DESCRIPTION AND PREDICTION:

KNOWLEDGE DISCOVERY IN UNIVERSITY DATABASES

MICHAEL KAM BARNGROVER

BOĞAZİÇİ UNIVERSITY

2017

DESCRIPTION AND PREDICTION:

KNOWLEDGE DISCOVERY IN UNIVERSITY DATABASES

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Management

by

Michael Kam Barngrover

Boğaziçi University

2017

DECLARATION OF ORIGINALITY

I, Michael Kam Barngrover, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature .....

Date .....

ABSTRACT

Description And Prediction:

Knowledge Discovery in University Databases

Data mining methods, including machine learning, have been applied for many years in business contexts and are now receiving a great deal of attention from educators and data scientists in higher education. Accurate, early prediction of whether a student is on track to graduate with distinction is a critical tool for administrators, educators, and advisers to ensure that at risk students are properly supported and students on the path to the highest success are able to stay on track. This study proposes a knowledge discovery in databases approach toward the development and evaluation of several prediction models to accurately predict with two years of academic data or less whether students will graduate with distinction. In developing the prediction models, several important factors of students' success are identified as well as additional insights about student experience.

ÖZET

Tanım ve Tahmin:

Üniversite Veritabanlarında Bilgi Keşfi

Makine öğrenmesini içeren veri madenciliği yöntemleri iş hayatında yıllardır uygulanmaktadır ve bugün bu yöntemler yükseköğretimdeki eğitimciler ve veri bilimcileri tarafından büyük ilgi görmektedir. Bir öğrencinin üstün başarı ile mezun olma yolunda olup olmadığına dair doğru ve erken tahmin, risk altındaki öğrencilerin uygun şekilde desteklenmesini ve en başarılı öğrencilerin bu yolda kalabilmesini garanti altına almak adına yöneticiler, eğitimciler ve danışmanlar için çok önemlidir. Bu çalışma, 2 yıl veya altındaki akademik veri ile öğrencilerin üstün başarı ile mezun olup olmayacağını doğru bir şekilde tahmin etmek için bazı tahmin modellerinin değerlendirilmesi ve geliştirilmesine yönelik veritabanlarında bir bilgi keşfi yöntemi sunmaktadır. Tahmin modellerinin geliştirilmesinde, öğrencilerin başarısının bazı önemli unsurları ile birlikte öğrenci deneyimi ile ilgili ek anlayışlar tanımlanmaktadır.

ACKNOWLEDGEMENTS

I wish to acknowledge my wife for her endless love and support during what has been a long and challenging process of research and writing. Being a foreign graduate student unsurprisingly fosters extra challenges, but having a partner in life makes a world of difference. Woe indeed upon the researchers who struggles alone.

I would also like to thank my adviser, Hayri Hoca, who has been patient with me over this long process. The freedom he has afforded me to explore various paths has perhaps made my process far longer, but correspondingly far richer for all the studies made. I know more now than I did when I started, and more now than if my thesis experience had been typical.

I want to thank the members of the committee for their time, attention, and valuable feedback. Writing comes voluminously to me, but unfortunately less often with concise clarity. It is with eternal gratitude that I receive and accept the feedback and criticism of anyone who endeavors to help me improve.

DEDICATION

I dedicate this research to all the classmates, educators, students, and administrators with whom I have studied, worked, or argued over my many years and many places. Education is equal parts art and science and exceeds the sum of its parts when everything goes right. The tragedy of civilized humanity is that there is never enough time and resources to ensure everyone have equal opportunities to learn. That my research now and in the future may support the more judicious management of learning resources has been and continues to be my motivation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDIX TABLES

LIST OF FIGURES

LIST OF APPENDIX FIGURES

ABBREVIATIONS

ANN	Artificial Neural Network
CRISP-DM	Cross Industry Standard Process for Data Mining
D3M	Data-Driven Decision Making
DT	Decision Tree
EDM	Educational Data Mining
KDD	Knowledge Discovery in Databases
K-NN	K-Nearest Neighbor
MLR	Multinomial Logistic Regression
ÖSYM	Measuring, Selection and Placement Center (Ölçme, Seçme, ve Yerleştirme Merkezi)
ÖYS	Central Entrance Examination (Öğrenci Yerleştirme Sınavı)
YÖK	Council of Higher Education (Yükseköğretim Kurulu Haberleri)

CHAPTER 1

INTRODUCTION

A great deal of attention has been directed toward the subject of data mining in recent years as a way for both for-profit and non-profit organizations to better understand their operations. Advancements in digital technologies, especially their increased accessibility to laypersons, have resulted in ever larger databases being maintained and utilized by organizations of every description. Universities are good examples of organizations that are likely to rest upon large stockpiles of historical data in which insights of potentially significant value lay waiting to be uncovered.

Universities across the world, particularly those ranking near the top of their disciplines, are in fierce competition not just to attract students but to deliver successful outcomes to students. Students are now able to judge universities on more than simple perceptions of prestige. Well informed students now expect a university to offer them evidence of a superior academic or social experience. Those universities with the data to measure their own performance possess both the knowledge to improve and the evidence to boast of it. Some universities collect and mine data specifically for the purpose of advertising, such as in alumni success and campus sentiment metrics. This puts added pressure on universities whose investments in their information management systems lag behind.

The increased impetus toward more sophisticated data management and mining comes not only from interuniversity competition and market positioning. More and more, market demand and cost concerns are driving universities toward greater utilization of Internet-based course options, which consequentially makes possible the

measurement of student engagement at levels of granularity not feasible in traditional lecture-based courses. More than offering an opportunity to track student engagement digitally, the lack of traditional instructor-observation measures of student engagement in these Internet-course environments necessitates the development of effective and robust digital metrics able to compensate for the lack of face-to-face assessment of student engagement.

Identifying early on in a student's career whether they are on a path to graduating with distinction or on the path to not graduating at all presents the university with an invaluable opportunity to intervene appropriately before it is too late to act. The purpose of this study is the development of a model of predicting academic outcomes, in particular graduate distinctions, for undergraduate management students of a public university in Turkey by following the CRISP-DM methodology, which is a context-appropriate framework for applying machine learning and data mining methods. A variety of methods exist that are suitable for predictive classification. This study seeks to determine which methods, from among decision tree, artificial neural network, and multinomial logistic regression, offer the best combination of accuracy and early deployment, and thus the most value to the university. A secondary objective is the discovery of valuable insights about the university's program and students' experiences during the iterative data processing and understanding phases of CRISP-DM.

This study to some extent follows upon earlier research conducted by Eda Guvenç, whose master's thesis was entitled "Student Performance Assessment in Higher Education Using Data Mining" (2001). In her thesis, Guvenç applied several data mining methods, in what was one of the earlier applications of data mining to educational data, to several cohorts of engineering students at Boğaziçi University, which is the same

Turkish university that is the focus of this study. While her thesis was composed when the field of Educational Data Mining (EDM) had not yet into a distinct discipline of its own, this study benefits from 16 years of further contributions to the literature. One key development is that the burden upon researchers has moved away from making the case for data mining applications in education to now determining best methods and proposing how to effectively integrate them into the educators' decision-making processes.

CHAPTER 2

LITERATURE REVIEW

Data mining in universities is an area of research that has gained growing interest in recent years leading to the establishment of Educational Data Mining as a distinct discipline. Central to this research are processes that best prepare data for data mining methods and for deployment in the decision-making processes of educators and university administrators.

2.1 An overview of knowledge discover and data mining

The first step taken by many organizations just beginning to approach their stockpiled data is to directly apply data mining methods, yet data mining initiatives are often frustrated by data that is collected and maintained without consideration of how it may be used later. Equally problematic is the misguided application of data mining methods that results in immaterial or misleading conclusions. Knowledge Discovery in Databases (KDD) is a conceptual framework that was developed to guide data scientists and data science practitioners in their efforts to provide organizations with meaningful insights and information and does so in part by position data mining as one step in a process. It places applications of data mining methods within a complete beginning-to-end process of knowledge-discovery and includes several steps which precede data mining, such as data preprocessing (Brachman and Anad, 1996; Fayad, Piatetsky-Shapiro, and Smyth, 1996). CRISP-DM, standing for Cross Industry Standard Process for Data Mining, is an example of such a KDD methodology that is popular among data scientists working with educational data.

Fayad et al. put forth their attempt at a unifying framework to ground in a common foundation the discussions held by KDD practitioners and researchers hailing from various disciplines. Their framework (Figure 1) has been widely accepted and established two overarching points, among others, which are that (i) KDD is an iterative series of steps each requiring human interaction and that (ii) data mining is itself only one step within the more complex KDD process. Each step, including the application of data mining, requires manual analysis, interpretation and, if necessary, modification and repetition. The latter point is important to distinguish between informed applications of data mining methods and blind ones, criticized in the literature as "data dredging" and resulting in what may be dangerously meaningless pattern discovery.

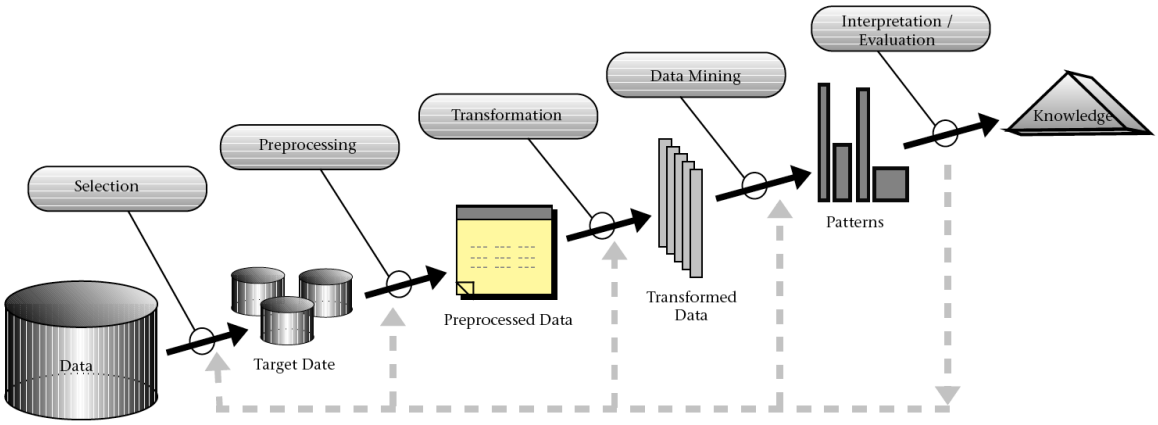

Fig. 1 The knowledge discovery in databases framework

Fayad et al., 1996

Data mining is the processes of exploring data, which often must be preprocessed, and identifying patterns, associations, structures, and other characteristics which represent meaningful information and insight. What makes derived information

and insight meaningful, and thus constitute knowledge, is their utility in facilitating Data-Driven Decision Making (D3M) in combination with other forms of decision making. There is an issue of terminology in the field of KDD/D3M/Data Mining, which will be discussed later in this paper. Contrary to Fayad et al.'s definition of data mining as a step within KDD, Baker (2010), in his chapter entitled Data Mining for Education in the International Encyclopedia of Education, explicitly stated that data mining and KDD were synonyms. Several researchers continue to refer the interchangeability of these terms (Romero, Romero, and Ventura, 2014).

Fayad et al. went on to diagnose knowledge discovery as a process guided by the pursuit of one of two goals, either (1) verification or (2) discovery (1996A & 1996B). Verification-driven knowledge discovery is conducted in order to find data that confirms an explicit or implicit hypothesis. The data mining step is repeated until the appropriate data is found. The result is little new knowledge because the bounds of the guiding hypothesis limit both the interpretation of discovered patterns and also the motive to innovate. Conversely, discovery-driven knowledge discovery is conducted with little guidance from the user and employs data mining tools to yield new and useful facts about the data. This knowledge discovery goal is further divided into (1) description (undirected) or (2) prediction (directed).

2.1.1 Data-driven decision making

D3M, as its name suggests, is defined simply as the use of data analysis to inform courses of action (Picciano, 2012). A general depiction of the decision making that incorporates data among internal and external factors is found in Figure 2. Importantly, D3M is intended to inform rather than replace the subjective elements of decision

making. This is important because attempts to rely exclusively upon data when making decisions causes problems such as (1) rejecting valuable decision factors not represented in the data, (2) marginalizing human experience, expertise, intuition, and (3) unnecessarily engendering resentment and resistance to D3M among a decision's less data-inclined stakeholders. The last problem is especially concerning in particularly risk-averse decision making environments such as universities, which often value consensus building among experts across many disparate disciplines.

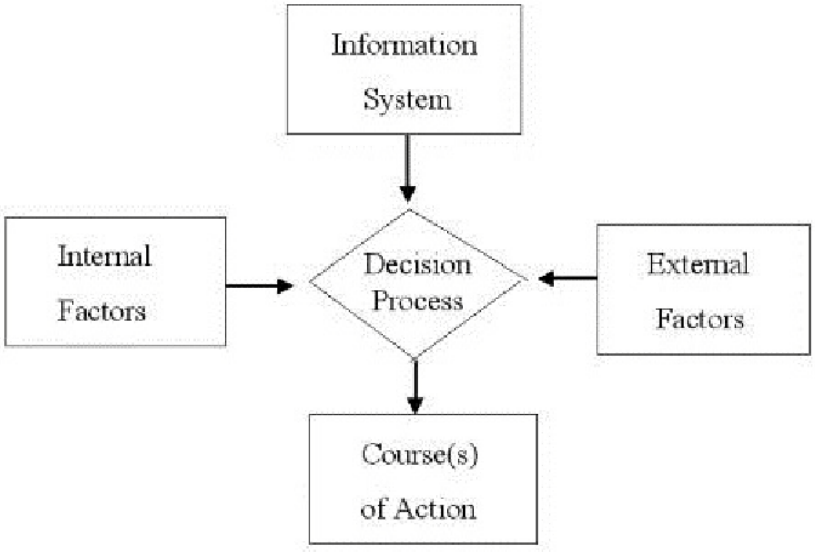

Fig. 2 The data driven decision making process

Picciano, 2012

Discussions of D3M feature terms, such as data warehousing, data disaggregation, data mining and analytics, the latter two of which have gained their own wider prominence in management discussion. Briefly, data warehousing refers to the database information system(s) that integrate, maintain, and store data. Data disaggregation refers to methods, frequently software, of breaking data files down into

desired sets which are appropriate for data mining. Data mining, as discussed previously, refers to searching data for patterns or other observations that provide information to better understand some phenomenon. Analytics is a term that has become so generalized in use that Picciano considered it almost synonymous with D3M. He observed that "analytics" often found its way into new jargon, such as "academic analytics" as coined by Goldstein and Katz (2005) who acknowledged its inclusion for lack of a better alternative. Similarly, Elias (2011) recognized four distinct categories of analytics that can be applied in a teaching context: learning analytics, action analytics, academic analytics, and web analytics. She defined learning analytics as the "sophisticated analytic tools are used to improve learning and education," and saw EDM as one of many subfields within it.

Elias juxtaposed Goldstein and Katz's broad definition of academic analytics as an application of business intelligence to education against Campbell and Oblinger's (2007) narrower definition of academic analytics as the marriage of statistical tools to the decision making of educators to improve student outcomes. By comparison, EDM, according to Campbell and Oblinger, focuses only on identifying patters in data rather than directly improving decision making.

Norris, Baer, Leonard, Pugliese, and Lefere (2008) added further distinction by introducing action analytics as the application of academic analytics for the production of actionable insights. They provided examples of several types of action analytics outputs: "service-oriented architectures, mash-ups of information/content and services, proven models of course/curriculum reinvention, and changes in faculty practice that improve performance and reduce costs".

Finally, IBM coined yet another terminological instance of analytics in their 2011 white paper entitled "Analytics for Achievement", wherein they described eight categories of instructional applications:

1. Monitoring individual student performance
2. Disaggregating student performance by selected characteristics such as major, year of study, ethnicity, etc.
3. Identifying outliers for early intervention
4. Predicting potential so that all students achieve optimally
5. Preventing attrition from a course or program
6. Identifying and developing effective instructional techniques
7. Analyzing standard assessment techniques and instruments (i.e. departmental and licensing exams)
8. Testing and evaluation of curricula.

Conceptualized in these ways, it is clear that analytics, whether or not it is qualified by learning, academic, or action, overlaps generously in discourse with both KDD and D3M when the latter two are also applied to learning and educational contexts.

2.1.2 Background of EDM

Picciano (2012) presented a concise summary of the historical development of learning analytics/EDM in American higher education. Following decades of technological development and innovation that saw the administrative functions and record keeping of American universities increasingly integrated into accessible information systems, university administrators were able to begin posing questions using data query languages and other decision-support systems. Rather than relying on cycles of annual studies,

decision making could now be data-informed on a regular basis and eventually in real time. Regional accrediting bodies in the US began to require universities and colleges to demonstrate their proficiency at managing their information systems and supporting their planning and decision making with data.

In the 1990s and early 2000s, courses and later entire degree programs began to appear online rather than in physical locations. Online learning demanded collecting and processing data of not only administrative functions but entire educational experiences. Instructor decisions must be made and justified in the absence of traditional channels of information such as in-class observation and must also be made in real time. The increasing use of online learning, either in augmentation or replacement of in-person instruction, provides strong pressure for learning analytics/EDM to provide ever more robust decision-support systems.

As an emergent subfield within data mining, EDM attracted a growing body of independent research in the early 2000s, culminating in 2008 with the establishment of an annual International Conference on Educational Data Mining and of the Journal of Educational Data Mining. The field has continued to grow and enjoy wider attention as educational institutions increasingly move into E-learning environments, particularly in developing countries where the presence and accessibility of educational data is often conducive to research.

2.1.3 Review of recent literature

Even before the formal establishment of EDM as a field of its own, the frequency of published researched applying elements of data mining to various types of raw data

produced by university systems had been increasing for several years, leading to the publication of several literature reviews.

Castro, Vellido, Nebot, and Mugica (2007) reviewed research conducted between 1999 and 2006, concluding that most EDM studies focused on classification and clustering problems related to E-learning platforms. Romero and Ventura (2007) presented a comprehensive survey of research published between 1995 and 2005 and summarized examples of various EDM methods, such as text mining, web mining, and visualization, which had been applied to both traditional classroom and E-learning-based problems. Even at that time, much of the more sophisticated statistical analyses were being applied to problems of E-learning systems rather than those of traditional classrooms.

Baker and Yacef (2009) composed their review in the aftermath of first EDM conferences and noted the corresponding increase in EDM research. Student models, models of domain knowledge, pedagogical support, and impacts of learning where identified as major targets of EDM research. Pena-Ayala (2014) concluded that the most common EDM approaches were directed at student modeling and assessment, particularly of performance, behavior, learning, and domain knowledge.

Romero and Ventura (2010) later returned to their earlier survey and further expanded it with the inclusion of 225 additional works. The identified eleven categories of EDM application: (1) analysis and visualization of data; (2) feedback for supporting instruction; (3) recommendations for students; (4) prediction of student performance; (5) student modeling; (6) detection of undesirable student behaviors; (7) grouping students; (8) social network analysis; (9) developing concept maps; (10) constructing courseware; (11) planning and scheduling.

Papamitsiou and Economides (2014), recognizing the competing terms of learning analytics and EDM, examined experimental case studies conducted between 2008-2013 and produced a SWOT analysis of the domain at that time. Among other conclusions, they found that increasing volumes of available education data was lowering the barrier to entry for researchers but that misinterpretation of results was common due to a focus on reporting rather than decision making. They classified case studies along six research objectives: (1) student modeling, (2) prediction of performance, (3) increase reflection and awareness, (4) prediction of dropout and retention, (5) improvement of assessment and feedback services, and (6) recommendation of resources.

Dutt, Ismail, and Herawan (2017) published a comprehensive review of published EDM papers focusing on applications of clustering algorithms between 1986 and 2016. They found that many clustering algorithms have been developed and utilized by researchers looking at a wide range of subjects but that no single clustering algorithm had been found to consistently appropriate across cases. Noting that different users would interpret the same data differently, they deemed the development of a unified clustering algorithm implausible for educational data. They concluded that the typically multi-level hierarchical and non-independent nature of educational data necessitates that researchers exercise great care when selecting clustering algorithms.

More specific research which contributed to or influenced the composition of this thesis and the conduct of its research is reviewed below. As the scope of this thesis' research is limited to administratively collected data from a predominantly traditional classroom environment, certain portions of the body of EDM literature were more practically related than others.

As this thesis looks at the performance of undergraduate management students in a leading public university in Turkey, it is important to consider earlier research conducted in this particular context. Fortunately, two such papers were identified, one a published research and the other an accepted master's thesis. Guruler, Istanbullu, and Karahasan (2010) used classification techniques to identify individual student characteristics of college freshmen and their association with future academic success. As mentioned in the introduction, Guvenç (2001) applied a selection of machine learning methods, most importantly affinity analysis, to several cohorts of engineering students at Boğaziçi University in Istanbul, Turkey.

Kotsiantis, Pierrakeas, and Pintelas (2003) applied five classification techniques to student demographic data of the Informatics degree program of a Greek distance learning university. These included decision tree, Perceptron-based learning, Bayesian Net, instance-based learning and rule-learning and were compared on the basis of their accuracy predicting student dropout. Naive Bayes, the algorithm chosen to represent the Bayes Net technique, performed best and a prototype web-based support tool was produced for the university using this algorithm.

