicity interchangeably to describe the characteristics of people from different ethnic groups who share a racial background, nationality, language, or religion. The concept of culture may, however, encompass systems of belief, knowledge, values, and behavior that are shared by a group (Scholes, 2020). Moreover, Culture represents the arts, laws, customs, capabilities, and individual habits of a particular group of people from a specific region or location (Tylor, 1871).

An individual's cultural background influences all aspects of their lives, including how they value and engage with their educational surroundings (Scholes, 2020). Students growing up in different cultural settings may approach education and learning differently. Therefore, the past quarter century has seen a growing interest in cross-national comparisons of student achievement (National Research Council, 2002).

In this chapter, we look deeper into the role of culture on students’ academic achievement, explain the relationship between the culture of where the academic predictions are held and the choice of predictive features, and discuss the vital role of culture on academic predictions.

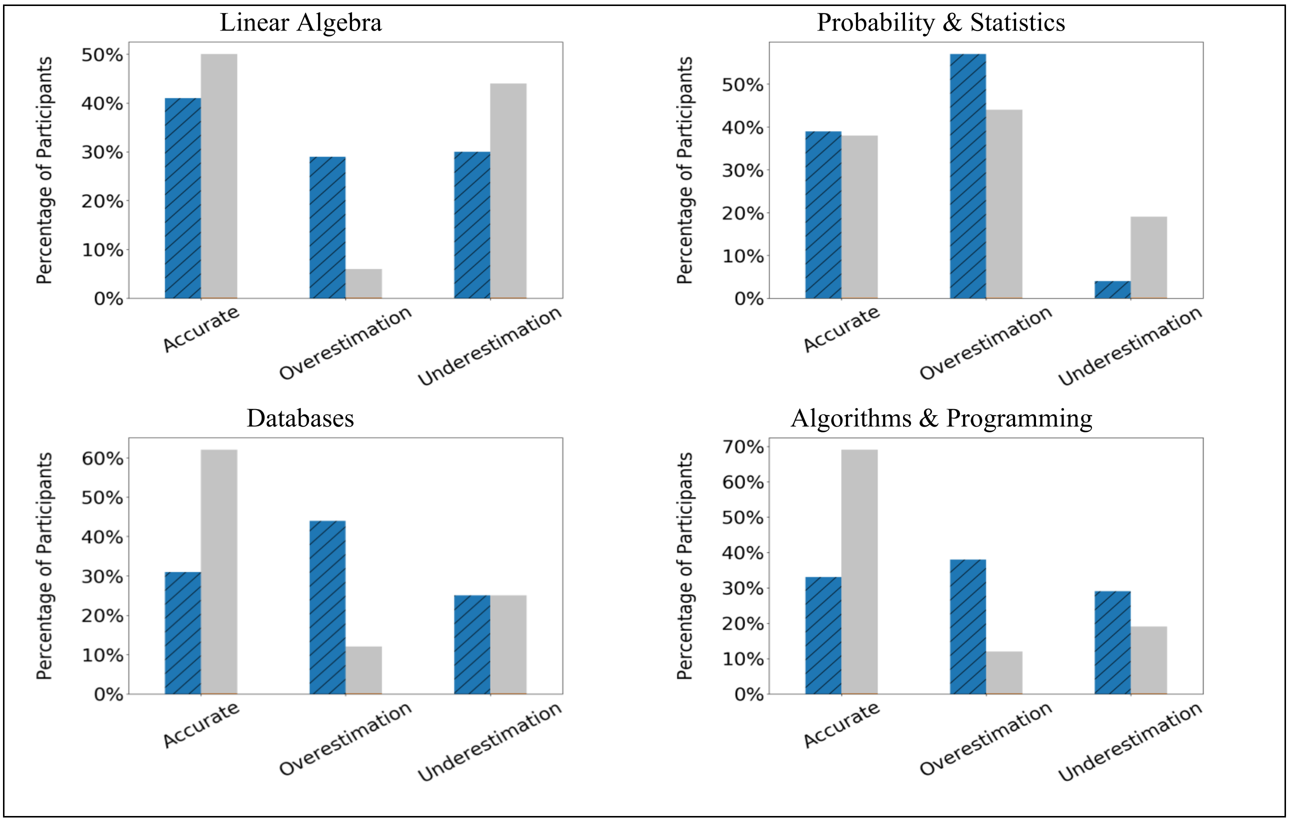
6.1 The Role of Culture on Students’ Academic Achievement

Many culture-related factors could affect students’ academic achievement. One of the important factors is how different students view themselves based on their culture. Students who are able to accurately evaluate their strengths and limitations will be able to utilize their strengths to their maximum potential. This will result in improved grades, perceived learning, improved presentation skills, and the development of assessment skills (De Grez, Valcke, and Roozen, 2012).

Individualistic cultures tend to have a more independent picture of themselves (they identify themselves on the base of their characteristics and see their characteristics as reasonably constant). As opposite to people from Individualistic cultures, people from collectivistic cultures usually have a view of themselves that is interdependent (they perceive themselves as connected to others and see themselves as being prone to change in different contexts) (Markus and Kitayama, 1991). There has been previous research into the concept of "overconfidence" or "over precision" in the context of cultural differences (Moore, Dev, and Goncharova, 2018). Research has demonstrated, for example, that people from China (collectivists) display a greater degree of overconfidence in their skills and knowledge than Americans (individualists) (Yates, Lee, and Bush, 1997; Yates et al., 1998). Yates, Lee, and Bush (1997) also stated that subjects in Asian cultures are more likely to be overconfident than those in Western cultures. Studies such as these indicate that students' cultural backgrounds affect how they view themselves, which, in turn, has an impact on how they approach learning.

By considering the impact of culture on academic achievement, we hypothesize that students' who overestimate their knowledge have lower chances of succeeding in higher education programs. To test our hypotheses, we investigated how students of different cultures asses themselves in terms of their knowledge and skills and how their assessment reflected on their achievement using an online survey (Alturki and Stuckenschmidt, 2021). The survey contains two parts: a self-assessment part and a test part. In the self-assessment section, students have been asked to evaluate themselves on five of the most important topics in the Business Informatics master's program offered by the university of Mannheim. Students are asked to take a short exam regarding the same topics in the test section. The participants in this study come from a variety of countries and cultures around the world and are seeking admission to the master's program in Business Informatics. They have been grouped into two groups: collectivistic students and individualistic students: 71 and 49, respectively. Comparing the results of the self-assessments and the test allows us to understand how students of different cultures view themselves.

It is apparent from our empirical study that collectivistic students are significantly more likely to overestimate their knowledge than individualistic students (Figure 22). Overall, overestimation is higher for collectivistic participants than for individual participants. In contrast to individualistic students, whose average overestimation is only 14.8%, collectivistic students had an average overestimation of 37.4%. In addition, it is observed that the individualistic participants are typically more accurate in their self-assessments, with an average accuracy of 60% on self-assessments.



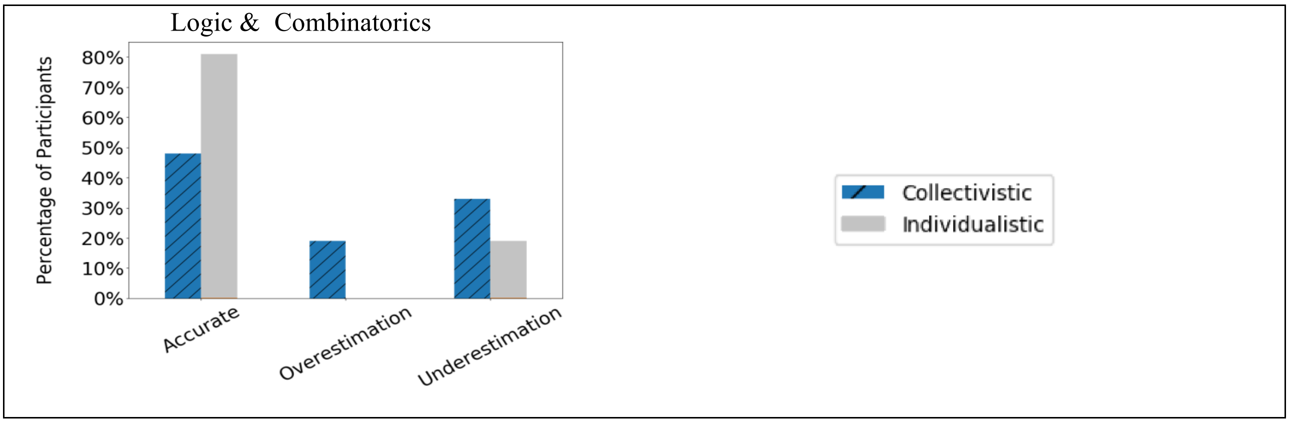


Figure 22: Relationship between the accuracy of SSA and culture (Alturki and Stuckenschmidt, 2021)

Table 20 gives a more detailed comparison between the students of the two cultures among different topics. To measure the association between the two nominal variables, we used Cramer's V. Values close to 1 imply a robust association between the attributes, and values close to 0 point to a weak one. The results further proves that there is a strong variation in how students of different cultures see themselves.

Table 20: Relationship between SSA and culture using Cramer's V (Alturki and Stuckenschmidt, 2021)

Topic	Collectivistic			Individualistic			Association (Cramer's V)
	Accurate	**Over-estimator**	**Under-estimator**	**Accurate**	**Over-estimator**	**Under-estimator**	
Linear Algebra	41%	29%	30%	50%	6%	44%	0.13
Probability & Statistics	39%	57%	4%	38%	44%	19%	0.17
Databases	31%	44%	25%	62%	12%	25%	0.22
Algorithms & Programming	33%	38%	29%	69%	12%	19%	0.22
Logic & Combinatorics	48%	19%	33%	81%	0%	19%	0.21
Average	**38.4%**	**37.4%**	**24.2%**	**60%**	**14.8%**	**25.2%**	**0.19**

6.2 Using Culture as A Feature for Performing Academic Predictions

As mentioned in chapter 3 (Literature Review), the selection of features for performing academic predictions are greatly affected by the cultural background of the countries where the research is conducted, especially with the choice of demographical features. While collectivistic cultures tend to choose features associated to the family of students, e.g., family support, family income, family size, and parents' qualifications, individualistic cultures tend to ignore such features and draw more focus on personal achievement. As the academic success factors differ from one nation to another, it is reasonable that the selection of the predictive features differs from one educational institution to another. However, some academic institutions enroll international students, which is common in western countries. When this is the case, more attention should be given to the selection of the predictive features.

Although there is a rich literature regarding the impact of culture on academic achievement, using it as a feature for performing academic predictions is still neglected. Drawing more attention to our specific study presented in the previous chapter (chapter 5), we found that culture considerably impacts academic predictions. In fact, culture was found to be the third most important feature for the performed academic predictions compared to a list of 12 features. To further interpret the reason behind the importance of such feature, we analyzed the students' performance in the Business Informatics master's program based on their culture (Figure 23). We found that most students not completing their degrees are the students of collectivistic cultures. In more detail, only 5% of the individualistic students enrolled in the master's program did not complete their master's degree. On the other hand, 17% of the collectivistic students did not succeed in completing their degree.

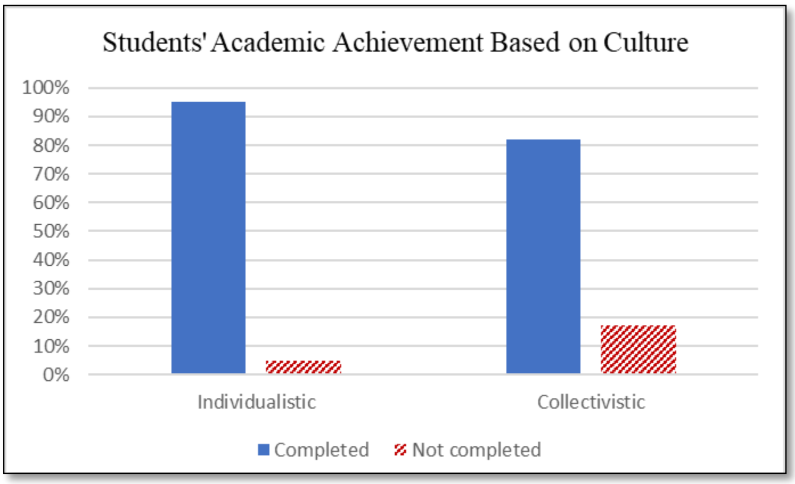


Figure 23: Students' academic achievement at the Business Informatics master's program based on culture

6.3 Conclusion

In this chapter, we investigated the role of culture on students' academic achievement and its significance on academic predictions. We tested how students of different cultures view themselves regarding the skills required to succeed in the Business Informatics master's program vs. their actual skills and knowledge. A significant difference has been observed between the perceptions of students from different cultures, with 0.19 average association. While the overestimation percentage is high for students with a collectivistic background, their individualistic peers happen to be more accurate. Being overconfident can negatively affect students' academic success as it reduces their performance over time (Moores and Chang, 2009). Such results prove the need for including culture in the academic predictions that are performed in universities that enroll students from diverse backgrounds.

CHAPTER 7: CONCLUSION

"*Education is our passport to the future, for tomorrow belongs to the people who prepare for it today.*" *-Malcolm X*

Investing in the youth through higher education can provide tremendous social and economic returns. Thus, it is crucial to reduce dropout rates and increase student success. Predicting a student's academic achievement at an early stage of their academic journey is one way of accomplishing this. Educational databases can provide valuable knowledge for a variety of purposes through EDM research, such as identifying students at risk of failing or dropping out of a degree, as well as discovering honorary students for allocating scholarships. Additionally, instructors may be able to determine the capabilities of each student, and accordingly, design teaching tasks based on those capabilities. For example, instructors may offer extracurricular learning materials to students facing difficulties, use different teaching strategies, or provide online tutoring videos to students who need them. This chapter presents a summary of the two studies that have been performed and represented in this thesis. Following that, we give short answers to the research questions that have been presented in chapter 1, section 1.3. Finally, we conclude by outlining general recommendations for future work.

