A MULTIFACETED DATA MINING APPROACH TO ANALYZING COLLEGE STUDENTS' PERSISTENCE AND GRADUATION

A Thesis submitted to the faculty of San Francisco State University
In partial fulfillment of the requirements for the Degree

Master of Science
In
Computer Science

by
Aparna Gopalakrishnan
San Francisco, California
May 2016
CERTIFICATION OF APPROVAL

I certify that I have read A Multifaceted Data Mining Approach to Analyzing College Students' Persistence and Graduation by Aparna Gopalakrishnan, and that in my opinion this work meets the criteria for approving a thesis submitted in partial fulfillment of the requirement for the degree Master of Science in Computer Science at San Francisco State University.

Hui Yang, Ph.D.
Associate Professor

Anagha Kulkarni, Ph.D.
Assistant Professor

Celia Graterol, M.P.H
Co-Director of Information Systems
Metro College Success Program
A MULTIFACETED DATA MINING APPROACH TO ANALYZING COLLEGE
STUDENTS’ PERSISTENCE AND GRADUATION

Aparna Gopalakrishnan
San Francisco, California
2016

This study describes a host of generalizable and data mining-based approaches to identify factors that contribute towards student persistence and graduation, using data from an academic program named Metro College Success Program at San Francisco State University, California. These approaches include (1) a visual analysis to identify bivariate relationships and to understand the flow of students in an educational institute, (2) an ensemble feature selection method to recognize factors that have a significant impact on a student’s persistence and graduation, (3) classification and prediction algorithms to predict whether a student will persist in a given semester and ultimately graduate, and (4) a variety of association patterns to help education practitioners gain further insights into factors that affect persistence and graduation. Our analysis reveals the following main insights: (1) most students who dropout do so in the fourth and seventh terms, (2) the educational level of a student’s mother, the ELM (Entry Level Mathematics) score and race are identified as the most influential factors in predicting a student’s third-term persistence, (3) Naïve Bayesian is the most suitable model for predicting graduation while AdaBoost and SVM models are most suited for predicting persistence (4) a student’s low ELM score and Pell eligibility (an indicator of socioeconomic status) together predict a lower rate of graduation. By collaborating with practitioners and focusing on generating human-interpretable results, the study helped identify bottlenecks to a student’s path towards graduation.

I certify that the abstract is a correct representation of the content of this thesis.

Chair, Thesis Committee

Date
ACKNOWLEDGEMENTS

Firstly, I want to thank my advisor Dr. Hui Yang for giving me the opportunity to work on a project that was academically challenging and whose results have an impact on future generations. Furthermore, I want to thank her for her guidance and support throughout the project. Secondly, I would like to thank Ms. Celia Graterol for helping me get access to the required data and for providing constant feedback on the analysis. Further, I would like to thank Dr. Anagha Kulkarni and Ms. Celia Graterol for being part of my thesis committee and for providing their time and valuable insights.

Last but not least, I want to thank my parents, family and friends who supported me throughout my graduate program. Special thanks to my husband Mr. Raghu Thricovil, whose motivation and support played a major role in helping me make it thus far.
TABLE OF CONTENTS

List of Tables ........................................................................................................................viii

List of Figures ..........................................................................................................................x

1. Introduction .....................................................................................................................1

2. Background and Related Work ......................................................................................8

3. Methods ...........................................................................................................................12

   1. Dataset Description ............................................................................................12
   2. Student Flowchart and Bivariate Visualization ................................................15
   3. Feature Ranking .................................................................................................19
       1. Correlation Feature Selection ...............................................................20
       2. Chi square ..............................................................................................20
       3. Fast Correlation based Filter .................................................................21
       4. Information Gain ....................................................................................21
       5. Kruskal Wallis .......................................................................................22
       6. Minimum-Redundancy-Maximum-Relevance ....................................22
       7. ReliefF .....................................................................................................22
   4. Classification and Prediction ............................................................................24
       1. Linear Support Vector MachineCorrelation Feature Selection ..........25
       2. K-Nearest Neighbours ...........................................................................25
       3. Adaboost ................................................................................................26
       4. Naïve Bayesian ......................................................................................26
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Description and possible values of the twelve student attributes</td>
<td>13</td>
</tr>
<tr>
<td>used in this study.</td>
<td></td>
</tr>
<tr>
<td>2. Description and possible values of the twelve student attributes</td>
<td>14</td>
</tr>
<tr>
<td>used in this study (continued).</td>
<td></td>
</tr>
<tr>
<td>3. Description of various datasets employed in this study.</td>
<td>15</td>
</tr>
<tr>
<td>4. Sample dataset to explain different metrics employed in association rules.</td>
<td>32</td>
</tr>
<tr>
<td>5. Description of various datasets employed in this study (copy).</td>
<td>38</td>
</tr>
<tr>
<td>6. Top five features selected by feature selection algorithm for the</td>
<td>53</td>
</tr>
<tr>
<td>third term persistence dataset.</td>
<td></td>
</tr>
<tr>
<td>7. Top five features selected by feature selection algorithm for the</td>
<td>54</td>
</tr>
<tr>
<td>graduation dataset.</td>
<td></td>
</tr>
<tr>
<td>8. Top five features selected by feature selection algorithm for the</td>
<td>54</td>
</tr>
<tr>
<td>Pell eligible students’ sub-dataset from the graduation dataset.</td>
<td></td>
</tr>
<tr>
<td>9. Top five features selected by feature selection algorithm for the</td>
<td>55</td>
</tr>
<tr>
<td>Non-Pell eligible students’ sub-dataset from the graduation dataset.</td>
<td></td>
</tr>
<tr>
<td>10. Top five features selected by feature selection algorithm for the</td>
<td>55</td>
</tr>
<tr>
<td>First generation eligible students’ sub-dataset from the graduation dataset.</td>
<td></td>
</tr>
<tr>
<td>11. Closed association rules on third term persistence dataset.</td>
<td>62</td>
</tr>
<tr>
<td>12. Contrast rules results on third term persistence dataset.</td>
<td>62</td>
</tr>
<tr>
<td>13. Inverse class rules results on third term persistence dataset.</td>
<td>63</td>
</tr>
<tr>
<td>14. Closed association rules on fifth term persistence dataset.</td>
<td>63</td>
</tr>
<tr>
<td>15. Contrast rules results on fifth term persistence dataset.</td>
<td>64</td>
</tr>
<tr>
<td>16. Inverse class rules results on fifth term persistence dataset.</td>
<td>64</td>
</tr>
</tbody>
</table>
17. Closed association rules on seventh term persistence dataset
18. Contrast rules results on seventh term persistence dataset
19. Inverse class rules results on seventh term persistence dataset
20. Closed association rules on graduation term persistence dataset
21. Contrast rules results on graduation term persistence dataset
22. Inverse class rules results on graduation term persistence dataset
23. Subset based frequent item sets on students still in school dataset
24. Accuracy and F1-measure of J48 model on various datasets
25. Decision tree based rules for third term persistence dataset
26. Input Data format for feature ranking and classification
27. Input Data format for association pattern mining
LIST OF FIGURES

Figures

1. Term-by-term flowchart of students from graduation dataset without attribute information .......................................................................................................................... 17
2. Gender distribution of different subsets of students .............................................. 18
3. Term-by-term flowchart of students from graduation dataset with attribute information (Term 1 through Term 4) ......................................................................................... 40
4. Term-by-term flowchart of students from graduation dataset with attribute information (Term 5 through Term 8) ......................................................................................... 41
5. Term-by-term flowchart of students from graduation dataset with attribute information (Term 9 through Term 12) ......................................................................................... 42
6. Flowchart of ELM-A students from graduation dataset ........................................ 43
7. Flowchart of ELM-B students from graduation dataset ........................................ 45
8. Flowchart of ELM-C students from graduation dataset ........................................ 46
9. Third term persistence distribution grouped by first generation value of the students .......................................................................................................................... 48
10. Distribution of third term persisted students in fifth term ................................... 48
11. Graduation distribution grouped by gender .......................................................... 49
12. Third term persistence distribution grouped by year .......................................... 50
13. Fifth term persistence distribution grouped by year ............................................ 51
14. Seventh term persistence distribution grouped by year ....................................... 52
15. Accuracy of the five classifiers using the persistence datasets .............................. 57
16. Accuracy of the five classifier using the graduation datasets .............................. 58
17. F1-measure of the five classifiers using the persistence datasets .......................... 60
18. F1-measure of the five classifiers using the graduation datasets .......................... 61
19. J48 pruned Decision tree for third term persistence dataset ................................... 70
20. J48 pruned Decision tree for fifth term persistence dataset ..................72
21. J48 pruned Decision tree for seventh term persistence dataset ..............73
22. J48 pruned Decision tree for graduation term persistence dataset ..........74
23. J48 pruned Decision tree for graduation term persistence dataset
   (continued 1) .................................................................................................75
24. J48 pruned Decision tree for graduation term persistence dataset
   (continued 2) .................................................................................................76
25. Screenshot of running the command to execute the FSP jar file ..............82
26. Screenshot of running the FSP jar file after selecting the algorithm .........83
27. Screenshot after completion of running the FSP jar file ............................83
28. Screenshot of running the classification algorithms for grad.csv dataset ....85
Chapter 1

Introduction

In the United States, the 4-year graduation rate of first-time, full-time students who started in 2007 was 39.4% while 5-year graduation rate, including 4-year graduation, for the same group was 55.1% [1]. Research is being done to find the reasons underlying student dropout. For example, studies done in other universities have shown that the likelihood that a student will dropout can be attributed to a number of reasons such as pre-university examination scores, type of financial support, father's level of education and so on [2]. A related outcome that practitioners try to optimize is persistence [3].

In this thesis, persistence is defined as a trait of a student where she continuously attends classes up to and including the said term in the college. A student is considered a dropout if she stops attending college at any term before graduation, and does not return in any of the subsequent terms. However, if a student stops attending in a term (for example, the 3rd term) and returns to class in a subsequent term (for example, the 5th term), the student is not considered as a dropout.

In the United States, only about 72.9% of first-time students who were enrolled in an educational program in 2012 persisted in 2013 [4]. Both persistence and graduation are often studied together [5] and it is therefore important to understand the relevant and
contributing factors that result in a college student's long-term persistence and eventually, timely graduation.

Based on research such as the ones mentioned above, universities and educational institutions worldwide are trying to improve the graduation rates of students using targeted support and intervention programs. Several intervention programs have been developed in various parts of the world that provide targeted support for select students based on traits such as past academic performance and socio-economic status. It however remains a challenging task to identify effective interventions that different students may need and when such interventions should be applied.

In this study, we address the following questions that are of great concern: (1) Identification of features pertinent to a student's persistence in the institution (2) Analyzing the risk of student's dropout so as to provide better support (3) Identification of demographic and behavioral patterns in students who have left or graduated and use the insight as a baseline to future student support.

To achieve this, we adopt a multi-disciplinary methodology by constantly collaborating with educational practitioners to (1) identify a list of important and practical problems, (2) employ both explorative and data mining approaches that aim to produce highly interpretable and actionable results, and (3) communicate with practitioners on a regular basis to ensure the relevance of the adopted algorithms and the validity of the findings, and ultimately to turn such findings into actionable results to improve existing educational practice.
To address the above questions, we adopt a multitude of analytical approaches, ranging from the straightforward yet insightful visual explorations (e.g., visualizing the bivariate relationship between a dependent variable such as entry-level-mathematics score and a target variable such as graduation) to predictive and interpretive data mining approaches such as classification and association pattern mining.

Specifically, the following four methods were adopted:

1. **Student flowchart and bivariate visualization:**
   The term-by-term student flowchart visualizes a student body's important milestones such as dropout, graduation and persistence in one chart to allow practitioners easily identify when maximum student attrition and graduation happened. Bivariate visualization on the other hand exposes the relationships between various attributes and the outcomes such as graduation. This will in turn provide insights on important problems such as dropout.

2. **Ensemble-based feature ranking:**
   This helps identify the most relevant features in the datasets and rank these features according to their impact on the target variable (e.g., graduation).

3. **Classification and predictive analysis:**
   In this analysis, predictive models are built using supervised learning algorithms to identify students at the risk of not graduating or not persisting. Along with finding a predictive model that meets the accuracy, precision and recall needs, this approach also validates if the features selected in the feature selection step indeed have the
predictive power for intervention.

(4) Association pattern analysis:

This analysis aids in identifying the patterns relating to the outcomes such as graduation and persistence [6]. First, interestingness measures are used to reduce the large number of rules to focus on rules that are impactful. Then, a novel approach to better study the class association rules is implemented to help practitioners understand the potential causality of the factors. This involves analyzing inverse and contrast rules to understand the balance and effect of related rules and using subset-based frequent patterns to better identify actionable insights to prevent student dropout. Decision trees based rules were also generated so as to compare and contrast with the association rules.