Golding and Donaldson (2006) studied possible indicators of future performance of undergraduate IT students at a university in Jamaica. Using stepwise linear regression, they concluded that students' performances in core first year courses were predictive of their performances in later years of the program.

Nguyen, Janecek, and Haddawy (2007) compared the predictive performance of decision tree and Bayesian network classifiers on student data, concluding that the former significantly outperformed the latter when applied to undergraduate and graduate student data for two universities in southeastern Asia.

Cortez and Silva (2008) endeavored to predict secondary student grades in mathematics and Portuguese courses by using past grades, demographic, social and other school-related data. They applied four classification techniques: decision tree, random forest, neural network, and support vector machine. Vandamme, Meskens, and Superby (2007) similarly applied three classification techniques, decision tree, neural networks, and linear discriminant analysis, to identify low, medium and high-risk classes of students in three French-speaking universities in Belgium.

Kabakchieva (2009) developed prediction models for student performance at a Bulgarian university. She applied decision tree, Bayesian, nearest neighbor, and rule learning classifiers to a selection of student data, including personal, university, and pre-university characteristics. Yehuala (2015) also applied decision trees and Naïve Bayes algorithms to build prediction models of student success at a university in Ethiopia.

Oskouei and Askari (2014) described several hundred high school and university students in Iran and then applied a variety of techniques, including classification and regression, to identify demographic characteristics with predictive value with regard to student academic performance.

Campagni, Merlini, Sprugnoli, and Verri (2015) applied several EDM methods to their study of the academic careers of graduate students at a university in Italy. The researchers proposed the concept of the ideal academic career, defined by both the sequence of course examinations and how quickly a student sat for their exams after the conclusion of the associated courses. Clustering techniques were applied after an extensive preprocessing phase and followed by a post-processing phase.

Zimmermann, Brodersen, Heinimann, and Buhmann (2015) studied how well the undergraduate achievements of computer science students could predict their graduate-

level performance at a university in Switzerland. They built eight linear regression models composed of different aggregations of student data and compared their predictive performance. Pro-processing analysis revealed that the strongest indicator of graduate school performance was a student's third year GPA. The researchers also sought to identify underlying subject matter constructions by applying factor analysis. Their expectation was to find clear mathematical and engineering constructions based upon their understanding of the university's computer science curriculum, but their results did not support this.

Asif, Merceron, Ali, and Haider (2017) applied classification and clustering methods to the performance of undergraduate IT students in a public university in Pakistan. They concluded that it was possible to accurately predict student's academic achievement in their four year program of study from a set of ten pre-university variables. They discovered two distinct student clusters along the vector of their course grading and attempted to identify a group of courses which were particularly good indicators of good or poor student performance. The result was an automated warning system which could issue alerts to students and educators when students in the low performing cluster triggered predictors of future poor performance.

2.2 Data mining in education

Several papers have focused on identifying common characteristics of research within EDM, including prevailing objectives and approaches. Baker (2010) identified five overarching categories of objective: prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgment. Each objective will be described in the sections below.

2.2.1 Prediction

Prediction is the most common goal in EDM research focused on improving administrative decision making. Understanding why some students achieve better student outcomes and while others do not is a core motivation for educators. Institutions of higher education frequently create their own classifications of students based on academic performance, such as honors and GPA, but also according to demographic and other types of student data. Such preexisting classes are often a natural starting point for prediction models. Creating new classes to address specific research questions is made easier by the continued growth in the variety and quantity of retrievable data cultivated in the information systems of modern universities.

Baker (2010) identifies three general types of prediction: (i) classification, (ii) regression, and (iii) density estimation. (i) Classification requires that the predicted variable be either a binary or categorical variable. Decision trees, logistic regression, and support vector machines are some of the most popular classification techniques. Because classification attempts to determine the proper membership of a data point in a predetermined class, it is known as a supervised learning technique. In (ii) regression, the predicted variable is a continuous variable. Several popular techniques are linear regression, neural networks, and support vector machine regression. The target variable in (iii) density estimation is a probability density function, which is a function of a continuous random variable which provides the probable location of that variable within a given interval.

2.2.2 Clustering

Clustering attempts to identify data points that naturally form groups and in doing so either reveal underlying insights about the data or simplify variable sets to make additional analysis easier. An optimally clustered data set will feature data points more similar to other points within their cluster than they are to points in other clusters (Baker, 2010). However, even ideal clustering results are subject to observer interpretation. Cluster analysis can be conducted either with or without a hypothesis guiding the interpretation of clusters. Because of this, clustering is generally referred to as an unsupervised learning technique.

Clustering techniques may be further distinguished by whether they are hierarchical or partitional in nature. The difference is that data points belong to only a single cluster in partitional clustering, whereas data points can otherwise belong to a hierarchy of clusters called a dendrogram. A dendrogram displays the hierarchy visually, making the nested relationships clearly observable. Clustering techniques are also differentiated by whether data points are assigned to clusters by degree of belonging or binarily, also known as Soft and Hard methods. Finally, clustering techniques can also be classified by their approach to implementing the algorithm: centroid-based clustering, graph-based clustering, grid-based clustering, density-based clustering, and neural networks-based clustering. Each approach possesses algorithms oriented toward combinations of hierarchical/partitional and Hard/Soft methods (Dutt et al., 2017).

2.2.3 Relationship mining

As its name suggests, the goal of relationship mining is to discover the strength of associations among variables in a data set. Associations may be among a number of

variables, between variables and a particular target variable, or between just two variables. There are four general classes of relationship mining techniques: (i) association rule mining, (ii) sequential pattern mining, (iii) correlation mining, and (iv) causal data mining (Baker, 2010). Association rule mining aims to establish if-then rules based on the relationship between variables, such as if a student fails a particular course then he or she will fail a related course. Association rule mining can also be used in sequential pattern mining, which focuses not only on the occurrences of a relationship but their sequential order as well. One application of sequential pattern mining is mapping students' course schedule decisions. The goal of correlational mining is the discovery of linear correlations between variables. In causal data mining, the objective is to determine causality by analyzing the covariance of the events of interest or by incorporating other sources of information about the events.

2.2.4 Discovery with models

Discovery with model is a method of EDM wherein a model, previously developed by one or more methods of knowledge discovery, is used in one of the other methods, such as prediction or relationship mining. The means of developing the model need not be the produce by machine learning, as human reasoning is equally viable. This method broadens the range of possible subjects able to be studied by incorporating into the analysis complex constructs and new variables derived from an original data set. Models validated and imported from other contexts, even other disciplines, can be brought into a data set, or generalizations made across time-series data can be modeled, validated, and then integrated back into the analysis of smaller units of time within the same data set

(Baker, 2007). Discovery with model is less an entirely new method of EMD and more a higher-level method of applying lower-level methods.

2.2.5 Distillation of data for human judgment

The final category Baker described is what he terms "distillation of data for human judgment", which refers to the steps necessary to visualize data from which human observers then derive insight directly. Despite continual advancement in the tools of knowledge discovery, humans continue to possess inferential abilities beyond what machine learning is yet capable. Even where machine learning is capable of recognizing such patterns, the resulting output may be indecipherable to less technically proficient observers, such as educators or school administrators. Distilling data for human judgment has the added value of increasing the likelihood that results are interpretable by other stakeholders.

Information visualization methods are generally applied to data for human observers to then (i) identify or (ii) classify data points. When the objective is identification, the goal of visualization methods is to render the data in such ways as to permit humans to easily recognize familiar patterns that are difficult to express formally. An example of such a pattern is the learning curve, which plots the number of learning opportunities and the performance of learning along x and y axes. Classification by human observation is performed by visually displaying sections of a data set and has been shown to speed up the development of prediction models (Baker & de Carvalho, 2008).

CHAPTER 3

METHODOLOGY

In this section, the steps taken by this research to prepare the data for analysis and analysis for interpretation are described. CRISP-DM, standing for Cross Industry Standard Process for Data Mining (Shearer, 2000), was the methodology selected for this research as its suitability in educational contexts had been established by other researchers (Delen, 2010). CRISP-DM consists of six-steps which map well to those outlined in Fayad et al.'s framework for KDD (1996). The steps begin with (i) understanding the context of the research questions, i.e. an undergraduate management bachelors program at a public university, (ii) collecting, studying and understanding the data, (iii) pre-processing, cleaning, and transforming the relevant data for analysis, (iv) selecting and developing models, (v) evaluating and assessing the models against the research questions, and (vi) suggesting a plan to deploying the models to improve decision-making processes at the subject university.

The first step was to understand the context of the research objective, i.e. an undergraduate management bachelors program at a public university in Turkey. The second step was to understand the nature of the available data, followed by preprocessing the data to prepare it for analysis. This stage typically takes the largest amount of time, and the point at which data has been sufficiently understood and prepared is a subjective determination. Even high quality data will often require preprocessing to address missing values and identify variables and potentially new variables that offer the highest value to the research objectives. Once data preprocessing was complete, two distinct phases of analysis: descriptive and modeling come next. In

addition to the tables and figures contained in the main text, additional supporting tables and figures are located in Appendix A and Appendix B respectively.

3.1 Understanding the context of the research objectives

The subject of this study is the undergraduate management program of Boğaziçi University and its students. The university, located in Istanbul, Turkey, is a public university to which students are allocated according to their preferences and results on the Central Entrance Examination (ÖYS). The university is known as one of the most prestigious universities in the country, and its graduates are considered among the most desirable in Turkey.

While EDM, which refers to data mining in educational contexts, has been long established, its adoption has been inconsistent among universities around the world. Turkey is an example of a university system that has less widely embraced these concepts, particularly among more established public universities like Boğaziçi University which possess strong traditions and an embedded decision-making culture. With Turkey's education sector experiencing several years of rapid growth driven by increased public and private investment in new universities across the country, traditional universities in Turkey are facing new challenges and opportunities.

These challenges include competing with increasingly respectable private universities for students that historically would have sought placement in one of a small number of historically prestigious public universities. Private universities may possess advantages with regard to funding attractive campus investments, offering higher wages to professors, and establishing more desirable foreign exchange agreements. By offering scholarships to the best and brightest of Turkey's high school students, it is imperative

that public universities, such as the one which is the focus of this study, leave no stone unturned in their efforts to provide the best possible student experiences and outcomes.

Before entering their department of study and formally beginning their degree studies, students must demonstrate their English language proficiency on an approved examination. The language proficiency requirement is not limited to students accepted directly from high school. Transfer students are also required to attend the preparatory program until they are able to pass an accepted proficiency exam.

Upon graduation, the university bestows graduate distinctions upon students who satisfy certain criteria of grade point average (GPA) and length of study criteria. These distinctions add considerable value to an already valuable degree. The university bestows High Honors upon students who graduate in less than eight semesters, not counting summer semesters, and with a final GPA of 3.5 or greater. The Honors distinction is bestowed upon students also graduating in less than eight semesters, but with final GPA between 3.0 and 3.49. These are the only two classes of graduate distinction.

3.2 Understanding the data

After understanding the nature of the problem or subject of interest, the second step of CRISP-DM, or any KDD project, is to understand the data that is available for study. Understanding the data also requires that the researcher identify what data is of value to the study and which is not and should be culled. This identification requires synthesizing knowledge of the study's objective with that of the available data. For this study, the data variables are first described and then appropriate constraints are developed to limit and define the data set for future exploration and application of EDM techniques.

3.2.1 Detailed description of the data

In this section, the student and course data which form the subject of analyses are described, and the means by which they were acquired is briefly explained. After consulting with several stakeholders to determine which authority within the university governed the database of student information, a formal request for data was submitted to the Registrar Office of Boğaziçi University, which approved the request.

Student data at Boğaziçi University was revealed to be stored in a number of forms determined by manner and frequency in which it is collected. The circumstance in which the data is recorded determines in which database silo it is stored and from which it must be retrieved. Prior to requesting a full set of data, a sample selection of data was obtained and studied. Ultimately, four data tables were requested and received, with each containing a number of variables of potential value to predicting students' attainment of graduate distinction. Four tables were ultimately received, closely matching those requested, and these are summarized in Table A1. All variables names were translated to make it linguistically consistent as well as to improve their readability.

What follows is a brief review of these initial variables. Each of these variables was specifically requested for its potential value in addressing achieving the research objective of predicting graduate distinction, but not all variables were found to be appropriate or usable. Particular issues encountered will be described while a more detail breakdown of preprocessing strategy and decision making will be reviewed in the Data Cleaning and Exploration sections of this study. Variables will be covered by table in which they appear while variables that appear in multiple tables will be described as part of the first table covered.

Beginning with the Demographic Table, StudentID is the encrypted student identifier produced by the database system for use in this research while maintaining student privacy. It is the variable that serves as the primary bridge between Demographic, Transcript, Course, and Exchange tables. The demographic data table also includes registered information on each student's nationality, sex, and the type and name of high school from which the student graduated.

The nationality variable includes only a single entry for each student, leaving cases of dual-nationality undocumented. Some Turkish citizens possess a second nationality, such as German or American, and it is not clear from the data which students chose to enroll with Turkish nationality while also possessing another. As the language of instruction at Boğaziçi University is that of a non-native language, it would be ideal to include students with a second nationality into analysis. However, was missing from the available data and thus was not a part of this study.

Each student is registered as either Male or Female at the time of registration, although the distinction between whether this variable is meant to represent the Sex or the Gender of the student is not clear. The Turkish label for this variable is Cinsiyet, which could be translated in English to either Sex or Gender. Matters of gender and sexual identity are a part of campus life for students, and Boğaziçi University's historic South Campus has featured gender-neutral public restrooms for several years, suggesting that gender identification is a subject of student life. Trans-gender issues have become more topical in Turkey as they have in many other parts of the world. For the purposes of this research, this variable will be translated and understood to represent the student's Sex, the anatomy of their reproductive system, at the time of registration rather than that student's assigned or self-identified gender.

Each student graduates from a single high school, although they may attend multiple high schools before that point. The registration system records the name and the type of high school from which the student graduates as HighSchoolName and HighSchoolType. High schools in Turkey are categorized by the type of curriculum they teach, which for example can instruct in a foreign language or can emphasize math and science, but these categorizations can be dynamic as the educational system in Turkey has been subject to a number of policy changes in recent years. The type of high school program from which a student graduates can either enhance or diminish the student's ability to choose subjects to study at university. The Turkish education system favors technical programs, and thus an analysis of academic performance by the type of high school a student has graduated from is an obvious area of interest. Unfortunately, as will be discuss in more detail in the Data Cleaning section, inconsistency in the recorded data makes this a difficult subject to study in practice.

Students wishing to attend university in Turkey must sit for placement exams which then determine to which university the student is eligible to apply. The years in which students sit for these exams are represented by the OSYMYear variable, and this is one way to distinguish between student cohorts. OSYM is the acronym of the Measuring, Selection and Placement Center, which is the body responsible for organizing university entrance examinations at the national level. The exam for which students must sit to earn eligibility varies by the type of high school program from which they've graduated and also varies slightly from year to year. Students are ranked by their exam scores and then matched to universities according to their university and subject preferences. Foreign students can also sit for certain alternative exams, such as the SAT for foreign students. Unfortunately, the data provided for exams and exams scores was

often missing and was inconsistent when present for the period of time reviewed, and so it was culled early on from the data set.

Graduate is a simple Yes/No Boolean value indicating whether the student has yet graduated. It is followed by FinalGPA and Honors, which reflect the most recent calculated GPA for the student and whether or not they have qualified for honors recognition. Honors values can be Honors, or High Honors, and are determined by a combination of student's GPA and length of time as an active student. High Honors is awarded to students whose GPA is 3.50 or higher, Honors for GPAs between 3.00 and 3.49. For both distinctions, students should complete their degrees in eight semesters or less with no disciplinary penalties (https://advising.boun.edu.tr/en/content/faq-undergraduate). Students failing either the GPA, number of semesters, or disciplinary checks were recoded as Normal for the purposes of this study.

Honors will be the target variable that this research will attempt to predict. As honors achievements are meant to be a testament to a student's success across the body of their academic career, it is valuable for universities to understand what factors may contribute to such success and to identify students who are likely to achiever or not achiever these distinctions.

UniversityEntrance denotes the semester in which students are registered to courses at the university, coded as the academic year and semester number. For example, students starting course in the fall of 2007 are coded 2007/2008-1. Academic years begin in the second half of a calendar year and continue until the beginning of the summer of the next calendar year. Semesters are coded as follows: fall is 1, spring is 2, and summer is 3. DepartmentEntrance is coded the same way and reflects the first semester when students register for courses in their degree program after having

satisfied the language proficiency requirement. Time between entering the university and the department may reflect that the student needed to spend time in English preparation courses or may have transferred to the department from another department within the university.

The Transcript Table aggregates student's course results that occur within a single semester as well as maintaining a running account of results from preceding semesters. Courses can either be graded on a letter grade or PASS/FAIL basis, and credits associated with courses are tabulated separately in this way. PASS/FAIL variables was culled from the data set as they do not contribute to GPA calculations and typically reflect course transfers from other universities or special seminars.

Total letter-graded course credits registered and earned by the student are summed for each semester and a running total is also calculated. Letter-grades were recorded to a numerical variable: AA=4.0, BA=3.5, BB=3.0, CB-2.5, CC-2.0, DC-1.5, DD-1.0, and F=0.0. To determine the number of points earned from a course, the numerical grade and credit value of the course are multiplied. This product is then summed and divided by the sum of course credits in a particular period of time resulting in a GPA calculation. GPAs are calculated for each semester and as a running total. This running total becomes the student’s FinalGPA upon graduation.

Semesters are numbered sequentially by the SemesterNumber variable, with 1 representing the student's first semester in their degree program. Thus, it does not count semesters spent in the English language preparatory program. Importantly, this variable does not count summer semesters even after the student has started their degree program. Summer semesters are assigned the same number value as the preceding spring semester. In this way, SemesterNumber is the value which can consistently keep track of

how long a student is actively studying in their degree program without being skewed by a student's decision to attend courses in summer. From it, academic years can be calculated.

The Course Table contains information on each CourseCode and StudentID pairing during the time period considered. Each course and student pairing is unique to the semester in which it occurred and the CourseSection which represents the identifier for multiple offerings of the same course in a single semester. CourseCredits reflects the credit value of each course.

Students may register multiple times for the same CourseCode depending on whether or not they pass or apply for grade forgiveness, which is possible if the students earn a DC or DD letter grade from a course. If a student successfully passes a previously failed course, the previous course is documented in the ReplacedCourse variable. Courses are typically replaced by the same course, but in some cases a course can be replaced by a different course. When a student successfully replaces the grade of a previous course, such as by retaking the course or by supplanting it with another accepted course, the replaced grade no longer contributes to the student's GPA.

The final table is the Exchange Table and contains information about foreign exchange programs in which students can participate. Most exchange programs are organized by agreements made between Boğaziçi University and a foreign university, but students may facilitate exchange opportunities on their own and then apply for course recognition by Boğaziçi University. Courses taken at other universities are credited to the student as PASS/FAIL courses, regardless of whether or not the student received a letter grade at the exchange university. This table does not contain information on individual courses, but rather tracks exchange program data by

StudentID and the Semester in which the exchange was registered. Each instance records the university's name and the country in which it is located.

The follow section will document the process of preparing the data in these tables for analysis.

3.3 Defining the data set

This section describes the efforts to construct the final data set and the reasons behind them. The initial data tables provided for this research included information on a total of 1036 students who entered the undergraduate degree program of Boğaziçi University's Department of Management beginning in fall 2007 until fall 2015. As this study ultimately focuses on building models able to predict students' graduate distinctions, it was critical to identify student and course data likely to contribute to those models.

3.3.1 Analysis of student nationalities

Student nationality was identified early on as a likely key characteristic to consider. As Boğaziçi University is a leading Turkish public university with English as its language of instruction, it attracts students from several countries in addition to Turkey. Students coming from other countries may possess potential language advantages. However, it was revealed that undergraduate management students were overwhelmingly Turkish nationals. As previously discussed, it is not possible to know if students registered with Turkish nationality also possess a second nationality. It is likewise not possible to know if these students from other countries are not members of the Turkish diaspora, though the significance of that would appear to be less material in this instance. Since the number of non-Turkish students is such a small percentage of the total students, it was

decided that it was best to cull these students from further analysis rather than trying to utilize nationality in future analysis. 1019 Turkish students remained.

3.3.2 Time period of the study

Understanding what constitutes student cohorts is an essential step in this study. Intuitively, time period is expected to be the major component of a cohort. In most cases, student cohorts could be built from the year in which they begin university or the year in which they sat for university entrance exams (OSYM). But this is not as straight forward in a case like Boğaziçi University, where most students spend some amount of time in English preparatory courses before beginning their degree studies. The student data set requested for this research was for all students who entered the Department of Management's undergraduate degree program in the Fall 2007/2008 academic year, yet this set included a number of students which first registered for courses at the university (as far back as 1971) as well as sat for entrance exams (as far back as 1999) over a long period of time. These were culled from the set, as were any students who entered the university prior to academic year 2005/2006 or who sat for entrance exams prior to 2005. This provided the beginning point for a data set.