7.1 Summary of the Empirical Studies

In this research, two distinctive studies are performed in which both have similar goals and objectives. Those studies have been represented in detail in chapter 4 and chapter 5. In chapter 4, the dataset of 300 undergraduate students has been collected from three departments of the Computer and Information Science College at Princess Nora University. Those records include pre-enrollment and post-enrollment features. The performance of eight data mining algorithms has been compared to develop three main predictive models within the first two years of the academic program. The algorithms that have been used are C4.5, Simple CART, LADTree, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Artificial Neural Networks, and Random Forest. We found that Naïve Base performs the best, followed by Random Forest. We also investigated the importance of each of the collected features on the overall predictions and found that the features that matter most to the predictions are student's GPA for each semester, the number of failed courses in the first and second year, their grade in the 'Database fundamentals', and 'Programming 1' core courses.

In chapter 5, the dataset of over 700 students' have been collected from the Faculty of Business Informatics and Mathematics at the University of Mannheim. Those records include demographics and post-enrollment features. We compared the performance of nine data mining algorithms in predicting students' academic achievement. Those algorithms are Logistic Regression, Naïve Bayes, K-nearest neighbor, Artificial Neural Networks, Support Vector Machine, Random Forest, Gradient Boosting, Light Gradient Boosting, and Extreme Gradient Boosting. We found that ensemble methods perform better than single classifiers with Random Forest being the best among them. We also investigated the importance of each collected feature on the overall predictions and found that the semesters' GPAs are the most significant features, followed by culture and distance.

Although each of the performed studies represented in chapter 4 and chapter 5 have its own limitations, the following are the limitations that both have in common:

- As both studies contained fewer than a thousand records, they are considered relatively small samples.
- Many other features may have influenced student success.
- Other data mining algorithms may have been investigated.
- Both studies did not evaluate the long-term impact of the predictive models.

7.2 Answers to the research questions

In the following, we present the research questions once again and report on the answers to the questions. However, prior to drawing inferences from the answers, it is essential to consider that the results of this study are unique to the universities where they have been conducted and might not be directly transferable or generalized to other institutions of higher education.

RQ 1. Is it possible to predict the students' final academic achievement in bachelor's and master's programs at an early stage?

- **Is it likely to have reasonably accurate predictions after the first and second semesters of the students' enrollment?**
  Performing academic predictions after the first and second semester can give quite good accuracies in the case of postgraduate programs where the studying duration is relatively short (Chapter 5, tables 15 and 16). However, this is not the case in undergraduate programs (chapter 4, table 10) where the study duration is four to five years. Certainly, the later the prediction is performed in the studying program, the higher the achieved accuracy.

- **Is it likely to have fairly accurate predictions using multi-class classification?**
  As a general observation, the more classes, the more challenging the prediction is. To go into more detail, dealing with binary-class problems can result in significantly high accuracies, as can be viewed in chapter 5, table 15. Dealing with 3 class problems can also produce good results. For instance, in both chapter 4 and chapter 5, we presented models that perform 3-class predictions (i.e., above average, average, and below average). In general, we were able to achieve good accuracy using three classes. On the other hand, the significance of the predictions was not as high when we used five classes (i.e., excellent, very good, good, acceptable, poor) (chapter 4, tables 9, 10, and 11). Since our focus is on horary students and students at risk of not completing a degree, performing 3-class predictions is sufficient enough for us.

RQ 2. What measurable aspects predict student academic achievement for bachelor's and master's programs in computer majors?

- **To what extent, if any, can student demographics predict academic achievement?**

  While we have not investigated the use of demographics in predicting undergraduate students (chapter 4), we have done so with the case of predicting postgraduate students (chapter 5). We found that age and gender do not have an impact on the prediction of the masters' students' academic achievemint. However, culture was found to have a significant impact on the prediction. That is since, students from individualistic cultures have been found to have higher chances of succeeding in the Business Informatics master's program.

- **To what extent, if any, can previous knowledge and GPA predict academic achievement?**

  Although previous grades from secondary school have been widely used for predicting bachelor's academic achievement (Chapter 3), we found that the achieved grade from secondary school does not have an impact on predicting the success of the undergraduate students at the College of Information and Computer Science at PNU. Unfortunately, we have not used the grades of a bachelor's degree in the case of predicting postgraduate students. However, we believe that this can be a significant feature as master's degrees are developed from undergraduate concepts, and master's students usually take courses that cover similar material as undergraduate courses, except stressing higher-level topics and theories.

- **To what extent can students' behavior after enrollment (e.g., registered academic credits) predict academic achievement?**

  When it comes to using the academic credits as a feature for performing academic predictions, the results show that it is not a significant factor for both the undergraduate and the postgraduate studies. This might be the case due to the fact that the students in both empirical studies have similar behavior in terms of choosing the number of registered courses.

- **To what extent does the distance from students' accommodation to university influence academic performance?**

  The distance from accommodation to campus has been investigated in the case of the postgraduate students (chapter 5). We have found that distance serves as an important feature for the prediction of academic success as we found that students who live far from campus are more likely to fail or dropout.

- **To what extent, if any, do academic language skills influence students' academic performance?**

  The academic language skill, which was English in our case, was investigated in the case of undergraduate students (chapter 4). According to our findings, it does not have a critical role in students' success at the Computer and Information Sciences College. That may be due to the nature of the courses, which are rather scientific and do not require a high level of English proficiency.

7.3 General Recommendations for Future Work

The scope of this research is limited to investigating the factors related to predicting students' achievement and finding methods to perform the predictions at the earliest stage possible in bachelor's and master's degree levels. Nevertheless, based on the achieved results, we strongly encourage instructors and decision-makers to consider using EDM to predict students' academic performance and tailor their learning experiences based on their individual needs. Further, we strongly encourage researchers to apply EDM studies across different universities in different parts of the world and compare their results. Moreover, results from this study provide directions for future work as follows:

Firstly, research with access to more comprehensive data may offer more conclusive results. Therefore, we suggest implementing the models on larger sets of data. We also recommend using the same prediction models in different master's and bachelor programs at the Universities where they have been conducted. This could give us more insights into whether the predictive models could be generalized and sufficiently work for other programs.

Although we achieved a high accuracy using only easy-to-collect attributes, other attributes may have a vital role. For instance, motivation (Stansfield, Mclellan, and Connolly, 2004) and socioeconomic status, which is known to predict educational achievement, particularly for Germany (OECD, 2018). Other aspects of a student's life can also affect their education, including obligations for work and family. Also, psychological factors, including learning style, self-efficacy (Ransdell, 2001; Riding and Rayner, 1998), interest and motivation, as well as the learning environment (Graaff et al., 2005) can all contribute to student learning and achievement. Even though the literature indicates that there are uncertainties concerning whether SSA can be beneficial educationally and whether it should be used for what purposes, it seems to us that it may be useful in predicting academic achievement. The significance level of SSA for predicting students' academic achievement should, therefore, be investigated in detail, Also, whether other factors impact the behavior of students when it comes to self-assessment.

To improve the accuracy and stability of a single learning algorithm, ensemble learning is recommended (Dietterich, 2000). They are also known for avoiding overfitting and improving predictions. Ensemble models have recently been used by researchers to predict student success. In the study that has been performed on bachelor's students (presented in chapter 4), we used random forests, which is a Bagging (or bootstrap aggregation) method. In the study that has been performed on master's students (presented in chapter 5), we used Random Forests, Gradient boosting, extreme gradient boosting, and light gradient boosting. However, other ensemble methods are worth exploring.

The results of this study can be used to design a recommender system that allows appropriate interventions for both the undergraduate students of the College of Information and Computer Science at PNU and the postgraduate students of the Business Informatics program at the University of Mannheim. Future studies could also examine the effects of providing tutoring classes to weak students who have failed or are at risk of failing courses. In addition, future research could focus on examining the effect of language skills in social science programs as well as the effect of orientation years in other bachelor's degree programs.

By applying the recommendations described above, we will be able to better understand EDM and the features that contribute to the accuracy of academic predictions. In turn, students' learning experience and quality of learning will be significantly improved.

REFERENCES

Abu Saa, Amjad. 2016. “Educational Data Mining & Students’ Performance Prediction.” *International Journal of Advanced Computer Science and Applications* 7 (5). https://doi.org/10.14569/IJACSA.2016.070531.

Abu Zohair, Lubna Mahmoud. 2019. “Prediction of Student’s Performance by Modelling Small Dataset Size.” *International Journal of Educational Technology in Higher Education* 16 (1): 1–18. https://doi.org/10.1186/S41239-019-0160-3/FIGURES/13.

Aiken, John M., Riccardo de Bin, Morten Hjorth-Jensen, and Marcos D. Caballero. 2020. “Predicting Time to Graduation at a Large Enrollment American University.” *PLoS ONE* 15 (11). https://doi.org/10.1371/JOURNAL.PONE.0242334.

Aina, Carmen, Eliana Baici, Giorgia Casalone, and Francesco Pastore. 2018. “The Economics of University Dropouts and Delayed Graduation: A Survey.” *Discussion Paper Series, Institute of Labour Economics*, IZA DP No. 11421, ISSN: 2365-9793.

Alemu Yehuala, Muluken. 2015. “Application of Data Mining Techniques for Student Success and Failure Prediction (The Case of Debre Markos University).” *International Journal of Scientific & Technology Research* 4 (4). ISSN: 2277-8616

Algarni, Abdulmohsen. 2016. “Data Mining in Education.” *International Journal of Advanced Computer Science and Applications* 7 (6). https://doi.org/10.14569/ijacsa.2016.070659.

Ali, Shoukat, Zubair Haider, Fahad Munir, Hamid Khan, and Awais Ahmed. 2013. “Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus.” *American Journal of Educational Research* 1 (8). https://doi.org/10.12691/education-1-8-3.

Alija, Sadri. 2013. “How Attendance Affects the General Success of the Student.” *International Journal of Academic Research in Business and Social Sciences* 3 (1). ISSN: 2222-6990

Aljohani, Othman. 2016. “Analyzing the Findings of the Saudi Research on Student Attrition in Higher Education.” *International Education Studies* 9 (8). https://doi.org/10.5539/ies.v9n8p184.

Alturki, Sarah, Ioana Hulpus, and Heiner Stuckenschmidt. 2020. “Predicting Academic Outcomes: A Survey from 2007 Till 2018.” *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-020-09476-0.

Alturki, Sarah, and Heiner Stuckenschmidt. 2021. “Assessing Students’ Self-Assessment Ability in an Interdisciplinary Domain.” *Journal of Applied Research in Higher Education*. https://doi.org/10.1108/JARHE-01-2021-0034.

Amudha, J., KP Soman, Y Kiran, and Y Kiran. 2011. “Feature Selection in Top-Down Visual Attention Model Using WEKA.” *International Journal of Computer Applications* 24 (4): 38–43. https://doi.org/10.5120/2955-3895

Anderson, Timothy. 2017. “Applications of Machine Learning to Student Grade Prediction in Quantitative Business Courses.” *Global Journal of Business Pedagogy* 1 (3): 13–22.

ISSN: 2574-0385

Arsad, P M, N Buniyamin, and J Ab Manan. 2014. "Students' English Language Proficiency and Its Impact on the Overall Student's Academic Performance: An Analysis and Prediction Using Neural Network Model." *WSEAS Transactions on Advances in Engineering Education* 11: 44–53. ISSN: 1790-1979

Arun, D. K., V. Namratha, B. V. Ramyashree, Yashita P. Jain, and Antara Roy Choudhury. 2021. "Student Academic Performance Prediction Using Educational Data Mining." *International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/ICCCI50826.2021.9457021.

Asif, Raheela, Agathe Merceron, Syed Abbas Ali, and Najmi Ghani Haider NED. 2017. "Analyzing Undergraduate Students' Performance Using Educational Data Mining." *Computers & Education* 113: 177–94. https://doi.org/10.1016/j.compedu.2017.05.007.

Aulck, Lovenoor, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. 2016. "Predicting Student Dropout in Higher Education." *Proceedings of the ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, New York, NY, USA. Retrieved 17 October, 2022 from https://arxiv.org/pdf/1606.06364.pdf.

Badr, Ghada, Afnan Algobail, Hanadi Almutairi, and Manal Almutery. 2016. "Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department." *Procedia Computer Science* 82: 80–89. https://doi.org/10.1016/j.procs.2016.04.012.

Baker, Ryan, Seiji Isotani, and Adriana Carvalho. 2011. "Mineração de Dados Educacionais: Oportunidades Para o Brasil." *Revista Brasileira de Informática Na Educação* 19 (02). https://doi.org/10.5753/rbie.2011.19.02.03.

Baker, Ryan S.J.d., and Kalina Yacef. 2009. "The state of educational data mining in 2009: A review and future visions". *Journal of Educational Data Mining*, 1(1): 3-17. http://doi.org/10.5281/zenodo.3554658

Baker, Ryan S J D. 2010. "Mining Data for Student Models." In: Nkambou, R., Bourdeau, J., Mizoguchi, R. (eds). *Advances in Intelligent Tutoring Systems*. Studies in Computational Intelligence, 308. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14363-2_16

Bamiah, Mervat Adib Bamiah, Sarfraz N. Brohi, and Babak Bashari Rad. 2018. "Big Data Technology in Education: Advantages, Implementations, and Challenges." *Journal of Engineering Science and Technology*, Special Issue on International Conference on Computer Science and Information Technology (ICCSIT 2018), 229–241. Retrieved 17 October, 2022 from https://jestec.taylors.edu.my/Special%20Issue%20ICCSIT%202018.htm

Bani-Salameh, Hisham N. 2018. "Teaching Language Effects on Students 'Performance." *Health Professions Education* 4 (1): 27–30. https://doi.org/10.1016/j.hpe.2017.01.005.