Many factors enable this kind of analysis to be feasible. Firstly, university systems hold an abundance of student data like demographics, enrollment data, course registration details, high-school academic information, and test scores. Secondly, more schools and universities are being diligent in collecting information, which when combined with the former factor, creates an explosion of valuable data. Thirdly, data mining in general and educational data mining in particular is becoming more sophisticated, resulting in better approaches to help in analyzing the data. Lastly, computing approaches such as Hadoop, programming languages such as R and Python and open-source data analysis packages such as scikit-learn [11] facilitate a quick and iterative approach to obtain valuable insights. Comparing with existing studies with an objective of understanding the underlying
factors resulting in a student's graduation or dropout, our study is unique in the following ways: (1) This study focuses on both persistence and graduation as outcomes for a student unlike most studies that focus on one or the other. (2) The analysis utilizes a multi-faceted approach which includes feature ranking, flowcharts, supervised learning algorithms and association patterns. The resultant analysis not only has the power to predict the outcomes but also offers explanatory causes using associative patterns. (3) Data-driven approach rooted in data mining is tempered with constant communication with domain experts. Collaboration with practitioners enhances the analysis and in turn this multi-disciplinary approach yields results that are highly interpretable.

To evaluate the above four approaches, we use students' data from the Metro College Success Program (Metro) at San Francisco State University in California from 2009 to 2013. The Metro College Success Program provides personalized support during the first two years of college through in-class academic support, tutoring and advising for students belonging to underrepresented minorities, first generation and low-income communities. Metro's goal is to help students gain the skills they need to succeed in college, find a meaningful career, and graduate in a timely manner. Before this study, the practitioners from Metro College Success Program have performed descriptive analysis of the data in collaboration with Academic Institutional Research (AIR) to understand dropout, persistence and graduation rates among the students enrolled in the program. However, this is the first time that Metro has partnered with the Computer Science department to perform educational data mining using Metro student data to understand what factors can be used
to predict graduation and persistence. There are a total of 651 students in the dataset. Each student record contains a total of 12 attributes (See Chapter 3.1 for details).

Through this study, the following main observations are made: (1) Using the student flowchart and bivariate visualization, we have found out that out of the students who leave the institution, most leave in the fourth and seventh terms. This has made these two terms crucial from the perspective of retaining a student. (2) Our ensemble-based feature ranking of seven algorithms has revealed that the education level of mother, ELM (standardized Entry Level Math test)) score and race as the top most relevant features with respect to students’ third term persistence. (3) Out of the five classification and prediction models employed for this study, Naïve Bayesian predicts graduation the best while AdaBoost and Linear SVM models are most suited for predicting persistence. Finally, (4) Association patterns indicate that ELM score and Pell eligibility (a low-income indicator) negatively affect a student’s potential to graduate.

By constantly collaborating with practitioners and focusing on the interpretability of the results, this study helps reveal potential inflection points for student dropout and avenues for intervention. This enables practitioners to predict potential issues and deal with them, which consequently helps ensure the success of college students. The results of the study have contributed actionable insights to the Metro College Success Program. These include:

- Identifying term 4 and term 7 as the terms where most students leave the institution. This insight combined with additional analysis performed by Metro leadership has supported fine tuning the services Metro provides. To address students dropping in
term 4, Metro is strengthening Math and English academic support in order to increase passing rates on remedial courses. And to address students dropping in term 7, Metro is piloting extending the intervention for one more year to a selected group of Metro upper division students.

- Using the results of feature ranking and classification algorithms, practitioners plan to better identify students that need more support during their program.
- Using association pattern based analysis, the study is helping identify students who might need additional Math or English tutoring.

In summary, the insights from this study can help practitioners design more effective and timely interventions and apply them to students who would benefit the most, all with the aim of improving the graduation rate.

The rest of the thesis is organized as follows: Chapter 2 deals with related work in the educational data mining field with the focus on researches about prediction of students’ drop-out and also on the association rules or pattern mining models in place. Chapter 3 discusses the proposed approach and the analysis carried out in our study. Chapter 4 presents the evaluation methodologies and the major results of this research. Chapter 5 describes the installation and details the instructions to run this project. Finally, Chapter 6 concludes this study.
Chapter 2

Related Work

Studies in the educational data mining field have focused on solving a variety of problems including predicting student's grades, identifying students who are most likely to dropout in their freshmen year, and characterizing a student's academic trajectories during their college years. Baker and Yacef [12] discussed trends and shifts of researches conducted in the field of Educational Data Mining (EDM). One approach included the use of statistics and web mining techniques to perform mining of distance learning data to understand behaviors in e-learning classes. Visualization techniques were used to understand social characteristics of computer-supported collaborative learning communities. This study also identified relationship mining methods including association rule mining, correlation mining, and sequential pattern mining and causal data mining as the most utilized EDM methods. For example, Kumar and Chadha [13] applied association rules to optimize the syllabi of educational programs.

Supervised learning models have been widely used in the prediction of graduation and dropout of students. In one research [14], four popular classification methods—artificial neural networks, support vector machines, decision trees and logistic regression were used to build prediction models to predict freshman attrition. Dekker et al. [7] adopted the following classification algorithms to predict whether freshman would dropout from a bachelor program: CART, C4.5, BayesNet, Simple Logistic and Random Forest. In the
paper, Dekker et al., discuss detecting the at risk group at an early stage so that they can prevent students from dropping out. Using classification model the paper was able to handle the following tasks:

1. identify success attributes specific to the program under study
2. identify the data that might result in a further increase of the prediction quality
3. modify the assessment process time-line, resulting in earlier prediction, even before the students entering the study.

The paper also mentions that insights into student success factors collected from teachers, education personnel and management will help to select appropriate measures to support the risk group, which would eventually decrease the dropout rate. In a study [15] carried out in Malaysia, the authors used support vector machine (SVM) for prediction and discovered that students who were in probation were most likely to be dismissed from the institution.

Lakkarju et al. detailed the implementation of a machine learning framework using classification algorithms to identify high school students who need interventions in order to graduate on time [8]. Another paper by Kotsiantis [16] focused on predicting a student's marks in an e-learning course. The author suggested that this could be used by tutors to help identify at-risk students and consequently implement early interventions. Feature selection is another notable analysis method used by researchers in the EDM field [17]. For example, feature selection has been used to eliminate irrelevant features from the dataset and also to rank the input features based on the order of importance with regard to its impact.
on the target outcome of interest. Features such as pre-university scores and gender, have been shown to have an impact on outcomes like graduation.

The paper [37] details how information visualization can allow us to understand large amounts of information at once. The objective of visualization as described in the paper is to “provide feedback to support course authors/teachers/administrators in decision making (about how to improve students’ learning, organize instructional resources more efficiently, etc.) and enable them to take appropriate proactive and/or remedial action.” The data visualization helps to uncover new, hidden, and interesting insights found in data.

Another paper [38], describes how colleges and universities are trying to address to the problem of collecting and analyzing unprecedented amounts of data related to student success. The paper describes the difficulty in managing, manipulating and processing enormous amount of student data. The authors of the paper created a framework using Sankey diagram to explore how students move into and out of the university and its departments over time. These diagrams help visualize the number of students who drop out or graduate along the way.

The approach taken in this study builds on past works such as the ones listed above but differs in a number of ways from the related work. Firstly in addition to graduation, this study focuses on persistence as an outcome as well. The latter has not been under much study even though it is clearly associated with graduation. This approach has enabled insights that inform on how various factors impact these outcomes differently. Our analysis has also found that a single model may not predict two related outcomes such as persistence
and graduation with the same effect. So, different predictive models are needed to predict different outcomes. Secondly, this study focuses on generating highly interpretable and usable results. High interpretability is key in fostering effective multi-disciplinary collaboration. Intuitive visualization techniques such as flowcharts and bivariate bar charts establish the foundation for interdisciplinary communication, which paves the way for applying more sophisticated data mining methods. Finally, this study takes a unique approach to using association rules to offer explanatory causes to the observed results. A student subgroup based associate pattern analysis is also employed to help infer the causes behind the different outcomes achieved by different student subgroups. Such information enables practitioners to not only design support programs for the general student body but also for segments of the population who might need help the most.
Chapter 3

Methods

In this chapter, we describe the four analytical approaches we have adopted to (1) identify the leading factors that potentially contribute to a student’s persistence in a given term and eventual graduation, and (2) infer the discriminating factors to help explain the sharp contrast in persistence and graduation among different student subgroups. These four approaches are bivariate correlation visualization and flow chart analysis, ensemble feature ranking, classification, and association pattern-based analysis.

To facilitate a more concrete discussion of the proposed analytical approaches, next we will first describe the specific dataset we used in our study. We then detail the four approaches. This however by no means indicates the specificity of our approaches. We reiterate here that the proposed approaches are applicable to typical educational datasets with similar goals to ours.

3.1. Dataset Description

The main dataset used in this study includes information on students who joined an academic program named Metro College Success Program at San Francisco State University between 2009 and 2013. Metro College Success Program primarily provides academic support for first-generation, low-income and/or historically underrepresented
students to succeed at a college level. Twelve attributes belonging to a total of 651 students in the above program are used for this analysis.

Table 1 Description and possible values of the twelve student attributes used in this study

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pell eligibility</td>
<td>Is the student eligible for Federal Pell grant?</td>
<td>{Yes, No}</td>
</tr>
<tr>
<td>2</td>
<td>First Generation</td>
<td>Is the student first generation college student?</td>
<td>{Yes, No}</td>
</tr>
<tr>
<td>3</td>
<td>EOP Status</td>
<td>Is the student part of the Educational Opportunity Program (EOP)?*?</td>
<td>{Yes, No}</td>
</tr>
<tr>
<td>4</td>
<td>ELM</td>
<td>The Entry Level Mathematics(ELM) score of the student</td>
<td>{A, B, C}</td>
</tr>
<tr>
<td>5</td>
<td>EPT</td>
<td>The English Placement Test (EPT) score of the student</td>
<td>{A, B, C}</td>
</tr>
<tr>
<td>6</td>
<td>Household income</td>
<td>The annual household income of the student's family</td>
<td>{Less than or equal to $30,000, $30,001-$50,000, $50,001-$70,000, $70,001 or higher, Unknown}</td>
</tr>
</tbody>
</table>
Table 2 Description and possible values of the twelve student attributes used in this study (continued)

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Education level of father</td>
<td>The education level of the student's father</td>
<td>{Postgraduate, Four year College Grad, Two year College Grad, Some College, High School Grad, Some High School, No High School, Unknown}</td>
</tr>
<tr>
<td>9</td>
<td>Race or ethnicity</td>
<td>The student's race or ethnicity</td>
<td>{African American, Asian, Hispanic/Latino, Native-Hawaiian, Two or more Races, White, Unknown}</td>
</tr>
<tr>
<td>10</td>
<td>Start term</td>
<td>The student's start term</td>
<td>{Fall, Spring}</td>
</tr>
<tr>
<td>11</td>
<td>Department</td>
<td>student's department</td>
<td>{Health, CAD (child and adolescent development), Science}</td>
</tr>
<tr>
<td>12</td>
<td>Gender</td>
<td>student's gender</td>
<td>{male, female}</td>
</tr>
</tbody>
</table>

The main dataset for the analysis is further categorized into four labeled subgroups as described in Table 3. These groups correspond to students who graduated or persisted in the third, fifth and seventh terms between 2009 and 2013, respectively. These three semesters are recommended by the education practitioners; however, the data mining approaches are applicable to subgroups of students who persisted in any given semester. Depending on the dataset, the outcome (also referred to as class label) would vary (See Table 3).
Table 3 Description of various datasets employed in this study.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Number of students</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Graduation dataset</td>
<td>Subset of students enrolled in 2009 and 2010</td>
<td>173</td>
<td>Graduated or Not graduated</td>
</tr>
<tr>
<td>2. Third term persistence dataset</td>
<td>Subset of students enrolled from 2009 through 2013</td>
<td>651</td>
<td>Third term persisted or third term not persisted</td>
</tr>
<tr>
<td>3. Fifth term persistence dataset</td>
<td>Subset of students enrolled from 2009 through 2012</td>
<td>455</td>
<td>Fifth term persisted or fifth term not persisted</td>
</tr>
<tr>
<td>4. Seventh term persistence dataset</td>
<td>Subset of students enrolled from 2009 through 2011</td>
<td>273</td>
<td>Seventh term persisted or seventh term not persisted</td>
</tr>
</tbody>
</table>

This study focuses on the following key attributes of the students, also referred to as features: Pell eligibility, first generation, sex/gender, start term, annual household income, race or ethnicity, EOP Status, ELM score, EPT score, education level of father or guardian2, education level mother or guardian1, and department. Table 1 details these twelve attributes along with their possible values.

3.2. Student Flowchart and Bivariate Visualization

Practitioners have access to large amount of data, but gaining insights out of this data is difficult in the raw format. Data visualization acts as a powerful and effective tool
that helps to learn about the data, form hypotheses, give directional information on relationships between factors and identify meaningful data mining problems.

The student flowchart is constructed as a graphical representation to understand how students move through the university over the years. Please see Fig. 1 for an example. A free online tool named draw.io was used to create the student flowchart. The tool has an intuitive interface to create flow charts using simple drag and drop. This helps with readability and enables practitioners gain an intuitive understanding of a given student body’s academic life, and their likelihood to dropout. For example, if 100 students joined the university in 2009, the student flowchart visually represents how many of these students left, graduated or persisted in the university in each term since their first term. At each term, the split of students along various important attributes are examined. At each term, the flowchart splits into three potential branches: students who graduated (if any), students who left (if any) and students who persisted (here persistence refers to students with status of persisted or open). Additionally, attribute information like first generation, Pell eligibility, elm score and gender are included in each node to provide additional insights.
Fig. 1 Term-by-term flowchart of students from graduation dataset without attribute information.
Visualizing the aggregated outcome of a student body on a term-by-term basis with the student flowchart is pivotal for two reasons. Firstly, the flowchart allows practitioners to easily spot patterns and anomalies w.r.t. the student body’s progress. For example, one can easily identify the terms with the most dropouts. This in turn can help practitioners design and implement preventative interventions in advance. Secondly, this flowchart naturally partitions the student body into different subgroups (e.g., the persisting vs. the non-persisting groups).