There remained the question as to which year to end the data set. To be able to predict students' graduation distinctions, students needed to have a reasonable opportunity graduate. Graduate distinctions require students to graduate in 8 semesters, or four years, yet it is important to differentiate between students who graduate in more than four years and those who do not graduate at all.

The average length of study for graduated students was determined to be 5.15 academic years spent in their departmental studies, with a standard deviation of 0.91.

This suggests that approximately 97.7% of graduated students graduate within seven academic years. In order then to be able to build a student set that does not penalize students for entering the department in later years would need to include only students who entered the department seven or more years prior to the latest semester for which there is data, Summer 2017. The result is a set of students entering the university's Department of Management during the four academic years beginning in 2007, 2008, 2009, and 2010. Figure 3 was plotted to visually confirm that the number of students who had graduated was acceptable.

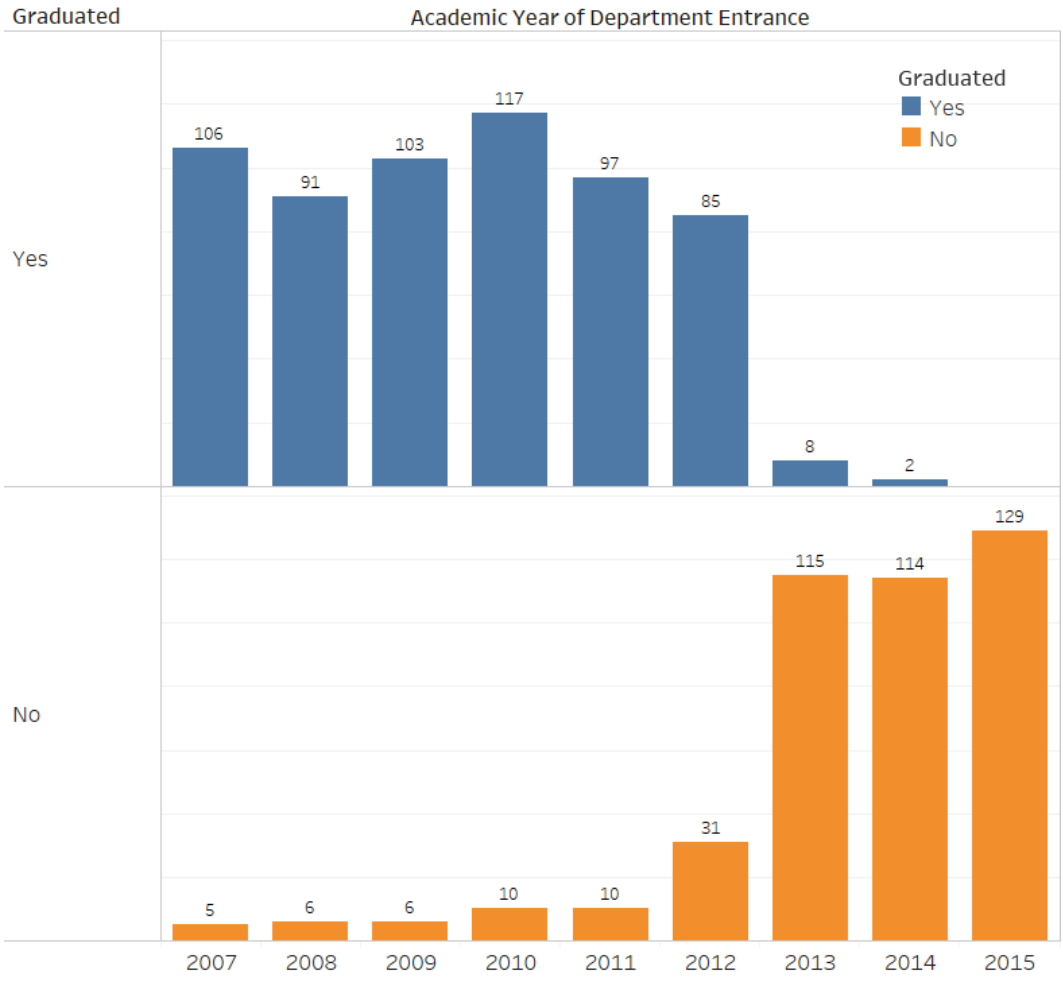

Fig. 3 Graduated and non-graduated students by year of department entry

3.3.3 Analysis of transfer students

The data set included one student who entered the university in the spring semester of the 2009/2010 academic year. Entering the university in a semester other than fall, the first semester of each academic year, indicates that the student transferred from another university. It was necessary to decide whether or not to remove transfer students from the data set prior to further analysis.

Because one of the research questions focused on predicting the achievement of Honors and High Honors, it was important to understand how many transfer students there were in the data set and the distribution of Honor/High Honors achievement among them. It became clear during the data cleaning and exploration phases that the original HonorsStatus variable was problematically inconsistent. This necessitated a corrected version of the variable be made, which is discussed in the Data Cleaning section of this paper.

Transfers students are fundamentally different than non-transfer students in that they must have already proven themselves to be successful students at another university to be able to then transfer to Boğaziçi University and also that their final GPA is calculated from approximately half or three quarters the number of courses. These differences lead to the expectation that transfer students are more likely to qualify for Honors or High Honors status. Consultation with university documentation and departmental authorities confirmed that transfer students are subject to the same criteria for Honors/High Honors recognition, and that the time spent in their previous university is counted.

A summary of these distributions of transfer students according to their graduation status is shown in Table 1, reflecting the new CorrectedHonors variable. The numbers indicated that a much higher percentage of transfer students achieve Honors and High Honors status than do non-transfer students, as well as that the total number of transfer students is a small fraction of the total number of students in the data set. While only approximately 10% of non-transfer students attain High Honors status, just over 35% did so among transfer students. High Honors was the highest frequency outcome observed among transfer students in the period of study. Despite the relatively small number of transfer students, this distribution confirmed that transfer students are more likely to achieve High Honors status than non-transfer students.

Table 1. Transfer and Non-Transfer Students by Graduate Distinctions

NonTransferStudent	Graduated	High Honors	36 Students
			9.78%
		Honors	107 Students
			29.08%
		Normal	205 Students
			55.71%
	Not Graduated	Non-Graduate	20 Students
			5.43%
Transfer Student	Graduated	High Honors	9 Students
			37.50%
		Honors	7 Students
			29.17%
		Normal	7 Students
			29.17%
	Not Graduated	Non-Graduate	1 Student
			4.17%

Because removing transfer students from the further analysis would reduce the cases of High Honors so significantly, and because they were ultimately equivalent students with respect to their status as students of the university, the decision was to retain transfer students in the data set. A new variable, TransferStatus, was created and included in the student clustering described later in this study.

3.3.4 Description of final data set

The final set resulted in a total of 392 students, all of which were Turkish nationals and who entered Boğaziçi University in or after the fall semester of the 2006/2007 academic year. Each student sat for the OSYM administered university entrance exam in or after 2005 and entered the Department of Management in and between the fall semester of the 2007/2008 academic ear and the spring semester of the 2010/2011 academic year.

3.4 Data cleaning

After understanding the nature of the data and the context of the research objectives, two variables were found to likely be important and also need extensive cleaning to correct erroneous values. In this section, the steps taken to study, clean, and prepare the data for EDM methods will be summarized. Preliminary analysis was conducted on the data in both Tableau and SPSS programs, visual analysis in the former and frequency analysis in the latter. Superfluous or otherwise low value variables were identified, which include variables for which data was either missing in the extreme or excessively inconsistent in form. The data was cleansed of errors determined to be typographical. Then, the basis for creating a subset of students was constructed with the objective to apply classification, or other EDM methods, to it. New calculated variables were created to

improve one critical variable, HonorsStatus, and also to provide additional information that was expected to be valuable to addressing the research questions. Following that, course data was analyzed and processes in similar ways. Several new variables were created from CourseCode values and course scheduling data. This was done to realize the latent classification potential in these types of original variables.

3.4.1 Cleaning the honors variable

Because predicting students' graduate distinction was the primary objective of this study, ensuring the integrity of the data in this variable was the highest prioritize in preliminary analysis and data cleaning. The frequency of HonorStatus values was observed and two observations stood out immediately: (i) no students who entered the university after the 2009/2010 academic year were awarded High Honors recognition and (ii) students could be awarded honors status without having graduated.

The first observation was surprising and without obvious explanation, while the second appeared counter-intuitive. Requests for explanation were made but unfortunately supporting information was received only for the latter case, clarifying those students with honors status despite having not graduated indicates the university's expectation that the student will graduate in the most recent semester, which was summer 2017. Unfortunately, this clarification was not entirely satisfactory as the expectation that a student will graduate at the end of a semester that had not yet concluded was considered unreliable.

To address these concerns, the decision was made to replace the provided HonorStatus variable with a new calculated field, labeled CorrectedHonors, which classified students according to the university's stated GPA and length of study criteria

for Honors/High Honors recognition. The new variable applies IF/THEN logic for each student's FinalGPA and their maximum value for their SemesterNumber variable. Possible values are: High Honors, Honors, and Non-Honors for students who graduated without distinction. Students who did not graduate where classified as Non-Graduate. Several checks were conducted to verify the accuracy of the new calculated variable's results. A comparison of the original and correct variables is shown in Figure 4.

An important observation from the new corrected variable was that there were students who were counted as Normal graduates in the provided data who would appear to have been eligible to receive Honors or High Honors status. These students may have been disqualified for disciplinary reasons, for which records were not included in this study. The disciplinary criterion of the Honors/High Honors achievement was deemed to be discountable as a factor in this study. There are also students who are recorded as having earned Honors status that would otherwise seem to have not satisfied the defined criteria. It contributes further to the impression that the university's provided HonorsStatus variable was problematic and unusable in its original form.

An outlier was identified in the form of a single student who entered the department in Fall 2008/2009, yet did not graduate while still technically being eligible for Honors recognition. The student was identified in the Semester data table as having only registered for the 3 semesters of the 2008/2009 academic year. The student participated in the preparatory program during the first two semesters, while they earned nine credits with a 3.0 SPA in the third semester, which was the summer semester for that academic year. The student does not appear again after that time and is presumed to have dropped out. Drop outs are rare in Turkish public universities because students attend university with minimal financial cost, so this particular student represents a rare

phenomenon in the data set. This particular student was not culled from the data set, but their status was recoded as Non-Graduate, since they would almost certainly be rendered ineligible for Honors status if they were to register at some point in the future and their apparent failure to graduate is judged to be material to this study.

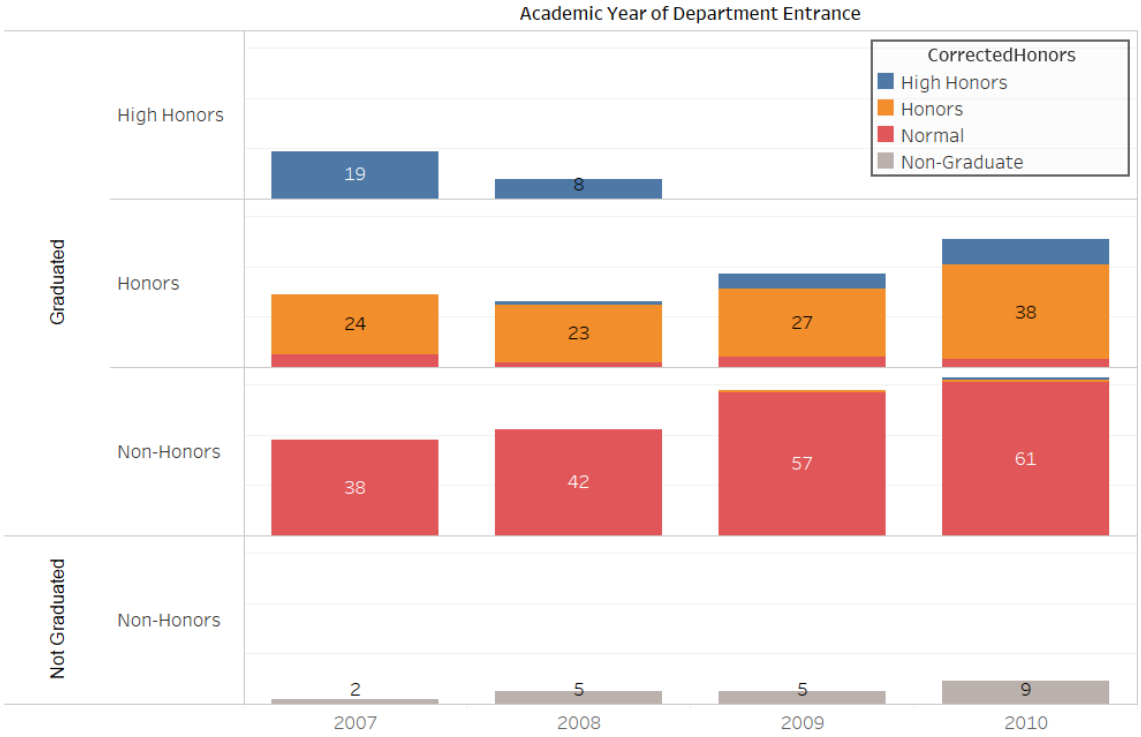

Fig. 4 Comparison of CorrectedHonors and original HonorStatus classification

3.4.2 Cleaning high school type values

A student's high school experience is accepted to be a major influencer of their later university experience. In Turkey, there are several types of high schools that vary by curriculum, criteria of acceptance, and language of instruction, which is perhaps most significant for the case of Boğaziçi University. The type of high school from which a student graduates was identified as a likely indicator of performance at the university.

Unfortunately, the data recorded for this variable was highly inconsistent and difficult to properly interpret. Changes made to Turkey's educational policies over the period of the study were one concern, but most problematic was simply the apparent lack of consistency in how the categories of high schools were recorded. Efforts were made to correct typographic errors and inconsistencies for which corrections could be made with confidence. The corrected values are listed in Table 2.

After this cleaning, there remained concern about the accuracy and consistency of the high school type classification. Online research was made into each of the documented high schools and this revealed as well that some of these high schools have closed in recent years and information that might support their classification was inaccessible at the time of this study. A more comprehensive and robust analysis of verified high school types remains a subject of potential value for future research.

Table 2. The Count and Percentage of Students by High Schools Types in Turkey

HighSchoolType	Count of Students	Percent of Total
Anadolu Lisesi Yabancı Dille Öğretim Yapan Öz	232	59.18%
Yabancı Dille Öğretim Yapan Öz	80	20.41%
Anadolu Öğretmen Lisesi	29	7.40%
Fen Lisesi	29	7.40%
Lise (Resmi ve Gündğz Öğretimi)	6	1.53%
Sosyal Bilimler Lisesi	5	1.28%
Özel Fen Lisesi	4	1.02%
Unrecorded	2	0.51%
Özel Lise	2	0.51%
Askeri Lise	2	0.51%
Lise Programı	1	0.26%

CHAPTER 4

DATA EXPLORATION

The cleaned and prepared data was extensively explored and descriptively analyzed. Insights derived from this exploration were considered of interest in their own right, in addition to the value they contributed to the development of the prediction models. Several new variables were created through the course of this exploration and many were incorporated into the prediction models, including high school clusters and normalized course scores.

Beginning with course data, this section contains analysis of subject categories and grade frequencies, the results of course clustering, and an investigation of seasonal variances in the grading of several difficult courses. In the following exploration of student academic data, insights are mined from English language preparatory data, students are clustered, which leads to the presentation of a new student success construct to replace the Honors graduate distinction used by the university, and then both the new student "Success" variable and its predecessor are tested against normalized student performance. This normalized performance analysis suggestions modifications to the university's criteria for Honors, which would also influence the Success variable used throughout the rest of the study. Next, data related to students' high schools are explored, augmented with additional research, and clustered to reveal interesting insights about the rural and urban backgrounds of students. Finally, data about student's participation in foreign exchange programs is explored leading to an actionable insight for academic advisers to consider.

4.1 Exploration of course data

The exploration of course data proceeds by categorizing courses by their subjects of study and the importance to the program's curriculum as inferred by the number of students enrolled in them. That is followed by a description of observed grading frequencies. The courses are then clustered according to aggregated student performances and other criteria in order to focus later analysis at meaningful sets of courses. Finally, the grading trends of courses are analyzed by season, to test the suggestion that course grading tends to be more generous in summer semester, as provided by student interviews.

4.1.1 Categorization of course by subject

A total of 531 unique course codes were registered to the 392 students in the final student set. The CourseCode variable was broken down into the subject and numerical components. For example, the Intro to Management course was identified by the course code AD 102, wherein AD marks the course as a Management course and 102 indicates the course is intended or required to be taken in a student's first academic year. Two new variables were created from this, labeled CourseSubject and CourseLevel. There were fifty-five values in the new CourseSubject variable and five values for CourseLevel, the latter representing courses for the four years of the bachelor program and graduate level electives.

Fifty-five course subjects were found to be too many to allow for effective descriptive analysis. This number was inflated by different but related course subjects, such as foreign language courses which possess different subject qualifiers for each language (FR for French, KR for Korean, GR for Germany). Similarly, different

engineering programs possess different course code qualifiers for each type of engineering discipline, such as a BM for biomedical engineering and IE for industrial engineering. Students in the final set registered to courses in seven engineering subjects. To reduce the number of course subjects into something more conducive to descriptive analysis, the CourseSubject variable was recoded into a new, higher-level variable, labeled CourseCategory.

Course categories were determined by reviewing the university department websites associated with the subject classifiers, such as EC for the Department of Economics. While some categories are largely composed of by a single subject, such as Economics (EC), Management (AD, MIS), or Sociology (SOC), other categories, such as Art, encompass a diversity of subjects and courses which are often elective. In cases where courses were elective and enrolled by few students, it was deemed acceptable to combine subjects such as literature, cinema, and fine art into a single course category, "Art" in this example.

The seventeen possible values for this new variable are: Management, Economics, Math, History, Humanities, Foreign Language, Turkish Language, Psychology, Philosophy, Physics, Sociology, Art, Education, Natural Science, Engineering, Tourism, and International Trade. The variety of study subjects found within the course data was indicative of a high degree of diversity in the academic experience of management students. There was likely a variety of distinctive student profiles during the period of study. This variable will be used in coming sections where the letter grade frequencies are discussed.

4.1.2 Course letter grade frequencies

The general distribution of letter grades was an important phenomenon to look at, as it would offer insight into the general experience of students studying in the management program. If, for example, AAs were very common and Fs were very uncommon, one would expect that the contribution of academic performance was a moderate or even mild stressor for students. If, however, students received Fs which great frequency, then it would be reasonable to expect academic stress to be a significant component of the student experience. Another expected inference was that as students matured and the program curriculum came to feature more elective options that the grades would begin to skew higher in later years of the program.

Figure 5 plots the frequency of letter grades for courses along the previously created CourseLevel variable, which indicates whether a course was intended to be a first year course, such as AD 101, or a third year course, such as AD 353. The results confirm the latter expectation that grading changes across time; the frequency of Fs drops as the courses become higher level. Students were almost twice as likely to earn AAs in their fourth year level courses than in their first year courses, and the frequency of Fs in fourth year courses were a quarter of what they were in first year courses. Other types of course grade outcomes, such as PASS, FAIL, and Incomplete were culled from this analysis. A set of 531 courses provided the basis for this and subsequent course analysis.

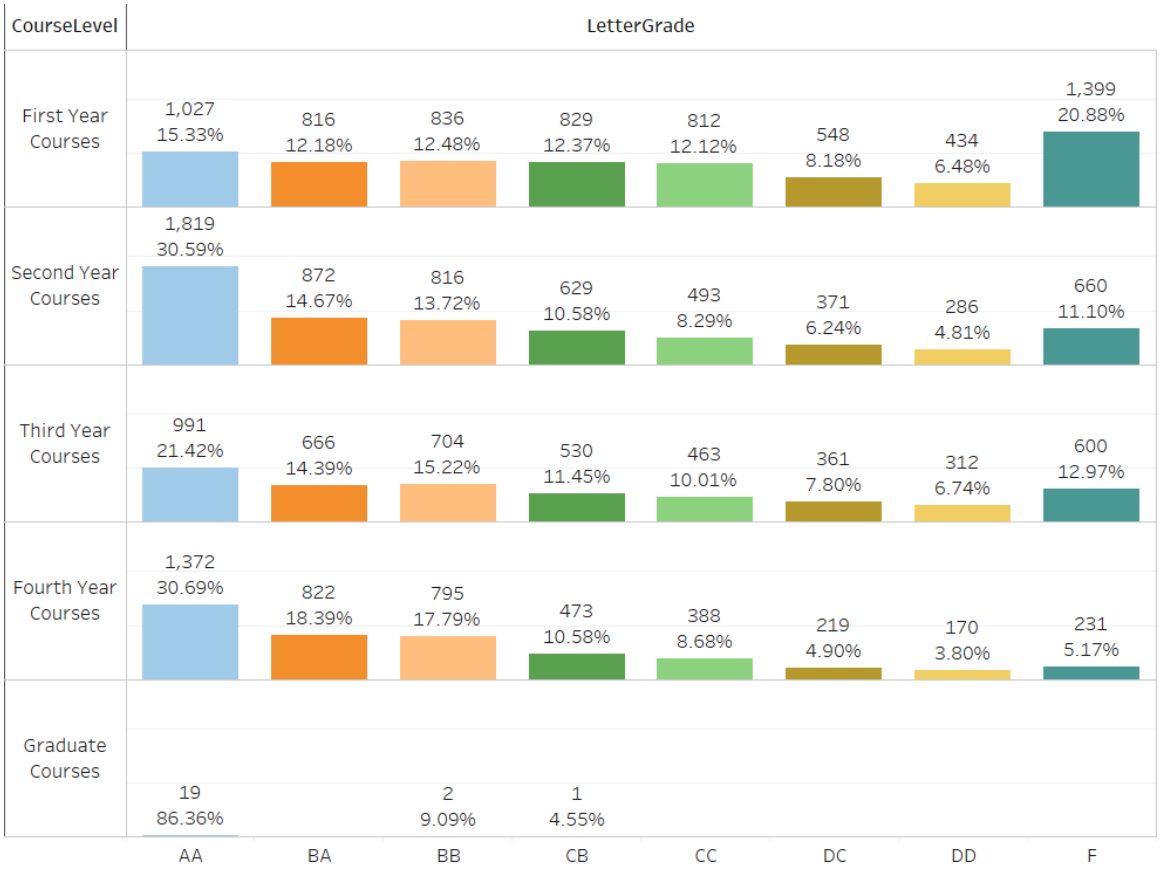

Fig. 5 Distribution of letter grades by course year

The results also clearly indicate that first year courses were by far the most challenging for students, as Fs were the most common grade received. These results include all letter grades issued in the study period, including students who earned multiple Fs in courses they repeated. The data suggested that students struggled through first year courses, experienced something of a reprieve in second year courses, before being challenged again by third year courses. Fourth year courses, as well as perhaps graduate level courses, appeared to be far less challenging to the students in this study.