Barba, Paula de, Gregor Kennedy, and Mary Ainley. 2016. "The Role of Students' Motivation and Participation in Predicting Performance in a MOOC." *Journal of Computer Assisted Learning* 32 (3): 218–31. https://doi.org/10.1111/JCAL.12130.

Barnes, Tiffany. 2005. "The Q-Matrix Method: Mining Student Response Data for Knowledge." *Proceedings of the AAAI-2005 Workshop on Educational Data Mining,*Pittsburgh, PA, USA.

Batini, Carlo, Politecnico Di Milano, and Andrea Maurino. 2009. "Methodologies for Data Quality Assessment and Improvement." *ACM Computing Surveys* 41 (3): 1-52. https://doi.org/10.1145/1541880.1541883.

Beck, Joseph E, and Jack Mostow. 2008. "How Who Should Practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students." In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds) Intelligent Tutoring Systems. Lecture Notes in Computer Science, 5091. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-69132-7_39.

Berens, Johannes, Simon Oster, Kerstin Schneider, and Julian Burghoff. 2019. "Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods." *Schumpeter School of Business and Economics* 11 (3): 0–32. https://doi.org/10.5281/ZENODO.3594771.

Bissell, Kelly, Ruan M. Lasalle, and Paolo Dal Cin. 2019. "The Cost of Cybercrime: Ninth Annual Cost of Cybercrime Study." ACCENTURE Report. Retrieved 17 October, 2022 from https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf.

Brahmeswara Kadaru, Bala, and Munipalli Umamaheswararao. 2017. "An Overview of General Data Mining Tools." *International Research Journal of Engineering and Technology* 04 (09): 930–936. ISSN: 2395-0072.

Braxton, John M., Jeffrey F. Milem, and Anna Shaw Sullivan. 2000. "The Influence of Active Learning on the College Student Departure Process: Toward a Revision of Tinto's Theory." *The Journal of Higher Education* 71 (5): 569. https://doi.org/10.2307/2649260.

Breiman, Leo. 2001. "Random Forests." *Machine Learning 2001 45:1* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brusilovsky, Peter, and Eva Millán. 2007. "User Models for Adaptive Hypermedia and Adaptive Educational Systems." In *The Adaptive Web*, 3–53. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_1.

Cai, Li, and Yangyong Zhu. 2015. "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era." *Data Science Journal* 14 (0): 2. https://doi.org/10.5334/dsj-2015-002.

Caie, Peter D., Neofytos Dimitriou, and Ognjen Arandjelović. 2021. "Precision Medicine in Digital Pathology via Image Analysis and Machine Learning." In *Artificial Intelligence and Deep Learning in Pathology*, 149–73. Elsevier. https://doi.org/10.1016/b978-0-323-67538-3.00008-7.

Calvet Liñán, Laura, and Ángel Alejandro Juan Pérez. 2015. "Educational Data Mining and Learning Analytics: Differences, Similarities, and Time Evolution." *RUSC. Universities and Knowledge Society Journal* 12 (3): 98. https://doi.org/10.7238/rusc.v12i3.2515.

Chang, Yuwen. 2008. "Gender Differences in Science Achievement, Science Self-Concept, and Science Values." In *The Proceedings of International Association for the Evaluation of Educational Achievement (IRC)*.

Chatti, Mohamed Amine, Anna Lea Dyckhoff, Ulrik Schroeder, and Hendrik Thüs. 2012. "A Reference Model for Learning Analytics." *International Journal of Technology Enhanced Learning* 4 (5/6): 318. https://doi.org/10.1504/IJTEL.2012.051815.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–57. https://doi.org/10.1613/JAIR.953.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.

Clark, Michael. 2013. "An Introduction to Machine Learning with Applications in R." Retrieved 17 October, 2022 from http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/ML_inR.pdf.

Cleve, Jürgen, and Uwe Lämmel. 2020. "Data Mining," Berlin, Boston: De Gruyter Oldenbourg. https://doi.org/10.1515/9783110676273.

Cohausz, Lea, Stuckenschmidt, Heiner, Alturki, sarah. Improving Acadmic Performance Prediction Using Temporal Event Features [unpublished manuscript]. Data and Web Science Group, University of Mannheim.

Costa, Evandro, Ryan S.J.d. Baker, Lucas Amorim, Jonathas Magalhães, and Tarsis Marinho. 2012. "Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações." *Jornada de Atualização Em Informática Na Educação* 1 (1): 1–29. Retrieved 17 October, 2022 from http://www.br-ie.org/pub/index.php/pie/article/view/2341.

Cutler, Adele, D. Richard Cutler, and John R. Stevens. 2012. "Random Forests." In *Ensemble Machine Learning*, 157–75. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-9326-7_5.

D. Buhmann, Martin. 2003. *Radial Basis Functions: Theory and Implementations*. 1st ed. Cambridge: Cambridge University Press.

Daniel, Ben. 2014. "Big Data and Analytics in Higher Education: Opportunities and Challenges." *British Journal of Educational Technology* 46 (5): 904–20. https://doi.org/10.1111/bjet.12230.

Daniel, Ben Keil. 2017. *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*. Springer. https://doi.org/10.1007/978-3-319-06520-5.

Daud, Ali, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D Lytras, Farhat Abbas, and Jalal S Alowibdi. 2017. "Predicting Student Performance Using Advanced Learning Analytics." In *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion*. Perth, Australia: ACM. https://doi.org/10.1145/3041021.3054164.

Davig, William B., and Judith W. Spain. 2003. "Impact on Freshmen Retention of Orientation Course Content: Proposed Persistence Model." *Journal of College Student Retention:*

Research, Theory & Practice 5 (3): 305–23. https://doi.org/10.2190/V6B4-PQAW-TTV0-CJCU.

Dekker, Gerben W, Mykola Pechenizkiy, and Jan M Vleeshouwers. 2009. "Predicting Students Drop Out: A Case Study." *International Working Group on Educational Data Mining*, 41–50. Retrieved 17 October, 2022 from http://www.win.tue.nl/~mpechen/research/edu.html.

Demetriou, Cynthia P, and Amy Schmitz-Sciborski. 2011. "Integration, Motivation, Strengths and Optimism : Retention Theories Past, Present and Future." In *Proceedings of the 7th National Symposium on Student Retention*, edited by R. Hayes, 300–312. Charleston, USA: Norman, OK: The University of Oklahoma.

Domingos, Pedro. 1995. "Rule Induction and Instance-Based Learning A Unified Approach." Retrieved 17 October, 2022 from https://pdfs.semanticscholar.org/3f9b/a769643cdc530e93f80cea49889415792099.pdf.

Domingos, Pedro M., and M. Pazzani. 1996. "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier." *ICML*.

Dutt, Ashish, Maizatul Akmar Ismail, and Tutut Herawan. 2017. "A Systematic Review on Educational Data Mining." *IEEE Access*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2017.2654247.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases." In: *The Fourteenth National Conference on Artificial Intelligence*, Providence, Rhode Island, USA.

Gaevic, Dragan, Dragan Djuric, and Vladan Devedic. 2006. *Model Driven Architecture and Ontology Development. Model Driven Architecture and Ontology Development*. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/3-540-32182-9.

Geng, Ming. 2006. "A Comparison of Logistic Regression to Random Forests for Exploring Differences in Risk Factors Associated with Stage at Diagnosis Between Black and White Colon Cancer Patients." Master's Thesis, University of Pittsburgh. Retrieved 17 October, 2022 from http://d-scholarship.pitt.edu/id/eprint/7034

Goni, Umar, wali S. Yagana, Ali Hajja Kaltum, and Bularafa Mohammed Waziri. 2015. "Gender Difference in Students' Academic Performance in Colleges of Education in Borno State, Nigeria: Implications for Counselling." *Journal of Education and Practice* 6 (32): 107–114. ISSN: 2222-1735

Graaff, Erik de., Gillian N. Saunders-Smits, Michael R. Nieweg. 2005. *Research and Practice of Active Learning in Engineering Education*. Pallas publication. ISBN: 9789085550914

Greller, Wolfgang, and Hendrik Drachsler. 2012. "Translating Learning into Numbers: A Generic Framework for Learning Analytics". *Journal of Educational Technology & Society*, 15 (3): 42–57. Retrieved 17 October, 2022 from http://www.jstor.org/stable/jeductechsoci.15.3.42

Grez, Luc De, Martin Valcke, and Irene Roozen. 2012. "How Effective Are Self- and Peer Assessment of Oral Presentation Skills Compared with Teachers' Assessments?" *Active Learning in Higher Education* 13 (2): 129–42.

https://doi.org/10.1177/1469787412441284.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. "Data Mining. Concepts and Techniques", 3rd Ed. The Morgan Kaufmann Series in Data Management Systems. Elsevier.

Harrington, Peter. 2011. "Machine Learning in Space: Extending Our Reach." *Machine Learning* 84 (3): 335–340. https://doi.org/10.1007/s10994-011-5249-4.

Hoel, Tore, and Weiqin Chen. 2018. "Privacy and Data Protection in Learning Analytics Should Be Motivated by an Educational Maxim—towards a Proposal." *Research and Practice in Technology Enhanced Learning* 13 (1): 20. https://doi.org/10.1186/s41039-018-0086-8.

Huang, Shaobo, and Ning Fang. 2013. "Predicting Student Academic Performance in an Engineering Dynamics Course: A Comparison of Four Types of Predictive Mathematical Models." https://doi.org/10.1016/j.compedu.2012.08.015.

Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2017. *Data Mining: Practical Machine Learning Tools and Techniques - Part II: More Advanced Machine Learning Schemes*. 4th Ed. Morgan Kaufmann.

Ibrahim, Zaidah, and Daliela Rusli. 2007. "Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression." In *21st Annual SAS Malaysia Forum*.

Jain, Akshay, Aditya Rawat, Arpit Arora, and Naresh Dhami. 2017. "Analysis of Various Decision Tree Algorithms for Classification in Data Mining." *International Journal of Computer Applications* 163 (8): 975–8887.

Jain, Priyank, Manasi Gyanchandani, and Nilay Khare. 2016. "Big Data Privacy: A Technological Perspective and Review." *Journal of Big Data* 3 (1): 25. https://doi.org/10.1186/s40537-016-0059-y.

Jeno, Lucas M, Anne G Danielsen, and Arild Raaheim. 2018. "A Prospective Investigation of Students' Academic Achievement and Dropout in Higher Education: A Self-Determination Theory Approach." *International Journal of Experimental Educational Psychology*. https://doi.org/10.1080/01443410.2018.1502412.

Jeong, Hogyeong, and Gautam Biswas. 2008. "Mining Student Behavior Models in Learning-by-Teaching Environments." *Educational Data Mining*, 127-136. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.217.3896.

Johnson, L., S. Adams, and M. Cummins. 2012. "The NMC Horizon Report: 2012 Higher Education Edition." NMC Horizon Report. Retrieved 10 September, 2022 from https://www.nmc.org/pdf/2012-horizon-report-HE.pdf.

Joseph, Manu. 2020. "The Gradient Boosters IV: LightGBM." Retrieved 10 September, 2022 from https://deep-and-shallow.com/2020/02/21/the-gradient-boosters-iii-lightgbm/.

Kabakchieva, Dorina. 2013. "Predicting Student Performance by Using Data Mining Methods for Classification." *Cybernetics and Information Technologies* 13 (1): 61–72. https://doi.org/10.2478/cait-2013-0006.

Kalmegh, S. 2015. "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and

RandomTree for Classification of Indian News." *International Journal of Innovative Science, Engineering & Technology* 2 (2): 438–446. ISSN: 2348-7968

Karakose, Turgut, Ramazan Yirci, Stamatios Papadakis, Tuncay Yavuz Ozdemir, Murat Demirkol, and Hakan Polat. 2021. "Science Mapping of the Global Knowledge Base on Management, Leadership, and Administration Related to COVID-19 for Promoting the Sustainability of Scientific Research." *Sustainability,* 13 (17): 9631. https://doi.org/10.3390/su13179631.

Karalar, Halit, Ceyhun Kapucu, and Hüseyin Gürüler. 2021. "Predicting Students at Risk of Academic Failure Using Ensemble Model during Pandemic in a Distance Learning System." *International Journal of Educational Technology in Higher Education* 18 (1). https://doi.org/10.1186/S41239-021-00300-Y.

Kaushal, Aarti, and Manshi Shukla. 2014. "Comparative Analysis to Highlight Pros and Cons of Data Mining Techniques-Clustering, Neural Network and Decision Tree." *International Journal of Computer Science and Information Technologies* 5 (1): 651–56.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." In *Advances in Neural Information Processing Systems 30*. Long Beach, CA. https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Kemper, Lorenz, Gerrit Vorhoff, and Berthold U. Wigger. 2020. "Predicting Student Dropout: A Machine Learning Approach." *European Journal of Higher Education* 10 (1): 28–47. https://doi.org/10.1080/21568235.2020.1718520.

Kim, Uichol. 1995. "Individualism and Collectivism A Psychological, Cultural and Ecological Analysis." http://eurasia.nias.ku.dk/publications/.

Kindervag, John, Stephanie Balaouras, Brian W. Hill, and Kelley Mak. 2012. "Control And Protect Sensitive Information In the Era of Big Data." Forrester Research Publication, Cambridge, MA, USA.

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. SAGE Publications Ltd. https://doi.org/10.4135/9781473909472.

Kotsiantis, S B, G E Tsekouras, and P E Pintelas. 2005. "Bagging Model Trees for Classification Problems." *LNCS* 3746: 328–37.

Kovačić, Zlatko J. 2010. "Early Prediction of Student Success: Mining Students Enrolment Data." In *Proceedings of Informing Science & IT Education Conference*, 647–65.

Kumar, Mukesh, A.J. Singh, and Disha Handa. 2017. "Literature Survey on Student's Performance Prediction in Education Using Data Mining Techniques." *International Journal of Education and Management Engineering* 7 (6): 40–49. https://doi.org/10.5815/ijeme.2017.06.05.