Once the overall behavior of the student body is established, bivariate data visualization is carried out for each of the student attributes in relation to the outcomes (graduation and persistence). To perform the analysis, a business intelligence and data analytics tool called Tableau [18] is used. This tool aids in obtaining quick insights on how the outcomes relate to the different features. For example, from Fig. 2, comparing the gender distribution of
the original student body and the graduated student body, it is evident that female students graduate much faster than male students.

Student flowcharts and bivariate analysis fulfill the role of helping identify relationships and dictated the data mining approaches used for the rest of the study. The bivariate analysis signals that certain features have more impact on the target variables than others. This insight has led to the use of an ensemble feature ranking approach in order to understand the relative degree of impact each of the features has on the outcomes. This is explained in the next Chapter 3.3. Furthermore, this information is used to predict student persistence and graduation as outlined in Chapter 3.4. The student flowchart, on the other hand, naturally partitions the student body into different subgroups. It also raises an important follow-up question: what causes the differences between these subgroups? Chapter 3.5, describes an association pattern-based solution to this question.

3.3. Feature ranking

Bivariate analysis described in Chapter 3.2 is crucial for practitioners to gain initial insights about the data. But the analysis demonstrates only a qualitative relationship between the outcomes and the factors and cannot facilitate comparison among different attributes based on their relationship with the class label (e.g., persistence in a given term or graduation). To address this issue, an ensemble feature ranking approach is adopted, which consists of a collection of seven feature selection/ranking algorithms. Given a
labeled dataset of m attributes x₁, x₂, ..., xₘ and its class label y (e.g., graduation), the goal of feature selection/ranking is to identify the most relevant attributes and rank them according to their impact on the class label. The main rationale behind an ensemble approach is to avoid potential biases introduced by a single algorithm. Next, a brief description of the seven algorithms included in our ensemble is listed below.

1) Correlation Feature Selection (CFS):

CFS [19] algorithm generates a measure that is based on the following hypothesis: “A good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other”. CFS computes feature-class and feature-feature correlations using symmetrical uncertainty and then chooses a subset of features using the Best First Search strategy. The CFS criterion is defined as follows:

\[
CFS = \max_{k} \left[ \frac{\sum_{j=1}^{k} r_{cf_j}}{\sqrt{\sum_{j=1}^{k} r_{ij}^2}} \right]
\]

(1)

The r_{cf} and r_{ij} variables are referred to as correlations [39].

2) Chi square:

Chi square [20] is used as a test of independence to gauge if the class label is independent of a particular feature. A test of independence assesses whether unpaired observations on two variables, expressed in a contingency table, are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality is related to the response). This helps to remove features that are irrelevant to the classification model [40].
3) Fast Correlation Based Filter (FCBF):

FCBF is a filter model feature selection algorithm that measures feature-class and feature-feature correlation. Given a dataset with N features and a class C, the algorithm finds a set of predominant features $S_{best}$ for the class variable. FCBF employs a two-step process: a relevance analysis and a redundancy analysis. In the relevancy analysis, the features are given a relevance score and irrelevant features are determined. At this step, the algorithm calculates the SU value for each feature, selects relevant features into $S_0$ list based on the predefined threshold $\delta$, and orders them in descending order according to their SU values. This is followed by a redundancy analysis where top predominant features are identified. At this step, the algorithm further processes the ordered list $S_0$ list to remove redundant features and only keeps predominant ones among all the selected relevant features [21].

4) Information Gain (IG):

IG [22] is a measure of dependence between the feature and the class label. Consider that $T$ denotes a set of training examples, each of the form $(x, y) = (x_1, x_2, x_3, ..., x_k, y)$ where $x_a \in \text{vals}(a)$ is the value of the $a^{th}$ attribute of example $x$ and $y$ is the corresponding class label. The expected information gain is the change in information entropy $H$ from a prior state to a state that takes some information. This is calculated as:

$$IG(T, a) = H(T) - H(T|a)$$  \hspace{1cm} (2)

This technique is one of the most popular feature selection techniques as it is easy
to compute and simple to interpret.

5) Kruskal Wallis(KW):

KW [23] test is a non-parametric method for testing whether samples originate from
the same distribution. It is used for comparing two or more independent samples of
equal or different sample sizes. It is important to realize that the Kruskal-Wallis
test is an omnibus test statistic and cannot tell which specific groups of the
independent variable are statistically significantly different from each other; it only
indicate that at least two groups were different[23].

6) Minimum-Redundancy-Maximum-Relevance (mRmR)

mRmR [24] selects features that are mutually far away from each other, while they
still have "high" correlation to the classification variable. Feature selection
identifies subsets of data that are relevant to the parameters used and is normally
called Maximum Relevance. These subsets often contain material which is relevant
but redundant and mRMR attempts to address this problem by removing those
redundant subsets. mRmR is an approximation to maximizing the dependency
between the joint distribution of the selected features and the class variable. mRMR
has been found to be more powerful than the maximum relevance selection[39].

7) ReliefF:

Relief [25] was a feature selection algorithm used in binary classification proposed
by Kira and Rendell in 1992. Its strengths are that it is not dependent on heuristics,
runs in low-order polynomial time, and is noise-tolerant and robust to feature
interactions, as well as being applicable for binary or continuous data. Kononenko et al. [26] suggested changes to the existing Relief algorithm in order to improve the reliability of the probability approximation, make it robust to incomplete data and generalized it to support multi-class problems. The resultant algorithm is referred to as ReliefF.

To perform feature ranking using the above seven algorithms, we use the open source software called feature selection pipeline [27]. To select top features not biased by any of the algorithms, the following approach is used. For each of these algorithms, the top five features are identified and an ensemble value is calculated as the average of the rank of the various features produced as output by the seven feature selection algorithms. Finally, the ensemble value is used to identify the top three features. Though this approach selects top three features, it could be easily extended to obtain ranks for all the features. Practitioners are currently interested in the top three features for each dataset. This ensemble approach reduces biases. Furthermore, it can be applied to subgroups in a student body for comparison purposes. For example if the first generation feature value is fixed to true, the feature selection algorithms ranks the remaining features against the graduation. This aids in understanding what are the most important features that practitioners can focus on to help first generation students graduate.
3.4. Classification and Prediction

The above feature selection task enables practitioners to gain valuable insights into the interplay between different attributes and the target variables of graduation and persistence. A more challenging question remains: can one predict whether a student will persist or graduate using these relevant attributes? Supervised learning algorithms such as classification are used in our study to address this problem. Such algorithms have been widely applied in industries ranging from high-tech, biomedical studies to higher education applications. In the context of our study, the classification results are intended to provide practitioners early warnings of students who might not persist or graduate in the near future, so that they can implement early interventions in a timely fashion.

Given a labeled dataset $D$ of $N$ records, where each record consists of $m$ attributes and one class label, the classification problem learns a model from the labeled training data and subsequently applies this model to predict the labels of unlabeled records. In this study, $m$ is 12 (see Table 1) and our class labels include graduation and persistence in third, fifth and seventh terms, respectively (see Table 2). These specific terms were chosen from among the 12 terms considered in the study based on the suggestion by the practitioners though this approach can be used to predict persistence in any term.

Clearly, we are dealing with a binary classification problem. We considered a total of five classifiers: Linear Support Vector Machine (SVM)[28], K-Nearest Neighbours (KNN)[29], AdaBoost[30], Naive Bayesian[31], and ExtraTrees Classifiers[33].
1. Linear Support Vector Machine (Linear SVM):

Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. In SVM model, the dataset is represented as points in space, so that the points of the separate categories are divided by a clear gap that is as wide as possible. New data is then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class. If the training data are linearly separable, we can select two parallel hyperplanes that separate the two classes of data, so that the distance between them is as large as possible [41].

2. K-Nearest Neighbours:

The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for both classification and regression. The input consists of the k closest training examples in the feature space. The output in k-NN classification is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). In the classification phase, k is a user-defined
constant, and an unlabeled vector (a query or test point) is classified by assigning the
label which is most frequent among the k training samples nearest to that query point.
A commonly used distance metric for continuous variables is Euclidean distance. k-
NN is a type of instance-based learning, or lazy learning, where the function is only
approximated locally and all computation is deferred until classification [42].

3. AdaBoost:
AdaBoost refers to a particular method of training a boosted classifier. A boost
classifier is a classifier in the form:
\[ FT(x) = \sum_{t=1}^{T} f_t(x), \text{ for } t = 1 \text{ to } T \]  \hspace{1cm} (3)
where each \( f_t \) is a weak learner that takes an object \( x \) as input and returns a real valued
result indicating the class of the object. AdaBoost is as it modifies the subsequent
weak learners in favor of those instances misclassified by previous classifiers.
AdaBoost is sensitive to noisy data and outliers. In some problems, however, it can
be less susceptible to the overfitting problem than other learning algorithms. The
individual learners can be weak, but as long as the performance of each one is slightly
better than random guessing, the final model can be proven to converge to a strong
learner [43]. In this study, the decision tree classifier is used as the weak learner.

4. Naïve Bayesian:
Naïve Bayes is a conditional probability model. Given a problem instance to be
classified, represented by a vector \( x = (x_1, \ldots, x_n) \) representing some \( n \) features
(independent variables), the algorithm assigns instance probabilities for each of \( K \)
possible outcomes or classes \[7\].

\[ p(C_k|x_1, \ldots, x_n) \quad (4) \]

An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters necessary for classification. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set \[44\].

5. Extra Trees Classifier:

An extremely randomized tree classifier \[33\]. Extra-trees differ from classic decision trees in the way they are built. When looking for the best split to separate the samples of a node into two groups, random splits are drawn for each of the max_features randomly selected features and the best split among those is chosen. When max_features is set 1, this amounts to building a totally random decision tree.

These classifiers are chosen to cover a wide range of learning methodologies from simple instance-based methods such as KNN, to probabilistic models such as logistic regression, to more advanced kernel-based learning such as SVM. This way, we can identify which learning methodology and classifier most fit the needs of the problem.

To evaluate the effectiveness of these algorithms, we adopt a similar approach as described by Lakkaraju et al. \[8\]. Specifically, given a labeled dataset, 20% of the data is randomly selected as the test-set and the remaining 80% as the training set. This is repeated for 20 rounds to eliminate potential bias. For each round, the following quantitative
measures are calculated for each of the five classifiers:

(1) Accuracy, defined as the percentage of testing samples classified correctly by the classifier model

(2) Precision, defined as the percentage of items correctly labeled as belonging to the positive class (e.g., graduated) [34]

\[
Precision = \frac{tp}{(tp+fp)}
\]  

where \(tp\) is true positive and \(fp\) is false positive.

(3) Recall, defined as the number of items correctly labeled as positive divided by the total number of elements that actually belong to the positive class (e.g. graduated) [34]

\[
Recall = \frac{tp}{tp+fn}
\]  

where \(tp\) is true positive and \(fn\) is false negative.

(4) F1-measure, which combines precision and recall and is the harmonic mean of precision and recall [34].

\[
F1 - measure = 2 \times \frac{precision \times recall}{precision + recall}
\]  

The average of the F-measure and accuracy over the 20 rounds is reported for all these five algorithms in Chapter 4.

3.5. Association pattern-based analysis

Classification and prediction help practitioners identify students who are more likely to
dropout but it does not help them understand the patterns observed in the dropout students. Association rule mining can help with this problem. For example, identifying that ELM score being 0 is a frequently occurring pattern in dropout students can function as a good indicator for practitioners to provide tutoring services to students with lower ELM score. This methodology also has led to analyzing the patterns in students who left the university, students who graduated and students who are persisting even after term 12 in the university. Such patterns are crucial for practitioners to identify students who need help at the right term in order to assist students to graduate on time.

Next, we describe the generic association pattern mining problem in the context of our academic dataset. We will then focus on the special types of association patterns we have designed to help practitioners perform cross-group comparison. For instance, these patterns will help address the question: while comparing the students who graduated within five years with those who did not, what are the patterns exhibited by the former group but not in the latter?

Let $D = \{s_1, s_2, \ldots, s_N\}$ be a collection of $N$ student records. Each record or transaction $s_i$ in the context of this study, consists of $m$ attribute-value pairs and a label, i.e., $s_i=<(a_{ij}, v_{ij}), \ldots, (a_{im}, v_{im}), \text{label}>$, where $a_{ij}$ and $v_{ij}$ are the $j$th attribute and corresponding value of the $i$th record (e.g., gender=female), and label is either graduation or persistence in a given term. Similar to the classic definition given by Agrawal et al. [6], the problem of association pattern mining is to extract association patterns in the format of $X \Rightarrow Y$, where $X$ and $Y \subseteq \{(a_i, v_i): 1 \leq i \leq m\} \cup \{$label$\}$ and $X \cap Y = \emptyset$ whose support[49] \(\text{supp}(X) = \)
The support of a frequent item set $X$ is defined as the ratio of the number of records containing $X$ to the total number of records. Similarly, the confidence of a rule $X \rightarrow Y$ is defined as the ratio of the support of the item set $X \cup Y$ to the support of the item set $X$. If both the support and confidence are above certain thresholds, then the rule is considered frequent and confident.