4.1.3 Analysis of courses by student performance characteristics

When the courses were analyzed individually by student performance five courses among the first year courses and three courses among the third were members stood out as particularly difficult courses. These courses were MATH 101, MATH 102, AD 131, HIST105, HUM 101, AD 311, AD 351, and AD 353. The concentration of these courses in these program years help to explain why these years were shown to be the most difficult for students.

Figure 6 presents course-specific grade averages in a colored heat map, which reveals these challenging courses (in shades of red) and the degree to which they appear to stand out from other courses. Most courses possess grade averages above 2.0, indicating that most students pass with a CC or higher grade. This course-specific grade averages were calculated by considering every grade issued, even Fs that students later replaced on their transcripts by retaking the course and earning passing grades. This method gives a better view of a course's actually grading experience than considering only the final course grades students graduate with. This method allows us to visually identify courses possessing high frequencies of failing grades.

The manner in which these difficult courses appeared to stand out so distinctly from other courses, even those within their same course level, supported the hypothesis that cluster analysis would reveal groups of difficult courses more alike each other than they are to courses with which they may share greater superficial similarity, such as subject or program year. To test this, a cluster analysis was performed and is described in the following section. Average student grades for courses with more than 160 registered students are found in Figure B1.

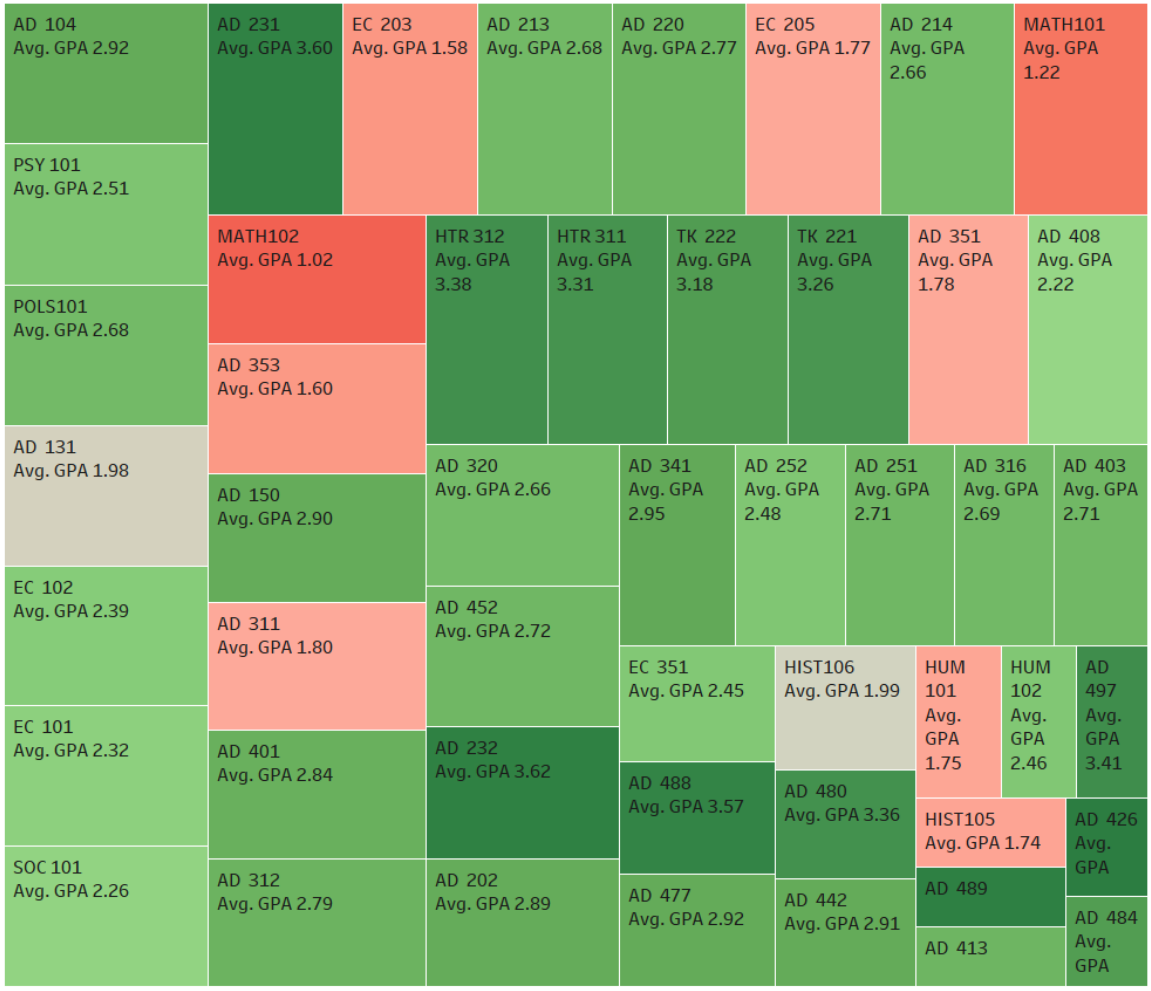

Fig. 6 Most commonly registered courses by average GPA
Box size indicates the number of students. Minimum number of registered students is 100.

4.1.4 Results of course clustering

Courses were clustered using the K-Nearest Neighbor method and a target of five clusters. The target number of clusters was determined by dendrogram analysis and the comparing the clarity of different numbers of clusters. The data set used for clustering contained the following variables for each course: CourseCode, CntStudents, CntStudents, AvgNumGrade, StdNumGrade, CntDs, CntFs, CntRepeats.. Rather than clustering solely on the basis of student grades, the number of students registered to

these courses was included in order to identify the difference between difficult elective courses and difficult required courses. PASS/FAIL classes, such as courses taken during a foreign exchange program, and instances of I (Incomplete) or W (Withdraw) grades were excluded from this data set. Mean average and standard deviation calculations cannot be calculated for classes not giving numerically translatable grades. The result was a set containing 531 distinct courses. The resulting course clusters are described below in Table 3. The list of variables used in course clustering is found in Table A2.

The largest cluster by number of courses, labeled Rare, contained 483 different course codes and typically saw only a few students registered per course. Students register for these courses at their discretion, and these courses featured the highest and tightest band of average grades. Students almost never failed or repeated these courses. Members of the Uncommon cluster where to similar to those of the Rare cluster in that students took registered for these courses either as electives or as options toward the satisfaction of program requirements, such as the first year history/humanities requirement which offers students the choice of courses.

The two smallest clusters stood out as targets for further analysis, descriptively labeled 1stYearMath and DifficultCommon. As the names suggest, 1stYearMath was composed of only MATH101 and MATH102 courses, which are required first year calculus courses, and DifficultCommon was composed of six other difficult courses to which most students registered. Keeping in mind that there are only 392 students in the data set, and not all students progressed to the point of registered for either MATH101 or MATH102, an average of 370 repeated enrollments was quite extraordinary. These two required math courses ultimately accounted for more Ds and Fs than almost all other enrolled courses combined.

Students who earn an F in a required course are obliged to retake it. Students who earn letter grades of DC or DD in a course are eligible to enroll in the course again and replace its grade, but the results reveal that few students take advantage of this opportunity to repeat courses in which they ear Ds. For all course clusters, the number of Fs and repeats are close to equal. Figure 7 shows the letter grade distribution of the majority of difficult courses, those which are members of the 1stYearMath and DifficultCommon course clusters. In addition to the high frequency of Ds and Fs, it is worth nothing that very few students receive As from these courses, indicating that the courses' difficulties are experienced consistently across the student body as opposed to particularly challenged students.

Students likewise often struggled to pass courses in the DifficultCommon cluster. This cluster featured two Economics courses, both in the second year, and four Management courses, two in the first year and two in the third year, in which students failed and repeated at significantly higher rates than in other courses.

Table 3. Final Course Clusters' Centers

Course Clusters					
Number of Courses	2	6	25	15	483
	Cluster Centers				
	1stYearMath	DifficultCommon	Common	Uncommon	Rare
AvgNumGrade	1.12	1.75	2.83	2.8	2.86
StdNumGrade	1.2	1.34	1.08	0.98	0.46
CntStudents	383	382	368	191	8
CntDs	128	116	44	24	1
CntFs	359	143	27	17	1
CntRepeat	370	143	32	18	1

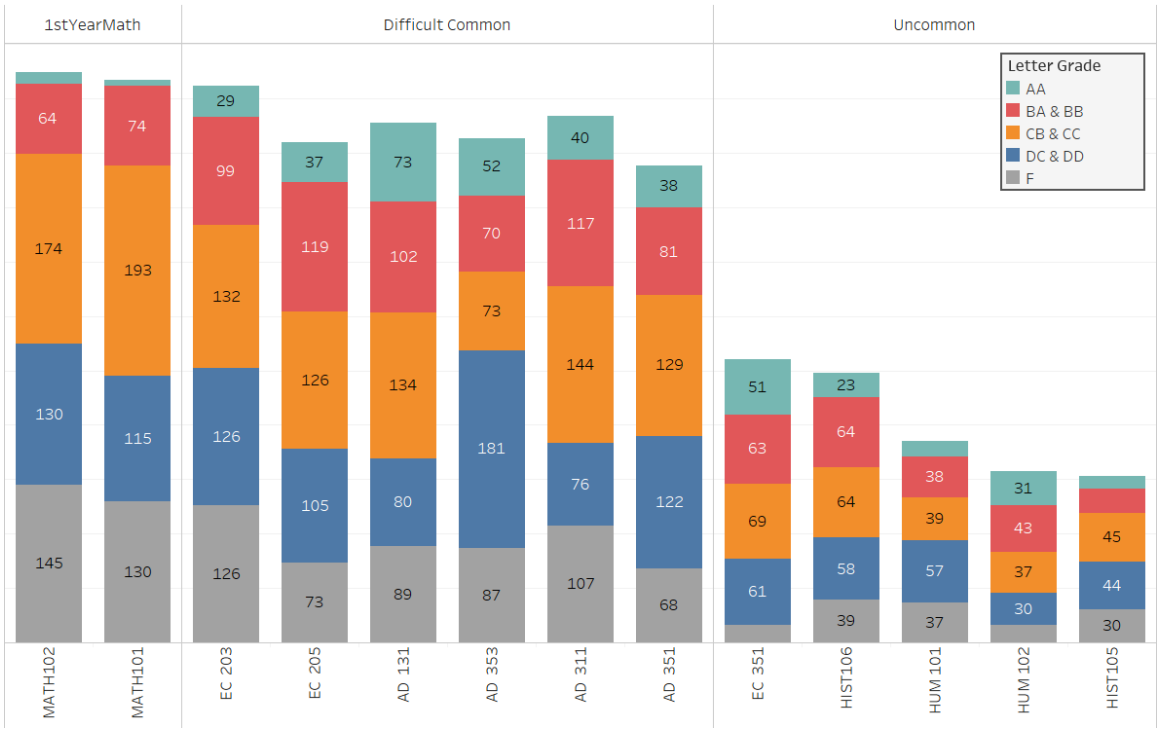

Fig. 7 Number of students by letter grade earned in the ten most difficult courses

4.1.5 Correlations between GPA and elective courses

The courses making up the Rare cluster indicate students' differing academic interests. Despite the relatively low number of students studying courses in some subject areas, it was of interest to analyze differences that students studying certain subjects might display. The final GPAs of students were analyzed according to the electives course subjects and are listed in Figure 8 along with other measures of student performance.

The average final GPA of the fourteen students who elected to take extra math courses was 3.25 (shown in dark green), which was well above the total average final GPA of 2.89 (shown as grey). Few other course groups were associated with significantly higher GPAs, while several were actually reflected of lower average GPAs.

Unexpectedly, the average final GPA of the twenty-two students of engineering courses, 2.73, was less than the average of all students, which was 2.89. It was expected that students with superior quantitative skills would perform better because of the value such skills would have in the program's difficult math, economics, and management operations courses. Generally, the average final GPAs were lower for most elective course categories that diverged from the clear business or social science subject areas of management, economics, international trade, tourism, and sociology, suggesting that students who take the courses less related to the program's primary study areas finish with lower GPAs.

Course Cluster	Course Category	Avg. Final Student GPA	Std. Final Student GPA	Cnt Students	Cnt Ds	Cnt Fs
Rare	Art	2.91	0.48	167	15	41
	Economics	2.97	0.50	151	35	3
	Education	2.75	0.35	39	0	1
	Engineering	2.73	0.46	22	0	8
	Foreign Language	2.94	0.45	337	34	66
	History	2.80	0.46	86	5	16
	International Tr	2.90	0.50	29	1	1
	Management	2.91	0.48	375	81	119
	Math	3.25	0.51	14	0	6
	Natural Science	2.73	0.41	128	22	28
	Philosophy	2.76	0.46	99	18	18
	Physics	2.84	0.50	62	12	13
	Psychology	2.83	0.45	121	47	11
	Sociology	2.93	0.46	33	5	7
	Tourism	2.91	0.44	31	4	1
	Turkish Language	2.74	0.44	16	1	5

Fig. 8 Characteristics of elective course subjects

4.1.6 Seasonal grading

During the planning stage of this research, it was suggested anecdotally by recently graduated undergraduate students that some students may intentionally fail certain

courses that they could take again in the summer in order to maximize the amount of attention they could devote to particular courses. The reason being was that courses offered in the summer are purportedly more leniently graded than these same courses are during fall and spring semesters. Analysis of the most difficult courses' scores revealed little evidence of this summer effect on grades. Two courses showed some variance in grading between fall semester and other seasons, but discussion with departmental authorities suggest that these effects were instructor-based and out-dated. Figure B2 depicts the frequency of grades for many of members of the most difficult course clusters.

4.2 Exploration of student academic data

In this section the varieties of students' academic experience are explored, beginning with a descriptive analysis of the English language preparatory program, in which most students were compelled to spend some number of semesters. Students will be clustered along various characteristics of their grade performances. The resulting insights from this clustering led to the proposal of a new construct to denote students' successful academic achievement upon graduation, which is described and then compared against the original Honors/High Honors construct.

4.2.1 Analysis of the English preparatory program

The influence of Boğaziçi University's foreign language of instruction was a key interest from the earlier planning stages of this study. It was reasoned that students who are better prepared linguistically to begin university-level studies were more likely to succeed in the first years of their studies, and thus more likely to earn graduate

distinctions of Honors and High Honors. Unfortunately, exam data for students was unavailable for this study, which is a noted weakness of this analysis.

The data available was limited to the type of proficiency exam a student succeeded in and the number of semesters, if any, the student spent enrolled in the university's English preparatory program. Rather than a student's passing proficiency score, how quickly a student was able to satisfy their language proficiency requirement was proposed as an indirect indicator of that student's linguistic preparedness for degree studied. Students who enter the university with high levels of proficiency are expected to either bypass the preparatory program all together.

A new variable was created to count the total number of preparatory semesters for each student. Yearly GPAs were calculated by summing the earned course points and earned credits for semesters occurring in the same academic year, and then dividing the former by the latter. Figure 9 shows the relationship between yearly GPA and the length of time student spent in the preparatory program.

The most important initial observation to make from Figure 9 is that students tended to overwhelming require two semesters in order to progress through the preparatory program, 67% of students in the period of study. Despite that, fifty-three students (13.59% of all students) were able to bypass the preparatory program completely, and their average GPA was higher in each of the first four years.

The second most important observation was the unmistakable decrease in average GPAs for each extra semester spent in the preparatory program. It holds consistent for the first four years of academic study and suggests that longer language preparatory work is directly correlated with poorer academic performance in the degree program. This is an intuitive result, but it remains striking in its consistency. Students

who spent more than three semesters in the preparatory program would be expected to struggle passing their classes and would have very little chance at achieving honors status.

	Cnt Prep Semester						
Yearly GPA	0	1	2	3	4	5	6
1stYGPA	Students: 53 13.59% 2.91 Avg.GPA 0.59 Std.GPA	Students: 22 5.64% 2.70 Avg.GPA 0.59 Std.GPA	Students: 262 67.18% 2.35 Avg.GPA 0.69 Std.GPA	Students: 23 5.90% 2.00 Avg.GPA 0.61 Std.GPA	Students: 22 5.64% 1.41 Avg.GPA 0.68 Std.GPA	Students: 5 1.28% 1.09 Avg.GPA 0.61 Std.GPA	Students: 3 0.77% 1.32 Avg.GPA 0.18 Std.GPA
2ndYGPA	Students: 53 13.62% 3.08 Avg.GPA 0.79 Std.GPA	Students: 22 5.66% 2.90 Avg.GPA 0.79 Std.GPA	Students: 261 67.10% 2.76 Avg.GPA 0.82 Std.GPA	Students: 23 5.91% 2.22 Avg.GPA 0.80 Std.GPA	Students: 22 5.66% 1.66 Avg.GPA 0.80 Std.GPA	Students: 5 1.29% 1.50 Avg.GPA 0.85 Std.GPA	Students: 3 0.77% 1.62 Avg.GPA 0.27 Std.GPA
3rdYGPA	Students: 50 12.99% 3.13 Avg.GPA 0.62 Std.GPA	Students: 22 5.71% 2.79 Avg.GPA 0.87 Std.GPA	Students: 260 67.53% 2.65 Avg.GPA 0.78 Std.GPA	Students: 23 5.97% 2.41 Avg.GPA 0.63 Std.GPA	Students: 22 5.71% 1.66 Avg.GPA 0.85 Std.GPA	Students: 5 1.30% 1.62 Avg.GPA 0.90 Std.GPA	Students: 3 0.78% 1.95 Avg.GPA 0.92 Std.GPA
4thYGPA	Students: 44 11.92% 3.25 Avg.GPA 0.64 Std.GPA	Students: 21 5.69% 2.90 Avg.GPA 0.88 Std.GPA	Students: 255 69.11% 2.81 Avg.GPA 0.86 Std.GPA	Students: 21 5.69% 2.29 Avg.GPA 0.69 Std.GPA	Students: 20 5.42% 2.17 Avg.GPA 0.80 Std.GPA	Students: 5 1.36% 1.57 Avg.GPA 0.94 Std.GPA	Students: 3 0.81% 2.15 Avg.GPA 0.80 Std.GPA
5thYGPA	Students: 7 6.73% 2.21 Avg.GPA 1.07 Std.GPA	Students: 11 10.58% 3.11 Avg.GPA 0.95 Std.GPA	Students: 62 59.62% 2.17 Avg.GPA 0.99 Std.GPA	Students: 11 10.58% 1.73 Avg.GPA 0.51 Std.GPA	Students: 9 8.65% 1.54 Avg.GPA 0.70 Std.GPA	Students: 2 1.92% 0.58 Avg.GPA 0.58 Std.GPA	Students: 2 1.92% 1.94 Avg.GPA 0.57 Std.GPA
6-10thYGPA	Students: 1 2.94% 3.50 Avg.GPA 0.00 Std.GPA	Students: 1 2.94% 3.25 Avg.GPA 0.00 Std.GPA	Students: 19 55.88% 1.63 Avg.GPA 1.02 Std.GPA	Students: 3 8.82% 0.72 Avg.GPA 0.44 Std.GPA	Students: 7 20.59% 1.07 Avg.GPA 0.75 Std.GPA	Students: 2 5.88% 0.74 Avg.GPA 0.74 Std.GPA	Students: 1 2.94% 2.31 Avg.GPA 0.00 Std.GPA

Fig. 9 Student's final GPA by academic year and number of prep semesters

Because these figures tracked transfer and non-transfer students, the total number of students fell from one year to the next. This typically reflects the number of students who graduated, which was highest after the fourth year, but also included a few students who dropped out of the program. This subject would benefit from deeper analysis and verification with a larger and more recent data set.