Kurgan, Lukasz A., and Petr Musilek. 2006. "A Survey of Knowledge Discovery and Data Mining Process Models." *The Knowledge Engineering Review* 21 (1): 1–24. https://doi.org/10.1017/S0269888906000737.

Kwon, Ohbyung, and Jae Mun Sim. 2013. "Effects of Data Set Features on the Performances of Classification Algorithms." *Expert Systems with Applications* 40 (5): 1847–57. https://doi.org/10.1016/j.eswa.2012.09.017.

Lakshmi, Devasena. 2012. "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set." *Bonfring International Journal of Man Machine Interface* 1 (1): 05–09. https://doi.org/10.9756/BIJMMI.1002.

Larose, Daniel T., and Chantel D. Larose. 2015. "Data Mining and Predictive Analytics." *Wiley Series on Methods and Applications in Data Mining*. 2nd Ed, Wiley Publications. ISBN: 9781118116197

Leonard, Cynthia. 2015. "The Importance of Securing Big Data: The Numbers Don't Lie." Datanami. August 6, 2015. https://www.datanami.com/2015/08/06/the-importance-of-securing-big-data-the-numbers-dont-lie/.

Liu, Ying. 2014. "Big Data and Predictive Business Analytics." *The Journal of Business Forecasting* 33 (4): 18–21. https://www.questia.com/library/journal/1P3-3601906361/big-data-and-predictive-business-analytics.

luhaybi, Mashael Al, Allan Tucker, and Leila Yousefi. 2018. "The Prediction of Student Failure Using Classification Methods: A Case Study." In *Computer Science & Information Technology*, 79–90. Academy & Industry Research Collaboration Center (AIRCC). https://doi.org/10.5121/csit.2018.80506.

Lukkarinen, Anna, Paula Koivukangas, and Tomi Seppälä. 2016. "Relationship between Class Attendance and Student Performance." *Procedia - Social and Behavioral Sciences* 228 (July): 341–47. https://doi.org/10.1016/j.sbspro.2016.07.051.

Ma, Li-Chen, Wooster, and Robert A. 2009. "Marital Status and Academic Performance in College." *College Student Journal* 13 (2): 106–11.

Madia, Kimberly. 2012. "Top Tips for Securing Big Data Environments ." IBM Big Data & Analytics Hub. September 14, 2012. https://www.ibmbigdatahub.com/blog/top-tips-securing-big-data-environments.

Markus, Hazel Rose, and Shinobu Kitayama. 1991. "Culture and the Self: Implications for Cognition, Emotion, and Motivation." *Psychological Review* 98 (2): 224–53. https://doi.org/10.1037/0033-295X.98.2.224.

Mazza, Riccardo, and Christian Milani. 2004. "GISMO: A Graphical Interactive Student Monitoring Tool for Course Management Systems." *T.E.L.'04 Technology Enhanced Learning International Conference. Milan*, 1–8. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.4480.

McMullen, Maram. 2014. "The Value and Attributes of an Effective Preparatory English Program: Perceptions of Saudi University Students." *English Language Teaching* 7 (7). https://doi.org/10.5539/elt.v7n7p131.

Meir, Ron, and Gunnar Rätsch. 2003. "An Introduction to Boosting and Leveraging." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2600: 118–83. https://doi.org/10.1007/3-540-36434-X_4.

Merceron, Agathe, and Kalina Yacef. 2010. "Measuring Correlation of Strong Symmetric

Association Rules in Educational Data." In: *Handbook of Educational Data Mining*, 1st Ed. CRC Press, Taylor & Francis Group. https://doi.org/10.1201/b10274

Mohamadian, Zahra, Sedighe Fallah, Ali Safdarian, and Zahra Jalali. 2015. "Online) An Open Access." *Indian Journal of Fundamental and Applied Life Sciences* 5 (S1): 1262–70.

Moore, Don A., Amelia S. Dev, and Ekaterina Y. Goncharova. 2018. "Overconfidence Across Cultures." *Collabra: Psychology* 4 (1): 36. https://doi.org/10.1525/collabra.153.

Moores, Trevor T., and Jerry Cha Jan Chang. 2009. "Self-Efficacy, Overconfidence, and the Negative Effect on Subsequent Performance: A Field Study." *Information and Management* 46 (2): 69–76. https://doi.org/10.1016/J.IM.2008.11.006.

Moura, José, and Carlos Serrão. 2015. "Security and Privacy Issues of Big Data." In *Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence*, edited by Noor Zaman, Mohamed Elhassan Seliaman, Mohd Fadzil Hassan, and Fausto Pedro García Márquez. https://doi.org/10.4018/978-1-4666-8505-5.ch002.

Nasrullah, Wildan Adji, Judi Prajetno Sugiono, Joan Santoso, and Agus Djaja Gunawan. 2021. "Predicting Student's Failure in Education Based on Dropout Status." *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, April, 183–88. https://doi.org/10.1109/EICONCIT50028.2021.9431905.

National Research Council. 2002. "Methodological Advances in Cross-National Surveys of Educational Achievement." *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington, DC: National Academies Press. https://doi.org/10.17226/10322.

Nguyen Thai Nghe, Paul Janecek, and Peter Haddawy. 2007. "A Comparative Analysis of Techniques for Predicting Academic Performance." In *2007 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports*, T2G-7-T2G-12. IEEE. https://doi.org/10.1109/FIE.2007.4417993.

Nithya, P, B Umamaheswari, and A Umadevi. 2016. "A Survey on Educational Data Mining in Field of Education." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 5 (1): 1–16.

OECD indicators. 2015. "Education at a Glance 2015 OECD INDICATORS." https://doi.org/10.1787/eag-2015-en.

Osmanbegović, Edin, Mirza Suljic, and Mirza Suljić. 2012. "Data Mining Approach for Predicting Student Performance." *Journal of Economics and Business* X (1).

Pal, Ajay Kumar, and Saurabh Pal. 2013. "Analysis and Mining of Educational Data for Predicting the Performance of Students." *International Journal of Electronics Communication and Computer Engineering* 4 (5): 2278–4209.

Peña-Ayala, Alejandro. 2014. "Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works." *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2013.08.042.

Peterson, B, and P S J D Baker. 2010. "Data Mining for Education." *International Encyclopedia of Education*. Elsevier.

Pinheiro, Rómulo, Gerald Wangenge-Ouma, Elizabeth Balbachevsky, and Yuzhuo Cai. 2015. “The Role of Higher Education in Society and the Changing Institutionalized Features in Higher Education.” In *The Palgrave International Handbook of Higher Education Policy and Governance*, 225–42. London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-137-45617-5_13.

Poon, Leonard K.M., Siu Cheung Kong, Michael Y.W. Wong, and Thomas S.H. Yau. 2017. “Mining Sequential Patterns of Students’ Access on Learning Management System.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10387 LNCS:191–98. Springer Verlag. https://doi.org/10.1007/978-3-319-61845-6_20.

Pradeep, Anjana, and Jeena Thomas. 2015. “Predicting College Students Dropout Using EDM Techniques.” *International Journal of Computer Applications* 123 (5): 26–34.

Prinsloo, Paul, and Sharon Slade. 2013. “An Evaluation of Policy Frameworks for Addressing Ethical Considerations in Learning Analytics.” In *ACM International Conference Proceeding Series*, 240–44. https://doi.org/10.1145/2460296.2460344.

Prinsloo, Paul, and Sharon Slade. 2015. “Student Privacy Self-Management: Implications for Learning Analytics.” In: *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15).* Association for Computing Machinery, New York, NY, USA, 83–92. https://doi.org/10.1145/2723576.2723585.

Raj, Vivek, and S. K. Manivannan. 2020. “Predicting Student Failure in University Examination Using Machine Learning Algorithms.” *International Journal of Innovative Technology and Exploring Engineering* 9 (5): 956–59. https://doi.org/10.35940/IJITEE.E2643.039520.

Ransdell, Sarah. 2001. “Predicting College Success: The Importance of Ability and Non-Cognitive Variables.” *International Journal of Educational Research* 35 (4): 357–64. https://doi.org/10.1016/S0883-0355(01)00032-5.

Riding, R. J., and Stephen Rayner. 1998. *Cognitive Styles and Learning Strategies : Understanding Style Differences in Learning and Behaviour*.

Riffai, M. M.M.A., Peter Duncan, David Edgar, and Ahmed Hassan Al-Bulushi. 2016. “The Potential for Big Data to Enhance the Higher Education Sector in Oman.” In *2016 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016*, 79–84. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICBDSC.2016.7460346.

Riyadh Economic Forum. 2011. “Technical Educational and Vocational Training, and Its Suitability for the Development Needs of the Workforce.” Riyadh, Saudi Arabia: Riyadh Economic Forum.

Romero, C, and S Ventura. 2007. “Educational Data Mining: A Survey from 1995 to 2005.” *ScienceDirect* 33 (33): 134–46. https://doi.org/10.1016/j.eswa.2006.04.005.

Romero, Cristobal, and Sebastian Ventura. 2013. “Data Mining in Education.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3 (1): 12–27. https://doi.org/10.1002/widm.1075.

Romero, Cristóbal, Sebastián Ventura, and Paul De Bra. 2004. "Knowledge Discovery with Genetic Programming for Providing Feedback to Courseware Authors." *User Modelling and User-Adapted Interaction* 14 (5): 425–64. https://doi.org/10.1007/s11257-004-7961-2.

Romero, Cristóbal, Amelia Zafra, Jose María Luna, and Sebastián Ventura. 2013. "Association Rule Mining Using Genetic Programming to Provide Feedback to Instructors from Multiple-Choice Quiz Data." *Expert Systems* 30 (2): 162–72. https://doi.org/10.1111/j.1468-0394.2012.00627.x.

Rotem, Nir, Gad Yair, and Elad Shustak. 2020. "Dropping out of Master's Degrees: Objective Predictors and Subjective Reasons." *Higher Education Research & Development,* 40 (5): 1070–1084. https://doi.org/10.1080/07294360.2020.1799951.

Rothman, Steven B. 2007. "Understanding Data Quality through Reliability: A Comparison of Data Reliability Assessment in Three International Relations Datasets." *International Studies Review* 9 (3): 437–456. https://doi.org/10.1111/j.1468-2486.2007.00698.x.

Scholes, Laura. 2020. "Social and Cultural Influences on Academic Achievement." *The Encyclopedia of Child and Adolescent Development*, January, 1–12. https://doi.org/10.1002/9781119171492.WECAD376.

Selwyn, Neil. 2011. *Education and Technology : Key Issues and Debates*. 1st ed. Continuum International Pub. Group.

Sembiring, Sajadin, M Zarlis, Dedy Hartama, Elvi Wani, and Program Magister. 2011. "Prediction of Student Academic Performance by an Application of Data Mining Techniques." In *International Conference on Management and Artificial Intelligence*, 6:110–14.

Shakeel, Khawar, and Naveed Anwer Butt. 2015. "Educational Data Mining to Reduce Student Dropout Rate by Using Classification." In *253rd OMICS International Conference on Big Data Analysis & Data Mining*. Lexington.

Sidi, Fatimah, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. 2012. "Data Quality: A Survey of Data Quality Dimensions." In *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*, 300–304. https://doi.org/10.1109/InfRKM.2012.6204995.

Silva, Carla, and José Fonseca. 2017. "Educational Data Mining: A Literature Review." *Advances in Intelligent Systems and Computing* 520: 87–94. https://doi.org/10.1007/978-3-319-46568-5_9.

Simsek, Ali, and Jale Balaban. 2010. "Learning Strategies of Successful and Unsuccessful University Students." *Contemporary Educational Technology* 1 (1): 36–45.

Smirani, Lassaad K., Hanaa A. Yamani, Leila Jamel Menzli, and Jihane A. Boulahia. 2022. "Using Ensemble Learning Algorithms to Predict Student Failure and Enabling Customized Educational Paths." Edited by Chenxi Huang. *Scientific Programming* 2022 (April): 1–15. https://doi.org/10.1155/2022/3805235.

Song, Yan-Yan, and Ying Lu. 2015. "Decision Tree Methods: Applications for Classification and Prediction." *Shanghai Archives of Psychiatry* 27 (2): 130–35. https://doi.org/10.11919/j.issn.1002-0829.215044.

Stansfield, Mark, Evelyn Mclellan, and Thomas Connolly. 2004. "Enhancing Student Performance in Online Learning and Traditional Face-to-Face Class Delivery." *Journal of Information Technology Education*. Vol. 3. http://2003.insite.nu.

Sturman, Michael C. 2003. "Searching for the Inverted U-Shaped Relationship Between Time and Performance: Meta-Analyses of the Experience/Performance, Tenure/Performance, and Age/Performance Relationships." https://doi.org/10.1016/S0149-2063_03_00028-X.

Survey Automation Software - EvaSys and EvaExam. 2019. https://en.evasys.de/main/home.html.

Szafran, Robert F, and Stephen F Austin. 2002. "The Effect of Academic Load on Success for New College Students: Is Lighter Better?" *NACADA Journal*. Vol. 22.

Tacoma, Sietske, Sergey Sosnovsky, Peter Boon, Johan Jeuring, and Paul Drijvers. 2018. "The Interplay between Inspectable Student Models and Didactics of Statistics." *Digital Experiences in Mathematics Education* 4 (2–3): 139–62. https://doi.org/10.1007/s40751-018-0040-9.

Tankard, Colin. 2012. "Big Data Security." *Network Security* 2012 (7): 5–8. https://doi.org/10.1016/s1353-4858(12)70063-6.

Tee, Sing What, Paul L. Bowen, Peta Doyle, and Fiona H. Rohde. 2007. "Factors Influencing Organizations to Improve Data Quality in Their Information Systems." *Accounting & Finance* 47 (2): 335–55. https://doi.org/10.1111/j.1467-629X.2006.00205.x.