Confidence can be interpreted as an estimate of the conditional probability, the probability of finding the right hand side (RHS) of the rule in transactions under the condition that these transactions also contain the left hand side (LHS). Note that the association patterns in the former format is commonly referred to as frequent item sets, whereas the latter is referred to as strong association rules.

Given the target problems studied in this work, we focus on extracting association rules whose right-hand-side only contains a label (e.g., graduated or not-graduated). Such rules are also referred to as class association rules. Furthermore, to reduce the redundancy in the resulting association patterns, we extract closed item sets [35] and the corresponding class association rules.

3.5.1. Identification of relevant class association rules

Even with the usage of closed association patterns, the resulting list of class association rules is still large and hinders direct application in practice. Moreover, many of the rules might not be ‘interesting’ and ‘relevant’. To address this issue, in addition to the commonly used interesting measurements of support and confidence, we use three other measurements to help identify the set of interesting and relevant rules:

a) Cosine($X \rightarrow Y$) or Cosine measure[46] calculated as follows:
Cosine\((X, Y) = \frac{\text{supp}(X,Y)}{\sqrt{\text{supp}(X) \cdot \text{supp}(Y)}}.\) \hspace{1cm} (8)

The range of this measure is \([0,1]\). A value around 0.5 indicates a weak or neutral correlation between X and Y, a value close to 1 indicates a strong positive correlation, and finally, a value close to 0 indicates a strong negative correlation. This equality shows that transactions not containing neither item X nor item Y have no influence on the result of \(\cosine(X \rightarrow Y)\).

b) Kulczynski measure: Kulczynski measure[47] is another measure used to find the interestingness of a rule and is calculated as:

\[\text{kulc}(A, B) = \frac{\text{sup}((A \cup B))}{2} \left( \frac{1}{\text{sup}(A)} + \frac{1}{\text{sup}(B)} \right)\] \hspace{1cm} (9)

If the value of this measure is near 0 or 1 then the rule is negatively or positively associated to each other respectively. A rule with kulc measure value equal to 0.5 can be considered as an uninteresting rule.

c) Imbalance Ratio (IR): IR[48] measures the imbalance of two item-sets X and Y. IR\((A,B)\) is calculated as:

\[\text{IR}(A, B) = \frac{\text{mod}(\text{sup}(A) - \text{sup}(B))}{\text{sup}(A) + \text{sup}(B) - \text{sup}(A \cup B)}\] \hspace{1cm} (10)

The range of IR measure is \([0,1]\). A value close to 0 indicates a perfectly balanced rule, a value close to 1 indicates a skewed rule and finally a value around 0.5 indicates a neutral rule.

Let's take a sample dataset to understand the meaning and importance of the above measures. For simplicity, we are using a single attribute and the class label.
Table 4: Sample dataset to explain different metrics employed in association rules

<table>
<thead>
<tr>
<th>ELM score</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Graduated</td>
</tr>
<tr>
<td>C</td>
<td>Not Graduated</td>
</tr>
<tr>
<td>B</td>
<td>Graduated</td>
</tr>
<tr>
<td>A</td>
<td>Graduated</td>
</tr>
</tbody>
</table>

Consider a rule \([ELM\, score,\, A] \Rightarrow\, Graduated\), for which the measures are calculated as below:

a) \(supp([ELM\, score,\, A] \Rightarrow\, Graduated) = \frac{\text{# of records containing } ([ELM\, score,\, A] \Rightarrow\, \text{Graduated})}{\text{Total # of records}}\)

\[= \frac{2}{4} = 0.5\]

i.e; 50% of the total records contain ELM score = A and graduated as the class label.

b) \(conf([ELM\, score,\, A] \Rightarrow\, \text{Graduated}) = \frac{supp([ELM\, score,\, A] \cup \text{Graduated})}{supp([ELM\, score,\, A])}\)

\[= \frac{2}{4} + \frac{2}{4} = 1\]

i.e; 100% of records that contain ELM score = A, also have graduated as the class label.

c) \(Cosine([ELM\, score,\, A],\, \text{Graduated}) = \frac{supp([ELM\, score,\, A],\, \text{Graduated})}{\sqrt{supp([ELM\, score,\, A])} \times supp(\text{Graduated})}\)

\[= \frac{2}{4} \div \sqrt{\frac{6}{16}} = 0.82\]

Cosine closer to 1 indicates strong positive correlation. So the above rule is strongly correlated.
d) $kulc(\text{[ELM score, A], Graduated})$

$$= \frac{supp(\text{[ELM score, A], Graduated})}{2} \left(\frac{1}{supp(\text{[ELM score, A]})} + \frac{1}{supp(\text{Graduated})}\right) = \frac{2}{8} \times \left(\frac{\frac{4}{2}}{\frac{4}{3}}\right) = 0.83$$

Kulc measure of 0.83 indicates that the rule $\text{[ELM score, A]} \Rightarrow \text{Graduated}$ is positively related and this implies that a student with ELM score A has a high chance of graduating.

e) $IR(\text{[ELM score, A], Graduated})$

$$= \frac{\text{mod}(supp(\text{[ELM score, A]}) - supp(\text{Graduated}))}{supp(\text{[ELM score, A]}) + supp(\text{Graduated}) - supp(\text{[ELM score, A], Graduated})} = \text{mod}(\frac{\frac{2}{4} - \frac{3}{4}}{\frac{2}{4} + \frac{3}{4} - \frac{2}{4}}) = 0.33$$

A value close to 0 for IR indicates a perfectly balanced rule. Hence we can say that the rule under consideration is a balanced rule.

The measure, cosine, kulc and IR are all null-invariant, in that they are not affected by null transactions. Additionally, by collectively considering IR, cosine and Kulczynski measures, we can identify rules that are not skewed by imbalanced datasets. To determine a proper threshold for each of these measures, we manually went through all the strong rules generated for one of the datasets as described in Chapter 3.1. As a result, for the persistence datasets, the following threshold values were selected: confidence $\geq 0.8$, $IR \leq 0.42$, cosine $\geq 0.6$ and Kulczynski $\geq 0.6$. For the graduation dataset, the following threshold values were selected: $IR \leq 0.42$, cosine $\geq 0.6$ and Kulczynski $\geq 0.6$. 
3.5.2. Inverse Class Rule Analysis

For each relevant rule, the inverse rule analysis would find the rule with the inverse outcome. In other words, for every relevant rule, $X \Rightarrow Y$ where $Y$ is the outcome in class association rule, an inverse class rule analysis, $X \Rightarrow Y'$ is analyzed where $Y'$ is the inverse of the class label. For example, if $[\text{ELM score-B} \Rightarrow \text{Graduated}]$ is a relevant rule, we will construct then its inverse class rule $[\text{ELM score-B} \Rightarrow \text{Not Graduated}]$ and consequently calculate its interesting measures. This analysis enables the comparison of these two rules to understand how the same feature values varies for different outcomes.

3.5.3. Contrast Rule Analysis

In the contrast rule analysis, the feature-value-contrast of the antecedent is considered and the rule is generated with the same consequent. To do this, one of the feature values in the antecedent is changed while keeping other values fixed namely for every relevant rule $X \Rightarrow Y$ where $Y$ is the outcome in class association rule and $X$ is a subset of attribute-value pairs, a contrast rule $X' \Rightarrow Y$ is analyzed, where $X'$ includes the feature values not in $X$. For example if $[(\text{ELM score, B}) \Rightarrow \text{Graduated}]$ is a relevant rule, then in contrast rules the relevance measures of $[(\text{ELM score, A}) \Rightarrow \text{Graduated}]$ and $[(\text{ELM score, C}) \Rightarrow \text{Graduated}]$ is computed. This would give the practitioners the ability to use these contrasting rules to understand how the different feature values vary for the same outcome.
3.5.4. Students subgroup-based frequent itemsets

In this analysis, association pattern mining is carried out on student subgroups that are identified by using the student flowchart as described in Chapter 3.2. Once the frequent patterns are identified in a subgroup, we also make sure that these patterns are unique by checking whether they are relevant patterns in other student subgroups and also the entire dataset.

The resulting patterns are valuable as they can not only characterize each subgroup but also provide vital clues on the root causes of students’ persistence, successful graduation and dropouts. For example, by looking at frequent item sets of students who dropped out of the program in terms 4 and 7, the terms with maximum dropout, practitioners can understand the common traits among the group of students who leave the institution. This can help identify possible root causes to dropouts and aid in designing interventions to prevent them. This approach can also be used to understand why students who persist may not be graduating. By looking at frequent item sets created from students who are still in school but haven’t graduated in six years, practitioners can identify what is holding back these students from graduating, even though they are committed to their education.

3.5.5. Comparison of Association Rules to Decision Tree based Rules

The class association rules obtained using Apriori algorithm were compared with the rules from the decision tree based classification model. C4.5 [50] builds decision trees from a set of training data based on the concept of information entropy. At each node of the tree,
C4.5 chooses the attribute that most effectively splits the dataset into subsets enriched in one class or the other. The splitting criterion is the normalized information gain or gain ratio. The attribute with the highest normalized information gain is chosen to make the split. The decision tree thus obtained, [51] has the class label forming the leaf node, and the conditions along the path forming the conjunction. This conjunction takes the following form:

\[
\text{if (condition1 and condition2 and condition3) then outcome.}
\]

These are referred to as decision rules and can be compared with the association rules to study any similarity or dissimilarity in patterns.

Association pattern-based analysis is made possible by a combination of the following steps:

1) Firstly, software created by Christian Borgelt [36] is used to generate closed itemsets which uses Apriori algorithm. Apriori algorithm uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules that highlight the general trends in the datasets [36].

2) Secondly, class association rules are generated from the closed item sets using python programming language and an open source library called Scikit-learn.
3) Thirdly, inverse and contrast rules are generated programatically using Python.

4) For subgroup based analysis, due to a limited number of students in the subgroups, we opt for generating frequent itemsets using the software created by Christian Borgelt [36]. The number of frequent itemsets are restricted by setting the support threshold to be greater than 0.6.

5) Finally, for decision tree based rules, the J48 algorithm was used. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool [53].
Chapter 4

Results

To understand and study the efficacy of the above analysis, the resulting outcomes were evaluated, studied and analyzed using the datasets described in Table 5 shown below. This table is identical to the table 3 shown in Chapter 3.1 under Dataset description. We next present our major findings obtained through the study. For each result, we will also discuss its potential impact in terms of how practitioners can use these results to better support students though their college education.

Table 5 Description of various datasets employed in this study.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Number of students</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Graduation dataset</td>
<td>Subset of students enrolled in 2009 and 2010</td>
<td>173</td>
<td>Graduated or Not graduated</td>
</tr>
<tr>
<td>2. Third term persistence dataset</td>
<td>Subset of students enrolled from 2009 through 2013</td>
<td>651</td>
<td>Third term persisted or third term not persisted</td>
</tr>
<tr>
<td>3. Fifth term persistence dataset</td>
<td>Subset of students enrolled from 2009 through 2012</td>
<td>455</td>
<td>Fifth term persisted or fifth term not persisted</td>
</tr>
<tr>
<td>4. Seventh term persistence dataset</td>
<td>Subset of students enrolled from 2009 through 2011</td>
<td>273</td>
<td>Seventh term persisted or seventh term not persisted</td>
</tr>
</tbody>
</table>
4.1 Student Flowchart and Visualization: Main Results

The student flowchart helps provide this insight by visualizing how students move through the terms and when they dropout. Fig. 3, 4 and 5 shows the snapshot of the resulting student flowchart that helps gain this insight. The flow chart charts the academic trajectories of the 173 students in the study for 12 terms. Additionally, at each node, information like Pell eligibility, first generation, gender and ELM scores are looked at. In the dataset analyzed, term 4 and term 7 are identified as the terms where most students leave the college. This result will help practitioners focus on understanding the needs of students in these terms and designing support programs and interventions that work best to avoid drop off. Additionally, by combining the data insights with their experience, the practitioners have developed a hypothesis that students leave in term 4 because they are not able to complete the Math and English remediation courses. As a result, providing additional support to students to complete the remedial courses become more important than other support programs as completion of remedial courses can have a sizable impact on the students to persist and eventually graduate. Additionally, preliminary analysis done by Metro practitioners suggest that Math remediation is more challenging than English remediation.
Fig. 3 Term-by-term flowchart of students from graduation dataset with attribute information (Term 1 through Term 4).
Fig. 4 Term-by-term flowchart of students from graduation dataset with attribute information (Term 5 through Term 8).
Fig. 5 Term-by-term flowchart of students from graduation dataset with attribute information (Term 9 through Term 12).
Additionally, flowcharts were constructed on certain important feature values identified.
in Chapter 4.2. Fig. 6, 7 and 8 represents flowchart of students from the graduation dataset with ELM score values of A, B and C. This was done to test the hypothesis that ELM scores of students bear an impact on the potential to graduate of students. Though the number of students in these sub-datasets is limited, we see a trend that students with ELM score value of A show a higher likelihood of graduating than when compared to students with ELM score B or C.

In the dataset, based on Figs. 6, 7 and 8: 21 students or 75% of students with an ELM score of A have graduated as of Spring 2015 and only 7% of students are still in school. The graduation rate for students with an ELM score of B is 61% while 12% of students are still in school. For students with ELM score of C, 47% for students have graduated while 22% are still in school. This suggests that a higher ELM score results in a higher likelihood of graduation.
Fig. 7 Flowchart of ELM-B students from graduation dataset.
Fig. 8 Flowchart of ELM-C students from graduation dataset.
The bivariate analysis helps practitioners confirm some of their hypotheses and observations. For example, there is a hypothesis among the practitioners that the students who are first generation college students are less likely to persist and hence need more support. The analysis confirms the hypothesis and quantifies the gap between first generation students and those who are not, is at a 5% difference in persistence. This result is illustrated in Fig. 9. Similarly, it is evident that the third term persistence of Pell eligible students is 6% lower than non-Pell eligible students.

The bivariate analysis has also generated new insights. The practitioners used to always focus on the persistence of the student population in only specific milestone terms (e.g., 3rd term persistence and 5th term persistence). In our analysis, in addition to looking at these specific terms' persistence, we also look at the persistence of subgroups in between these specific terms. For example, Fig. 10 shows that only 8% of students who persisted in the third term did not persist in the 5th term. Such information can further help practitioners identify the terms that need the most attention.
Fig. 9 Third term persistence distribution grouped by first generation value of the students

Fig. 10 Distribution of third term persisted students in fifth term
There are similar insights for graduation as well. Fig. 11 confirms and quantifies the perception that female students are more likely to graduate than male students. This gap, as illustrated by Fig. 11, is observed to be about 17%. The analysis also reveals some surprising results. Though there is a perception that the education level of parents is important in predicting a student’s success, practitioners are surprised to learn that if a student’s mother is a graduate from a 4-yr college grad or postgraduate, and the likelihood of graduation is increased by over 15%.
The third, fifth and seventh term persistence of the students were studied for all the available attributes. Fig. 12, 13 and 14 shows the persistence distribution by year. The third term, fifth term and seventh term persistence seems to be lower in 2009 compared to the other years. This may also be because of the fact that 2009 dataset is much smaller than that of other years. This analysis would help practitioners to identify variation in persistence over the years. Fig. 12 also shows that the third term persistence increased in 2010 when compared to 2009 and is remaining stable from 2010 through 2012.

Educational practitioners informed us that one of the major reason for around 10% of students to leave the institution is due to not completing remediation courses. Students automatically dropped-out from the university if they do not complete the remediation
courses in the first year of their program. Over 70% of Metro students are required to take remedial courses.

At term 5, around 13% of students do not persist. This percentage is inclusive of the third term persisted students. Fifth term persisted students seems to become stable in 2010 and 2011 as shown in Fig. 13.

Fig. 14 shows high percentage of students not persisting in the seventh term. This indicate that students find it harder to persist in seventh term. One of the plausible hypothesis is that Metro program being a 2 year program provides support to the students until the fourth term. This lack of continued support may explain the lack of persistence increasing from about 10% in the third term to about 23% in the seventh term.
4.2 Feature Ranking: Main Results

Seven feature ranking algorithms are used to identify the top important features in the various datasets. Since these algorithms take different approaches to calculating feature importance, the results vary for each. Table 5 below outlines the top five features selected by each of the algorithms used for the third term persistence dataset. From that data in Table 6, education level of mother, ELM score and race are found as the most important features in the third term persistence dataset.
Table 6 Top five features selected by feature selection algorithm for the third term persistence dataset

<table>
<thead>
<tr>
<th>CFS</th>
<th>Chi square</th>
<th>FCBF</th>
<th>IG</th>
<th>KW</th>
<th>mRmR</th>
<th>ReliefF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Race</td>
<td>Edu. level of mother</td>
<td>ELM</td>
<td>Edu. level of mother</td>
<td>ELM</td>
<td>Race</td>
</tr>
<tr>
<td>2</td>
<td>Income</td>
<td>ELM</td>
<td>First Generation</td>
<td>Edu. level of father</td>
<td>EPT</td>
<td>Income</td>
</tr>
<tr>
<td>3</td>
<td>EPT</td>
<td>Edu. level of father</td>
<td>-</td>
<td>ELM</td>
<td>Income</td>
<td>EPT</td>
</tr>
<tr>
<td>4</td>
<td>ELM</td>
<td>Income</td>
<td>-</td>
<td>Income</td>
<td>First Generation</td>
<td>ELM</td>
</tr>
<tr>
<td>5</td>
<td>Edu. level of father</td>
<td>Race</td>
<td>-</td>
<td>Race</td>
<td>Race</td>
<td>Edu. level of father</td>
</tr>
</tbody>
</table>

Table 7 outlines the features with the top ranks in the graduation dataset. Education level of mother, ELM score and Race are identified as the top most relevant features according to the seven feature selection algorithms.

Feature ranking is also used to identify the most relevant features in subsets of data. This helps isolate nuances for students sharing a particular trait. For example, Table 8 gives the top five features identified by the seven feature ranking algorithms for Pell eligible students from the graduation dataset.

From Table 8, ELM score, EPT score and Gender come up as the top features for this subset. First-generation was expected to show up as an important feature in the ranking but since the dataset consists of Metro College Success Program students who are predominantly first-generation, this did not show up in the results.
Table 7 Top five features selected by feature selection algorithm for graduation dataset

<table>
<thead>
<tr>
<th></th>
<th>CFS</th>
<th>Chi Square</th>
<th>FCBF</th>
<th>IG</th>
<th>KW</th>
<th>mRmR</th>
<th>ReliefF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Edu. of mother</td>
<td>ELM score</td>
<td>ELM score</td>
<td>Edu. of mother</td>
<td>Edu. of mother</td>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>EOP Status</td>
<td>Edu. of mother</td>
<td>Gender</td>
<td>ELM score</td>
<td>Race</td>
<td>EOP Status</td>
<td>EPT score</td>
</tr>
<tr>
<td>3</td>
<td>Dpt.</td>
<td>Edu. of father</td>
<td>Pell eligibility</td>
<td>Edu. of father</td>
<td>Pell eligibility</td>
<td>Dpt.</td>
<td>ELM score</td>
</tr>
<tr>
<td>4</td>
<td>Gender</td>
<td>EPT score</td>
<td>-</td>
<td>Race</td>
<td>EOP Status</td>
<td>Gender</td>
<td>Dpt.</td>
</tr>
<tr>
<td>5</td>
<td>Race</td>
<td>Race</td>
<td>-</td>
<td>EPT score</td>
<td>EPT score</td>
<td>Race</td>
<td>Pell eligibility</td>
</tr>
</tbody>
</table>

Table 8 Top five features selected by feature selection algorithm for the Pell eligible students' sub-dataset from the graduation dataset

<table>
<thead>
<tr>
<th></th>
<th>CFS</th>
<th>Chi Square</th>
<th>FCBF</th>
<th>IG</th>
<th>KW</th>
<th>mRmR</th>
<th>ReliefF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gender</td>
<td>ELM score</td>
<td>ELM score</td>
<td>ELM score</td>
<td>ELM score</td>
<td>Gender</td>
<td>EPT score</td>
</tr>
<tr>
<td>2</td>
<td>ELM score</td>
<td>EPT score</td>
<td>Gender</td>
<td>EPT score</td>
<td>ELM score</td>
<td>ELM score</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>EPT score</td>
<td>Edu. of father</td>
<td>-</td>
<td>Edu. of mother</td>
<td>Edu. of mother</td>
<td>EPT score</td>
<td>Race</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>Edu. of mother</td>
<td>-</td>
<td>Edu. of father</td>
<td>Gender</td>
<td>-</td>
<td>Department</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>Income</td>
<td>-</td>
<td>Income</td>
<td>Income</td>
<td>-</td>
<td>Gender</td>
</tr>
</tbody>
</table>
Feature ranking on the non-pell eligible students' sub-dataset as shown in Table 9 identifies race as the topmost important feature followed by gender and cohort. It is surprising to see that ELM score is the most important feature for pell eligible students for graduation whereas it is not even in the top five important features for non-pell eligible students' dataset.

Table 9 Top five features selected by feature selection algorithm for the Non-Pell eligible students' sub-dataset from the graduation dataset

<table>
<thead>
<tr>
<th></th>
<th>CFS</th>
<th>FCBF</th>
<th>IG</th>
<th>KW</th>
<th>mRmR</th>
<th>ReliefF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cohort</td>
<td>Race</td>
<td>Race</td>
<td>Race</td>
<td>Cohort</td>
<td>Education level mother</td>
</tr>
<tr>
<td>2</td>
<td>Gender</td>
<td>Gender</td>
<td>Education level father</td>
<td>Cohort</td>
<td>Gender</td>
<td>Start term</td>
</tr>
<tr>
<td>3</td>
<td>Start term</td>
<td>-</td>
<td>Education level mother</td>
<td>Gender</td>
<td>Start term</td>
<td>Education level father</td>
</tr>
<tr>
<td>4</td>
<td>Race</td>
<td>-</td>
<td>EPT</td>
<td>Start term</td>
<td>Race</td>
<td>Race</td>
</tr>
<tr>
<td>5</td>
<td>EPT</td>
<td>-</td>
<td>Gender</td>
<td>First Generation</td>
<td>EPT</td>
<td>Gender</td>
</tr>
</tbody>
</table>

Table 10 Top five features selected by feature selection algorithm for the First generation eligible students' sub-dataset from the graduation dataset

<table>
<thead>
<tr>
<th></th>
<th>CFS</th>
<th>FCBF</th>
<th>IG</th>
<th>KW</th>
<th>mRmR</th>
<th>ReliefF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Race</td>
<td>Race</td>
<td>Race</td>
<td>EPT</td>
<td>Race</td>
<td>EPT</td>
</tr>
<tr>
<td>2</td>
<td>ELM</td>
<td>EPT</td>
<td>EPT</td>
<td>ELM</td>
<td>ELM</td>
<td>Race</td>
</tr>
<tr>
<td>3</td>
<td>EPT</td>
<td>-</td>
<td>ELM</td>
<td>Race</td>
<td>EPT</td>
<td>ELM</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>Education level mother</td>
<td>Gender</td>
<td>-</td>
<td>Cohort</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>Education level father</td>
<td>Start term</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The first generation students' sub-dataset shows that race, EPT and ELM scores are important for graduation as given in Table 10. This further shows the need for English and
Math tutoring classes for first generation college students to help these students to graduate on time.

Additionally a similar feature ranking approach on an extended graduation dataset found that seventh term persistence, fifth term persistence, and third term persistence respectively are the most important factors influencing a student’s graduation. This validates the known presumption that persistence is very vital for graduation.

Feature ranking helps identify important features in each dataset. Practitioners can utilize such knowledge as a baseline to target students who need the most support to graduate and persist in a university.

4.3 Classification and Prediction: Main Results

The datasets that the prediction models are deployed on can be split into two groups. The first group includes three graduation datasets: graduation dataset including third term persistence value as an additional attribute, graduation dataset including third, fifth and seventh term persistence value as additional attributes, and graduation dataset including third, fourth, fifth, and sixth and seventh term persistence value as additional attributes. The second group includes three persistence datasets: third term persistence dataset, fifth term persistence dataset and seventh term persistence dataset. The results of these measures are shown in Fig. 15, 16, 17 and 18.
Fig. 15: Accuracy of the five classifiers using the persistence datasets.
Fig. 16: Accuracy of the five classifiers using the graduation datasets.
To evaluate the supervised classifier models, we mainly look at the metrics F1-measure (a harmonic mean of precision and recall) and accuracy. Fig. 15 and Fig. 16 show the accuracy of the various models for the persistence and graduation datasets. Naïve Bayesian is identified as the model that gave the better overall accuracy for the graduation datasets with lower error bars while AdaBoost and Linear SVC (i.e., Linear SVM) models, in addition to having almost the same accuracy, demonstrated significantly higher in accuracy compared to other models.

Fig. 17 and Fig. 18 show the F1-measure for the persistence and graduation datasets. These results further validate the observations noted in the accuracy result. Naïve Bayesian model scores the better overall F1-measure for the graduation datasets. This leads to the conclusion that Naïve Bayesian performs better than the other classifier models to predict graduation. The possible reason Naïve Bayesian model works better on the graduation dataset could be that it is less likely to overfit. This is particularly important given that all these datasets are relatively small.

For the persistence datasets, Adaboost and SVM models perform better than the other models. The Adaboost model learns from a set of weak learners and by iteratively learning from incorrectly classified examples. This may be the reason why an ensemble model like Adaboost yields better accuracy, precision and recall for the persistence datasets. The SVM model works best when the attribute values of transactions in different class labels can be separated by a clear gap as wide as possible. During association rule based analysis of the persistence datasets, we notice that the rules have high support and confidence. This
indicates that features are aligned closely with the class label, which could possibly cause attribute values of different class labels to be separated farther in space. This is possibly why SVM works better than the other models for predicting persistence.

Fig. 17: F1-measure of the five classifiers using the persistence datasets.
4.4 Association Pattern-based Analysis: Main Results

The results of association rules mining are shown in the Tables below. The aim is to identify balanced rules that are correlated (either positively or negatively) for the insight to be actionable. If a rule is balanced and positive, practitioners can use it to reinforce behaviors. If a rule is balanced and negative, practitioners can use it to discover possible issues and design interventions or increase support.
Table 11 Closed association rules on third term persistence dataset

<table>
<thead>
<tr>
<th>Rule</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Cosine</th>
<th>IR</th>
<th>Kule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>['First-Gen'] =&gt; Third-term-Persisted</td>
<td>0.65</td>
<td>0.89</td>
<td>0.80</td>
<td>0.17</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>['Female'] =&gt; Third-term-Persisted</td>
<td>0.64</td>
<td>0.90</td>
<td>0.80</td>
<td>0.19</td>
<td>0.80</td>
</tr>
<tr>
<td>3</td>
<td>['Pell-Yes'] =&gt; Third-term-Persisted</td>
<td>0.63</td>
<td>0.88</td>
<td>0.79</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td>4</td>
<td>['EOP-Status-No'] =&gt; Third-term-Persisted</td>
<td>0.58</td>
<td>0.90</td>
<td>0.76</td>
<td>0.27</td>
<td>0.77</td>
</tr>
</tbody>
</table>

In Table 11, which shows the closed association rules on the third term persistence dataset, the result indicates that a student who is first generation or Pell eligible, and persisting in third term is a balanced and positive correlated rule. This is interesting since first generation students generally tend to need more support or show lower chance of persisting. The other important rules show that female students and students who are not in educational opportunity programs and persist in third term is also a balanced and positively correlated rule.

Table 12 Contrast rules results on third term persistence dataset

<table>
<thead>
<tr>
<th>Rule</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Cosine</th>
<th>IR</th>
<th>Kule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>['First-Gen'] =&gt; Third-term-Persisted</td>
<td>0.65</td>
<td>0.89</td>
<td>0.80</td>
<td>0.17</td>
<td>0.81</td>
</tr>
<tr>
<td>~1</td>
<td>['Not-first-Gen'] =&gt; Third-term-Persisted</td>
<td>0.25</td>
<td>0.92</td>
<td>0.50</td>
<td>0.68</td>
<td>0.60</td>
</tr>
<tr>
<td>2</td>
<td>['Female'] =&gt; Third-term-Persisted</td>
<td>0.64</td>
<td>0.90</td>
<td>0.80</td>
<td>0.19</td>
<td>0.80</td>
</tr>
<tr>
<td>~2</td>
<td>['Male'] =&gt; Third-term-Persisted</td>
<td>0.26</td>
<td>0.89</td>
<td>0.51</td>
<td>0.65</td>
<td>0.59</td>
</tr>
<tr>
<td>3</td>
<td>['Pell-Yes'] =&gt; Third-term-Persisted</td>
<td>0.63</td>
<td>0.88</td>
<td>0.79</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td>~3</td>
<td>['Pell-No'] =&gt; Third-term-Persisted</td>
<td>0.26</td>
<td>0.93</td>
<td>0.52</td>
<td>0.67</td>
<td>0.61</td>
</tr>
<tr>
<td>4</td>
<td>['EOP-Status-No'] =&gt; Third-term-Persisted</td>
<td>0.58</td>
<td>0.90</td>
<td>0.76</td>
<td>0.27</td>
<td>0.77</td>
</tr>
<tr>
<td>~4</td>
<td>['EOP-Status-Yes'] =&gt; Third-term-Persisted</td>
<td>0.32</td>
<td>0.89</td>
<td>0.56</td>
<td>0.58</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Table 13: Inverse class rules results on third term persistence dataset

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Cosine</th>
<th>IR</th>
<th>Kulc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>['First-Gen'] = Third-term-Persisted</td>
<td>0.65</td>
<td>0.89</td>
<td>0.80</td>
<td>0.17</td>
<td>0.81</td>
</tr>
<tr>
<td>1'</td>
<td>['First-Gen'] = Third-term-Not-Persisted</td>
<td>0.08</td>
<td>0.11</td>
<td>0.29</td>
<td>0.83</td>
<td>0.44</td>
</tr>
<tr>
<td>2</td>
<td>['Female'] = Third-term-Persisted</td>
<td>0.64</td>
<td>0.90</td>
<td>0.80</td>
<td>0.19</td>
<td>0.80</td>
</tr>
<tr>
<td>2'</td>
<td>['Female'] = Third-term-Not-Persisted</td>
<td>0.07</td>
<td>0.10</td>
<td>0.27</td>
<td>0.82</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>['Pell-Yes'] = Third-term-Persisted</td>
<td>0.63</td>
<td>0.88</td>
<td>0.79</td>
<td>0.18</td>
<td>0.80</td>
</tr>
<tr>
<td>3'</td>
<td>['Pell-Yes'] = Third-term-Not-Persisted</td>
<td>0.08</td>
<td>0.12</td>
<td>0.31</td>
<td>0.83</td>
<td>0.46</td>
</tr>
<tr>
<td>4</td>
<td>['EOP-Status-No'] = Third-term-Persisted</td>
<td>0.58</td>
<td>0.90</td>
<td>0.76</td>
<td>0.27</td>
<td>0.77</td>
</tr>
<tr>
<td>4'</td>
<td>['EOP-Status-No'] = Third-term-Not-Persisted</td>
<td>0.06</td>
<td>0.10</td>
<td>0.25</td>
<td>0.79</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Table 14 Closed association rules on fifth term persistence dataset

<table>
<thead>
<tr>
<th>Rule Number</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Cosine</th>
<th>IR</th>
<th>Kulc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>['First-Gen'] = Fifth-term-Persisted</td>
<td>0.60</td>
<td>0.84</td>
<td>0.77</td>
<td>0.14</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>['Pell-Yes'] = Fifth-term-Persisted</td>
<td>0.59</td>
<td>0.84</td>
<td>0.77</td>
<td>0.14</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>['EOP-Status-No'] = Fifth-term-Persisted</td>
<td>0.58</td>
<td>0.83</td>
<td>0.76</td>
<td>0.15</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>['Female'] = Fifth-term-Persisted</td>
<td>0.60</td>
<td>0.86</td>
<td>0.78</td>
<td>0.15</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Table 15: Contrast rules results on fifth term persistence dataset

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Cosine</th>
<th>IR</th>
<th>Kulc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>['First-Gen'] = Fifth-Term-Persisted</td>
<td>0.597</td>
<td>0.836</td>
<td>0.768</td>
<td>0.137</td>
<td>0.771</td>
</tr>
<tr>
<td>~1</td>
<td>['Not-first-Gen'] = Fifth-Term-Persisted</td>
<td>0.249</td>
<td>0.869</td>
<td>0.506</td>
<td>0.633</td>
<td>0.582</td>
</tr>
<tr>
<td>2</td>
<td>['Pell-Yes'] = Fifth-Term-Persisted</td>
<td>0.593</td>
<td>0.835</td>
<td>0.765</td>
<td>0.142</td>
<td>0.768</td>
</tr>
<tr>
<td>~2</td>
<td>['Pell-No'] = Fifth-Term-Persisted</td>
<td>0.253</td>
<td>0.871</td>
<td>0.511</td>
<td>0.628</td>
<td>0.585</td>
</tr>
<tr>
<td>3</td>
<td>['Female'] = Fifth-Term-Persisted</td>
<td>0.601</td>
<td>0.858</td>
<td>0.781</td>
<td>0.154</td>
<td>0.785</td>
</tr>
<tr>
<td>~3</td>
<td>['Male'] = Fifth-Term-Persisted</td>
<td>0.244</td>
<td>0.816</td>
<td>0.486</td>
<td>0.606</td>
<td>0.553</td>
</tr>
<tr>
<td>4</td>
<td>['EOP-Status-No'] = Fifth-Term-Persisted</td>
<td>0.581</td>
<td>0.833</td>
<td>0.757</td>
<td>0.153</td>
<td>0.76</td>
</tr>
<tr>
<td>~4</td>
<td>['EOP-Status-Yes'] = Fifth-Term-Persisted</td>
<td>0.264</td>
<td>0.876</td>
<td>0.523</td>
<td>0.616</td>
<td>0.594</td>
</tr>
</tbody>
</table>

Table 16: Inverse class rules results on fifth term persistence dataset

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Cosine</th>
<th>IR</th>
<th>Kulc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>['First-Gen'] = Fifth-Term-Persisted</td>
<td>0.60</td>
<td>0.84</td>
<td>0.77</td>
<td>0.14</td>
<td>0.77</td>
</tr>
<tr>
<td>1'</td>
<td>['First-Gen'] = Fifth-Term-Not-Persisted</td>
<td>0.12</td>
<td>0.16</td>
<td>0.35</td>
<td>0.75</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>['Pell-Yes'] = Fifth-Term-Persisted</td>
<td>0.59</td>
<td>0.84</td>
<td>0.77</td>
<td>0.14</td>
<td>0.77</td>
</tr>
<tr>
<td>2'</td>
<td>['Pell-Yes'] = Fifth-Term-Not-Persisted</td>
<td>0.12</td>
<td>0.17</td>
<td>0.35</td>
<td>0.74</td>
<td>0.46</td>
</tr>
<tr>
<td>3</td>
<td>['EOP-Status-No'] = Fifth-Term-Persisted</td>
<td>0.58</td>
<td>0.83</td>
<td>0.76</td>
<td>0.15</td>
<td>0.76</td>
</tr>
<tr>
<td>3'</td>
<td>['EOP-Status-No'] = Fifth-Term-Not-Persisted</td>
<td>0.12</td>
<td>0.17</td>
<td>0.36</td>
<td>0.74</td>
<td>0.46</td>
</tr>
<tr>
<td>4</td>
<td>['Female'] = Fifth-Term-Persisted</td>
<td>0.60</td>
<td>0.86</td>
<td>0.78</td>
<td>0.15</td>
<td>0.79</td>
</tr>
<tr>
<td>4'</td>
<td>['Female'] = Fifth-Term-Not-Persisted</td>
<td>0.10</td>
<td>0.14</td>
<td>0.30</td>
<td>0.72</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Similar to Table 11, the closed association rules on the fifth term persistence dataset, the result indicates that a student who is first generation or Pell eligible, and persisting in fifth term is a balanced and positive correlated rule. Also, female students and students who are not in educational opportunity programs persisting in fifth term is also obtained as a balanced and positively correlated rule.
Closed association rules on seventh term persistence dataset along with the first generation and Pell eligible students persisting in seventh term rules, also gives health students persisting in seventh term as one of the positively balanced rule. Health students persisting in seventh term is in line to the results from Table 20.
In Table 20, which shows the closed association rules on graduation dataset, students in Health department and female students graduating are two important rules. Also, one of the surprising findings is Pell eligible students not graduating. This observation along with
results from Table 10 suggests that Pell eligible students have no difficulty in persisting in third term but they finally find it harder to graduate.

Table 22 Inverse class rules results on graduation term persistence dataset

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Cosine</th>
<th>IR</th>
<th>Kule</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>['Health'] =&gt; Graduated</td>
<td>0.45</td>
<td>0.55</td>
<td>0.65</td>
<td>0.26</td>
<td>0.66</td>
</tr>
<tr>
<td>1'</td>
<td>['Health'] = Not-Graduated</td>
<td>0.37</td>
<td>0.45</td>
<td>0.63</td>
<td>0.44</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>['Female'] =&gt; Graduated</td>
<td>0.44</td>
<td>0.61</td>
<td>0.69</td>
<td>0.17</td>
<td>0.69</td>
</tr>
<tr>
<td>2'</td>
<td>['Female'] = Not-Graduated</td>
<td>0.28</td>
<td>0.39</td>
<td>0.50</td>
<td>0.33</td>
<td>0.52</td>
</tr>
<tr>
<td>3</td>
<td>['EOP-Status-No'] =&gt; Graduated</td>
<td>0.43</td>
<td>0.61</td>
<td>0.68</td>
<td>0.15</td>
<td>0.68</td>
</tr>
<tr>
<td>3'</td>
<td>['EOP-Status-No'] = Not-Graduated</td>
<td>0.27</td>
<td>0.39</td>
<td>0.50</td>
<td>0.32</td>
<td>0.51</td>
</tr>
<tr>
<td>4</td>
<td>['Pell-Yes'] =&gt; Not-Graduated</td>
<td>0.34</td>
<td>0.47</td>
<td>0.61</td>
<td>0.36</td>
<td>0.63</td>
</tr>
<tr>
<td>4'</td>
<td>['Pell-Yes'] = Graduated</td>
<td>0.38</td>
<td>0.53</td>
<td>0.60</td>
<td>0.16</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 21 represents the contrast rules generated on the closed association rules of graduation dataset. These rules helps to evaluate the importance of the rules in Table 21. For example, Pell eligible students not graduating is one of the important rules in Table 21. Table 22 looks at non-Pell eligible students not graduating. The latter seems to be a negatively associated rule but more balanced. Furthermore from Table 22, Pell eligible students graduating rule is analyzed and this shows that the rule is positively correlated and very balanced.
In subset based frequent item sets, it is noted that ELM score being C is one among the frequent items occurring in the following datasets, namely, students who left in the 4th term dataset, students who left in the 7th term dataset, students who left any term dataset and students who are still in school dataset. It is also found that in graduated students' dataset percentage of pell eligible students are much lower than when compared to the actual dataset. Also, it is found that ELM score being C and EPT score being C are two frequently occurring patterns in the students who are still in school dataset. Table 23 shows these patterns along with their support in the actual dataset. The identification of these pattern are very important since practitioners can use this to help students who have lower ELM score and/or EPT score by providing additional tutoring services for English or Math courses.

To compare association rules with decision tree based rules, a Weka implementation of decision trees is employed. The J48 model, with 5-fold cross-validation (CV) split is utilized for this analysis. The J48 model which is based on the C4.5, a supervised learning decision tree model with 5-fold cross-validation (CV) split was utilized for this analysis. In 5-fold CV, the data is split into 5 parts: each part is held out for testing and the remaining four parts are used for training the model. This process is repeated 5 times such that each part from the split is used only once in testing and four times in training. The main datasets
namely third term, fifth term, seventh term persistence and graduation datasets were analyzed for patterns. Please find below the accuracy and F1-measure of the J48 model for the different datasets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Accuracy</th>
<th>F1-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third term persistence dataset</td>
<td>88%</td>
<td>0.94</td>
</tr>
<tr>
<td>Fifth term persistence dataset</td>
<td>84%</td>
<td>0.91</td>
</tr>
<tr>
<td>Seventh term persistence dataset</td>
<td>78%</td>
<td>0.87</td>
</tr>
<tr>
<td>Graduation dataset</td>
<td>54%</td>
<td>0.63</td>
</tr>
</tbody>
</table>

The accuracy and F1-measure of the J48 model is higher than Naïve Bayesian, ExtraTrees and K Neighbours, and is lower than Adaboost and Linear SVC models for persistence datasets. Additionally, given that all the datasets are relatively small, the rules generated by J48 model are prone to overfitting [54]. Overfitting is caused when decision trees are too specific to the training data that they may not perform well on real-world instances. To lower the effect of overfitting, pruning is performed on the J48 model results. Pruning involves removing any nodes (including the subtrees of that node) that have no negative effect on the accuracy of the decision tree.

Figure 19 shows the J48 pruned tree for the third term persistence datasets. The feature values connected by the dotted line are the nodes at the same level of the decision tree. For example, Dpt (Department) = Science, Dpt = Health, Dpt = CAD are all nodes at the same
level in the tree. The root node for this tree is the feature “income”.

Fig. 19 J48 pruned Decision tree for third term persistence dataset.

The J48 decision tree is constructed on the basis of gain ratio. For the decision tree on
the third term persistence dataset, the feature income has the highest gain ratio among all
attributes. Hence the initial partition of the dataset is done according to income, and
subsequently, the attribute with the highest gain ratio in a given partition is used to grow
each subtree. So in the third term persistence dataset, in the subtree with Income-Unknown,
Pell eligible is the next feature with the highest information gain. The path from the root to
the leaf node is a decision rule with class label as the outcome. For example, given below
is one of the decision rules:

\( \text{if } \text{Income} = \text{Income-Unknown and Pell eligible} = \text{Pell-No} \) then outcome is persisted

Some of the feature values have two numbers in parenthesis next to them. Here, the first number is the total number of instances reaching the leaf and the second number is the number of those instances that are misclassified. Few of the relevant decision rules from the above decision tree are as follows:

<table>
<thead>
<tr>
<th>Rule #</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>If Income-Less-than-or-equal-to-$30000 then Persisted</td>
</tr>
<tr>
<td>Rule 2</td>
<td>If Income = Income-Unknown and Pell eligible = Pell-No then Persisted (123.0/11.0)</td>
</tr>
<tr>
<td>Rule 3</td>
<td>If Income = Income-Unknown and Pell eligible = Pell-Yes and Department = Health1 and ELM = ELM-B then Persisted (26.0/3.0)</td>
</tr>
</tbody>
</table>

Each rule generated by the decision tree is disjoint, in the sense that a student who is part of Rule 1 will not appear in Rule 2 and Rule 3.

These decision rules for third term persistence dataset is compared to association rules for the same from Table 11, and it is noticed that there are no common rules in both these analysis. All decision rules for the third term persistence dataset has income feature in the rule whereas income feature was not prominent as a relevant rule in association rules analysis for the same dataset. But few of the association rules are subsets of the decision rules. For example, when we compare the decision rule 3 and the association rule \( [\text{Pell-} \)
Yes') \Rightarrow \text{Third-term-Persisted} \text{ for third term persistence dataset, it is observed that this association rule is actually the superset of the decision rule 3.}

**Fig. 20** J48 pruned Decision tree for fifth term persistence dataset.

```plaintext
<table>
<thead>
<tr>
<th>Income = Income-Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pell eligible = Pell-No</td>
</tr>
<tr>
<td>Dpt = Health</td>
</tr>
<tr>
<td>ELM = ELM-C</td>
</tr>
<tr>
<td>EPT = EPT-C</td>
</tr>
<tr>
<td>Race = Hispanic: Not Persisted (10.0/4.0)</td>
</tr>
<tr>
<td>Race = Asian: Not Persisted (4.0/2.0)</td>
</tr>
<tr>
<td>Race = African-American: Persisted (1.0)</td>
</tr>
<tr>
<td>Race = Unknown: Persisted (3.0)</td>
</tr>
<tr>
<td>Race = Two-or-More-Races: Not Persisted (2.0/1.0)</td>
</tr>
<tr>
<td>Race = Race-White</td>
</tr>
<tr>
<td>EOP = EOP-Status-Yes: Persisted (3.0/1.0)</td>
</tr>
<tr>
<td>EOP = EOP-Status-No: Not Persisted (2.0)</td>
</tr>
<tr>
<td>Race = Native-Hawaiian: Persisted (0.0)</td>
</tr>
<tr>
<td>EPT = EPT-A: Not Persisted (3.0)</td>
</tr>
<tr>
<td>EPT = EPT-B</td>
</tr>
<tr>
<td>Start term = Fall: Not Persisted (4.0/1.0)</td>
</tr>
<tr>
<td>Start term = Spring: Persisted (3.0)</td>
</tr>
<tr>
<td>ELM = ELM-B</td>
</tr>
<tr>
<td>EPT = EPT-C</td>
</tr>
<tr>
<td>Start term = Fall</td>
</tr>
<tr>
<td>Edu. Mother = Mother-No-High-School: Persisted (1.0)</td>
</tr>
<tr>
<td>Edu. Mother = Mother-Some-High-School: Not Persisted (0.0)</td>
</tr>
<tr>
<td>Edu. Mother = Mother-Some-College: Not Persisted (2.0)</td>
</tr>
<tr>
<td>Edu. Mother = Mother-High-School-Grad: Not Persisted (2.0)</td>
</tr>
<tr>
<td>Edu. Mother = Mother-4-Yr-College-Grad: Persisted (1.0)</td>
</tr>
<tr>
<td>Edu. Mother = Mother-Edu-Unknown: Not Persisted (0.0)</td>
</tr>
<tr>
<td>Edu. Mother = Mother-Postgraduate: Not Persisted (0.0)</td>
</tr>
<tr>
<td>Edu. Mother = Mother-2-Yr-College-Grad: Not Persisted (0.0)</td>
</tr>
<tr>
<td>Start term = Spring: Persisted (9.0/1.0)</td>
</tr>
<tr>
<td>EPT = EPT-A: Persisted (3.0)</td>
</tr>
<tr>
<td>EPT = EPT-B: Persisted (8.0)</td>
</tr>
<tr>
<td>ELM = ELM-A: Persisted (9.0/2.0)</td>
</tr>
<tr>
<td>Dpt = CAD</td>
</tr>
<tr>
<td>Gender = Female</td>
</tr>
<tr>
<td>EOP = EOP-Status-Yes: Not Persisted (6.0/1.0)</td>
</tr>
<tr>
<td>EOP = EOP-Status-No: Persisted (11.0/3.0)</td>
</tr>
<tr>
<td>Gender = Male: Not Persisted (3.0)</td>
</tr>
<tr>
<td>Dpt = Science: Not Persisted (10.0)</td>
</tr>
<tr>
<td>Pell eligible = Pell-Yes</td>
</tr>
<tr>
<td>Income = Income-Less-than-or-equal-to-$30000: Persisted (108.0/3.0)</td>
</tr>
<tr>
<td>Income = Income-$70001-or-higher: Persisted (41.0/2.0)</td>
</tr>
<tr>
<td>Income = Income-$30001-$50000: Persisted (67.0/4.0)</td>
</tr>
<tr>
<td>Income = Income-$50001-$70000: Persisted (37.0)</td>
</tr>
<tr>
<td>Income = Income-$70001-or-higher: Persisted (41.0/2.0)</td>
</tr>
</tbody>
</table>
```
Similar to decision rules on third term persistence dataset, the decision rules on fifth and seventh term persistence also have Income feature as the splitting attribute. It is observed that the decision rules for these datasets do not have any common rules with the association rules for the respective datasets. But it is noticed that the rules for seventh term in the path

\textit{If Income = Income-Unknown and Pell eligible = Pell-Yes and ELM = ELM-C}, majority of these students do not persist which supports the findings from the student flowchart based analysis in figure 8 that shows that students with ELM score of C find it harder to graduate.

Fig. 21 J48 pruned Decision tree for seventh term persistence dataset.

```
Income = Income-Unknown
| Pell eligible = Pell-Yes
| | ELM = ELM-C
| | EPT = EPT-C
| | | Race = Hispanic: Not Persisted (13.0/2.0)
| | | Race = Asian: Persisted (4.0/1.0)
| | | Race = African-American: Not Persisted (2.0/1.0)
| | | Race = Unknown: Persisted (4.0/1.0)
| | | Race = Two-or-More-Races: Not Persisted (3.0/1.0)
| | | Race = Race-White: Not Persisted (4.0/1.0)
| | | Race = Native-Hawaiian: Not Persisted (0.0)
| | EPT = EPT-A: Not Persisted (3.0)
| EPT = EPT-B
| | EOP = EOP-Status-Yes: Persisted (2.0)
| | EOP = EOP-Status-No: Not Persisted (8.0/3.0)
| ELM = ELM-B: Persisted (29.0/7.0)
| ELM = ELM-A: Persisted (11.0/3.0)
| Pell eligible = Pell-No: Persisted (64.0/13.0)
Income = Income-Less-than-or-equal-to-$30000: Persisted (55.0/1.0)
Income = Income-$70001-or-higher: Persisted (16.0/1.0)
Income = Income-$30001-$50000: Persisted (41.0/1.0)
Income = Income-$50001-$70000: Persisted (14.0)
```
Fig. 22 J48 pruned Decision tree for graduation term persistence dataset.

```
ELM = ELM-B
  dept = CAD: Graduated (9.0/1.0)
  dept = Health
    Edu. Father = Father-2-Yr-College-Grad: Graduated (2.0)
    Edu. Father = Father-Some-High-School: Not-Graduated (7.0/1.0)
    Edu. Father = Father-Unknown
      EOP = EOP-Status-No: Graduated (2.0)
      EOP = EOP-Status-Yes: Not-Graduated (2.0)
    Edu. Father = Father-Postgraduate: Graduated (0.0)
    Edu. Father = Father-Some-College: Graduated (6.0/1.0)
  Edu. Father = Father-4-Yr-College-Grad
    EPT = EPT-B: Not-Graduated (2.0)
    EPT = EPT-C: Not-Graduated (2.0)
    EPT = EPT-A: Graduated (2.0/1.0)
  Edu. Father = Father-No-High-School
    EPT = EPT-B: Graduated (3.0)
    EPT = EPT-C: Not-Graduated (3.0/1.0)
    EPT = EPT-A: Graduated (0.0)
  Edu. Father = Father-High-School-Grad
    Pell = Pell-No
      Start term = Fall: Graduated (3.0/1.0)
      Start term = Spring: Not-Graduated (2.0)
    Pell = Pell-Yes: Graduated (4.0)
```
Fig. 23 J48 pruned Decision tree for graduation term persistence dataset (continued 1)

```
ELM = ELM-C
Pel = Pell-No. Graduated (26.0/8.0)
Pell = Pell-Yes
  | Race-Unknown = Hispanic/Latino: Not-Graduated (23.0/5.0)
  | Race-Unknown = Race-Unknown
  | dept = CAD: Not-Graduated (4.0/1.0)
  | dept = Health: Graduated (12.0/2.0)
  | Race-Unknown = Asian-Only
  | Edu. Father = Father-2-Yr-College-Grad: Not-Graduated (0.0)
  | Edu. Father = Father-Some-High-School: Not-Graduated (5.0)
  | Edu. Father = Father-Unknown: Graduated (2.0)
  | Edu. Father = Father-Postgraduate: Not-Graduated (0.0)
  | Edu. Father = Father-Some-College: Not-Graduated (0.0)
  | Edu. Father = Father-4-Yr-College-Grad: Not-Graduated (0.0)
  | Edu. Father = Father-No-High-School: Graduated (2.0)
  | Edu. Father = Father-High-School-Grad: Graduated (4.0/2.0)
  | Race-Unknown = White-Only
  | Edu. Mother = Mother-Some-College: Graduated (1.0)
  | Edu. Mother = Mother-No-High-School: Not-Graduated (0.0)
  | Edu. Mother = Mother-Unknown: Not-Graduated (2.0)
  | Edu. Mother = Mother-Postgraduate: Not-Graduated (0.0)
  | Edu. Mother = Mother-4-Yr-College-Grad: Not-Graduated (0.0)
  | Edu. Mother = Mother-Some-High-School: Graduated (1.0)
  | Edu. Mother = Mother-High-School-Grad: Not-Graduated (3.0)
  | Edu. Mother = Mother-2-Yr-College-Grad: Not-Graduated (0.0)
  | Race-Unknown = Two-or-More-Races: Not-Graduated (5.0/2.0)
  | Race-Unknown = African-American-Only
  | Edu. Mother = Mother-Some-College: Graduated (2.0)
  | Edu. Mother = Mother-No-High-School: Graduated (0.0)
  | Edu. Mother = Mother-Unknown: Not-Graduated (4.0/1.0)
  | Edu. Mother = Mother-Postgraduate: Graduated (0.0)
  | Edu. Mother = Mother-4-Yr-College-Grad: Graduated (0.0)
  | Edu. Mother = Mother-Some-High-School: Graduated (0.0)
  | Edu. Mother = Mother-High-School-Grad: Graduated (0.0)
  | Edu. Mother = Mother-2-Yr-College-Grad: Graduated (0.0)
  | Race-Unknown = Native-Hawaiian: Not-Graduated (0.0)
```
Fig. 24 J48 pruned Decision tree for graduation term persistence dataset (continued 2)

Unlike the persistence datasets, the graduation datasets have ELM scores as the first splitting criterion. The decision rule \( \text{if ELM score}=A \text{ and gender}=\text{female} \text{ then persisted} \), can be considered as a subset of the association rule \( \text{gender}=\text{Female} \Rightarrow \text{persisted} \). The majority of students who follows the decision rule path \( \text{ELM score}=A \) graduate from the university which supports the students flowchart analysis from fig 5.

We notice from the decision tree rules-based analysis that the attributes in the rules are a subset of the consequent attributes from class association rules. Class association rules tend to be more generic than decision rules. Instances satisfying one association rule can satisfy another class association rule for the same dataset while that is not the case for decision rules. Decision rules yield distinct path leading to one of the class label or outcome.

To summarize, there are no matching rules between decision tree generated rules and association rules for the different datasets. But some association rules are a superset of the
decision rules. For example, the association rule ['Pell-Yes'] \Rightarrow \text{Third-term-Persisted} for third term persistence dataset, is actually the superset of the decision rule 3 from Table 23. It is also observed that some rules generated from decision trees corroborate some of the findings from analyses like the student flowchart for seventh term persistence and graduation dataset. Furthermore, in addition to not having an exact match, some association rules like ['First Generation' = Yes] \Rightarrow \text{Seventh-term-Persisted}, seems to have no connection to the decision rules. Overall, the decision tree analysis supports the student flowchart analysis and the association rules analysis to a certain extent.
Chapter 5

Installation

This chapter describes how to use any generic dataset to perform analysis discussed in the study. The chapter starts with the system requirements, followed by the input formats, instructions on how to run the programs and finally concludes with screenshots of the steps to further illustrate the process.

5.1. System Requirements

The following software and packages need to be installed to run the analysis: Java 1.7, Python 2.7+, Scikit-learn, and Panda. For data visualization, an interactive business intelligence tool named Tableau was used, even though any data visualization tool including excel can be substituted. Flowcharts were constructed with the help of a free online diagram software known as Draw.io. For feature selection the FSP jar file, implemented by Suchi Vora as part of her Master’s thesis work at San Francisco State University, found in the github repository (https://github.com/SuchiVora/Feature-Selection-Pipeline) is to be used. To generate the closed itemsets, the apriori executable created by Christian Borgelt [36] is required. All the tools and packages above are available for Windows, Macintosh and Unix distributions except for Tableau which is not available on Unix. Open Office may be used in its place for data visualization on Unix distributions.
5.2. Dataset Formats

The input dataset format depends on the analysis to be performed. The various data format needed for each analysis are:

1. Dataset for data visualization:

Metro College Success Program uses Salesforce as a backend to store the data. Data from Salesforce can be connected directly to Tableau desktop using Salesforce login credentials. You can also export out the required data from Salesforce into Excel or CSV. Tableau can take input data from multiple data sources like Excel, CSV, MySQL etc. Hence any input format that Tableau permits can be used for data visualization.

2. Dataset for feature ranking and classification:

<table>
<thead>
<tr>
<th>Edu level of mother</th>
<th>Edu level of father</th>
<th>First Generation</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>no school</td>
<td>no school</td>
<td>First Generation</td>
<td>Not Graduated</td>
</tr>
<tr>
<td>college grad</td>
<td>4yr college grad</td>
<td>First Generation</td>
<td>Graduated</td>
</tr>
<tr>
<td>high school grad</td>
<td>postgraduate</td>
<td>First Generation</td>
<td>Graduated</td>
</tr>
<tr>
<td>unknown</td>
<td>no school</td>
<td>First Generation</td>
<td>Graduated</td>
</tr>
</tbody>
</table>

For feature ranking, the FSP jar file[38] accepts the data in the csv file format. The classification algorithms also accepts data in a similar format. Each csv file should have a header row followed by the data rows. The last column in the csv file should
be the class label or target variable. The column name of the class label should be "class". All the feature values in the dataset used in this study is categorical. The sample dataset is shown in Table 26. For classification the input file format is again the csv but the dataset is split into the attribute values and class labels. Create these files in separate folders but provide the same name for the attribute value file and class label file of a particular dataset. Also, ensure that the class labels are in binary.

3. Dataset for association pattern mining:

Table 27 Input Data format for association pattern mining

|-------------------------------|-------------------------------|----------------------|---------------|

The first step of the association pattern mining is running the data through the apriori algorithm. Though the algorithm takes a csv file as input similar to the feature ranking analysis, the format is different than above. No header rows is needed for the dataset. All data rows should have their column name appended to them. This would help in correct identification of feature values in the closed item sets results, if two columns have identical feature values. Additionally, any space
needs to be replaced by "-", to prevent apriori algorithm considering feature value with spaces as two different items. The sample dataset is shown in Table 27.

5.3. Instructions to run the programs

The following section details the run instructions for the feature ranking, classification and prediction and the association pattern mining algorithms.

1. Feature Ranking:

   a) Download the FSP jar from the following github repository

   \[https://github.com/SuchiVora/Feature-Selection-Pipeline\]

   b) Create two folders namely /data and /results in the same location as the jar file

   c) Store the csv dataset for feature ranking file in the /data folder

   d) In command prompt, navigate to the location where the FSP jar is stored

   e) Run the following command:

   \[java -jar feature_selection.jar [name of the dataset] [csv dataset file]\]

   For example: \[java -jar feature_selection.jar GradResults grad.csv\]

   f) The program will prompt the user to choose a feature selection algorithm. Select the algorithm that one would like to run.

   g) After the algorithm is selected and the enter button is hit, the results gets stored in the /results folder.
h) The output in /results/performanceresults.txt includes accuracy, precision, recall and F1-measure. The output in /results/[Feature-selection-algo-name]_optimalfeatures.csv includes the list of features that were identified as optimal. Screenshots of running FSP jar file for grad.csv dataset is given below:

Fig. 25 Screenshot of running the command to execute the FSP jar file
Fig. 26 Screenshot of running the FSP jar file after selecting the algorithm

Fig. 27 Screenshot after completion of running the FSP jar file

2. Classification and Prediction:

a) Create a folder to store your input and output files.
For example: C:/Users/User1/Anaconda/Results where Results is the new folder created by the user.

b) Inside this folder create two sub-folders “/trainingset” and “/truelabels”

c) Place the classification dataset without the class labels into the “/trainingset” subfolder and the truelabels in this “/truelabels” subfolder. For example if the dataset is grad.csv then the following would be the path to the grad.csv file:

C:/Users/User1/Anaconda/Results/trainingset/grad.csv

C:/Users/User1/Anaconda/Results/truelabels/grad.csv

d) Download the project from the following github repository:

https://github.com/AparnaThricovil/data-mining

e) Open the command prompt and run the python script “python ClassificationYearly.py”.

The program will ask for the path to the input file. Specify the path to the “Results” folder (folder created in Step 1). The program would automatically navigate to the datasets folder created inside of results to access the input csv file.

f) The program will execute and all the results would be stored inside C:/Users/User1/Anaconda/Results/alloutputfiles.
g) There two main output results files. The accuracy.txt file includes the accuracy of all the five classifiers for the given datasets. The metrics.csv file contains the F1-measure, Precision and Recall values of the five classifiers on the given dataset. Screenshot of running the classification algorithms for grad.csv dataset is given below:

![Fig. 28 Screenshot of running the classification algorithms for grad.csv dataset](image)

3. Association pattern mining

a) Download and the apriori python file from Christian Borgelt’s website (http://www.borgelt.net/apriori.html).

b) Place the input to the apriori algorithm in the required format at the desired location.

c) In a command prompt, navigate to the path where the apriori file is stored and run the algorithm as shown below to create the closed item sets:
apriori -tc -m2 [path to input csv file] -k"," [path to output file]

For example:

apriori -tc -m2 C:Users/User1/results/input/Grad.csv -k"," C:Users/User1/results/output/grad.txt.

This example generates the closed item sets for the input dataset.

d) In a command prompt, navigate to the path where the apriori file is stored and run the algorithm as shown below to create the frequent item sets for the sub-group based dataset:

apriori -ts [path to input csv file] -k"," [path to output file]

e) Download the project from the following github repository:

https://github.com/AparnaThricovil/data-mining. This is the same repository as the one used in the classification step. Keep the apriori results and the initial input files in the two folders under a root folder. For e.g; the input files can be in /AprioriResults/InputFiles folder and output from apriori in /AprioriResults/AprioriOutput folder structure. To generate closed association rules, inverse and contrast rule for the dataset, run the following python script:

python associationRules.py.
f) This would prompt an input from the user on the class variable for which rules needs to be generated. The following 2 inputs are needed from the user:

a) A number input corresponding to: 1. Graduated or not, 2. Third term persisted or not 3. Fifth term persisted or not and 4. Seventh term persisted or not.

b) Path containing both the input and the apriori results files. This path from the e.g, mentioned in Step e would be /AprioriResults.

Depending on the options chosen, the program generates class association rules, inverse and contrast rules on the closed item sets.

g) If the path to input files was specified as C:/Users/User1/Anaconda/AprioriResults, then output files will be created inside this folder with the folder name AprioriOutput.

h) For decision tree-based analysis, use the dataset to run the J48 classifier in Weka. Select 5-fold cross-validation split. The pruned decision tree is available in Weka’s UI.

Three output files one for each of class association rules, inverse and contrast rules on the closed item sets will be stored in the output folder. The output files will be in csv format and will include the values of the measures like support, IR, Kulc, confidence and cosine.
Chapter 6

Conclusion and Future Work

This study presents a multifaceted approach that uses various data mining techniques combined with observations and insights from years of practitioner's experience to identify potential factors that are causing students to drop out of college. The work described here is done in collaboration with the Metro College Success Program at San Francisco State University in California. The main goal of this program is to increase the graduation rates of first-generation, low-income, underrepresented minority students. To this end, persistence was thought to be a critical factor for graduation. The approaches taken in this study (student flowchart, bivariate visualization, feature ranking, classification and prediction, association pattern based analysis and evaluation criteria) are generalizable to institutions across the world. While the factors, such as test scores, and data points might vary based on location, these approaches can be replicated with sizable impact. Additionally, as this study focuses on interpretability for practitioners, the results can enable immediate action and impact.

The Metro program can use these results to design interventions to help improve the rate of graduation of students. For example, one of the results suggests that students with low remedial scores, especially in mathematics, drop out of college. In the future, practitioners can use this result to justify the need for additional support on remedial course completion. Similarly, Metro can provide more support for Pell eligible students as the results show
that these students do not graduate as much as the average population. The other major findings of this study include:

(1) Using the student flowchart and bivariate visualization, the study found out that out of the students who leave the institution, most leave in the fourth and seventh terms. This has made these two terms crucial from the perspective of retaining a student.

(2) The ensemble-based feature ranking of seven algorithms has revealed that the education level of mother, ELM (standardized Entry Level Math test) score and race as the top most relevant features with respect to students' third term persistence.

(3) Out of the five classification and prediction models employed for this study, Naïve Bayesian predicts graduation the best while AdaBoost and Linear SVM models are most suited for predicting persistence.

(4) Association patterns indicate that ELM score and Pell eligibility negatively affect a student's potential to graduate.

Additionally, the analysis done in this study can be expanded in the future in a number of ways:

1. Firstly, a comparative analysis based on a comparison group can be undertaken to understand the statistical significance of the results. Metro College Success Program is currently collecting data on a comparison group which will enable this analysis.

2. Current analysis is based on data from 651 students from 2009 to 2013. With each year, data available for this study is growing rapidly as Metro College Success
Program expands to more students in San Francisco State University. This study can be repeated with additional data to help draw conclusions, especially for subgroups which may be too small as of now.

3. This study developed a prediction algorithm that can be used to predict the persistence and graduation of students. The accuracy of this algorithm has been tested with standard cross-validation techniques but it has not been tested with newly generated data. A further study can validate the algorithm by running it on additional data and examining the accuracy.

4. The study results can be enriched by using more pertinent features in the analysis. One such feature that Metro practitioners have recently identified is probation information. Some of the other potentially important features that can be added include: high school education information, college attendance data, data on the academic and non-academic advising hours attended, physico-social factors, etc.

In the long run, we hope data-driven studies such as this work will motivate more and more educational institutions around the world to utilize data mining techniques to improve their understanding of the underlying issues, thereby more effectively helping their students reach their educational goals.
References


48. Tianyi Wu, Yuguo Chen and Jiawei Han, "Re-examination of interestingness measures in pattern mining: a unified framework", Data Mining Knowledge Discovery 2010.