4.2.2 Results of student clustering

While the CorrectedHonors variable classifies students on the basis of their course performance and the speed with which they complete the program, it only considers final outcomes. Students may not arrive at these final outcomes in the same way, and the differences in how students register or progress through the program may reveal valuable insights for departmental decision makers.

To illuminate possible fundamental differences in the academic path the students take, a k-means cluster analysis was performed. Prior to the k-means analysis, a dendrogram was created by a hierarchical cluster analysis. Combining visual analysis with the results of several iterations of the k-means analysis, six was identified as the best number of clusters. In actuality, two large clusters were observed, with three smaller groups of distinct student groups and a single unique outlier remaining. This outlier was kept as part of the cluster set for its instructive value. Those clusters were labeled and are described below. Full cluster definitions are listed in Table A3.

Cluster 1: Possible Honors Graduates (175 students, 44.64%)

Students in this cluster, which is the largest, tended to graduate in four years and tend to not earn Ds or Fs in their courses. They typically did not register for summer courses. They tended to have lighter course loads in their fourth year. 36% of students in this cluster participated in foreign exchange programs. 64% of these students were female.

Cluster 2: Non-Honors Graduates (161 students, 41.07%)

Students belonging to this cluster tended to graduate between four and five years, and typically needed to repeat a few required courses. Ds and Fs were not very common, except for the 1stYearMath and DifficultCommon clusters of courses. These students

registered for the summer semester more regularly the students of Cluster 1, but not more than a few courses in total. 62% of these students were male.

Cluster 3: Challenged Students (26 students, 6.63%)

Students of this cluster struggled to pass their courses and graduate from the university. Only 62% of these students graduated from the program, and those that did tended to require six to seven years to do so, not including the time they spent in the language preparatory program. These students, which are overwhelmingly male (73%), were more likely than the other groups to spend three semesters in the English language preparatory program before entering the degree program. It is likely that some of these students were working rather than actively attending courses with the intent to graduate. Male Turkish students enrolled in university are able to postpone compulsory military service (military service is not compulsory for females) and this is another possible reason to explain why some of these students did not progressing at a more typical rate.

Cluster 4: High Performing Students (19 students, 4.85%)

This group of students were academic successful and included a number of transfer students (21%). They were characterized by high final GPAs (average 3.41), but occasionally needed more than 4 years to graduate. This would have disqualified these students from achieving Honors or High Honors status, despite their high academic performance. These students registered for and earned points from large numbers of elective courses, as indicated by their cluster's RarePoints center value. 68% of these students were female and almost half participated in a foreign exchange program.

Cluster 5: Failed Students (10 students, 2.55%)

This cluster of students did not graduate and typically did not progress beyond the first few years of the program. As the university does not typically expel students for failure to graduate, these students may continue to register or be registered for new semesters without earning many passing grades. These students were predominantly male (80%), and the same suggested reasons for why the students of Cluster 3 were slow to complete their studies may be applicable here.

Cluster 6: Extended-Study High-Performing Graduate (1 student)

This single student is unusual enough that he is well and truly an outlier. While his case does not tell us much about the other students or even about normal students in general, he was a curious example of a student who averaged almost an AA over five years of full course loads, including high Summer course loads. In the end, this student registered for over 100 credits from Rare courses and averaged almost a 4.00 GPA in those classes. For perspective, the full undergraduate management degree program's curriculum required 144 credits during the period of the study. This student would not have been eligible for High Honors recognition because of the length of his study, but his unique case might cause one to question the purpose of such academic recognition if it excludes students with academic performance of this nature.

The key insight from this cluster analysis, aside from confirming that students take different paths through the program, is that there are the university's Honors/High Honors criteria seems to be cutting out some of the university's most academic successful students. If spending an extra semester or two in the language preparatory program does not disqualify a student from achieving recognition for their success in the degree program, then perhaps neither should spending an extra semester or two to

pursue extended study in specialized electives and graduate-level courses disqualify a student from such recognition. In the following section, a new student success classification will be presented and described.

4.2.3 Definition of a new graduate distinction

In the previous section, students were clustered according to several characteristics of their university experience. It was observed that some students, who attain relatively high levels of academic success, as indicated by their GPAs at the time of their graduation, would technically be ineligible for earning Honors/High Honors status because they spent longer than eight semesters, not counting summer semesters, in the degree program.

This would cause problems for prediction, because these non-honors graduates would appear to be Honors or High Honors students by most measures of academic performance which would be included in the model. This kind of contradictory classification would make the work of the machine learning algorithms more difficult and almost certainly result in less accurate results. Practically, the result would also lead to incorrectly identifying successful students as less successful ones, which would provide value to the department should the prediction models be deployed.

Revised classification rules were created for this study to improve the performance of and value derived from prediction modeling. The new rules are described below and led to the creation of a new variable, labeled Success. These rules primarily focused on increasing the length of time a student can to more accurately differentiate between students who are struggling and those who are performing well in their courses. The length of potential study was increased from eight semesters to ten for

non-transfer students and from six to eight semesters for transfer students. The GPA criteria was kept from the university's current Honors classification system. Additionally, this study proposes to acknowledge those students who continue to study in excess of this semester limit while maintaining a GPA above 3.5. The full definition is outlined below:

Class 1: High Achiever

For non-transfer students who graduate in five years or less after entering the Department of Management, while finishing with a GPA at or above 3.5.

For transfer students who graduate in four years or less after entering the Department of Management, while finishing with a GPA at or above 3.5.

Class 2: Achiever

For non-transfer students who graduate in five years or less after entering the Department of Management, while finishing with a GPA between 3.0 and 3.49.

For transfer students who graduate in four years or less after entering the Department of Management, while finishing with a GPA at or above 3.5.

Class 3: Graduate

For all students who graduate with a final GPA below 3.0, regardless of the length of time they studied.

Class 4: Late Achiever

Not all students attend university for the specific purpose or graduating and entering the workforce. Some students are keen to extend their studies for one of several reasons, and the presence in classes is not seen as detrimental to university. This study proposes acknowledging these students who are motivated to pursue extended independent study if they are able to do so with high academic success.

For non-transfer students who graduate after more than five years of study after entering the Department of Management, while finishing with a GPA at or above 3.5, and transfer students who graduate after more than four years of departmental study while also maintaining a GPA above 3.5.

As extended study allows a student to earn credits from elective courses, which have been observed to be graded higher on average than required courses in the first years of the program, a 3.5 GPA criterion is considered to the best compromise.

Class 5: Non-Graduate

All other students are collected in this class, including both those still actively studying after many years and those who dropped-out of the university.

Compared with the university's original criteria, the new Success construct classified most students equivalently, seen in Figure 10. However, as intended, it did identify several students for graduate distinction that the previous Honors construct ignored. In total, five additional students were classified as members of the highest achievement class, along with ten extra students in the lower achievement class (Honors in the old system and Achievers in the new one). These students would have been discounted on the basis of studying for one or one half extra year.

There were no students in the data set that satisfied the criteria for Late Achievers, suggesting that such a class of distinction may not be needed. A larger data set would help to confirm the practical value of recognizing late graduates with very high GPAs.

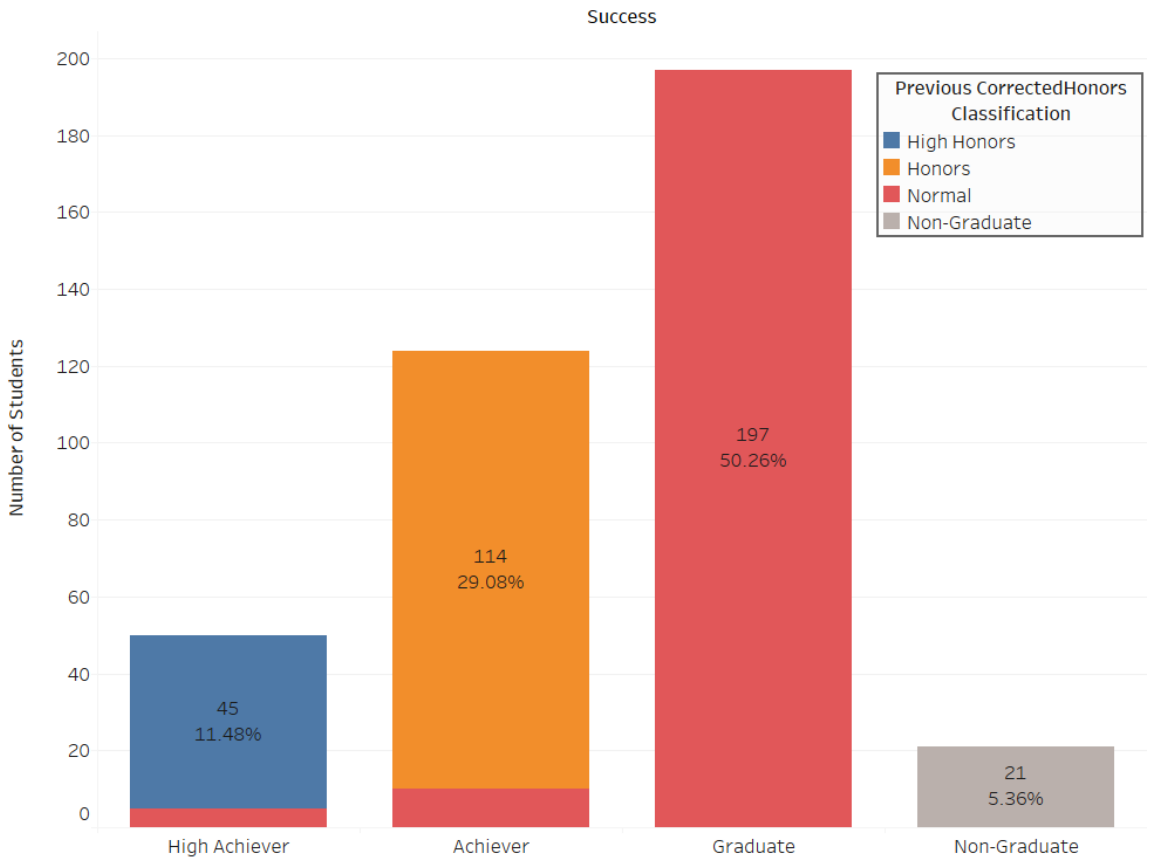

Fig. 10 Comparison of CorrectedHonors and new Success classifications

4.2.4 Verification of grade distinctions by grading normalization

During the analysis of grade frequencies, it was observed that the average grades given in certain classes are higher than others. Particularly, many elective courses that students enrolled in were likely to be graded closer to AA than were many required courses. Because of this, it was possible that students with similar final GPAs might have reached that result with different levels of difficulty. The final GPA would not accurately reflect the degree to which a student outperformed his or her peers. Normalizing course grades makes it possible to consider a student's earned grade in the context of all grades earned from a course and deriving a numerical metric.

Students' grades were normalized for each course contained in the four course clusters: 1stYearMath, DifficultCommon, Common, and Uncommon. Each of these courses was enrolled by between 110 and 391 students. The Rare course cluster was excluded because these courses were enrolled by 100 or fewer students and the majority of these were enrolled by less than ten students in the data set. Normalizing scores would have been problematic with so few data points. All courses were letter graded, so there were no issues trying to translate PASS/FAIL to this scale.

Each student's normalized scores were calculated for each course by subtracting from their highest numeric grade earned in that particular course the arithmetic mean of all numeric grades in that course and dividing that difference by the standard deviation of all numeric grades in the course. The result is a numerical variable with a range above and below zero and indicative of a student's performance relative to other students in that course. If every student in a course were to earn an AA, then each student's normalized score should be zero

Because this method of normalization utilizes standard deviation, the range of normalized scores is not limited to a one and negative one. For certain courses in which the most common grade given is an F, such as the 1stYearMath courses, students earning an AA earn normalized scores exceeding 1.0. Conversely, a score below negative one would result in cases in which a student fails a course from which most students earn high letter grades. The full list of courses along with the arithmetic mean and standard deviations of their total grades can be found in Figure B1.

To test the extent to which course selection plays a role in the GPA scores of students and their graduate distinction, the average of normalized course scores and final GPA was scatter plotted for each student. As expected, there is a strongly positive

correlation between these two variables, particularly as the values rise, displayed in Figure 11. Color coding illustrates the explicit effect of GPA on graduate distinctions, but also that normalized scores suggests that some students either benefitted or were penalized by the relative difficulty of their course selections. Three students were able to attain Achiever classification despite actually performing worse on average than the whole of the student set, as indicated by the three blue boxes below the X axis.

Many students shared similar normal score averages but varied widely by GPA and thus whether or not they earned a graduate distinction, as indicated by the mix of brown crosses and blue squares between 0.2 and 0.4 on the average of normalized scores axis.

Following the trend line in Figure 11 suggests that GPAs above 3.15 accurately indicated the students performing best academically relative to their peers, but GPAs beneath that level were less consistently correlated with students' average normalized academic performance. As the Success and Honors classification systems differ only by the length of study criterion and both utilize GPA as a key criterion, the observed correlation with averaged normalized scores was consistent between them. The results suggest that the university consider raising the minimum GPA threshold for graduate distinction from 3.0 to 3.15 and from 3.5 to 3.6 to ensure that the graduate distinctions bestowed upon students are fair.

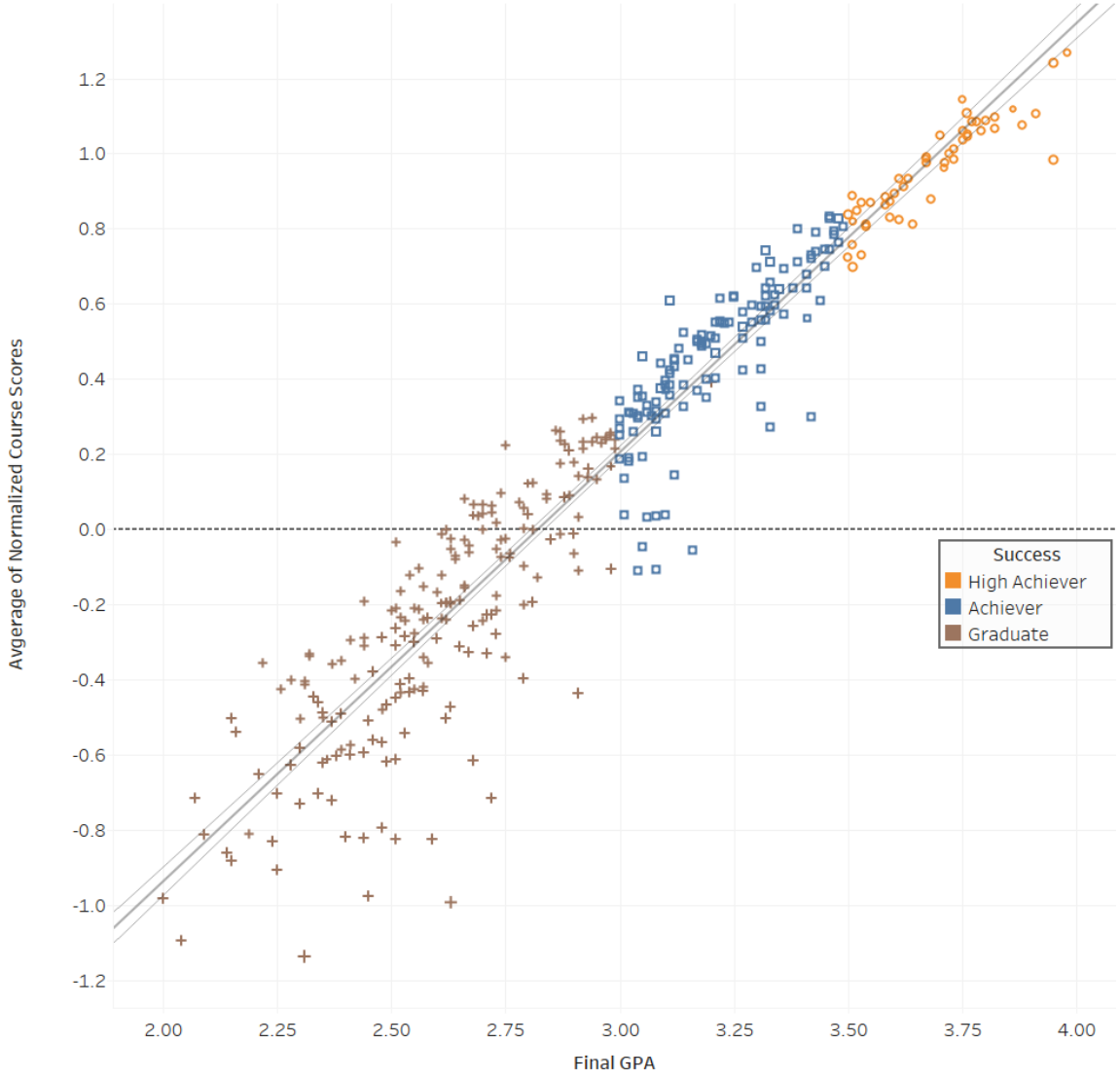

Fig. 11 Correlation between students' average normalized scores and final GPA

4.3 Exploration of high school data

A student's high school was expected to hold some value in predicting whether or not that student would succeed at Boğaziçi University, as thus whether or not they would attain graduate distinctions. There were a total of 174 high schools contributing students to Boğaziçi University's Department of Management undergraduate program during the period of the study. In this section, the general characteristics of high schools are

reviewed, particularly their locations and the type of their curriculum. Later the correlation between high school and graduate distinctions are described.

4.3.1 Distribution of students and high schools in Turkey

The largest sources of students to Boğaziçi University's undergraduate degree in management during this study period were typically the largest population centers in Turkey: Istanbul, Ankara, Izmir, Bursa, etc... (although not in that order). Ankara is the second largest city by population in Turkey, yet was only the fifth largest source of students, which may have been due to the number of prestigious universities located there. Boğaziçi University received the highest percentage of its students from its own city of Istanbul. As public universities are allocated students on the basis of national examination rather than selecting them, this distribution suggests that high schools in Istanbul are more successful at preparing students for university entrance examination than are other parts of country.

With a population of over eleven million, Istanbul is the largest city in Turkey and the region. For students coming from smaller cities or rural environments, it would be reasonable to assume that moving to the big city could influence their academic experience. Alternatively, an observed difference in a student's academic outcomes along the city size criterion could indicate differences in the educational preparedness provided by institutions in these locales.

Tables 4 and 5 reveal the frequencies of graduate distinctions by city size. As no other city in Turkey is close to Istanbul in population, and because it served as the basis for comparison, it was left alone. Cities with urban areas of more than one million residents were combined in the Large Cities group: Ankara, Izmir, Bursa, Adana,

Gaziantep. Cities with less than one million residents in a central urban area were grouped under Small to Medium Cities.

Table 4. Count of Students from Turkish Cities of Different Sizes

High School Cities	Number of High Schools	Number of Students	Percent of All Students
Istanbul	43	152	0.3887
Large Cities	42	103	0.2634
Small to Medium Cities	89	136	0.3478

Table 5. Frequency of Graduate Distinction by City Size

Success		High School Cities		
		Istanbul	Large Cities	Small to Medium Cities
High Achiever	% of Students	66.00%	22.00%	12.00%
	Number of High Schools	16	9	6
	Number of Students	33	11	6
Achiever	% of Students	46.77%	26.61%	26.61%
	Number of High Schools	27	23	25
	Number of Students	58	33	33
Graduate	% of Students	28.93%	26.40%	44.67%
	Number of High Schools	28	23	66
	Number of Students	57	52	88
Non-Graduate	% of Students	19.05%	33.33%	47.62%
	Number of High Schools	3	7	10
	Number of Students	4	7	10

The results are clear that Istanbul provides the lion's share of students earning High Achiever and Achiever outcomes, approximately 64% and 47% respectively. This is despite Istanbul having provided 40% of all students in the period of the study.

Students coming from large cities were balanced in their students' frequency of graduate distinctions, but students from Small to Medium Cities made up the largest share of students not earning academic distinction and those not graduating at all.

Additional analysis, reflected in Figure 12, suggests that language proficiency is the key disadvantaging criteria for students from these small and medium-sized cities. Very few of these students are able to bypass the prep program or to complete it in a single semesters, and these are students that are shown to average higher GPA, the key criterion for graduate distinctions.