Tomar, Divya, and Sonali Agarwal. 2013. "A Survey on Data Mining Approaches for Healthcare." *International Journal of Bio-Science and Bio-Technology* 5 (5): 241–66. https://doi.org/10.14257/ijbsbt.2013.5.5.25.

Tylor, Edward. 1871. "Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom." *Primitive Culture* 1. https://openlibrary.org/books/OL6946625M/Primitive_culture.

Villwock, Rosangela, Andressa Appio, and Aldioni Adaiani Andreta. 2015. "Educational Data Mining with Focus on Dropout Rates." *IJCSNS International Journal of Computer Science and Network Security* 15 (3).

Wait, Isaac W., and Justin W. Gressel. 2009. "Relationship Between TOEFL Score and Academic Success for International Engineering Students." *Journal of Engineering Education* 98 (4): 389–98. https://doi.org/10.1002/j.2168-9830.2009.tb01035.x.

Wang, Richard Y, and Diane M Strong. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems* 12 (4): 5–33.

Warren, Samuel D, and Louis D Brandeis. 1890. "The Right to Privacy." *Law Review*. Vol. 4. http://links.jstor.org/sici?sici=0017-811X%2818901215%294%3A5%3C193%3ATRTP%3E2.0.CO%3B2-C.

Watkins, D., and J. Hattie. 1985. "A Longitudinal Study of the Approaches to Learning of Australian Tertiary Students. Human Learning." *Journal of Practical Research &*

Applications 4 (2): 127–41.

Westin, Alan F. 1968. *Washington and Lee Law Review Privacy And Freedom*. 1st ed. Vol. 25. https://scholarlycommons.law.wlu.edu/wlulr/vol25/iss1/20.

Wolf, M. A., R. Jones, S. Hall, and B Wise. 2014. "Capacity Enablers And Barriers For Learning Analytics: Implications For Policy And Practice." www.all4ed.org.

Yadav, Surjeet Kumar, Brijesh Bharadwaj, and Saurabh Pal. 2011. "Data Mining Applications: A Comparative Study for Predicting Student's Performance." *International Journal of Innovative Technology and Creative Engineering* 1 (12): 13–19.

Yadav, Surjeet Kumar, and Saurabh Pal. 2012. "Data Mining: A Prediction for Performance Improvement of Engineering Students Using Classification." *World of Computer Science and Information Technology Journal (WCSIT)* 2 (2): 51–56.

Yassein, Nawal Ali, Rasha Gaffer, M Helali, and Somia B Mohomad. 2017. "Citation: Yassein NA, Helali RGM, Mohomad SB (2017) Predicting Student Academic Performance in KSA Using Data Mining Techniques." *J Inform Tech Softw Eng* 7 (5): 213. https://doi.org/10.4172/2165-7866.1000213.

Yates, J.Frank, Ju-Whei Lee, and Julie GG. Bush. 1997. "General Knowledge Overconfidence: Cross-National Variations, Response Style, and Reality." *Organizational Behavior and Human Decision Processes* 70 (2): 87–94. https://doi.org/10.1006/OBHD.1997.2696.

Yates, J.Frank, Ju-Whei Lee, Hiromi Shinotsuka, Andrea L Patalano, and Winston R Sieck. 1998. "Cross-Cultural Variations in Probability Judgment Accuracy: Beyond General Knowledge Overconfidence?" *Organizational Behavior and Human Decision Processes* 74 (2): 89–117. https://doi.org/10.1006/OBHD.1998.2771.

Yess, J. P. 2009. "Influence of Marriage on the Scholastic Achievement of Community College Students. Humanities, Social Sciences and Law." *American Journal of Educational Research.* 4 (2): 103–18.

York, Travis T, Charles Gibson, and Susan Rankin. 2015. "Defining and Measuring Academic Success - Practical Assessment, Research &amp; Evaluation" 20 (5).

Yu-Wei, Chiu (David Chiu). 2015. *Machine Learning with R Cookbook : Explore over 110 Recipes to Analyze Data and Build Predictive Models with the Simple and Easy-to-Use R Code*. Packt Publishing.

Yue, Hongtao, and Xuanning Fu. 2017. "Rethinking Graduation and Time to Degree: A Fresh Perspective." *Research in Higher Education* 58 (2): 184–213. https://doi.org/10.1007/S11162-016-9420-4/TABLES/5.

Zhao, Yijun, Qiangwen Xu, Ming Chen, and Gary M Weiss. 2020. "Predicting Student Performance in a Master of Data Science Program Using Admissions Data." In *Proceedings of The 13th International Conference on Educational Data Mining*, 325–33.

Zimmermann, Judith, Kay H. Brodersen, Jean-Philippe Pellet, Elias August, and Joachim M Buhmann. 2011. "Predicting Graduate-Level Performance from Undergraduate Achievements." In *Proceedings of the 4th International Conference on Educational Data Mining*. July 6-8, Eindhoven, The Netherlands

APPENDICES

Appendix A: Examples of EDM studies based on their method and application

Table 21: Classification of EDM studies based on their method and application

Author/ Year	Study description	DM Method	DM Application
Pal (2012)	Predict which students will likely drop out within the first year in a university program.	Prediction/ Relationship mining	Student modeling/ Scientific research
Dekker, Pechenizkiy, & Vleeshouwers (2009)	Predicting the students drop out in an Electrical Engineering program.	Prediction/ Relationship mining	Student modeling/ Scientific research
Macfadyen & Dawson (2010)	Finding out which LMS tracking data features correlate significantly with students' academic performance and investigating which student's prediction model is more effective.	Prediction/ Relationship mining	Student modeling/ Scientific research/ Pedagogical support
Al luhaybi, Tucker, & Yousefi (2018)	Grouping students according to their final results and predicting which students' are at risk of failure in computer science core courses.	Clustering/ Prediction/ Relationship mining	Student modeling/ Scientific research
Alturki, Alturki, & Stuckenschmidt (2021)	Predict honorary students in a bachelor's degree program.	Prediction/ Relationship mining	Student modeling/ Scientific research
Varghese et al., (2011)	Model students based on their attendance, internal mark assessment, seminar assessment, class assignment assessment, and gained marks to assist in formulating the schedule for internal assessments and the curriculum.	Clustering/ Relationship mining	Student modeling/ Scientific research
Pavlik, Cen, Wu, & Koedinger (2008)	Determining a skill model by considering the covariation of individual features that could subsequently be used to produce improvements to a Cognitive Tutor.	Clustering/ Relationship mining	Domain modeling/ knowledge structures/ Scientific research
Hung et al. (2017)	Identifying at-risk online students from the tenth week of their studying program.	Clustering	Student modeling/ Scientific research/ Pedagogical support
Harwati, Alfiani, & Wulandari, (2015)	Mapping students using the K-mean Cluster algorithm to find hidden patterns and classifying students based on their demographics and attendance.	Clustering/ Relationship mining	Student modeling/ Scientific research

Barnes, Bitzer, & Vouk (2005)	Compare the q-matrix method with factor analysis and k-means cluster analysis for fitting and understanding data using data collected from online settings.	Clustering	Domain modeling and knowledge structures/ Scientific research
Baker (2007)	Exploring whether state or trait features are better predictors of how much students' are likely gaming the system.	Relationship mining	Scientific research
Ali et al. (2013)	Examining the features influencing academic achievements of graduate students.	Relationship mining	Scientific research
Beck & Mostow (2008)	Relating students' achievement to the amount of each type of pedagogical support a student has taken to find out how effective each type of pedagogical support is for inhancing learning.	Relationship mining	Pedagogical support
Baker, Corbett, & Wagner (2006)	Displaying sub-sections of a dataset in a text format and labeling it by human coders to be the foundation for developing a predictor.	Distillation of data for human judgment	Student modeling/ Scientific research
Hershkovitz & Nachmias (2008)	Providing an intensive representation of students' behavior in a specific period to produce a conceptual framework and a tool for determining the motivation of online students.	Distillation of data for human judgment	Student modeling/ Scientific research/ Pedagogical support
Gobert, Hershkovitz, Baker, Wixon, & Pedro (2013)	Studying the motivations related to carelessness to build an automated predictor of students' carelessness.	Discovery with models	Scientific research

Appendix B: Prerequisite Courses for succeeding the Buiness Informatics master's program

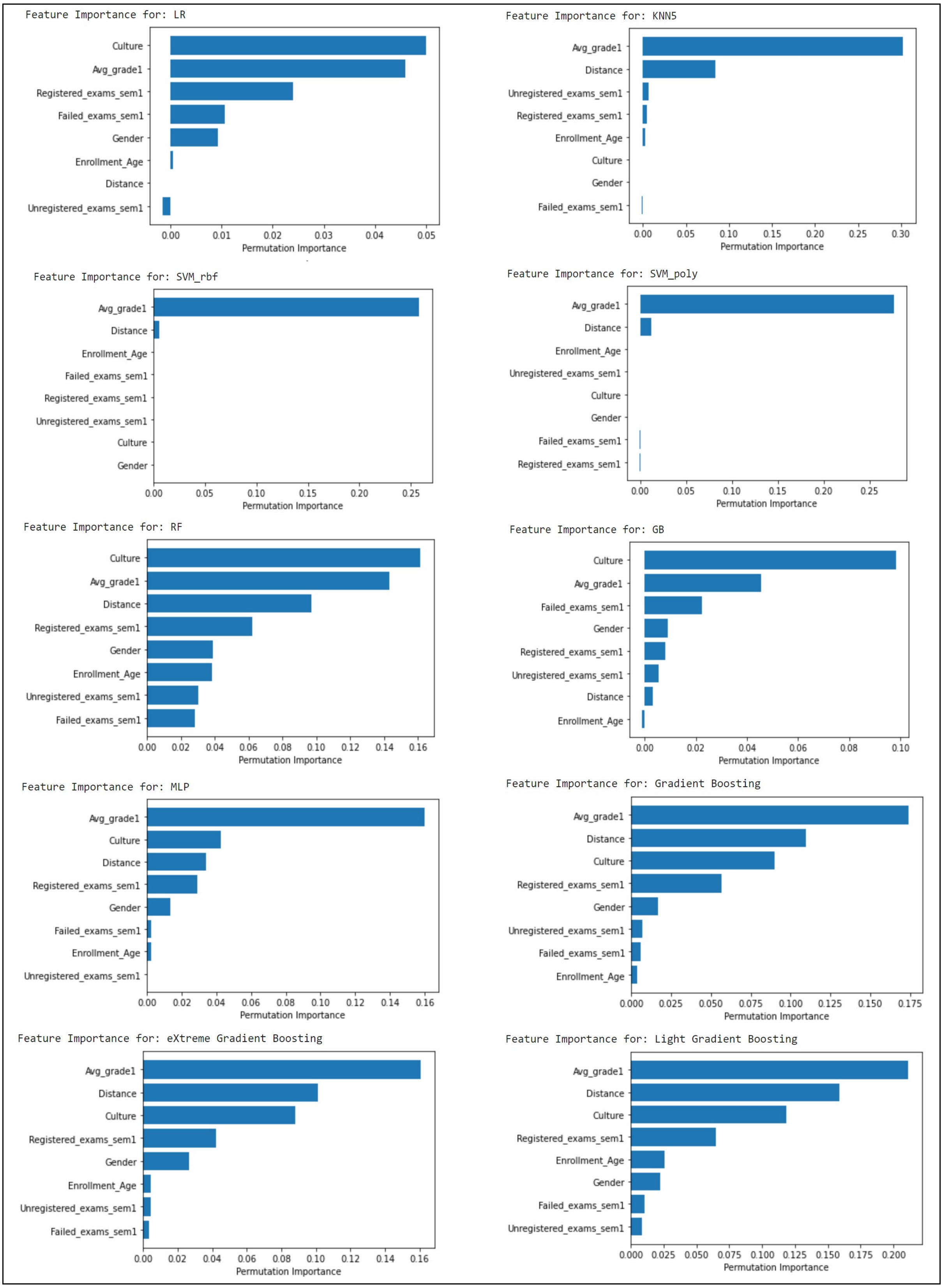
The applicants of the Business Informatics master program at the University of Mannheim need to have prior knowledge in some areas before entering the program. For instance, to be accepted into some courses, the student needs to complete a similar course in the same or a related subject, at a lower grade level. Having that type of knowledge will enable him/her to understand the courses in a less problematical manner. The following table represents the prerequisites that are required for different courses.