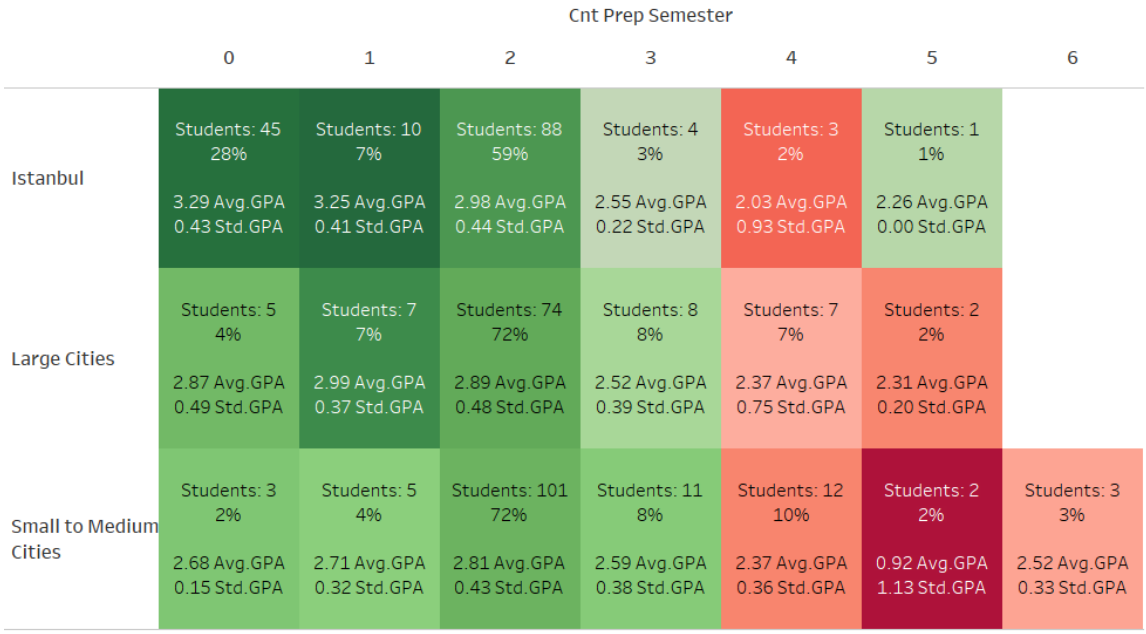

Fig. 12 Count of students by city size and count of semesters in prep program. Color indicates average final GPA.

Further investigation is needed confirm what other factors in addition to language proficiency which might disadvantage students from these small cities. Once

these factors are identified, the department or university would be able to plan interventions or develop additional preparatory curriculum.

Figure 13 is a map of this distribution of students and the ratio of their graduate distinctions. It shows how varied were the sources of students during these years. It also makes clear the extent to which High Achiever, indicated by the orange pie slice, is an outcome limited almost exclusively to students of cities found in Turkish western regions. Like other large countries, different parts of Turkey feature different ethnic and regional cultures. Because of the large amount of internal migration driving Istanbul's growing size, much of this national cultural diversity was likely to be found within the cohort of students coming from Istanbul. Despite this fact, further study is needed to better understand why student outcomes seemed limited for students coming from much of the country during the period of the study.

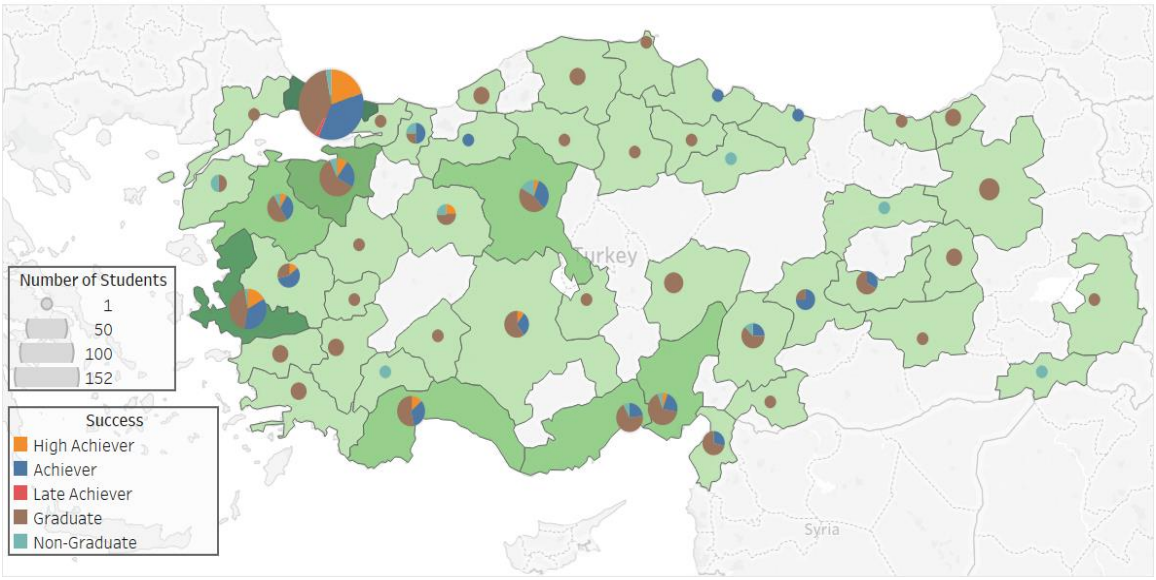

Fig. 13 Students by high school city and ratio of graduate distinctions

4.3.2 Results of high school clustering

As previously mentioned, data indicating the type of high school that a student graduated from was inconsistent and difficult to interpret. To some extent, this is due to changing national education policies. More consistently recorded were the names of the high schools and this provided a basis for identifying high schools, particularly the more prestigious, foreign language high schools in Istanbul that might contribute disproportionately to both the number of students and to graduate distinctions.

High schools were clustered by the number of students they contributed resulting in four groups of high schools: Very Rare, Occasional, Feeder, and Major Feeder. The cluster centers are shown in Table A4. The hypothesis that certain prestigious high schools disproportionately contribute students to the university is supported by looking at the names of the high schools that compose the MajorFeeder and Frequent clusters. Those six high schools are listed in Figure 14, along with their high school type and the number of students from these high schools who achieve the different Success classification. While no available measure of prestige exists to compare the high schools in this list, most of the high schools are well known to educators in Turkey by their names and reputations. Four of the six high schools are located in Istanbul, while the last two are located in the cities of Bursa and Izmir, respectively.

The three Major Feeder high schools stood out for frequencies of their students' graduate distinctions. Perhaps owing to the school's pervasive English curriculum, 78% students coming from Özel Amerikan Robert Lisesi were classified as Achievers or High Achievers. Kadıkoy Anadolu Lisesi had a similarly high percentage of students (71%) earning this distinction. Conversely, only 26% of Bursa Anadolu Lisesi's students

earned these distinctions. Compared to the other high schools in this cluster, only Bornova Anadolu Lisesi comes close to such a high frequency of students graduating from the university program without graduate distinction but with only half as many student cases. Deeper research into the characteristics of these high schools and the experiences of the students coming from them would likely generating valuable insights about these students.

				Success			
Cluster	High School Name	High School Type		High Achiever	Achiever	Graduate	Non-Graduate
Major Feeder	OZEL AMERIKAN ROBERT LISESI	Yabancı Dille Öğretim Yapan Öz	Number of Students	8	10	4	1
			% of Students	34.78%	43.48%	17.39%	4.35%
	BURSA ANADOLU LISESI	Anadolu Lisesi Yabancı Dille Öğretim Yapan Öz	Number of Students	2	3	14	
			% of Students	10.53%	15.79%	73.68%	
	KADIKOY ANADOLU LISESI	Anadolu Lisesi Yabancı Dille Öğretim Yapan Öz	Number of Students	5	7	3	2
			% of Students	29.41%	41.18%	17.65%	11.76%
Feeder	ISTANBUL ERKEK LISESI	Anadolu Lisesi Yabancı Dille Öğretim Yapan Öz	Number of Students	2	4	6	
			% of Students	16.67%	33.33%	50.00%	
	BORNOVA ANADOLU LISESI	Anadolu Lisesi Yabancı Dille Öğretim Yapan Öz	Number of Students	2	2	7	
			% of Students	18.18%	18.18%	63.64%	
	USKUDAR H.AVNI SOZEN ANADOLU LISESI	Anadolu Lisesi Yabancı Dille Öğretim Yapan Öz	Number of Students		5	4	
			% of Students		55.56%	44.44%	

Figure 14 Distribution of graduate distinctions by feeder high schools

4.4 Exploration of foreign exchange data

The final insight produced by data exploration was of possible effects that participation in exchange programs may have on student achievement. Foreign exchange programs represent valuable opportunities for students to experience another culture, including academic culture, but the experience may be disruptive to the student's experience. In this section, information available about students' participation in foreign exchange

programs will be explored. The data available is limited, as only eighty-five of the 392 students in the data set participated in exchange opportunities.

78% of students in the data set did not participate in a foreign exchange program. Of those that did, almost all went to European schools, particularly Germany, the Netherlands, and France. The total numbers of students are too low to draw confident conclusions about possible relationships between exchange country and a student's final GPA. Further research involving a larger data set or augmented with qualitative data may elucidate meaningful insight.

Of greater value to advisers in the department than the country in which a student goes on exchange is when the student goes on exchange. To test whether the timing of the exchange was correlated with any effect on student GPA, several yearly academic GPAs of students participating in exchange programs were placed in Table 6. The students where then separated according to the academic year in which they took their exchange. All but a single transfer student took their foreign exchange opportunities in their third or fourth academic years.

The average GPA of students who participated in an exchange in their third academic year was noticeably lower than both the preceding and succeeding academic years. While earlier observations noted that the third academic year tended to be more challenging for students with regard to grading, the amount of the difference exceeded expectation as a similar drop was not observed in students who took their exchange program in their fourth year. This analysis did not consider whether students went on exchange in the fall or spring semester.

While the data does not present an obvious explanation for this drop in average GPA, courses taken in foreign institutions are credited as PASS/FAIL courses, even if

they are accepted to replace required courses at Boğaziçi University. Because of this, (i) students miss out on the opportunity to contribute graded courses to their GPA since PASS/FAIL courses cannot be calculated into a GPA. Third year exchange students may thus experience the effects of difficult third year courses more strongly.

This observed third year difference would benefit from validation with a larger data set. If confirmed, this represents an issue worthy of further study, if only to provide guidance for advisers of students considering an exchange program.

Table 6. Average Final and Yearly GPAs by Year of Exchange Programs

Exchange Year	Number of Students	Avg. FinalGPA	Avg.2nd YGPA	Avg.3rd YGPA	Avg.4th YGPA
2nd Year	1	3.41	2.52	3.59	
3rd Year	51	3.23	3.29	2.94	3.36
4th year	33	3.06	2.97	3.16	3.21

CHAPTER 5

PREDICTION MODELING

The primary goal of this study is to develop models and evaluate machine learning methods that lead to the accurate prediction of whether students would achieve graduate distinctions on the basis of a student's demographic and early academic performance. The Tableau and R Studio programs were used to prepared the models and apply the machine learning methods. This section will briefly describe the machine learning classification methods used, decision tree, neural network, and multinomial logistic regression, as well as the nature of the two classes of prediction models built upon the insights gleaned from data exploration. Finally, the results of each combination of method and model will be interpreted, evaluated, and compared.

5.1 Description of machine learning methods used

Decision trees are algorithms that work by identifying the variables more useful for classification and building a tree-like series of hierarchical decisions. This method has become particularly popular among researchers because the output is easy to understand and converts naturally to classification rules. For this reason, decision trees feature prominently in the EDM literature. These algorithms operate creating hierarchical nodes and splitting theses nodes according to the IF/THEN conditions applied to the variable of greatest predictive importance. If the algorithm is unconstrained, the number of nodes can become very high and result in perfect classification results on training data, but poor performance on testing or validation data.

Such a result is of little practical value and thus it is important to prune the tree of low importance nodes. The following issues are faced by most decision tree algorithms:

- Balance of tree structure and pruning
- Choosing splitting attributes
- Ordering of splitting attributes
- Number of splits to take
- Stopping criteria

Artificial neural networks (ANN) are algorithms that attempt to simulate the behavior of biologic neurons and are frequently applied for prediction. A group of neurons work in parallel toward an objective while communicating across links. ANNs identify a set of nodes, smaller in number than the number of inputs variables, which exist in one or more of hidden layers. One key manner in which ANNs differ is in whether or not they employer single or multiple hidden layers. This study utilizes the NNET package of R Studio, which is a simple feed-forward algorithm utilizing a single hidden layer, in addition to an input and an output layer. The hidden layers are composed of neurons that combine their inputs and generate an output that is passed on to subsequent layers.

The final method used in this study is multinomial logistic regression, multinomial because the dependent variable, Success, has four possible values: High Achiever, Achiever, Graduate, Non-Graduate. This method calculates the effect of a particular variable on the log-odds of a particular case belonging to the classes of the target variable. Nominal input variables are coded as binary, “dummy” variables, and the results are balanced somewhere in between decision tree and neural networks in terms of how easy they are to interpret.

5.2 Description of the GPA and normalized score models

Models capable of predicting whether or not students are on track to graduate with distinctions in the first two years of a student's studies are the objective and potential contributions of this study. Identifying students who are on track for success or failure is of critical importance to academic advisers as well as departmental stakeholders. The earlier a reasonable prediction can be made, the greater the odds are of a student being able to benefit from assistance and support. The target variable was the previously created Success variable, which represented the new Achiever/High Achiever graduate distinction classification that replaced the CorrectedHonors.

Transfer students were removed from the prediction models because their experiences varied, having entered the program after having already completed some amount of the required course work at another university. This brought the total number of students available for training and testing to 368.

Two distinct model constructs were created, one featuring GPA and course count data and the other featuring averaged normalized scores. These variables were calculated for four of the five course clusters: 1stYearMath, DifficultCommon, Common, and Uncommon. The Rare course cluster was omitted because normalize scores could not be calculated for these courses and such courses very rarely appear in a student's first two years.

The period between fall and spring semesters are intuitive periods when advisers and students are able to come together to determine future course schedules and address issues or concerns. The periods following the end of each of the first two fall and spring semesters, accounting four periods in total, were selected as the most valuable periods

for prediction modeling. These represented the half year, one year, one and a half year, and two year points in each student's degree program. Academic data was aggregated at each of these four periods of time for both model constructs.

In addition to the academic data, the models included the following demographic variables: High School Cluster, High School Cities, Sex. The high school variables had earlier been found to possess correlation with graduate distinctions. The motivation to include the Sex variable was to explore whether it held any importance for these models.

In total, two types of models were built with four sets of aggregated data and applied to three machine learning methods. The result was twenty-four predictive models that were compared by each of these three dimensions.

Because GPA serves as one of the critical criteria for graduate distinction, it was expected that the GPA derived models would outperform the Normalized Score models. However, the results of the exploration of Grading Normalization suggested that normalization might perform at least as well as the GPA model. An equivalent or superior performance would strengthen the case that the normalization of student grades presents a robust alternative to more traditional measures of academic performance.

Each model was trained on 70% of the original data set and tested on the remaining 30%, resulting in 258 and 110 cases respectively. The 70/30 split was maintained the same percentage of graduate distinction classes in both partitions to ensure a viable test set in light of the low frequencies of some of the classes.

5.3 Description of decision tree methods

This study applied the RPART algorithm for its decision tree analysis. The algorithm comes as part of the RPART package for R Studio. The maximum depth of the trees

were preset to five to control for over-fitting and to keep the rule lists as short as possible. The balance between prediction accuracy and rule complexity was an important consideration when constructing these models to maximize the ease of adoption and implementation by educators. The minimum number of cases required for the splitting of a node was 10. The eight models that were trained and tested are briefly summarized below. Confusion matrices for all decision trees are found in Tables C1 through C8 in Appendix C.

Table 7 lists the importance scores of each variable contained in the models. The importance of variables was more evenly distributed in the Normalized Score models compared to the GPA models. GPA models drew more value from students' GPA in the most common group of courses, rather than the most difficult. It is intuitive that students who get high grades in the largest cluster of courses should be more likely to graduate with distinction, but by ignoring other course clusters this model does not appear to be considering students holistically. Comparatively, Normalized Score models seemed to be predicting on a more balanced assessment of students' relative performance across all course clusters.

The best performing models were built with one year of data and successfully predicted graduate distinctions 74% of the time with the GPA model and 73% with the Normalized Score model. Confusion matrices results for both are shown in Tables C2 and C6 in Appendix C. These results and those of the other decision trees reveal that Normalized Score models were more successful at classifying High Achievers than are the GPA models, able to correctly predict ten of twelve cases in most three of the four models, but less successfully able to distinguish between Achiever and Graduate

distinctions. This observation makes sense in light of the incongruence between GPA and Normalized Scores as discussed earlier and shown in Figure 11.

Non-graduates, as the least frequent class, proved to be the most difficult class of students to predict, as none of the models succeeded well. Most models commonly misclassified non-graduates as graduates, along with confusion achievers and high achievers for on another. Intuition suggests that students tend not to differentiate themselves sufficiently enough after only two years to permit more accurate prediction. However, when these classes were simplified and prediction was made binary, the predictive accuracy of these models exceeded 80% in most of the decision trees. Clearly, these models struggled more often with the subtle differences between students on the cusp between two distinctions, such as Achiever and High Achiever, but were capable of making the simpler prediction of whether a student would graduate with or without distinction.

The rule lists created by the one year GPA model decision tree is listed in Table 8. There were a total of twelve terminal nodes as a result of this tree, one each leading to Non-Graduate classification, two to High Achiever, four to Achiever, and the final five leading to Graduate. The first node in this tree, posing the condition with the greatest predictive importance, was whether or not students' GPA in the Common course cluster was at or above 3.0. Because this is the largest group of courses to which almost students registered, it was expected that this course cluster would provide the initial condition. Students who have averaged a BB or better in this cluster of required courses were likely to maintain that and graduate with Achiever or High Achiever distinction. Students who ended their first year of studies with less than a 3.0 GPA from Common cluster courses had very little chance to graduate with distinction. After CommonGPA,

GPA in the 1stYearMath courses provided the next most important conditions for node splitting.

Table 7. Variable Importance for Decision Tree Models

Model Type	GPA Model	Avg	2Year	1.5Year	1Year	0.5Year
GPA	CommonGPA	42	32	39	46	51
	1stYearMathGPA	20	17	21	23	19
	UncommonGPA	14	10	12	15	17
	DifficultCommonGPA	11	21	15	8	0
	CommonCntCourse	4	7	7	1	0
	High School Clusters	2	0	0	2	7
	High School Cities	2	1	0	2	5
	DifficultCommonCntCourse	2	0	6	0	0
	1stYearMathCntCourse	1	0	1	2	
	Sex	0	0	0	1	0
Normalized Score	UCNormPerf	31	20	40	36	26
	CNormPerf	20	16	7	14	42
	1stYearMathNormPerf	26	27	31	26	20
	DCNormPerf	13	29	10	10	2
	High School Cities	5	4	4	7	4
	High School Cluster	5	3	8	3	7
	Sex	1	0	0	5	0

An example prediction is: a student who finishes his or her first year with a GPA at or above 3.0 in their Common cluster courses, 2.6 in 1stYearMath courses, and 3.1 in courses in the Uncommon cluster are 100% likely to graduate as High Achievers. Conversely, finishing the first year with a GPA in Common courses of between 1.8 and 3, while having a GPA in 1stYearMath courses of less than 2.6 means the student is predicted to graduate without distinction with 91% probability.

5.4 Description of neural network

Eight neural networks were created from the two groups of four data sets comprised of GPA and normalized scores respectively. The NNET algorithm contained in the eponymous package for R Studio was used, which is feed-forward algorithm utilizing a single hidden layer. Nine nodes was selected as an appropriate number because it fell in between the number of inputs and outputs, which is a general guideline when making this selection, and for its performance increase over either eight or ten nodes. Softmax modeling was applied along with a decay rate of 0.0005, the latter of which assisted in avoiding over-fitting to the training data. The results of the four neural network models were again tested on the 30% test sample. All prediction results are shown in confusion matrix format in Tables D1 through D8 in Appendix D.

Neither GPA nor Normalized Score models were particularly successful predicting classes of graduate distinction at with one year or less of aggregated data. After a year's worth of data, the GPA derived model demonstrated superior total predictive performance, 65%, as seen in Table D2, and was also less likely to misclassify students as High Achievers or Achievers, with 60% and 67% accuracy when assigning those classes. However, both models were excessively conservative when classifying the Achiever class, misclassifying many of these students as Graduate.

Normalized Score models performed very poorly with the neural network method and one reason was likely that these models possess a relatively small number of variables, less even than the GPA models, and this limited the number of possible connections with which to build a network. Neural networks appear to be the least appropriate method given the variety of nature of variables available for this study.

Table 8. Rule Table for One Year Period of Study - DT with GPA Model

	Rule Definition
Rule 1	Success = "High Achiever", probability = 100%
	WHERE CommonGPA >= 3 AND 1stYearMathGPA >= 1.3 AND UncommonGPA >= 3.1
Rule 2	Success = "High Achiever", probability = 92%
	WHERE CommonGPA >= 3.7
Rule 3	Success = "Achiever", probability = 80%
	WHERE CommonGPA >= 3 AND 1stYearMathGPA < 2.6 AND 1stYearMathGPA >= 1.4
Rule 4	Success = "Achiever", probability = 75%
	WHERE CommonGPA < 3 AND 1stYearMathGPA >= 1.3 AND CommonGPA >= 2.6 AND UncommonGPA >= 2.2
Rule 5	Success = "Achiever", probability = 72%
	WHERE CommonGPA >= 3 AND 1stYearMathGPA >= 2.6 AND CommonGPA < 3.1
Rule 6	Success = "Achiever", probability = 60%
	WHERE CommonGPA >= 3 AND 1stYearMathGPA < 2.6 AND 1stYearMathGPA < 1.4 AND CommonGPA >= 3.1
Rule 7	Success = "Graduate", probability = 91%
	WHERE CommonGPA < 3 AND 1stYearMathGPA < 1.3 AND CommonGPA >= 1.8
Rule 8	Success = "Graduate", probability = 83%
	WHERE CommonGPA >= 3 AND 1stYearMathGPA < 2.6 AND 1stYearMathGPA < 1.4 AND CommonGPA < 3.1
Rule 9	Success = "Graduate", probability = 83%
	WHERE CommonGPA < 3 AND 1stYearMathGPA >= 1.3 AND CommonGPA < 2.6
Rule 10	Success = "Graduate", probability = 79%
	WHERE CommonGPA < 3 AND 1stYearMathGPA < 1.3 AND CommonGPA >= 1.8 AND CommonGPA < 1.5
Rule 11	Success = "Graduate", probability = 71%
	WHERE CommonGPA < 3 AND 1stYearMathGPA >= 1.3 AND CommonGPA >= 2.6 AND UncommonGPA < 2.2
Rule 12	Success = "Non-Graduate", probability = 80%
	WHERE CommonGPA < 3 AND 1stYearMathGPA < 1.3 AND CommonGPA >= 1.8 AND CommonGPA >= 1.5

5.5 Multinomial logistic regression

The NNET package was again used, as it contained the Multinom logistic regression algorithm for multinomial classification. Simple highest probability classification was utilized when predicting case classes.

In these models, the course count variables were removed from the GPA models. This was done after several trials where it was found they offered little value. Their relatively low importance to the decision tree models was another factor in making the decision to remove them from the multinomial logistic regression models. This left a total of thirteen inputs for the Normal Score models and eight for the GPA models.

All prediction results for MLR models are shown in confusion matrix format in Tables E1 through E8 in Appendix E. As with the neural network and decision tree methods, GPA models generally performed better than the Normalized Score model with the multinomial logistic method. Table E2 lists the confusion matric results for the most accurate model with one year of data, which utilized the GPA model and was accurate 72% of all cases. Performance was similar to that of decision tree and better than neural networks. Non-graduate remained difficult to accurately predict; Achiever, while classified correctly 63% of the time, was otherwise most likely to be misclassified as Graduate.

Table 9 below shows the list of coefficients produced by this model. The intercept was the Graduate class of the graduate distinction variable, Success. Additionally, the intercepts included Feeder, Istanbul, and Male values for the High School Cluster, High School Cities, and Sex variables. As with the decision tree and neural network methods, variables associated with student performance in the Common course cluster was a significant indicator of Achiever and High Achiever classification.

The one year aggregation with GPA model was found to be possess the best combination of predictive accuracy and level of aggregation. In this model, an increase of one in the average GPA for Common courses led to an increase of 2.7 and 4.7 in the log likelihood of graduating as an Achiever or High Achiever. GPAs in 1stYearMath courses were also found to be significant predicators of all three classes, but with especially strong increase of 4.02 in log-odds of High Achiever classification.

Table 9. Coefficient List for One Year Period of Study - MLR with GPA Model

	Dependent Variable:		
	High Achiever	Achiever	Non-Graduate
High.School.ClusterMajor Feeder	0.483	2.372	14.519***
	(0.954)	(1.939)	(0.642)
High.School.ClusterOccassional	0.462	-0.714	13.719***
	(0.905)	(1.883)	(0.532)
High.School.ClusterVery Rare	1.051	1.48	13.583***
	(0.859)	(1.859)	(0.558)
High.School.CitiesLarge Cities	-0.718	-1.086	0.19
	(0.534)	(0.979)	(0.846)
High.School.CitiesSmall to Medium Cities	-0.681	-1.746	-0.084
	(0.529)	(1.240)	(0.867)
SexK	-0.015	-1.45	-0.204
	(0.428)	(0.907)	(0.658)
1stYearMathGPA	1.465***	4.020***	-1.486**
	(0.345)	(0.844)	(0.672)
DifficultCommonGPA	0.151	1.182**	-0.048
	(0.191)	(0.536(	(0.303)
CommonGPA	2.697***	4.733***	-0.598
	(0.583)	(1.444)	(0.588)
UncommonGPA	0.216	1.763***	0.207
	(0.213)	(0.563)	(0.401)
RareGPA	0.052	-0.216	0.004
	(0.129)	(0.255)	(0.224)
Constant	-11.424***	-31.922***	-13.750***
	(1.816)	(5.964)	(0.796)
Akaike Inf. Crit.	360.255	360.255	360.255
Note:	*p<0.1; ** p<0.05; ***p<0.01		

However, unlike decision tree or neural networks, high school cluster was found to be an important variable. For logistic regression, the multinomial classes of high school cluster were recoded as dummy variables with values of one or zero. The same was also done for sex and high school city, though these variables were not found to have significant influence. Of the twenty non-transfer students who failed to graduate, none came from Feeder high schools. This meant that whether or not a student came from the other three high school clusters, Major Feeder, Occasional, and Very Rare, was significant. Being a graduate from a high school in these clusters represented an increase of 13.5 to 14.5 in the log-odds of not graduate at all. The full list of variables by their importance is found in Table E9.

5.6 Evaluation of models

In summary, the models tested reflected three machine learning methods and two model constructs with four different levels of academic data aggregation for a total of twenty-four trained and tested models. Of those twenty-four, the linear regression and decision tree methods performed best, in terms of total prediction accuracy and earliest deployment, with the GPA model and one year of aggregated data. None of the models found good results with neural networks, suggested that the amount and variety of data available was insufficient to gain the best value from this method. Because GPA is a key criterion of graduate distinction, it was expected that these models would perform better than Normalized Scores. However, it was observed that the difference was not so great and that further development of normalized score models may result in better performance.

5.7 Model deployment

The deployment of any of these models depends on their ability to be adopted easily into the existing decision making processes of the university and its management department. Presently, academic advisers and other decision makers make a large number of decisions at the beginning of each semester, such as deciding on student requests and decided which at-risk students toward which to devote their limited resources. Students may request to overload or to underload their schedules or to add a minor field of study to their degree. Identifying whether a student is on track to graduate with distinction could contribute to the decision to approve the student's request.

As the models' prediction accuracies are less than perfect, they are not suitable, by themselves, to replace intuition and personal experience. However, some of the models offered predictive potential sufficient after one year to provide advisers and educators with an objective tool for quickly assessing a cohort and predicting students' future performance. The models are deployable as semesterly reports, or could be expanded to include on-going course grades and provide real-time assessments to advisers. Student classification results could even be deployed to prioritize or pre-review student petitions according to the student's predicted graduate distinction.

CHAPTER 6

CONCLUSION

In this study, a process of prediction modeling was developed and tested with a sample of student data provided by Boğaziçi University. Decision tree, artificial neural network, and multinomial logistic regression methods were tested on a series of eight models representing different points in time in a student's academic career, and the accuracy of their predictions were compared. Multinomial logistic regression and decision tree were found to predict student's membership in one of the four classes of graduate distinction with the highest degree of total accuracy, exceeding 70% accuracy with one year worth of academic data and a GPA model. When classification was limited to two classes of either achiever or non-achiever, both decision tree and multinomial logistic regression were able to predict with better than 80% accuracy using the GPA model and one year worth of data.

Models built using GPA data was found to perform better than models utilizing normalized scores. However, the gap in performance between the two types of models with the decision tree and multinomial logistic regression methods was often narrow and future research may consider ways to develop the normalized score models further. Because GPA is a criterion of the predicted target, graduate distinction, it was mildly surprising that normalized score models performed as well as they did, comparatively, despite possessing fewer variables.

In addition to the prediction models, this study made two other contributions in the form of a new classification system for graduate distinction and a normalized score assessment for students. Both are presented as improvements upon current policies,

specifically the university's Honors/High Honors distinctions and GPA-based assessments of student academic performance, that address inconsistencies or bias therein.

The results of this study confirm that CRISP-DM is an effective methodology for building prediction models from education data as well as for uncovering meaningful and actionable insights about student experiences. A number of insights about the university's undergraduate management program and subjects for future research were uncovered during the course the study.

Finally, this study also confirmed the importance of prioritizing the collection, management, and utilization of student data. Information systems work on the GIGO principle: Garbage In, Garbage Out. Several variables were rendered useless for analysis purposes or required extensive correction and development. More and better questions from departmental stakeholders will lead to more efficient and higher value knowledge collection, leading to higher value knowledge discovery.

While some of the insights born from this study may not lead to immediately actionable conclusions, each insight drawn feeds back into the KDD cycle. The practical value of applying the KDD methodologies, such as CRISP-DM, lies in both the outputs and the cycle of iterative improvements. Organizations of all types, such as Boğaziçi University and its Department of Management, reap the greatest value from their data by incorporating knowledge discovery as deeply into their decision making culture as possible, ensuring continuous cycles of improvement.

Future research may further develop a normalized score metric as an alternative to GPA as the key academic performance metric. Early steps toward that goal were taken in the course of this study, but successful development and deployment would

depend upon working with department or university stakeholders to establish a consensus conceptualization of academic success. The consequences of emphasizing a new metric over GPA could be disruptive and similar efforts undertaken in past years at universities met with strong resistance, particular from less-quantitative disciplines. Similar resistance might be expected at any university with strong traditions and empowered faculties. GPA, as currently implemented, is very much a subjective measure of a student's experience, ignoring the performance of other students and the effects of course selection. Peer-comparative measures may provide universities with a more internally valuable tool for fairly assessing students.

Another possible area of follow-up research is to look at undergraduates who proceed to enroll in graduate programs at the same university. The ready availability of historical data on these students would be expected to provide a fertile field in which to mine insights. A starting point might be a replication of Zimmermann, Brodersen, Heinimann, and Buhmann's (2015) study on the relationship of undergraduate and graduate student performance.

However, the most obvious focus of future attention is the performance of students in the English language preparatory program. It was unfortunate that direct measures of language proficiency were not available for this research, as it is essential to answers questions about the effect of student preparedness on academic performance. This study attempted to infer language proficiency indirectly from the number of semesters spent in the preparatory program and the type of a student's high school, but this was a less than ideal surrogate measure. Confirming a negative or neutral effect of language proficiency on student performance would likely prove valuable to developers of course and program curricula.

APPENDIX A

DESCRIPTIVE DATA TABLES

Table A1. List of data variable received from university

	Variable Name	Description	Variable Type
Demographic	StudentID	Encrypted Student ID	Number
	Nationality	Student's Registered Nationality	String
	Sex	Male or Female	String
	HighSchoolName	Name of Student's Graduating High School	String
	HighSchoolType	The Category of High School	String
	OSYMYear	Year the Sat for University Qualification Exams	Number
	OYSOOBP	Qualification Exam Score	Number
	YerlestirmeSira	Qualification Exam Ranking	Number
	Graduated	Whether Student Has Graduated	String
	FinalGPA	GPA at time of Latest Calculation	Number
	Honors	If Graduated, Did Student Recent Honors with Degree	String
	UniversityEntrance	Semester Student Entered the University	String
	DepartmentEntrance	Semester Student Entered Department, Usually after Language Preparation Courses	String
Transcript	StudentID	Encrypted Student ID	String
	Semester	Academic Year and Semester	String
	RegisteredCredits	Accumlated Credits from Courses Registered	Number
	EarnedCredits	Accumulated Credits Earned from Courses Registered	Number
	GPA	Grade Point Average	Number
	SPA	Semester Point Average	Number
	SemesterRegisteredCredits	Credits Registered in Semester	Number
	SemesterEarnedCredits	Credits Earned in Semester	Number
	SemesterTotalPoints	Course Grade * Credits Earned that Semester	Number
	TotalPoints	Accumulated Credits Earned * Course Grade	Number
	SemesterNumber	Number Semester in Student's Career	Number
	SemesterStatus	Whether the student is in probation or other type of special semester situation	String
Course	StudentID	Encrypted Student ID	String
	Semester	Academic Year and Semester	String
	CourseCode	University's Course Identifier	String
	CourseSection	Identifier for Courses with Multiple Sessions in Semester	Number
	RegistrationType	Whether Course is Normal or a Repeat or for No Credit	String
	LetterGrade	Letter Grade Awarded for Course	W, P
	ReplacedCourse	Course Code of Replaced Course	String
	CourseCredits	Number of Course Credits Awarded	Number
Exchange	StudentID	Encrypted Student ID	String
	Semester	Academic Year and Semester	String
	UniversityName	Name of University Where Student Participated in Exchange Program	String
	Country	Country of the Exchange University	String

Table A2. List of variables included in course clustering.

Variable Name	Description	Variable Type
CourseCode	The course's alphanumeric code	String
CntStudents	The number of students registered to this course	Numeric
AvgNumGrade	The mean average of numerical grades earned by students in the course	Numeric
StdNumGrade	The standard deviation of numerical grades earned by students in the course	Numeric
CntDs	The count of all DCs and DDcearned by students in the course	Numeric
CntFs	The count of all Fs earned by students in the course	Numeric
CntRepeats	The number of times students enrolled in the course after the first time	Numeric

Table A3. Student clustering results and the centers of each cluster

	Student Clusters					
	1	**2**	**3**	**4**	**5**	**6**
Number of Students	175	161	26	19	10	1
Percent of Total Students	44.64%	41.07%	6.63%	4.85%	2.55%	0.26%
Variable	**Cluster Centers**					
FinalGPA	3.30	2.63	2.25	**3.41**	1.34	3.95
Number of Fall & Spring Semesters	8.0	8.4	13.3	**8.2**	6.1	10.0
Percent Graduated	100%	99%	**62%**	100%	**0%**	100%
Percent Transfered	6%	6%	0%	21%	10%	0%
Semesters in Preparatory Program	1.6	2.1	2.8	1.3	2.1	2.0
Percent Foreign Exchange	36%	11%	0%	42%	0%	0%
Percent Male	37%	62%	**73%**	**32%**	**80%**	100%
1stYRegisteredCredits	42.5	44.1	43.7	51.1	38.7	44.0
2ndYRegisteredCredits	43.4	43.8	39.9	44.8	27.0	49.0
3rdYRegisteredCredits	33.3	41.9	42.0	33.2	11.1	55.0
4thYRegisteredCredits	26.8	34.1	43.2	29.4	7.1	41.0
5thYRegisteredCredits	.6	8.0	**44.6**	7.8	6.5	27.0
6+YRegisteredCredits	.0	.6	**61.2**	0.0	6.2	0.0
SummerCredits	7.9	**16.4**	**34.1**	10.7	7.4	**31.0**
Points-1stYearMath	19.3	15.9	12.5	22.4	5.6	32.0
Points-DifficultCommon	50.6	36.5	33.3	54.7	6.0	70.5
Points-Common	233.9	186.0	161.0	232.5	39.6	243.5
Points-Uncommon	78.1	63.4	42.0	63.2	6.7	90.0
Points-Rare	92.5	79.6	88.2	**214.1**	8.2	450.5
Repeats-1stYearMath	.3	2.2	9.6	.7	3.9	1.0
Repeats-DifficultCommon	.4	2.6	11.8	.5	2.4	0.0
Repeats-Common	.4	2.0	11.5	.4	7.7	2.0
Repeats-Uncommon	.2	.9	3.0	0.0	.7	0.0
Repeats-Rare	.2	1.1	4.3	.2	1.1	0.0

Table A4. Results of High School Clustering by Number of Students.

Cluster	Average Students
Major Feeder	19.67
Feeder	10.67
Occassional	4.91
Very Rare	1.08

APPENDIX B

EXPLORATORY FIGURES

Course Cluster	Course Code	GPA	Std.Dev GPA	CntStudents	CntDs	CntFs	Avg. Final GPAs
1stYearMath	MATH101	1.22	1.2	383	124	305	2.65
	MATH102	1.02	1.2	382	132	413	2.68
Difficult Common	AD 131	1.98	1.4	388	82	137	2.74
	AD 311	1.80	1.4	379	77	164	2.79
	AD 351	1.78	1.3	373	122	104	2.81
	EC 205	1.77	1.3	384	106	146	2.79
	AD 353	1.60	1.3	382	181	133	2.80
	EC 203	1.58	1.3	386	126	173	2.79
Common	AD 232	3.62	0.7	346	10	4	2.95
	AD 231	3.60	0.8	386	12	8	2.90
	HTR 312	3.38	0.9	376	18	8	2.91
	HTR 311	3.31	1.0	375	14	18	2.89
	TK 221	3.26	1.1	374	34	15	2.89
	TK 222	3.18	1.1	374	30	19	2.89
	AD 341	2.95	0.9	315	19	9	2.92
	AD 104	2.92	1.2	390	44	34	2.84
	AD 150	2.90	1.1	379	15	33	2.84
	AD 202	2.89	1.0	333	28	17	2.94
	AD 401	2.84	1.0	378	51	8	2.92
	AD 312	2.79	1.0	377	51	14	2.88
	AD 220	2.77	1.2	384	39	40	2.86
	AD 452	2.72	1.1	368	77	12	2.90
	AD 251	2.71	1.0	295	34	15	2.84
	AD 213	2.68	1.2	385	41	38	2.86
	POLS101	2.68	1.0	389	40	26	2.83
	AD 320	2.66	1.1	371	49	26	2.89
	AD 214	2.66	1.3	383	72	36	2.90
	PSY 101	2.51	1.2	389	68	45	2.77
	AD 252	2.48	1.4	301	30	69	2.77
	EC 102	2.39	1.2	388	86	49	2.79
	EC 101	2.32	1.2	388	83	58	2.79
	SOC 101	2.26	1.0	388	80	32	2.82
	AD 408	2.22	1.2	372	83	44	2.90
Uncommon	AD 426	3.68	0.4	110	0	0	3.12
	AD 489	3.63	0.4	121	0	0	2.75
	AD 488	3.57	0.6	237	2	1	2.96
	AD 497	3.41	0.7	148	2	3	2.83
	AD 480	3.36	0.8	206	7	4	2.83
	AD 477	2.92	1.2	237	30	18	2.84
	AD 442	2.91	0.9	204	9	6	2.80
	AD 413	2.76	1.0	121	12	4	2.85
	AD 403	2.71	1.1	255	22	16	2.94
	AD 316	2.69	1.2	272	25	33	2.92
	HUM 102	2.46	1.2	154	31	15	2.86
	EC 351	2.45	1.2	245	62	15	2.96
	HIST106	1.99	1.3	235	58	55	2.86
	HUM 101	1.75	1.3	177	58	43	2.82
	HIST105	1.74	1.3	141	45	40	2.82

Fig. B1 Academic variables, including average final GPA of registered students

Courses with more than 140 students registered students.