Table 22: Master courses of the Business Informatics program and their prerequisites

Prerequisite	Master Course
Linear Algebra	1- Algorthmics (CS 550) 2- Data Mining & Matrics (IE 673) 3- Heigher Level Computer Vision 4- Image processing 5- Text Analtics (IE 661) 6- Information Retreavel & web Search (IE 663)
Statistics	1- Algorthmics (CS 550) 2- Transactions Systems 3- Anfrageobtimierung (IE 630) 4- Hot Topics in Machine Learning (IE 674) 5- Data Mining 6- Methods & Theories in IS (IS 541)
Probability Theory	1- Information Retrievel & web Search 2- Text Analytics (IE 661) 3- Decision support (IE 560)
Databases	1- Transaction systems 2- Anfrageobtimierung (IE 630) 3- Large-scale data management
Algorthmen & Datenstrukturen	1- Algorthmics (CS 550) 2- Database system || (CS 530)
Programming Skills	1- Information Retreavel & web Search (IE 663) 2- Web mining (IE 671) 3- Semantic web technology 4- Data mining || (IE 500) 5- Web data integration (IE 670) 6- Database system || (CS 530) 7- Image Processing 8- Heigher Level Computer vision 9- Self-organized Systems (IS 627) 10- Pervasive Computing (IS 625) 11- System SW (IS 553)
Basic knowledge of propositional & first-order logic	1- Decicion support (IE 560)
Management of Enterprise Systems	1. Business Intelligence & Management Support System (IS 602) 2. Product Management & Product Design for SW (IS 629)

Appendix C: Results of using permutation feature importance function on different algorithms

Model 1: Predicting the academic status after the 1st semester feature importance


Feature Importance for: LR

Feature Importance for: KNN5

Feature Importance for: SVM_rbf

Feature Importance for: SVM_poly

Feature Importance for: RF

Feature Importance for: GB

Feature Importance for: MLP

Feature Importance for: Gradient Boosting

Feature Importance for: eXtreme Gradient Boosting

Feature Importance for: Light Gradient Boosting

Figure 24: Feature importance for predicting the academic status after the 1st semester

Model 2: Predicting the academic status after the 2nd semester feature importance

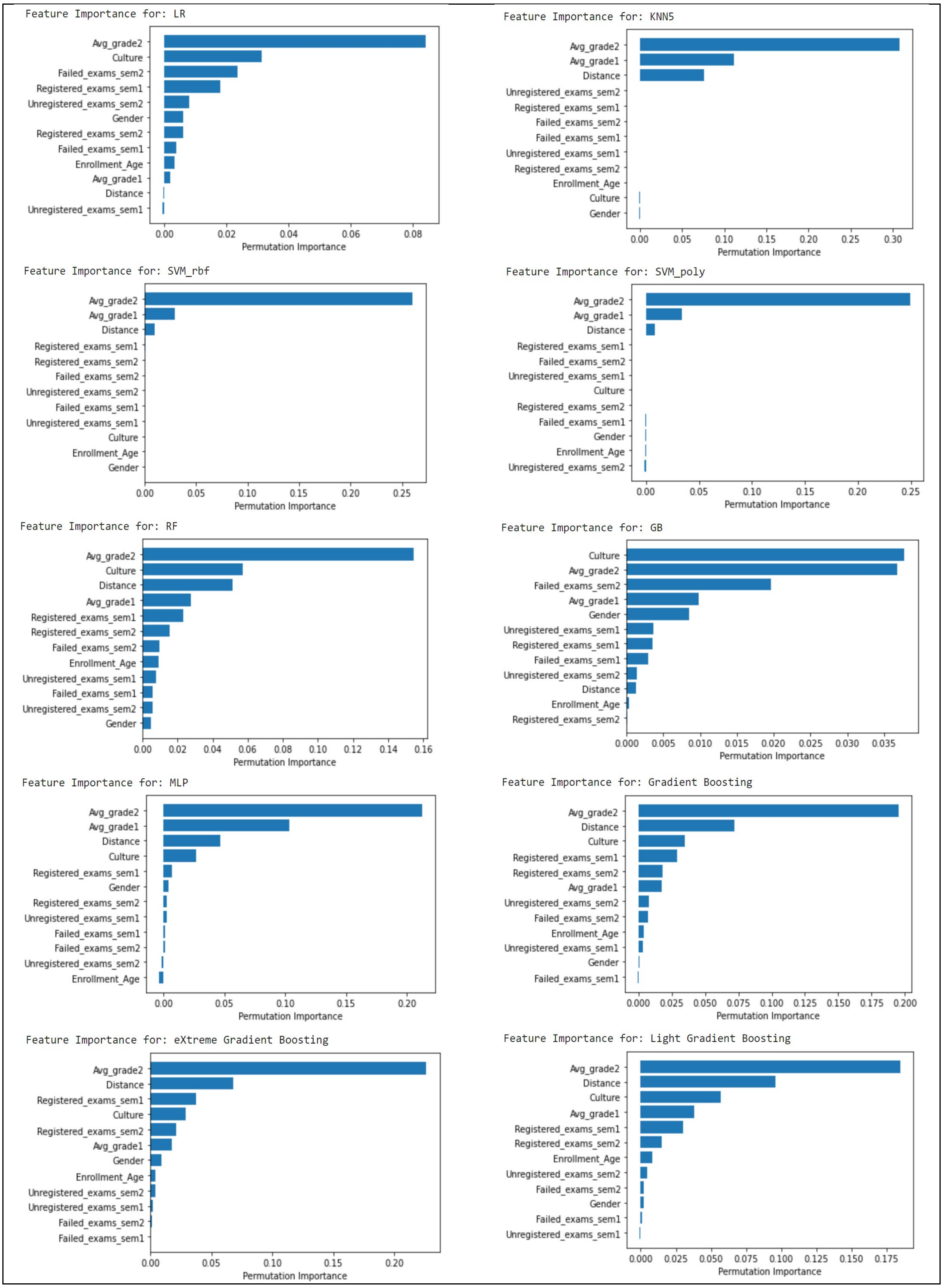


Figure 25: Feature importance for predicting the academic status after the 2nd semester

Model 3: Predicting the academic grade after the 1st semester feature importance

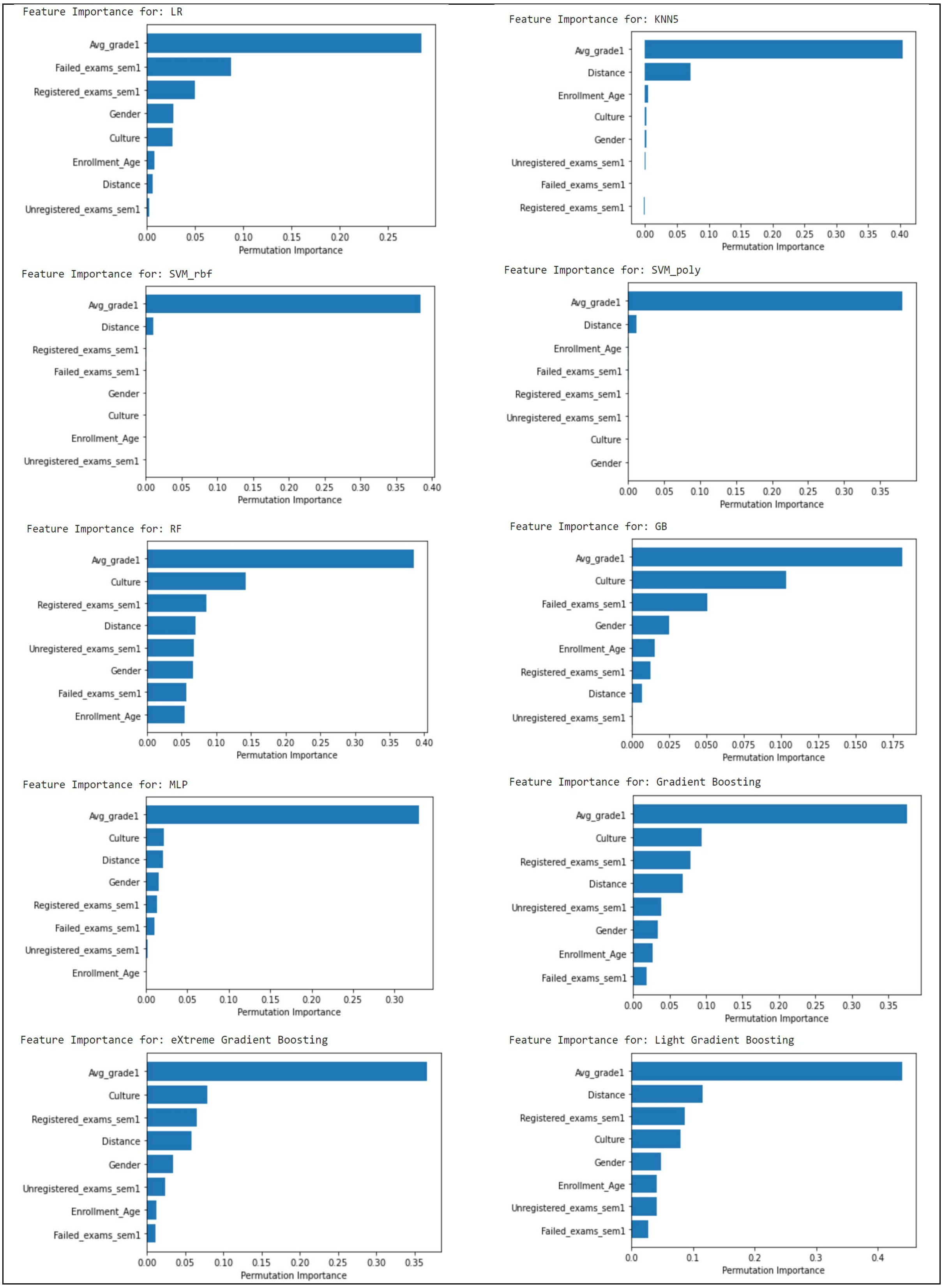


Figure 26: Feature importance for predicting the academic grade after the 1st semester

Model 4: Predicting the academic grade after the 2nd semester feature importance

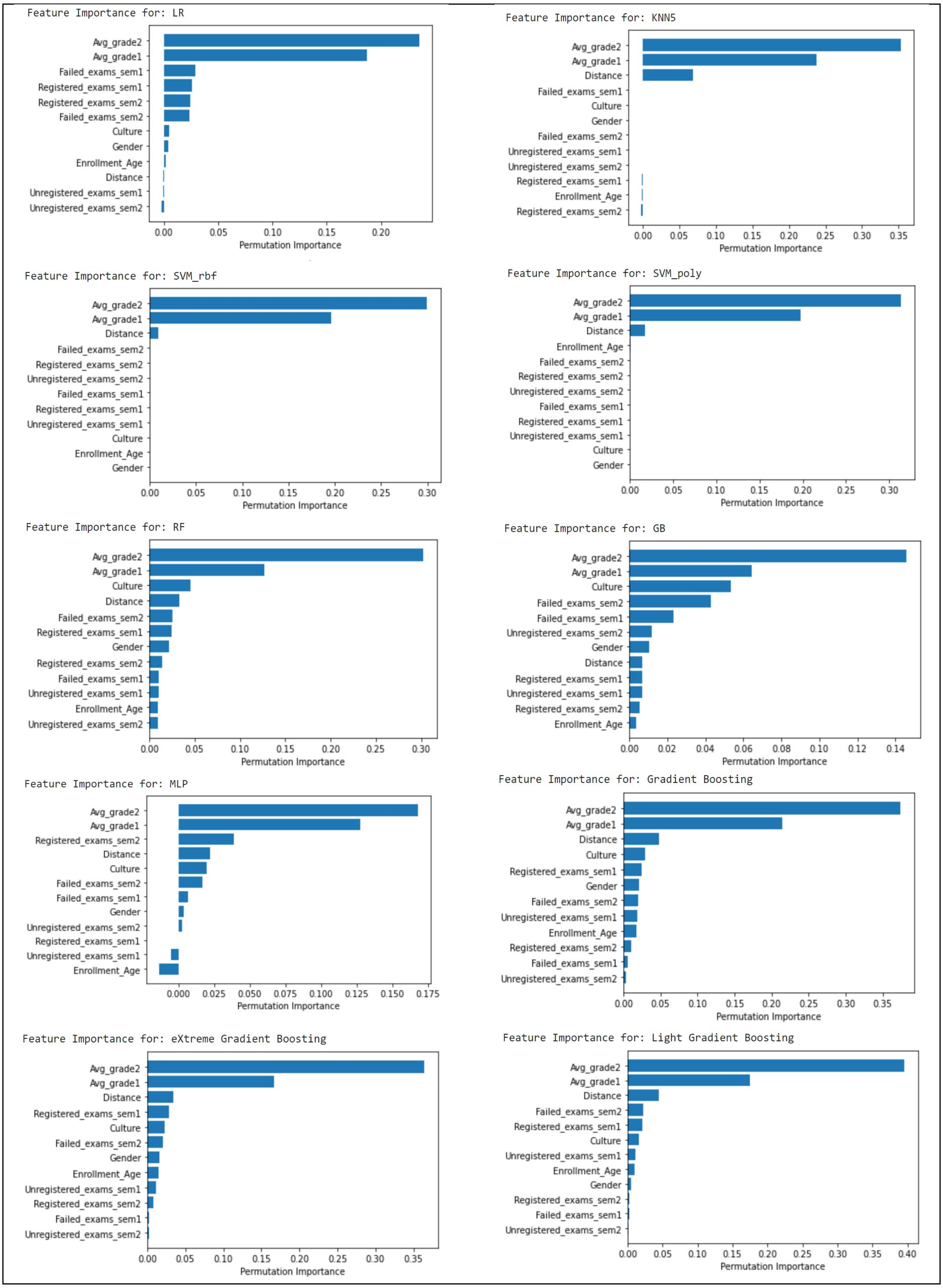


Figure 27: Feature importance for predicting the academic grade after the 2nd semester

Appendix D: Relationship between academic achievement and other features in the master’s program

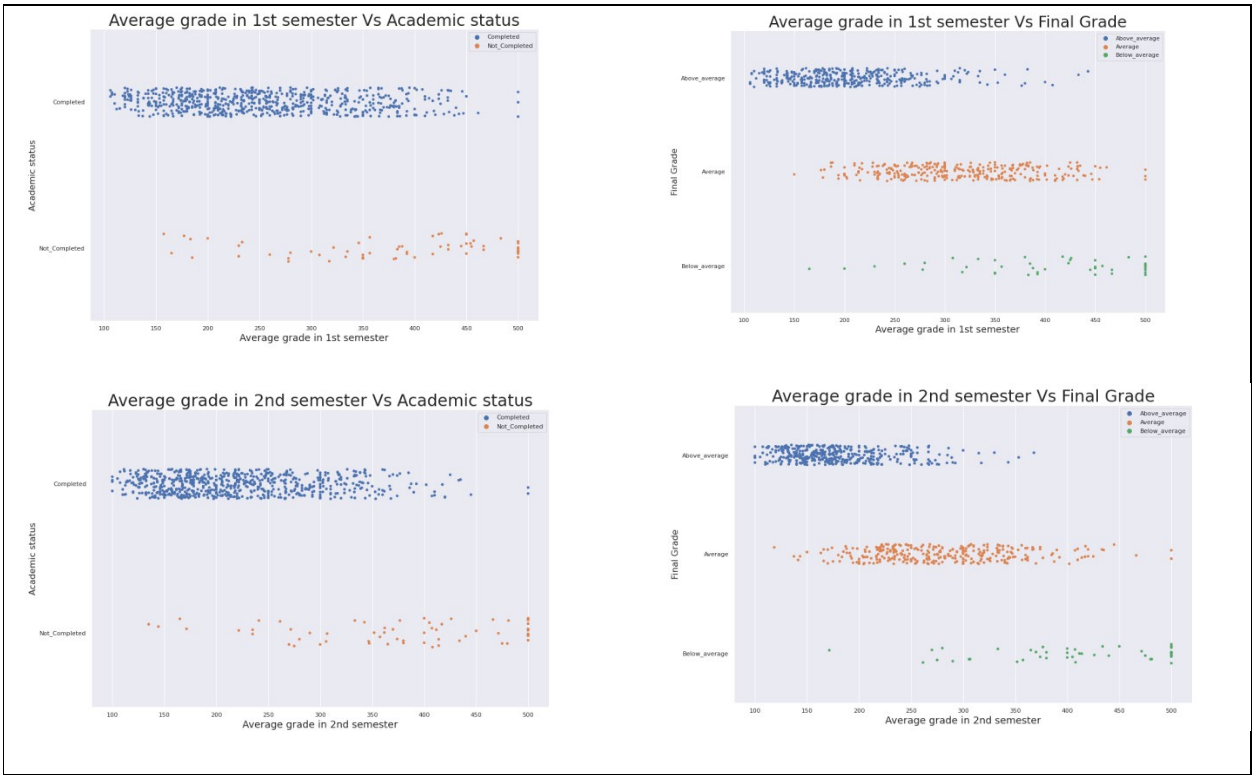


Figure 28: Relationship between academic achievement and each semester’s grade in the Business Informatics master’s program

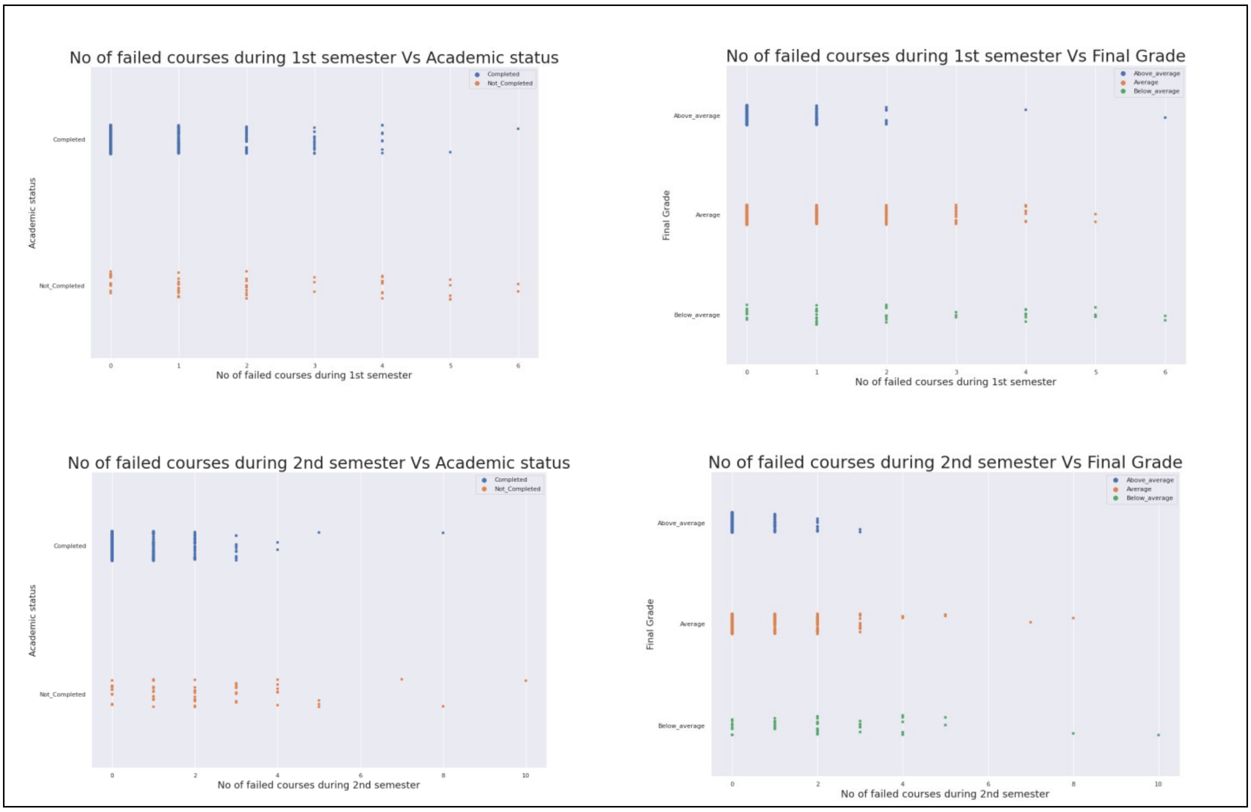
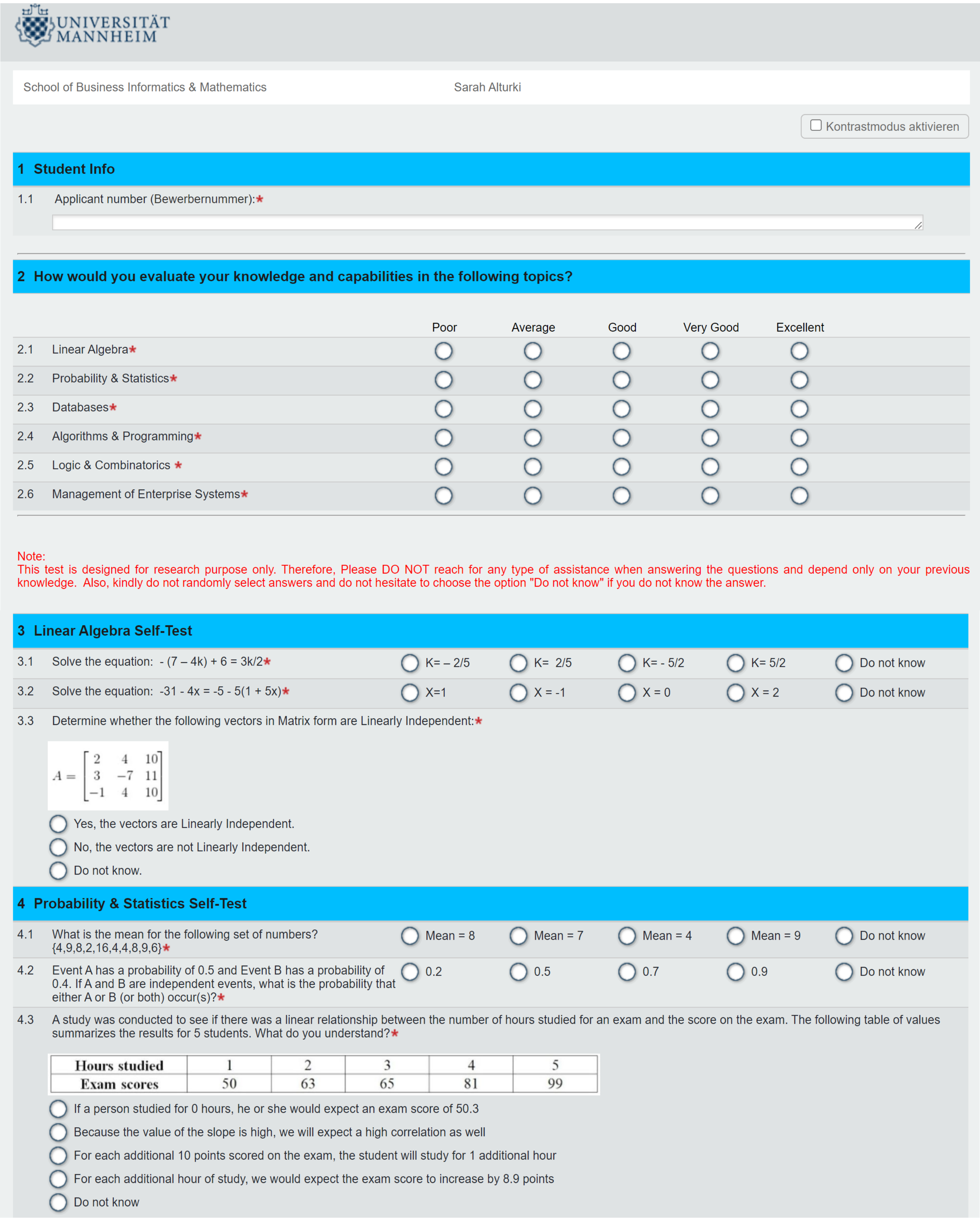


Figure 29: Relationship between academic achievement and the number of failed courses each semester in the Business Informatics master’s program

Appendix E: Self-assessment and Self-test Survey



UNIVERSITÄT MANNHEIM

School of Business Informatics & Mathematics   Sarah Alturki

□ Kontrastmodus aktivieren

1 Student Info

1.1 Applicant number (Bewerbernummer):*

2 How would you evaluate your knowledge and capabilities in the following topics?

		Poor	Average	Good	Very Good	Excellent
2.1	Linear Algebra*	○	○	○	○	○
2.2	Probability & Statistics*	○	○	○	○	○
2.3	Databases*	○	○	○	○	○
2.4	Algorithms & Programming*	○	○	○	○	○
2.5	Logic & Combinatorics *	○	○	○	○	○
2.6	Management of Enterprise Systems*	○	○	○	○	○

Note:
This test is designed for research purpose only. Therefore, Please DO NOT reach for any type of assistance when answering the questions and depend only on your previous knowledge. Also, kindly do not randomly select answers and do not hesitate to choose the option "Do not know" if you do not know the answer.

3 Linear Algebra Self-Test

3.1	Solve the equation: - (7 – 4k) + 6 = 3k/2*	○ K= – 2/5	○ K= 2/5	○ K= - 5/2	○ K= 5/2	○ Do not know
3.2	Solve the equation: -31 - 4x = -5 - 5(1 + 5x)*	○ X=1	○ X = -1	○ X = 0	○ X = 2	○ Do not know

3.3 Determine whether the following vectors in Matrix form are Linearly Independent:*

$$A = \begin{bmatrix} 2 & 4 & 10 \\ 3 & -7 & 11 \\ -1 & 4 & 10 \end{bmatrix}$$

○ Yes, the vectors are Linearly Independent.

○ No, the vectors are not Linearly Independent.

○ Do not know.

4 Probability & Statistics Self-Test

4.1	What is the mean for the following set of numbers? {4,9,8,2,16,4,4,8,9,6}*	○ Mean = 8	○ Mean = 7	○ Mean = 4	○ Mean = 9	○ Do not know
4.2	Event A has a probability of 0.5 and Event B has a probability of 0.4. If A and B are independent events, what is the probability that either A or B (or both) occur(s)?*	○ 0.2	○ 0.5	○ 0.7	○ 0.9	○ Do not know

4.3 A study was conducted to see if there was a linear relationship between the number of hours studied for an exam and the score on the exam. The following table of values summarizes the results for 5 students. What do you understand?*

Hours studied	1	2	3	4	5
Exam scores	50	63	65	81	99

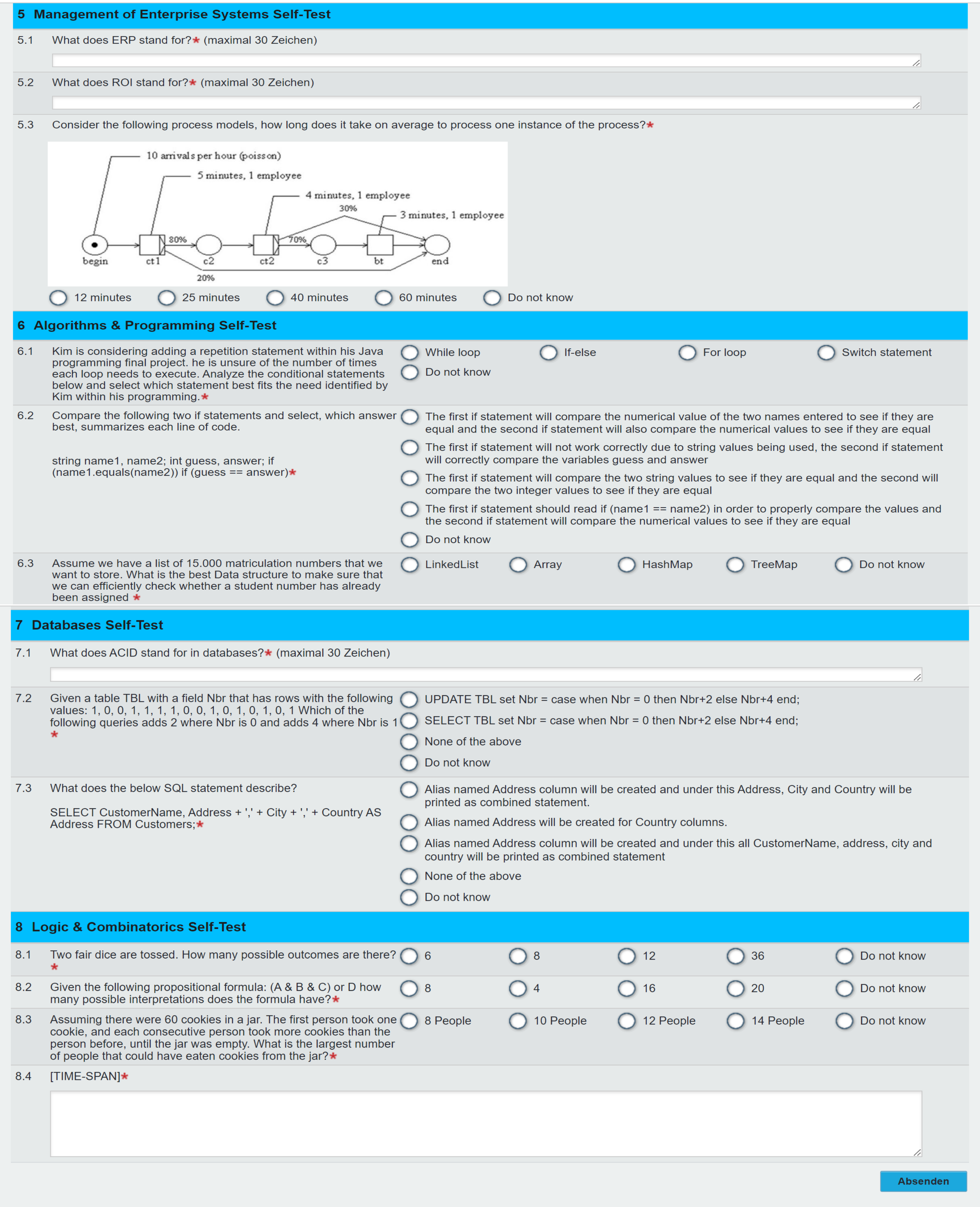
○ If a person studied for 0 hours, he or she would expect an exam score of 50.3

○ Because the value of the slope is high, we will expect a high correlation as well

○ For each additional 10 points scored on the exam, the student will study for 1 additional hour

○ For each additional hour of study, we would expect the exam score to increase by 8.9 points

○ Do not know



5 Management of Enterprise Systems Self-Test

5.1 What does ERP stand for?* (maximal 30 Zeichen)

5.2 What does ROI stand for?* (maximal 30 Zeichen)

5.3 Consider the following process models, how long does it take on average to process one instance of the process?*

○ 12 minutes ○ 25 minutes ○ 40 minutes ○ 60 minutes ○ Do not know

6 Algorithms & Programming Self-Test

6.1 Kim is considering adding a repetition statement within his Java programming final project. he is unsure of the number of times each loop needs to execute. Analyze the conditional statements below and select which statement best fits the need identified by Kim within his programming.*

○ While loop ○ If-else ○ For loop ○ Switch statement ○ Do not know

6.2 Compare the following two if statements and select, which answer best, summarizes each line of code.

string name1, name2; int guess, answer; if (name1.equals(name2)) if (guess == answer)*

○ The first if statement will compare the numerical value of the two names entered to see if they are equal and the second if statement will also compare the numerical values to see if they are equal

○ The first if statement will not work correctly due to string values being used, the second if statement will correctly compare the variables guess and answer

○ The first if statement will compare the two string values to see if they are equal and the second will compare the two integer values to see if they are equal

○ The first if statement should read if (name1 == name2) in order to properly compare the values and the second if statement will compare the numerical values to see if they are equal

○ Do not know

6.3 Assume we have a list of 15.000 matriculation numbers that we want to store. What is the best Data structure to make sure that we can efficiently check whether a student number has already been assigned *

○ LinkedList ○ Array ○ HashMap ○ TreeMap ○ Do not know

7 Databases Self-Test

7.1 What does ACID stand for in databases?* (maximal 30 Zeichen)

7.2 Given a table TBL with a field Nbr that has rows with the following values: 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1 Which of the following queries adds 2 where Nbr is 0 and adds 4 where Nbr is 1 *

○ UPDATE TBL set Nbr = case when Nbr = 0 then Nbr+2 else Nbr+4 end;

○ SELECT TBL set Nbr = case when Nbr = 0 then Nbr+2 else Nbr+4 end;

○ None of the above

○ Do not know

7.3 What does the below SQL statement describe?

SELECT CustomerName, Address + ',' + City + ',' + Country AS Address FROM Customers;*

○ Alias named Address column will be created and under this Address, City and Country will be printed as combined statement.

○ Alias named Address will be created for Country columns.

○ Alias named Address column will be created and under this all CustomerName, address, city and country will be printed as combined statement

○ None of the above

○ Do not know

8 Logic & Combinatorics Self-Test

8.1 Two fair dice are tossed. How many possible outcomes are there? *

○ 6 ○ 8 ○ 12 ○ 36 ○ Do not know

8.2 Given the following propositional formula: (A & B & C) or D how many possible interpretations does the formula have?*

○ 8 ○ 4 ○ 16 ○ 20 ○ Do not know

8.3 Assuming there were 60 cookies in a jar. The first person took one cookie, and each consecutive person took more cookies than the person before, until the jar was empty. What is the largest number of people that could have eaten cookies from the jar?*

○ 8 People ○ 10 People ○ 12 People ○ 14 People ○ Do not know

8.4 [TIME-SPAN]*

Appendix F: Relationship between SSA and different attributes

Relationship between the accuracy of SSA and gender

Table 23: Relationship between the accuracy of SSA and gender using Cramer’s V (Alturki and Stuckenschmidt, 2021)

Topic	Males			Females			Association (Cramer's V)
	Accurate	**Over-estimator**	**Under-estimator**	**Accurate**	**Over-estimator**	**Under-estimator**	
Linear Algebra	48%	33%	19%	38%	16%	46%	0.26
Probability & Statistics	31%	65%	4%	50%	42%	8%	0.19
Databases	31%	50%	19%	42%	28%	30%	0.18
Algorithms & Programming	31%	42%	27%	46%	24%	30%	0.15
Logic & Combinatorics	60%	19%	21%	48%	6%	46%	0.26
Average	**40.2%**	**41.8%**	**18%**	**44.8%**	**23.2%**	**32%**	**0.21**

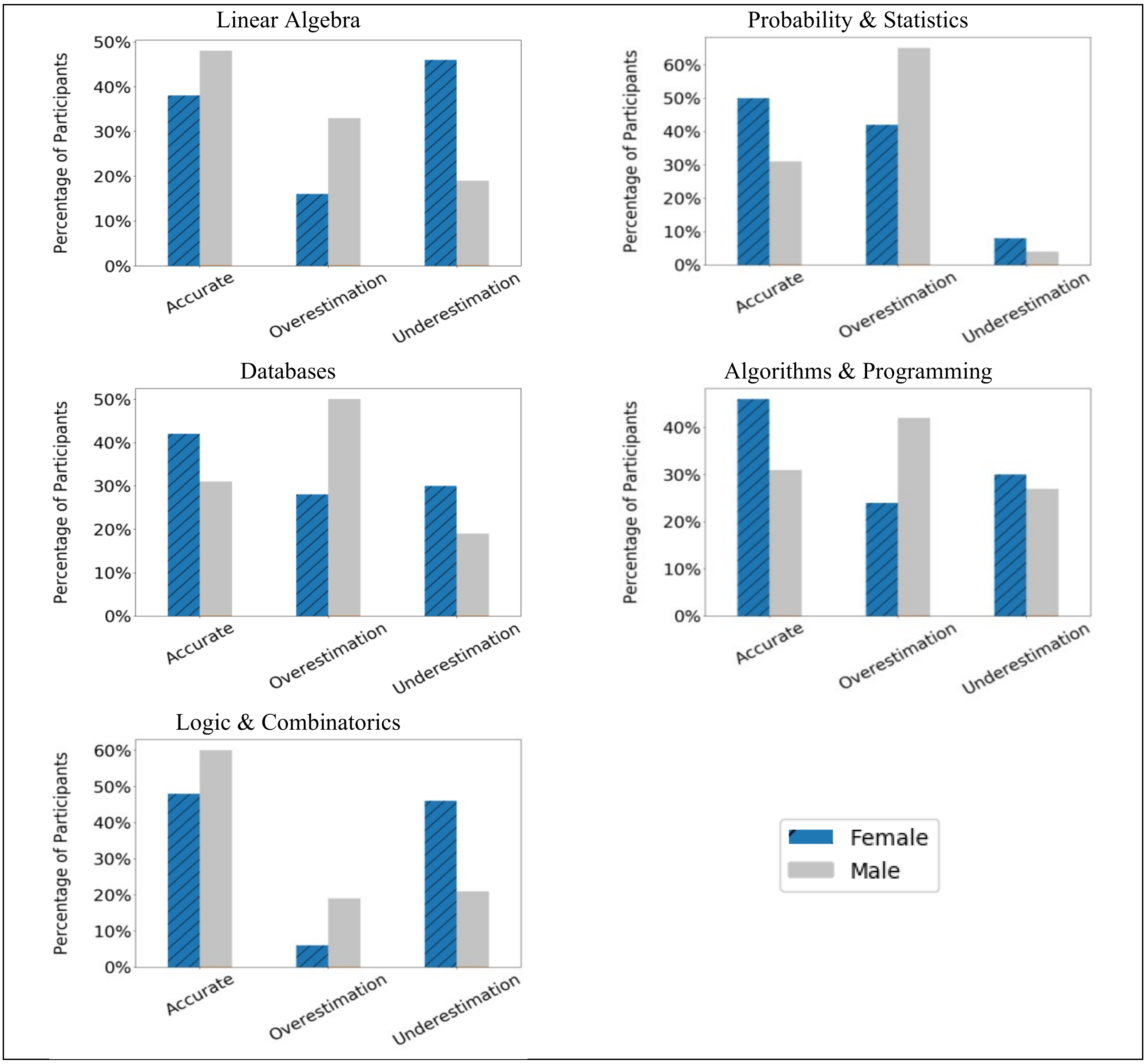


Figure 30: Relationship between the accuracy of SSA and gender (Alturki and Stuckenschmidt, 2021)

Relationship between the accuracy of SSA and the achieved GPA:

Table 24: Relationship between the accuracy of SSA and GPA using ANOVA (Alturki and Stuckenschmidt, 2021).

Topic	Correlation (ANOVA)	
	F-value	**P-value**
Linear Algebra	4.65	0.01
Probability & Statistics	6.04	0.00
Databases	6.53	0.00
Algorithms & Programming	6.23	0.00
Logic & Combinatorics	5.29	0.01

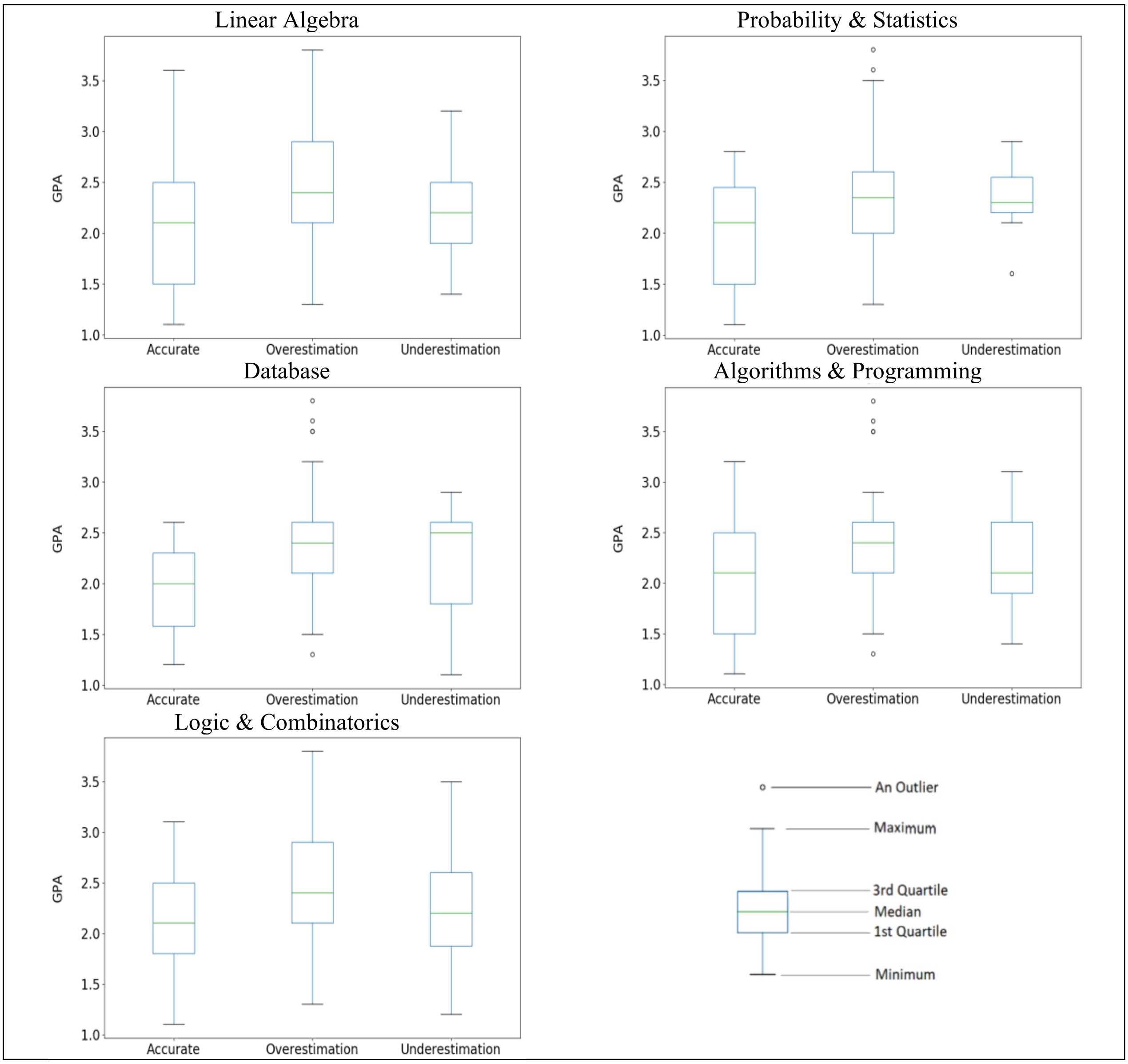


Figure 31: Relationship between the accuracy of SSA and GPA (Alturki and Stuckenschmidt, 2021).