CourseCluster	CourseCode	Season	AA		BA		BB		CB		CC		DC		DD		F	
			#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
1stYearMath	MATH101	Fall	5	1.21%	13	3.15%	30	7.26%	46	11.14%	73	17.68%	38	9.20%	36	8.72%	172	41.65%
		Spring	2	0.87%	13	5.68%	14	6.11%	35	15.28%	30	13.10%	17	7.42%	13	5.68%	105	45.85%
		Summer			3	4.48%	5	7.46%	5	7.46%	6	8.96%	8	11.94%	12	17.91%	28	41.79%
	MATH102	Fall	6	2.40%	5	2.00%	11	4.40%	27	10.80%	17	6.80%	15	6.00%	19	7.60%	150	60.00%
		Spring	4	1.05%	12	3.14%	18	4.71%	42	10.99%	61	15.97%	36	9.42%	34	8.90%	175	45.81%
		Summer	3	1.80%	11	6.59%	7	4.19%	13	7.78%	17	10.18%	9	5.39%	19	11.38%	88	52.69%
Difficult Common	AD 131	Fall	37	30.33%	8	6.56%	12	9.84%	11	9.02%	5	4.10%	9	7.38%	9	7.38%	31	25.41%
		Spring	41	9.93%	39	9.44%	45	10.90%	58	14.04%	60	14.53%	44	10.65%	20	4.84%	106	25.67%
	EC 203	Fall	19	4.81%	28	7.09%	42	10.63%	34	8.61%	64	16.20%	48	12.15%	36	9.11%	124	31.39%
		Summer	10	6.06%	11	6.67%	19	11.52%	16	9.70%	18	10.91%	23	13.94%	19	11.52%	49	29.70%
	EC 205	Fall	2	2.90%	4	5.80%	3	4.35%	7	10.14%	4	5.80%	5	7.25%	6	8.70%	38	55.07%
		Spring	25	6.43%	27	6.94%	59	15.17%	52	13.37%	53	13.62%	53	13.62%	30	7.71%	90	23.14%
		Summer	10	13.16%	7	9.21%	19	25.00%	7	9.21%	3	3.95%	7	9.21%	5	6.58%	18	23.68%
	AD 353	Fall	31	11.83%	18	6.87%	20	7.63%	11	4.20%	18	6.87%	34	12.98%	60	22.90%	70	26.72%
		Spring	21	8.50%	15	6.07%	17	6.88%	18	7.29%	26	10.53%	44	17.81%	43	17.41%	63	25.51%
	AD 311	Fall	27	7.38%	42	11.48%	47	12.84%	64	17.49%	32	8.74%	27	7.38%	16	4.37%	111	30.33%
		Spring	13	7.39%	11	6.25%	17	9.66%	26	14.77%	22	12.50%	24	13.64%	10	5.68%	53	30.11%
	AD 351	Fall	13	4.92%	8	3.03%	26	9.85%	29	10.98%	44	16.67%	40	15.15%	40	15.15%	64	24.24%
		Spring	23	11.06%	19	9.13%	28	13.46%	28	13.46%	28	13.46%	18	8.65%	24	11.54%	40	19.23%
		Summer	2	100.00%														

Fig. B2 Seasonal grading for the most difficult course clusters

APPENDIX C

DECISION TREE

Table C1. Half Year Period of Study - DT with GPA Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**2**	10	0	0	12	*17%*	*83%*
Achiever	1	**26**	8	0	35	*74%*	
Graduate	0	12	**45**	0	57	*79%*	*79%*
Non-Graduate	0	1	3	**2**	6	*33%*	
Prediction Accuracy	*67%*	*53%*	*80%*	*100%*	*110*	*68%*	*81%*

Table C2. One Year Period of Study - DT with GPA Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**8**	4	0	0	12	*67%*	*83%*
Achiever	4	**23**	8	0	35	*66%*	
Graduate	0	6	**50**	1	57	*88%*	*90%*
Non-Graduate	0	0	6	**0**	6	*0%*	
Prediction Accuracy	*67%*	*70%*	*78%*	*0%*	*110*	*74%*	*87%*

Table C3. One and a Half Year Period of Study - DT with GPA Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**6**	5	1	0	12	*50%*	*77%*
Achiever	4	**21**	10	0	35	*60%*	
Graduate	0	7	**50**	0	57	*88%*	*89%*
Non-Graduate	0	0	6	**0**	6	*0%*	
Prediction Accuracy	*60%*	*64%*	*75%*	*NA*	*110*	*70%*	*84%*

Table C4. Two Year Period of Study - DT with GPA Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**8**	4	0	0	12	*67%*	*83%*
Achiever	5	**22**	8	0	35	*63%*	
Graduate	0	7	**50**	0	57	*88%*	*89%*
Non-Graduate	0	0	6	**0**	6	*0%*	
Prediction Accuracy	*62%*	*67%*	*78%*	*NA*	*110*	*73%*	*86%*

Table C5. Half Year Period of Study - DT with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**8**	3	1	0	12	*67%*	*77%*
Achiever	13	**12**	10	0	35	*34%*	
Graduate	4	8	**44**	1	57	*77%*	*81%*
Non-Graduate	0	0	4	**2**	6	*33%*	
Prediction Accuracy	*32%*	*52%*	*75%*	*67%*	*110*	*60%*	*79%*

Table C6. One Year Period of Study - DT with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**10**	1	1	0	12	*83%*	*74%*
Achiever	3	**21**	11	0	35	*60%*	
Graduate	0	8	**49**	0	57	*86%*	*87%*
Non-Graduate	0	0	6	**0**	6	*0%*	
Prediction Accuracy	*77%*	*70%*	*73%*	*NA*	*110*	*73%*	*82%*

Table C7. One and a Half Year Period of Study - DT with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**10**	1	0	1	12	*83%*	*60%*
Achiever	5	**12**	18	0	35	*34%*	
Graduate	0	9	**48**	0	57	*84%*	*86%*
Non-Graduate	0	0	3	**3**	6	*50%*	
Prediction Accuracy	*67%*	*55%*	*70%*	*75%*	*110*	*66%*	*75%*

Table C8. Two Year Period of Study - DT with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**10**	1	1	0	12	*83%*	*79%*
Achiever	3	**23**	9	0	35	*66%*	
Graduate	1	10	**46**	0	57	*81%*	*83%*
Non-Graduate	0	0	6	**0**	6	*0%*	
Prediction Accuracy	*71%*	*68%*	*74%*	*NA*	*110*	*72%*	*81%*

APPENDIX D

NEURAL NETWORK

Table D1. Half Year Period of Study - NN with GPA Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**6**	6	0	0	12	50%	68%
Achiever	3	**17**	13	2	35	49%	
Graduate	2	10	**40**	5	57	70%	79%
Non-Graduate	0	1	3	**2**	6	33%	
Prediction Accuracy	55%	50%	71%	22%	110	59%	75%

Table D2. One Year Period of Study - NN with GPA Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**7**	4	1	0	12	58%	79%
Achiever	5	**21**	8	1	35	60%	
Graduate	0	13	**41**	3	57	72%	79%
Non-Graduate	0	0	4	**2**	6	33%	
Prediction Accuracy	58%	55%	76%	33%	110	65%	79%

Table D3. One and a Half Year Period of Study - NN with GPA Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**11**	1	0	0	12	92%	70%
Achiever	5	**16**	14	0	35	46%	
Graduate	0	5	**47**	5	57	82%	90%
Non-Graduate	1	0	4	**1**	6	17%	
Prediction Accuracy	65%	73%	72%	17%	110	68%	82%

Table D4. Two YearPeriod of Study - NN with GPA Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**8**	4	0	0	12	67%	83%
Achiever	6	**21**	8	0	35	60%	
Graduate	0	10	**44**	3	57	77%	81%
Non-Graduate	1	1	3	**1**	6	17%	
Prediction Accuracy	53%	58%	80%	25%	110	67%	82%

Table D5. Half Year Period of Study - NN with Normal Score Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**5**	5	2	0	12	42%	72%
Achiever	5	**19**	10	1	35	54%	
Graduate	2	13	**35**	7	57	61%	73%
Non-Graduate	0	2	4	**0**	6	0%	
Prediction Accuracy	42%	49%	69%	0%	110	54%	73%

Table D6. One Year Period of Study - NN with Normal Score Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**10**	1	1	0	12	83%	77%
Achiever	9	**16**	9	1	35	46%	
Graduate	0	14	**38**	5	57	67%	78%
Non-Graduate	0	0	4	**2**	6	33%	
Prediction Accuracy	53%	52%	73%	25%	110	60%	77%

Table D7. One and a Half Year Period of Study - NN with Normal Score Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**9**	2	0	1	12	75%	74%
Achiever	5	**19**	11	0	35	54%	
Graduate	0	18	**35**	4	57	61%	71%
Non-Graduate	0	0	4	**2**	6	33%	
Prediction Accuracy	64%	49%	70%	29%	110	59%	73%

Table D8. Two Year Period of Study - NN with Normal Score Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**9**	2	1	0	12	75%	72%
Achiever	7	**16**	12	0	35	46%	
Graduate	1	19	**31**	6	57	54%	68%
Non-Graduate	0	0	3	**3**	6	50%	
Prediction Accuracy	53%	43%	66%	33%	110	54%	70%

APPENDIX E

MULTINOMIAL LOGISTIC REGRESSION

Table E1. Half Year Period of Study - MLR with GPA Model

	Prediction					Correct Classification of Cases	Binary Prediction Results: A/HA vs. G/NG
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	5	7	0	0	12	42%	74%
Achiever	2	21	12	0	35	60%	
Graduate	1	7	47	2	57	82%	86%
Non-Graduate	0	1	4	1	6	17%	
Prediction Accuracy	63%	58%	75%	33%	110	67%	81%

Table E2. One Year Period of Study - MLR with GPA Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**8**	4	0	0	12	*67%*	*79%*
Achiever	3	**22**	10	0	35	*63%*	
Graduate	1	7	**48**	1	57	*84%*	*87%*
Non-Graduate	0	0	5	**1**	6	*17%*	
Prediction Accuracy	*67%*	*67%*	*76%*	*50%*	*110*	*72%*	*84%*

Table E3. One and a Half Year Period of Study - MLR with GPA Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**7**	5	0	0	12	*58%*	*79%*
Achiever	1	**24**	10	0	35	*69%*	
Graduate	0	9	**47**	1	57	*82%*	*86%*
Non-Graduate	0	0	5	**1**	6	*17%*	
Prediction Accuracy	*88%*	*63%*	*76%*	*50%*	*110*	*72%*	*83%*

Table E4. Two YearPeriod of Study - MLR with GPA Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**10**	2	0	0	12	*83%*	*81%*
Achiever	4	**22**	9	0	35	*63%*	
Graduate	0	7	**48**	2	57	*84%*	*89%*
Non-Graduate	0	0	4	**2**	6	*33%*	
Prediction Accuracy	*71%*	*71%*	*79%*	*50%*	*110*	*75%*	*85%*

Table E5. Half Year Period of Study - MLR with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**3**	8	1	0	12	*25%*	*68%*
Achiever	5	**16**	14	0	35	*46%*	
Graduate	1	9	**47**	0	57	*82%*	*84%*
Non-Graduate	0	0	6	**0**	6	*0%*	
Prediction Accuracy	*33%*	*48%*	*69%*	*NA*	*110*	*60%*	*77%*

Table E6. One Year Period of Study - MLR with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**10**	1	1	0	12	*83%*	*77%*
Achiever	6	**19**	10	0	35	*54%*	
Graduate	0	10	**46**	1	57	*81%*	*84%*
Non-Graduate	0	0	4	**2**	6	*33%*	
Prediction Accuracy	*63%*	*63%*	*75%*	*67%*	*110*	*70%*	*81%*

Table E7. One and a Half Year Period of Study - MLR with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**10**	2	0	0	12	*83%*	*79%*
Achiever	3	**22**	10	0	35	*63%*	
Graduate	0	10	**46**	1	57	*81%*	*84%*
Non-Graduate	0	0	3	**3**	6	*50%*	
Prediction Accuracy	*77%*	*65%*	*78%*	*75%*	*110*	*74%*	*82%*

Table E8. Two Year Period of Study - MLR with Normal Score Model

	Prediction					*Correct Classification of Cases*	*Binary Prediction Results: A/HA vs. G/NG*
	High Achiever	Achiever	Graduate	Non-Graduate	Students		
High Achiever	**9**	3	0	0	12	*75%*	*74%*
Achiever	4	**19**	12	0	35	*54%*	
Graduate	1	10	**45**	1	57	*79%*	*83%*
Non-Graduate	0	0	3	**3**	6	*50%*	
Prediction Accuracy	*64%*	*59%*	*75%*	*75%*	*110*	*69%*	*79%*

Table E9. Variable Importance for Scores for Multinomial Logistic Regression

Model Type	GPA Model	Normalized	Avg	2Year	1.5Year	1Year	0.5Year
GPA	High.School.ClusterMajor Feeder	100%	16	14.18	16.28	16.6	16.28
	High.School.ClusterVery Rare	89%	14	12.6	14.86	14.21	14.86
	CommonGPA	87%	14	19.04	14.03	8.03	14.03
	High.School.ClusterOccassional	87%	14	14.45	12.63	15.1	12.63
	1stYearMathGPA	42%	7	5.34	7.27	6.97	7.27
	DifficultCommonGPA	16%	3	4.33	2.22	1.38	2.22
	High.School.CitiesSmall to Medium Cities	16%	2	2.75	2.42	2.34	2.42
	UncommonGPA	15%	2	2.2	2.81	1.77	2.81
	High.School.CitiesLarge Cities	13%	2	2.86	1.81	1.99	1.81
	SexK	11%	2	1.12	2.16	1.64	2.16
	RareGPA	6%	1	2.42	0.58	0.37	0.58
Normal Score Model	High.School.ClusterMajor Feeder	100%	16	14.86	16.42	16.45	16
	High.School.ClusterOccassional	94%	15	14.3	15.5	16.27	13.78
	High.School.ClusterVery Rare	92%	15	13.69	15.04	15.37	14.57
	CommonNormPerf	79%	13	18.92	11.42	11.45	8.83
	1stYearMathNormPerf	48%	8	7.44	9.57	8.18	5.67
	DifficultCommonNormPerf	31%	5	6.9	5.97	3.01	3.7
	UncommonNormPerf	28%	4	4.27	7.06	4.83	1.67
	High.School.CitiesSmall to Medium Cities	20%	3	4.65	4.7	2.48	1.04
	SexK	13%	2	1.41	3.99	1.72	0.92
	High.School.CitiesLarge Cities	10%	2	2.44	2.05	1.24	0.83

REFERENCES

Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting students' performance using ID3 and C4.5 classification algorithms. *International Journal of Data Mining & Knowledge Management Process*, *3*(5), 39-52. doi: 10.5121/ijdkp.2013.3504

Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data Mining: A prediction for Student's Performance Using Classification Method. *World Journal of Computer Application and Technology*, *2*(2), 43-47. doi: 10.13189/wjcat.2014.020203

Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006, December). Mining student data using decision trees. Paper presented at the *International Arab Conference on Information Technology (ACIT'2006)*, Irbird, Jordan. Retrieved from http://acit2k.org/ACIT/

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*(2017), 177-194. doi: 10.1016/j.compedu.2017.05.007

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2), 17. doi: 10.20429/ijsotl.2010.040217

Baker, R.S.J.d. (2010). Data Mining for Education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International Encyclopedia of Education 3rd edition*, (pp. 112-118). Oxford, UK: Elsevier.

Baker, R.S.J.d., & de Carvalho, A. M. J. A. (2008). Labeling Student Behavior Faster and More Precisely with Text Replays. In Baker, R.S.J.d., Barnes, T., & Beck, J.E. (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 38-47). Montreal, Canada. Retrieved from http://www.educationaldatamining.org/

Baker, R.S.J.d., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining, 1* (1), 3-17. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM

Baradwaj, B. K., & Pal, S. (2011). Mining Educational Data to Analyze Students' Performance. *International Journal of Advanced Computer Science and Applications (IJACSA), 2*(6), 63-69. Retrieved from http://thesai.org/Publications/IJACSA

Beck, J., & Woolf, B. P. (2000, June). High-Level Student Modeling with Machine Learning. In Gauthier, G., Frasson, C., & VanLehn, K., (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems* (pp. 584-593). Montréal, Canada: Springer-Verlag Berlin Heidelberg.

Brachman, R. J., & Anand, T. (1996). The process of knowledge discovery in databases. In U. Fayyad, G. Piatetsky-Shapirao, P. Smyth, & R. Uthurusamy (Eds.), A*dvances in knowledge discovery and data mining* (pp. 37-57). Menlo Park, CA: AAAI Press.

Campagni, R., Merlini, D., Sprugnoli, R., & Verri, M. C. (2015). Data mining models for student careers. *Expert Systems with Applications*, *42*(13), 5508-5521. doi: 10.1016/j.eswa.2015.02.052

Campbell, J. P., & Oblinger, D. G. (2007). Academic analytics. *EDUCAUSE review*, *42*(4), 40-57. Retrieved from https://www.educause.edu/

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In L.C. Jain, R. A. Tedman & D. K. Tedman (Eds.), *Evolution of teaching and learning paradigms in intelligent environment* (pp. 183–221). New York: Springe-Verlag.

Chemers, M. M., Hu, L. T., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational psychology*, *93*(1), 55-64. doi: 10.1037/0022-0663.93.1.55

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In A. Brito, & J. Teixeira (Eds.), *Proceedings of 5th Annual Future Business Technology Conference* (pp. 5-12). Porto, Portugal: EUROSIS.

Delavari, N., Phon-amnuaisuk, S., & Beikzadeh, M. (2008). Data Mining Application in Higher Learning Institutions. *Informatics in Education*, *7*(2), 31–54. Retrieved from https://www.mii.lt/informatics_in_education/

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506. doi: 10.1016/j.dss.2010.06.003

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, *5*, 15991-16005. doi: 10.1109/ACCESS.2017.2654247

Elias, T. (2011). *Learning Analytics: Definitions, Processes and Potential*. Unpublished Manuscript. Retrieved from http://www.learninganalytics.net

Erdoğan, Ş. Z., & Timor, M. (2005). A data mining application in a student database. *Journal of aeronautics and space technologies*, *2*(2), 53-57. Retrieved from http://www.springer.com/gp/

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-54. Retrieved from http://www.aaai.org/Magazine/magazine.php

Felton, J., & Koper, P. T. (2005). Nominal GPA and real GPA: a simple adjustment that compensates for grade inflation. *Assessment & Evaluation in Higher Education*, *30*(6), 561-569. doi: 10.1080/02602930500260571

García, E., Romero, C., Ventura, S., & De Castro, C. (2011). A collaborative educational association rule mining tool. *The Internet and Higher Education*, *14*(2), 77-88. doi: 10.1016/j.iheduc.2010.07.006

Golding, P., & Donaldson, O. (2006, October). Predicting academic performance. In *Frontiers in education conference, 36th Annual ASEE/IEEE Frontiers in Education Conference.* (pp. 21-26). San Diego, CA: IEEE. doi: 10.1109/FIE.2006.322661

Goldstein, P. J., & Katz, R. N. (2005). *Academic analytics: The uses of management information and technology in higher education* (Vol. 8). Boulder, CO: EDUCAUSE Center for Applied Research

Goyal, M., & Vohra, R. (2012). Applications of data mining in higher education. *International Journal of Computer Science Issues*, *9*(2), 113. Retrieved from http://www.ijcsi.org/

Guruler, H., Istanbullu, A., & Karahasan, M. (2010). A new student performance analysing system using knowledge discovery in higher educational databases. *Computers & Education*, *55*(1), 247-254. doi: 10.1016/j.compedu.2010.01.010

Guvenc, E. (2001). *Student performance assessment in higher education using data mining* (Unpublished master's thesis). Boğaziçi University. Istanbul, Turkey.

Hershkovitz, A., & Nachmias, R. (2008). Developing a log-based motivation measuring tool. In Baker, R.S.J.d., Barnes, T., & Beck, J.E. (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (99–106). Montreal, Canada. Retrieved from http://www.educationaldatamining.org/

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469-478. doi: doi.org/10.1016/j.chb.2014.04.002

IBM Software Group. (2001). Analytics for achievement [White Paper]. Retrieved June 21, 2017, from ftp://public.dhe.ibm.com/software/data/sw-library/cognos/pdfs/whitepapers/wp_analytics_for_achievement.pdf

Johnson, V. E. (1997). An alternative to traditional GPA for evaluating student performance. *Statistical Science*, *12*(4), 251-269. Retrieved from http://www.imstat.org/sts/

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, *13*(1), 61-72. doi: 10.2478/cait-2013-0006

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

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003, September). Preventing student dropout in distance learning using machine learning techniques. In Watada, J. (Eds.), *Proceedings of the International Conference on*

Knowledge-Based and Intelligent Information and Engineering Systems (pp. 267-274). Kitakyushu, Japan. doi: 10.1007/s10462-011-9234-x

Kumar, V., & Chadha, A. (2011). An empirical study of the applications of data mining techniques in higher education. *International Journal of Advanced Computer Science and Applications*, *2*(3), 80-84. Retrieved from http://thesai.org/Publications/IJACSA

Larkey, P. D. (1997). [An Alternative to Traditional GPA for Evaluating Student Performance]: Comment: Adjusting Grades at Duke University. *Statistical Science*, *12*(4), 269-271. Retrieved from http://www.imstat.org/sts/

Laugerman, M. R., & Shelley, M. (2013, June). A structural equation model correlating success in engineering with academic variables for community college transfer students. Paper presented at *ASEE Annual Conference and Exposition*, Atlanta, GA. Retrieved from http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1003&context=stat_las_conf

McAfee, A., Brynjolfsson, E., & Davenport, T. H. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68. Retrieved from https://hbr.org/

McKenzie, K., & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher education research & development*, *20*(1), 21-33. doi: 10.1080/07924360120043621

Nghe, N. T., Janecek, P., & Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. Paper presented at *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual* (pp. T2G-7-T2G-12). Milwaukee, WI. doi: 10.1109/FIE.2007.4417993

Norris, D., Baer, L., Leonard, J., Pugliese, L., & Lefrere, P. (2008). Action analytics: Measuring and improving performance that matters in higher education. *EDUCAUSE review*, *43*(1), 42. Retrieved from https://www.educause.edu/

Oskouei, R. J., & Askari, M. (2014). Predicting Academic Performance with Applying Data Mining Techniques (Generalizing the results of two Different Case Studies). *Computer Engineering and Applications Journal*, *3*(2), 79-88. Retrieved from http://comengapp.ilkom.unsri.ac.id/index.php/comengapp

Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3-12. Retrieved from https://ideas.repec.org/s/tuz/journl.html

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society*, *17*(4), 49. Retrieved from http://www.ifets.info/

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462. doi: 10.1016/j.eswa.2013.08.042

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks*, *16*(3), 9-20. Retrieved from http://olc.onlinelearningconsortium.org/publications/olj_main

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146. doi: 10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618. doi: 10.1109/TSMCC.2010.2053532

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008, June). Data mining algorithms to classify students. In Baker, R.S.J.d., Barnes, T., Beck, J.E. (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 187-191). Montreal, Canada. Retrieved from http://www.educationaldatamining.org/

Sacín, C. V., Agapito, J. B., Shafti, L., & Ortigosa, A. (2009). Recommendation in Higher Education Using Data Mining Techniques. In Barnes, T., Desmarais, M., Romero, C., & Ventura, S. (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining* (pp. 190–199). Cordoba, Spain. Retrieved from http://www.educationaldatamining.org/

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, *5*(4), 13-22.

Shelley, M. C., & Yildirim, A. (2013). Transfer of learning in mathematics, science, and reading among students in Turkey: A study using 2009 PISA data. *International Journal of Education in Mathematics, Science and Technology*, *1*(2), 83. Retrieved from http://ijemst.com/home.html

Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications*, *39*(10), 9468-9476. doi: 10.1016/j.eswa.2012.02.112

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics*, *15*(4), 405-419. doi: 10.1080/09645290701409939

Yadav, S. K., & Pal, S. (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. Retrieved from http://www.wcsit.org/

Yehuala, M. A. (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), 91-94. Retrieved from http://www.ijstr.org/

Zimmermann, J., Brodersen, K. H., Heinimann, H. R., & Buhmann, J. M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining*, *7*(3), 151-176. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM