Using a Latent Class Forest to Identify At-Risk Students in Higher Education

Kevin Pelaez
San Diego State University
kpelaez10@gmail.com

Maureen Guarcello
San Diego State University
mguarcello@sdsu.edu

Juanjuan Fan
San Diego State University
jjfan@sdsu.edu

Richard A. Levine
San Diego State University
rlevine@sdsu.edu

Mark Laumakis
San Diego State University
mlaumakis@sdsu.edu

Higher education institutions often examine performance discrepancies of specific subgroups, such as students from underrepresented minority and first-generation backgrounds. An increase in educational technology and computational power has promoted research interest in using data mining tools to help identify groups of students who are academically at-risk. Institutions can then implement data-informed decisions to help promote student access, increase retention and graduation rates, and guide intervention programs. We introduce a latent class forest, a latent class analysis and a random forest ensemble that will recursively partition observations into groups to help identify at-risk students. The procedure is a form of model-based hierarchical clustering that relies on latent class trees to optimally identify subgroups.

We motivate and apply our latent class forest method to identify key demographic and academic characteristics of at-risk students in a large enrollment, bottleneck introductory psychology course at San Diego State University (SDSU). A post hoc analysis is conducted to measure the efficacy of Supplemental Instruction (SI) across these groups. SI is a peer-led academic intervention that targets historically challenging courses and aims to increase student performance. In doing so, we are able to identify populations that benefit most from SI to guide program recruitment and help increase the introductory psychology course success rate.

Keywords: Latent Class Forest, supplemental instruction, clustering, at-risk students, Latent Class Analysis

1. Introduction

An increase in educational technology and the advancement of data mining and machine learning tools have prompted researcher interest in analyzing patterns within data from educational environments. Educational data mining builds on these ideas by exploring learning analytics, student performance, and educational research to improve school quality (Davies et al., 2017; Papamitsiou and Economides, 2014; Romero and Ventura, 2010). Cluster analysis is a specific data mining tool that creates homogeneous groups based off an observation's features. This analytics tool helps instructors, advisors, and administrators identify at-risk students and recommend appropriate academic and social intervention programs, make appropriate school-wide decisions, and increase the overall quality of education (Braxton, 2000; Romero and Ventura, 2010). A student is considered at-risk if they are less likely to succeed in an academic course or program compared to their counterparts (Braxton, 2000; Romero and Ventura, 2010).

This has prompted our research interest in using machine learning methods within educational settings. In particular, educational researchers are often interested in identifying key factors that increase students' risk of failing a course (Baker and Yacef, 2009; Farsides and Woodfield, 2003; Gray et al., 2014; Gray et al., 2016; Jayaprakash et al., 2014; Romero and Ventura, 2010). For example, rather than using an algorithm to predict whether a particular student will pass a class or not, Tsai et al. (2011) apply clustering algorithms to the outcome of whether undergraduate students will pass a computer literacy class or not. Using these clusters, they extract patterns, such as gender, nationality, and college name, that characterize the risk groups. They then use these patterns to create decision rules that can serve as guidelines for assessing the proficiency levels of future students. Similarly, Chan and Bauer (2014) use hierarchical clustering to create three clusters characterized by the risk levels in the groups (low, medium, and high-risk). Using these clusters, they identify cognitive, social, and academic factors that were unique to the low-risk group and that could be used to help identify, at the beginning of a semester, students who are at risk of failing a course.

We are motivated by the goal of using educational data mining methods, like cluster analysis, to identify at-risk student subgroups. Of particular interest is latent class analysis (LCA), a model-based clustering approach that depends on the expectation-maximization (EM) algorithm to assign likelihoods of class membership. To our knowledge, there are no articles that use LCA specifically to cluster or identify students who are at risk of failing a course in educational settings. However, LCA is often compared to k-means clustering because of its popularity for clustering tasks and because they are both divisive clustering algorithms with similar goals (Schreiber and Pekarik, 2014; Xu, 2011).

For example, Xu (2011), Maull et al. (2010), and Talavera and Gaudioso (2004) all compare LCA to other clustering algorithms within educational settings. Overall, the authors find that LCA outperformed other methods. In particular, LCA provided more meaningful and interpretable results and was not sensitive to variable types. This is of interest when working with diverse data that include categorical, numerical, and other types of variables. Outside of educational applications, Brusco et al. (2016) run simulations, and Schreiber and Pekarik (2014) consider an application, both comparing LCA to other clustering algorithms. The authors note that LCA is computationally expensive, but provides more meaningful and interpretable results when considering real-world applications.

In this paper, we introduce a novel latent class forest (LCF) data mining method. LCF embeds latent class analysis machinery within a random forest ensemble approach to identify at-risk students. In general, LCA outperforms other distance-based clustering algorithms, like k-means clustering, since it provides diagnostic statistics to identify the appropriate number of clusters, provides stronger interpretation ability, and easily incorporates variables at different scales, avoiding variable selection bias and sensitivity to variable types (Xu, 2011; Maull et al., 2010). That said, LCA selects a large number of groups when the sample size and number of features is large and may not converge properly due to the expectation-maximization (EM) procedure (Xu, 2011; Van den Bergh et al., 2017). To overcome these issues and provide stronger classifiers, we introduce a random forest procedure which gains strength by creating an ensemble average of potentially suboptimal trees (Breiman, 2001) and creates a distance matrix over which observations may be easily clustered. Of particular note, the random forest produces stronger and uncorrelated trees that generate more accurate classifications relative to a single tree (Breiman, 2001). Random forest is also competitive among machine learning algorithms generally (Caruana and Niculescu-Mizil, 2006; Caruana et al., 2008; Fernández-Delgado et al., 2014).

We applied the LCF approach to identify at-risk students in a large enrollment, high challenge Introductory Psychology course (PSY 101) at San Diego State University (SDSU). Historically, repeatable and failing grades in this course range from 19% to 40%. Demographic and performance data about the PSY 101 students were used to build the LCF model. We then related the LCF clusters with course performance. The course grade is used as a proxy to determine student success in the class. The model may thus be applied to identify at-risk students in future offerings prior to course enrollment.

Additionally, LCF is used to study the impact of a new Supplemental Instruction (SI) program offered in PSY 101. Of particular interest in this post hoc analysis is the effect of SI on the at-risk subgroups identified by LCF. Like many other intervention programs on campus, SI is resource intensive and constrained by available classroom space. SDSU is thus not able to offer SI in all high challenge courses, and the voluntary nature of the program does not necessarily support all students in a given course. Our first goal with LCF is to create clusters that are categorized by their risk levels to identify any key demographic or academic information that may increase the risk of a student failing PSY 101. Second, we will identify groups of students that will benefit most from SI to guide resource allocation decisions and understand how to best meet the needs of all students.

2. Background: Latent Class Analysis

Xu (2011) performed a study to measure the performance of LCA and k-means clustering methods on educational data sets. Overall, Xu (2011) concluded that LCA outperforms k-means in terms of ability to provide educational interpretation, stability of the algorithm, and sensitivity to input types. In particular, LCA allows for a mixed type of inputs and does not require input standardization to appropriately quantify distances, in contrast to an application of k-means clustering. Xu (2011) used a Davies-Boudlin index to measure the intra- and inter-cluster dispersion and found that LCA is more consistent and provides stronger clusters in all cases considered. Similarly, Maull et al. (2010) used both k-means and LCA to cluster the curriculum planning behavior of teachers. Maull et al. (2010) found that the clustering methods present different interpretations of the data and ultimately concluded that LCA outperforms k-means clustering. Talavera and Gaudioso (2004) used LCA to cluster students using learning management systems. The paper identified six groups of student behavior profiles that provide stronger conclusions, categorized as pass, mixed, or fail, that were not observed using other clustering methods.

Outside of educational settings, Brusco et al. (2016) provided a simulation comparison of k-means, k-median, and LCA as three popular clustering algorithms in psychological research. Overall, the best performance measured by accuracy, interpretation, and efficiency was the k-median method, followed closely by LCA. This conclusion was primarily due to the computationally intensive algorithm required by the LCA method. However, LCA provides diagnostic statistics that help find the optimal number of subgroups, a feature that is not available in the other methods. Schreiber and Pekarik (2014) compared LCA with k-means and hierarchical clustering to understand characteristics of museum visitors. This study found that LCA outperformed both k-means and hierarchical clustering by offering a richer interpretation, providing diagnostic statistics, and avoiding selection bias. Furthermore, the authors have suggested that the current advances in computational efficiency and large data will facilitate the process of using LCA.

As this literature suggests, the k-means algorithm suffers from a number of weaknesses: the total number of groups must be specified, the algorithm is sensitive to outliers and variable type (e.g., number of levels in a multi-category input), inputs must be standardized, and there is no clear approach to identify key characteristics of the clusters. While LCA overcomes these drawbacks, the iterative algorithm requires substantive computational expense to converge and LCA may select a large number of groups if the sample size and number of features are large.

The foundation of the LCF approach we propose in this paper is a hierarchical mixture model. Houseman et al. (2008) introduced a recursively partitioning mixture model (RPMM) algorithm to deal with high-dimensional data analysis problems. Houseman et al. (2008) recursively partitioned the data in a similar process to a decision tree algorithm (see e.g., James et al. (2013), Chapter 8), but used a mixture model using all the inputs to guide the choice of the split rule at each node and an adjusted Bayesian information criterion (BIC) as the stopping rule for tree growth. Unlike traditional clustering approaches, including finite mixture models, this RPMM can optimally select the number of clusters, is more stable, and tends to provide stronger interpretations (Houseman et al., 2008). In particular, Houseman et al. (2008) found that RPMM outperforms all other nonparametric clustering approaches and is at least as reliable as traditional mixture model methods in terms of computational efficiency and interpretation ability.

Koestler et al. (2010) built on the Houseman et al. (2008) RPMM and introduced a semi-supervised RPMM (SS-RPMM). Instead of using all the covariates in the hierarchical model for each split, SS-RPMM strategically selects covariates that have the largest absolute Cox-score to help identify the strongest predictors. Koestler et al. (2010) finds that the SS-RPMM outperforms other semi-supervised methods in terms of efficiency and interpretation.

Van den Bergh et al. (2017) also introduced a recursively partitioning hierarchical clustering method, coined latent class trees (LCT). Similar to Houseman et al. (2008), Van den Bergh et al. (2017) used mixture models to partition the data until the splitting and stopping criteria no longer indicated a better fit. The paper compared a BIC, Akaike information criterion (AIC), and total bivariate residual (TBVR) to evaluate a split. Van den Bergh et al. (2017) found that the BIC yields fewer clusters since BIC has the largest penalty on the number of parameters and sample size, thus terminating the tree earlier and selecting a smaller number of classes. That said, Van den Bergh et al. (2017) suggested exploring different stopping criteria, such as parent and child node size, the convergence of class likelihoods, and other quality of split measures. The latent class forest ensemble we propose is motivated by the LCT approach of Van den Bergh et al. (2017) and the RPMM approach of Houseman et al. (2008).

3. Data

Data were collected from three semesters of PSY 101: Introductory Psychology at SDSU, Fall 2015, Spring 2016, and Fall 2016. The data were deidentified before beginning the analysis to protect the confidentiality of the students. All three semesters were taught by the same instructor, covered the same curriculum, and instituted analogous assessments. PSY 101 is an introductory level course at SDSU that many freshmen and sophomores will take to satisfy their general education and pre-major requirements. This course has been identified by the California State University (CSU) system as a bottleneck course due to a high D, F, or withdrawal rate. In this PSY 101 data set, about 15% of the students earned a D or an F. As a result, many students have to repeat the class for a higher grade, potentially disturbing their four-year graduation plan.

Learning management system (LMS) grade book, demographic, and admissions data were collected for all enrolled PSY 101 students. Since we were interested in identifying at-risk students before they enrolled in PSY 101, we used demographic, admissions, and academic data prior to the start of the semester for this analysis (see Table 1). There were 2017 complete observations, with 171 observations missing partial data and 52 observations missing all the demographic data. The missing values for the 171 observations were imputed (see Section 3.3.) and the 52 incomplete observations were discarded.

3.1. Response and Supplemental Instruction

The response measure used to identify an at-risk student was the final course letter grade the student received in the class. We created two categories: Pass (P) and D, F, or withdrawal (DFW). P corresponds to a grade of C- or higher, and DFW corresponds to a grade of D+ or below as well as withdrawal. This is a standard measure in the California State University system for identifying success in a course.

Supplemental Instruction (SI) was created at the University of Missouri-Kansas City (UMKC) in 1973 as an academic assistance program of facilitated, peer-led study sessions to increase the students' interactions with class content (UMKC, 2017). SI started at San Diego State University (SDSU) in the Fall of 2015 and followed the traditional model created at UMKC. This program targets high challenge, large enrollment courses with high DFW rates. SI sessions are led by undergraduate students who were previously successful in the course, receiving a B+ grade or higher. These SI leaders are trained in active learning strategies, attend class with the students, and encourage students to become actively involved in their education through various cognitive and learning skills (UMKC, 2017). Since SI targets courses, not students, the SI sessions are voluntary, non-remedial, and free for all students (UMKC, 2017).

Furthermore, every undergraduate major at SDSU is "impacted" for both first-time first years and upper division transfers (CSU, 2017b; CSU, 2017a; Guarcello, 2015). Impacted majors are degree programs for which the number of applications received by the SDSU campus is larger than the number of available spaces (CSU, 2018). This bottleneck has affected the university at the course level, e.g., PSY 101, and is slowing down student progress towards graduation. As a result, the CSU began exploring bottleneck courses in 2013 and identified four types of bottlenecks: student readiness, amount of time and when a course was offered, facilities, and advising and scheduling (Guarcello, 2015; Smith and Hanley, 2013). SI aims to reduce these bottleneck effects by increasing retention, improving student academic success, and increasing graduation rates (UMKC, 2017).

3.2. Student Demographics

Students self report their demographic information at the time they apply to the university. Categorical inputs include an honors indicator that notes if a student was accepted into the honors program at SDSU, an indicator noting if the student lived on campus in the residence halls, and self-reported first-generation status. The first-generation no college (FGNC) indicator notes whether any of the student's parents had any college experience. Similarly, the first-generation some college (FGSC) indicator notes whether a student's parents had some college but did not obtain a college degree.

We also have information about whether students participated in the Educational Opportunity Program (EOP) or Compact for Success. The SDSU EOP serves first-generation and low-income students, predominantly from underrepresented minority backgrounds. The California State University system defines underrepresented minority students (URM) as students who are Black or African-American, Latinx/Hispanic, or American Indian/Native American. EOP provides support through academic, financial, and counseling services to promote the recruitment and retention of these populations (SDSU, 2018b). A cohort of incoming students is also selected to participate in a five-week Summer Bridge program. They participate in various college readiness workshops, can earn up to six units of college credit towards their math or English course requirements, and begin to familiarize themselves with the campus and college life.

The Compact for Success program is a partnership program between SDSU and the Sweetwater Unified School District aimed to provide students the Math and English skills necessary to succeed in college. Students initially join in the 7th grade and participate in various college readiness activities and rigorous courses until they graduate high school (SDSU, 2018a). Students in the program are guaranteed admission to SDSU if they have satisfied the math and English proficiency requirements for entrance to the university. Most of the students in this program are also from URM backgrounds, the majority of which are identified as Latinx. Ultimately, the goal of this program is to lower the achievement gap and improve retention and graduation rates of students from underprivileged backgrounds (SDSU, 2018a).

Table 1: Description of variables in the PSY 101 SI study.

Input	Label	Class		Description
Response - Pass/DFW				
Grade	DFW	325	(15%)	DFW and P indicator.
	Pass	1863	(85%)	
Treatment				
Supplemental Instruction	Did Not Attend	1589	(73%)	Supplemental Instruction participation.
	Attended	599	(27%)	
Categorical Inputs				
Gender	Female - 0	1391	(64%)	Self-identified gender.
	Male - 1	797	(36%)	
Honors	No - 0	2105	(96%)	Enrollment in the Honors program.
	Yes - 1	83	(4%)	
Disabled	No - 0	2156	(99%)	Registration with disability services.
	Yes - 1	32	(1%)	
EOP	No - 0	1978	(90%)	Participation in the Educational Opportunity Program (EOP).
	Yes - 1	210	(10%)	
Dorm	No - 0	1200	(55%)	On-campus housing in the dorms.
	Yes - 1	988	(45%)	
Student Level	First Year - 0	1194	(55%)	Student level at SDSU.
	Sophomore - 1	730	(33%)	
	Junior - 2	169	(8%)	
	Senior - 3	95	(4%)	
Major Status	Pre-Major - 0	476	(22%)	Major status.
	Major - 1	1712	(78%)	
Ethnicity	Afr Am - 0	92	(4%)	Self-identified ethnicity.
	Asian - 1	100	(5%)	
	Fil - 2	199	(9%)	
	Intr - 3	96	(4%)	
	Mex Am - 4	520	(24%)	
	Mult - 5	166	(8%)	
	Nat Am - 6	17	(1%)	
	Oth Hisp - 7	115	(5%)	
	Other - 8	56	(3%)	
	PI, Nat HW - 9	9	(<1%)	
	SE Asian - 10	58	(3%)	
	White - 11	760	(35%)	
FGNC	No - 0	1848	(84%)	First-generation, parents had no college.
	Yes - 1	340	(16%)	
FGSC	No - 0	1421	(65%)	First-generation, parents had some college.
	Yes - 1	767	(35%)	
Compact	No - 0	1899	(87%)	Participation in SDSU Compact for Success scholarship program.
	Yes - 1	289	(13%)	
Numeric Inputs				
Total GPA	TotGPA (sd)	2.09	(1.76)	Total college GPA.
Age	Age (sd)	19	(2)	Age in years.
SAT	SAT (sd)	1111	(154)	SAT composite score.
HS GPA	HS GPA (sd)	3.62	(0.63)	High school GPA.
HS Graduation Year	HSGradYr (sd)	2015	(2)	High school graduation year.

Note: We divided the variables into the output (response), categorical inputs, and numeric inputs. For the categorical variables, we present the number of students and in parentheses percentage break down in each category. For the numeric variables, we report the average and in parentheses standard deviation.

In terms of numerical inputs, the students' total college grade point average (GPA), age, SAT composite scores, high school GPA, and high school graduation year were included. SDSU stores only transfer units, not letter grades for transferred or advance placement (AP) classes. These courses thus do not affect the GPA.

As one would expect, this set of inputs will exhibit correlations (e.g., EOP and first-generation or ethnicity, student level and dorm, age and high school graduation year, etc.). Since latent class analysis and random forest do not require inputs to be uncorrelated, all of the inputs were included in the model and analysis.

3.3. Missing Values

Of the 171 observations missing partial data, there were 139 students who were missing only SAT scores, five missing only high school GPA, and 27 missing both high school GPA and SAT scores. The missing data may have arisen from two mechanisms. First, transfer students are not required to submit information about high school GPA or their SAT score. Second, students may have taken the ACT instead of the SAT. For this study, we did not have information about whether a PSY 101 student transferred into the university nor ACT scores. We are thus not able to confirm these missing data mechanisms. The missing values on these 171 students were imputed using a classification and regression (CART) univariate imputation method in the multiple imputation by chained equations (mice) imputation algorithm (the `mice` package in `R`; see van Buuren 2013, for details). Consequently, 2188 total observations were used in the LCF application.

4. Method

In this paper, we introduce a latent class analysis (LCA) and random forest ensemble that will recursively partition observations into two groups to build a latent class forest, extending the RPMM method proposed by Houseman et al. (2008) and the LCT method by Van den Bergh et al. (2017). LCA is an unsupervised, model-based clustering procedure in which multivariate mixture models are used to group observations based on "latent" variables that are characterized by the likelihood of them existing in a group, rather than a distance measure like in k-means clustering. The Expectation-Maximization (EM) algorithm is used to estimate the parameters of the different groups and is repeated until a convergence criterion is met on the likelihood function. This binary decision process is repeated for every internal node of the latent class tree until a set of stopping criteria are met, similar to the RPMM of Houseman et al. (2008).

Unlike the method proposed by Houseman et al. (2008), a random forest procedure is applied to build a collection of latent class trees. Instead of selecting one variable to create a split like in a traditional decision tree, the model will randomly select $\sqrt{p}$ inputs to use in the EM algorithm and assign observations to the left or right node based on where they are most likely to exist (Liaw et al., 2002). Furthermore, the method in this paper does not use BIC as a stopping or splitting criterion, since a random forest does not require pruning and uses a bootstrapped portion of the data to build each tree instead of the entire data set (Breiman, 2001). This method also allows for continuous and numerical inputs. After creating the forest, the algorithm quantifies the distances between every pair of observations based on how similar the terminal nodes are, and then clusters those distances. Unlike a decision tree or random forest, LCA is unsupervised, and thus the splits created are independent of the treatment or response.

In this section, we divide the description of this approach into three parts. First, we detail the process of building a latent class tree and forest. Second, we provide details about the EM algorithm. Finally, we detail the computation of the distance matrix and the clustering portion of our algorithm.

4.1. A Forest of Latent Class Trees

A classical decision tree is created by recursively splitting observations based on observation characteristics (inputs). There are three types of nodes in a decision tree: root node which initiates the tree, internal node which is characterized by a decision rule and split into two child nodes, and a terminal node which has no child nodes. The binary decision rule in the root node and internal nodes are defined by a cut point on a given input. For example, an internal node may split observations according to an SAT score being above or below 800. We refer the reader to Chapter 8 of James et al. (2013) for more details. A similar process is applied to create a tree in the latent class forest, described in Algorithm 1. However, unlike the classic decision tree, the internal node of a latent class tree creates a binary split by identifying one of two classes or clusters to which each observation is most likely to belong.

Algorithm 1 Part 1 - Building Trees

```
 1: Impute missing values
 2: for each of $T$ trees do
 3:     Randomly sample a proportion $\Delta$ of the data with replacement (bootstrap data)
 4:     while an internal node exists do
 5:         for each internal node do
 6:             Randomly select $\sqrt{p}$ inputs as splitting variables
 7:             EM algorithm assigns observations into two groups using the $\sqrt{p}$ inputs
 8:             if child nodes have at least $k$ observations then accept the new child nodes
 9:             else label node as terminal
10:             end if
11:         end for
12:     end while
13: end for
```

A few items of note about the LCF Algorithm 1. The loops entail building a forest of trees (**Line 2**) and checking to grow internal nodes (**Lines 4-12**). In **Line 2**, the user selects an appropriate T number of trees for the forest. At least 500 trees are generally recommended to balance computational efficiency and accuracy (Breiman, 2001; Shalabh, 2009). In our applications, we obtain a bootstrap data set of $\Delta = 2/3$ of the original data sample size in **Line 3**, as recommended by Breiman (2001). The bootstrap sample is a randomly selected subset of the data taken with replacement. The tree is grown using this data subset. In **Line 6**, we follow the recommendation of Breiman (2001) and round down the $\sqrt{p}$ for the random subset of variables selected to use in the EM algorithm, where p is the total number of inputs. **Lines 6-9** describe the process to build a single tree.

The EM algorithm in **Line 7** guides the splitting criterion for every internal node. We perform the iterative EM algorithm for a maximum of 1000 iterations or until the change in likelihood estimates differ by less than a tolerance level of 10^{-8}. Generally speaking, this process consists of two steps after initial parameters are selected. First, for a given internal node, the *expectation* step provides soft assignments to the left and right child nodes using the currently estimated parameters from the iterative process. Second, the *maximization* step uses those assignments to maximize a new weighted likelihood and return new estimates to continue the iterative process. Since LCA makes the assumption that local independence holds, we are able to treat each input as independent inputs (Collins and Lanza, 2010).

This process is repeated until the EM stopping rule is achieved. Using the final estimates, we assign each observation a class membership based on where they are more likely to exist, creating the left and right child nodes. This is repeated for every internal node until the tree stopping criteria are met.

Initially, we considered the AIC and BIC, similar to Van den Bergh et al. (2017), Koestler et al. (2010), and Houseman et al. (2008). Even though Van den Bergh et al. (2017) found that AIC and BIC built large trees relative to other criteria (e.g., total bivariate residual and entropy), over 90% of the trees had under five splits using the AIC and BIC. In an effort to follow Breiman (2001)'s recommendation of growing large trees in a random forest, we also explored using terminal node size as stopping criteria. In fact, although they did not explore this, Van den Bergh et al. (2017) also state that using the size of the terminal node could be used as a stopping criterion when building a latent class tree.

To identify the appropriate number of observations required in the terminal node, k, we considered values from five to 70 by increments of five. We ultimately decided to require $k > 30$ and did not specify a limit on the number of terminal nodes. However, it is worth noting that the inferences would be the same for a model with $k > 30$. Furthermore, the time difference between both models (166 minutes for $k > 30$ and 149 minutes for $k > 40$; on a dual-core, 2.7 *GHz* MacBook Pro with 8 *GB* RAM) is practically insignificant. See the appendix for more details.

4.2. EM Algorithm

The EM algorithm is an iterative process to obtain the maximum likelihood estimate for parameters characterizing an assumed model for the observed data. The estimation procedure replaces a potentially difficult likelihood maximization with a sequence of easier optimization problems that can help identify the global maximum (Casella and Berger 2002, Section 7.2). The process consists of two steps after initial parameters are selected. First, the *expectation* step provides soft assignments to each group by using the currently estimated parameters from the iterative process.

In this application, the finite mixture model of an observation x is expressed as the sum of weighted distributions,

$$p(X|\theta) = \sum_{k=1}^{K} \pi_k f_k(X|\theta_k),$$

where K is the number of total classes, $1 \leq k \leq K$, θ_k are the corresponding parameters for the kth distribution, π_k is the prior unconditional probability that an observation is a member of class k, and $\sum_{k=1}^{K} \pi_k = 1$. Estimates for π_k are provided for the first iteration. Furthermore, we are interested in estimating the memberships of two classes in each internal node of a tree. Therefore $K = 2$ so $k \in \{1, 2\}$. Categorical variables with two levels are assumed to have a Bernoulli distribution characterized by a single probability parameter. Multi-level categorical variables, with say L levels, are assumed to have a multinomial-categorical distribution characterized by L probability parameters. For example, the dorm indicator follows a Bernoulli distribution and the ethnicity follows a multinomial-categorical distribution. Continuous variables are assumed to have a Gaussian distribution characterized by a mean parameter and standard deviation parameter.

Similar to Houseman et al. (2008), Koestler et al. (2010), and Van den Bergh et al. (2017), the likelihood function for N is then expressed as

$$L(\theta|X) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f_k(X_i|\theta_k),$$

where N indicates the total number of students. Our log-likelihood is expressed as

$$l(\theta) = \sum_{i=1}^{N} \left\{ \sum_{k=1}^{K} \pi_k f_k(X_i|\theta_k) \right\}.$$

The MLE $\hat{\theta}_k$ can be expressed as the parameter that maximizes the log-likelihood, or $\hat{\theta}_k = argmax_\theta \; l(\theta)$. That is, we are interested in estimating the parameters θ_k and π_k in $l(\theta)$. The MLE cannot be obtained in closed form since we cannot easily take the partial derivatives over the sum in the log-likelihood. Fortunately, the EM algorithm builds on the assumption that the group assignments are known.

Let $z_{k,i} \in \{0, 1\}$ indicate if an observation is in the kth group, stored in the variable Z. Using the marginal distribution of X and the conditional distribution of Z given x, we can compute the class memberships $\gamma_{k,i}$ for each class and observation. Here $\gamma_{1,i} = 1 - \gamma_{2,i}$ since $K = 2$ in our setting. The probability of an observation x_i existing in class $Z_i = k$ is then expressed as

$$\gamma_{k,i} = E\left[Z_i|X_i, \theta_k\right] = P(Z_i = k|X_i, \theta_k). \tag{1}$$

In other words, $\gamma_{k,i}$ is the soft assignment for each group in this iteration of the expectation step. In terms of assignment, x_i is assigned to the group k with the highest $\gamma_{k,i}$.

Second, the *maximization* step uses these group assignments to maximize a new weighted likelihood and return new estimates to feed back into the expectation step and continue the iterative process. Using the computed $\gamma_{k,i}$ for each observation and class we can calculate estimates for the parameters θ_k. To maximize each parameter, we will take the partial derivatives with respect to a particular parameter for each input.

The new weighted proportions for categorical inputs in the k^{th} group are

$$\hat{p}_{k,l} = \frac{\sum_{i=1}^{N} \gamma_{k,i} I[x_i = l]}{\sum_{i=1}^{N} \gamma_{k,i}},$$

where l is one of the categories of that input. For example, student level has four categories (first year, sophomore, junior, senior) so $l \in \{1, 2, 3, 4\}$. Thus $\hat{p}_{k,l}$ is the weighted proportion for the l^{th} category in group k.

The weighted means and variances for the numerical inputs in the k^{th} group are

$$\begin{aligned} \hat{\mu}_k &= \frac{\sum_{i=1}^{N} x_i \gamma_{k,i}}{\sum_{i=1}^{N} \gamma_{k,i}} \\ \hat{\sigma}_k^2 &= \frac{\sum_{i=1}^{N} (x_i - \mu_k)^2 \gamma_{k,i}}{\sum_{i=1}^{N} \gamma_{k,i}}. \end{aligned}$$

The maximum likelihood estimate of the proportion of observations in class k is updated at each iteration as

$$\hat{\pi}_k = \frac{\sum_i^N \gamma_{k,i}}{N}.$$

As an iterative, optimization method, the EM algorithm guarantees convergence to a local maximum, not necessarily a global maximum. Therefore, it is important to begin the algorithm with different initial parameters. The convergence criterion is to repeat the iterative process until the class memberships in (1) do not change beyond a specified tolerance from one iteration to the next. Furthermore, the LCA model used in this application assumes local independence which states that the observed inputs are conditionally independent (Collins and Lanza, 2010).

4.3. Forest Cluster

After creating the latent class trees, we created a distance matrix for each pair of observations defined by how similar the terminal nodes are to each other. Algorithm 2 details this process. Similarity is determined by a chi-squared test on a 2×2 table of response against terminal node membership (**Line 4**). In our application, a p-value of one would indicate exact DFW proportions in both terminal nodes. The closer the p-value is to zero, the greater the difference between the two nodes, or larger "distance." These pairwise distances are stored in a distance matrix D_t for each LCT t. The distance matrices, and thus pairwise distances, are added across all trees in the forest (D, **Line 8**). We then use a distance-based clustering method on the matrix D.

Algorithm 2 Part 2 - Defining and Clustering Distances

```
1: for each tree $t$ do
2:     for each pair of terminal nodes, $(i, j)$ do
3:         if $i$ and $j$ are in the same terminal node then assign $D_t(i, j) = 1$
4:         else Compute the chi-squared test p-value and store in $D_t(i, j)$
5:         end if
6:     end for
7: end for
8: Sum the $n \times n$ distance matrices $D_t,\ t = 1, \ldots, T$
9: Cluster the aggregated matrix $D$
```

In this step of our application, we use the Hartigan and Wong (1979) k-means clustering method. In clustering these distances rather than the input data, we overcome the downfalls of k-means clustering mentioned by Xu (2011) (see Section 2.). In fact, k-means clustering is a natural choice for clustering distances since it is a widely used and simple algorithm (Schreiber and Pekarik, 2014; Xu, 2011). It is common practice to use the "kink" method to select the appropriate number of groups, where we pick the model at the "kink" after which the rate of change of variance within groups slows. Other methods include exploring AIC, BIC, silhouette, gap statistics, and cluster plots (Friedman et al., 2001). In our application, we use a combination of expert opinion by administrative decision-makers and these measures to identify an appropriate number of groups. We emphasize that this choice needs to focus on ease of interpretation and model simplicity. In particular, we do not necessarily select the model (or number of groups here) with the largest or smallest objective criterion, as the RPMM or SS-RPMM methods do

(Houseman et al., 2008; Koestler et al., 2010). Instead, we pick the model that would provide the best feedback to educators and administrators. We will provide more practical details in the next section.

5. Results

Recall that we wish to identify at-risk subgroups in the PSY 101 course, and study the impact of Supplemental Instruction in these subgroups. We apply LCF in Algorithm 1 to the 2188 students in the PSY 101 data set described in Section 3. In our application, key administrative stakeholders suggested a three-category low/medium/high-risk grouping from which they could most easily make decisions on resource allocations and programmatic refinements. Nevertheless, the variance within each group, BIC, AIC, silhouette, and gap statistic in addition to the Davies-Bouldin, Dunn, and Calinski-Harabasz indices were explored for up to twenty groups to assess this choice. There was no clear or consistent number of groups suggested by these measures. The silhouette plot, Davies-Bouldin index, and Dunn index suggested only two groups. The p-values in Table 2 might suggest two groups depending on the α-level cut-off chosen (e.g., 5% vs. 10% say). The gap statistic did not show much of an improvement beyond five clusters. The other measures suggested a large number of 10 or more clusters. The cluster plots showed the greatest separation in the three group model. We also found the conclusions drawn from models with three to five groups were similar. The choice for number of clusters thus seemed to come down to a two group or three group model. Given leanings of our administrative stakeholders, particularly that resource allocations would most easily be argued and justified using a delineation into low/medium/high-risk, we selected the three group model.

Table 2 shows the size and proportion of students who earned a DFW grade for each of the three groups identified. The p-values come from chi-squared tests comparing the proportion of DFWs in the group to the rest of the class. The table also shows SI participation in each group. We assigned a risk categorization to these groups, being “High,” “Mid,” or “Low” based on the DFW rates. We identified a High category with the highest DFW percentage of 28%, a Mid category with a DFW percentage of 12%, and a Low category with a DFW percentage of 6%.

Table 2: Description of the three LCF clusters.

Group			Response			SI	
Number	Risk	Size	DFWs	%	p-value	Attended	%
1	High	677	187	28%	0.004	178	26%
2	Mid	816	99	12%	0.066	159	19%
3	Low	695	39	6%	0.077	262	38%

Note: For each cluster, the number of students, DFW rate, and SI attendance rates are reported. The p-values compare the DFW proportions in the group to the rest of the class.

Overall, students who were commuters, FGNC, FGSC, URM and members of EOP and Compact Scholars are concentrated in the higher risk groups (see Table 3). Conversely, students who were white, not commuters, FGNC, FGSC, nor likely to be members of EOP or Compact are more concentrated in the low-risk groups. Furthermore, lower average SAT scores, and lower average high school GPAs are concentrated in the higher risk groups, whereas the lower risk groups have higher average SAT scores and higher average high school GPAs.

5.1. Clustering

Here, we will expand on the results of the three cluster groups. A detailed table with descriptive statistics for these interest groups is shown in Table 10 in the appendix. In general, students from historically underrepresented and underserved populations in higher education (URM, first-generation, EOP, Compact), lower academic performance (SAT, high school GPA), and students who do not live in campus dorms are concentrated in the high-risk group. The correlation between these variables could be explained by the demographic and academic inputs characterizing students in these groups. For example, a student who is Mexican-American and from San Diego may be more likely to have participated in Compact for Success in high school, EOP at SDSU, live with their parents during college, be a first-generation student, and have lower SAT scores.

Particularly, EOP and Compact Scholar students are overrepresented in the higher risk groups. These two programs recruit students who are first-generation, from a lower socioeconomic status, or from ethnic populations that are underrepresented in higher education. This pattern translates to what we see in the distribution of the ethnicity and underrepresented minority groups (URM) and students of color. In fact, about 54% of those in the high-risk group are URM students, whereas 18% of those in the lower risk group are URM students. Conversely, white students comprise 35% of the students in the PSY 101 data set but comprise 12% in Group 1 and 56% in Group 3. In terms of the parents' college experience, there was an overrepresentation of students who had parents with no college experience in the high-risk group for both FGNC and FGSC students.

Table 3: Descriptive statistics for the three cluster groups and class. See Table 1 for more detailed descriptions.

Input	Description	PSY 101 Data Set		Group 1		Group 2		Group 3	
Size	Group Size	2188		677		816		695	
Risk				High		Mid		Low	
Response - Pass/Fail Counts									
Grade	Fail	325	(15%)	187	(28%)	99	(12%)	39	(6%)
	Pass	1863	(85%)	490	(72%)	717	(88%)	656	(94%)
Categorical Variables									
Gender	F	1391	(64%)	429	(63%)	462	(57%)	500	(72%)
	M	797	(36%)	248	(37%)	354	(43%)	195	(28%)
Honors	No	2105	(96%)	677	(100%)	795	(97%)	633	(91%)
	Yes	83	(4%)	0	(0%)	21	(3%)	62	(9%)
Disabled	No	2156	(99%)	671	(99%)	802	(98%)	683	(98%)
	Yes	32	(1%)	6	(1%)	14	(2%)	12	(2%)
EOP	No	1978	(90%)	532	(79%)	769	(94%)	677	(97%)
	Yes	210	(10%)	145	(21%)	47	(6%)	18	(3%)
Dorm	No	1200	(55%)	548	(81%)	501	(61%)	151	(22%)
	Yes	988	(45%)	129	(19%)	315	(39%)	544	(78%)
Student Level	First Year	1194	(55%)	351	(52%)	398	(49%)	445	(64%)
	Sophomore	730	(33%)	168	(25%)	332	(41%)	230	(33%)
	Junior	169	(8%)	78	(12%)	71	(9%)	20	(3%)
	Senior	95	(4%)	80	(12%)	15	(2%)	0	(0%)
Major Status	Pre Major	476	(22%)	197	(29%)	151	(19%)	128	(18%)
	Major	1712	(78%)	480	(71%)	665	(81%)	567	(82%)
Ethnicity	Afr Am	92	(4%)	41	(6%)	32	(4%)	19	(3%)
	Asian	100	(5%)	18	(3%)	31	(4%)	51	(7%)
	Fil	199	(9%)	78	(12%)	86	(11%)	35	(5%)
	Intr	96	(4%)	63	(9%)	27	(3%)	6	(1%)
	Mex Am	520	(24%)	285	(42%)	177	(22%)	58	(8%)
	Mult	166	(8%)	34	(5%)	80	(10%)	52	(7%)
	Nat Am	17	(1%)	6	(1%)	10	(1%)	1	(<1%)
	Oth Hisp	115	(5%)	31	(5%)	41	(5%)	43	(6%)
	Other	56	(3%)	7	(1%)	22	(3%)	27	(4%)
	PI, Nat HW	9	(<1%)	4	(1%)	4	(<1%)	1	(<1%)
	SE Asian	58	(3%)	26	(4%)	17	(2%)	15	(2%)
	White	760	(35%)	84	(12%)	289	(35%)	387	(56%)
FGNC	0	1848	(84%)	458	(68%)	709	(87%)	681	(98%)
	1	340	(16%)	219	(32%)	107	(13%)	14	(2%)
FGSC	0	1421	(65%)	280	(41%)	528	(65%)	613	(88%)
	1	767	(35%)	397	(59%)	288	(35%)	82	(12%)
Compact	0	1899	(87%)	469	(69%)	737	(90%)	693	(100%)
	1	289	(13%)	208	(31%)	79	(10%)	2	(<1%)
Numeric Variables									
Age		19.33	(1.76)	20	(2.64)	19.25	(1.29)	18.78	(0.52)
Total GPA		2.09	(1.45)	2.01	(1.29)	2.08	(1.42)	2.17	(1.63)
SAT		1111	(154)	962	(142)	1130	(92)	1234	(89)
HS GPA		3.62	(0.63)	3.33	(0.96)	3.69	(0.37)	3.82	(0.26)
HS Graduation Year		2015	(1.72)	2014	(2.54)	2015	(1.32)	2015	(0.55)

Academically, we see some differences in the honors participation, SAT scores, and high school GPA scores. Specifically, students in the high-risk groups had low honors participation, below average SAT scores, and lower high school and college GPAs. It is important to note that a student may be waived from the English or Math placement exam if they score at least a 550 on the respective SAT subsection of the test. If a student does not meet the SDSU proficiency requirements for English or Math, then they will have to take one or more remedial math courses before enrolling in a course that meets their graduation requirements. As a result, this can increase the time that it takes a student to graduate from SDSU. However the average SAT score for those in the highest risk group was 962, suggesting that the average student in this group scored under a 550 on at least one of the SAT subsections. In fact, about 90% of those in Group 1 had under 1100, 39% of those in Group 2, and 4% of those in Group 3.

These patterns were also observed in the high school GPA. In particular, a GPA of 3.77 is an A- average. About 68% of those in Group 1, 55% of those in Group 2, and 42% of those in Group 3 had under a 3.77 high school GPA. Similarly, a GPA of 2.0 is a C average. About 10% of those in Group 1 had under a 2.0 total college GPA, suggesting that they have potentially earned more C- or DF grades than passing grades in their SDSU coursework. A letter grade of C or better is required for admission to a number of major programs at SDSU, particularly STEM majors. If a student does earn a C- in one of these classes, then it may also increase the time it takes for students to graduate from SDSU if not deter these students from pursuing a STEM degree.

5.2. Comparison to k-means clustering, RPMM, and SS-RPMM

As mentioned in Section 4, the LCF ensemble builds on the RPMM model and incorporates k-means clustering machinery within a random forest algorithm. Since the goal of the LCF algorithm is to cluster students into groups categorized by their risk of failure, we compared LCF to k-means clustering, RPMM, and SS-RPMM using the same data set. In all four methods, we identified three risk groups and explored the cluster size, interpretability, and risk levels across the three groups. We also explored three different cluster validity indices to help identify the best model.

Table 4 presents cluster size across the four methods. The LCF model had the most balanced sample size across the risk groups. The RPMM model had the second smallest range, with the smallest group containing 501 students (most at-risk) and the largest containing 997 students (mid risk). The SS-RPMM and k-means clustering methods presented unbalanced groups in that one or two risk groups had relatively small sample sizes. Although the goal of course is for SI to impact as many students as possible, k-means identifying a cluster of 2047 students over-generalizes this impact in that we are not able to assess the impact of potentially important factors of at-risk students that we were able to identify with LCF (e.g., EOP, on-campus living, student level, FGNC, Compact for Success program, etc.).

In terms of interpretability and inferences, we explore patterns across the three cluster groups created by each method. It is important to note that these patterns are subjective and dependent on the researcher and application. In this case, we looked for clear patterns, increase or decrease in means or percentages, from the most at-risk group to the least at-risk group. This helps identify any characteristics that could be related to the students' risk level.

Table 5 displays that the LCF method showed the greatest differentiation across risk groups, with 11 inputs showing a differentiated increase or decrease as measured by its practical im-

Table 4: Student cluster size for the four methods.

		Risk Level		
		High (n)	Mid (n)	Low (n)
Method	LCF	677	816	695
	k-means	2047	9	50
	RPMM	501	997	690
	SS-RPMM	848	971	369

plications. For example, an increase in the SAT scores from 1100 to 1120 is not practically important, whereas an increase in the GPA from 3.00 to 3.50 may be noteworthy. The k-means method had three inputs, the RPMM method had two inputs, and the SS-RPMM method had two inputs displaying clear patterns across the risk groups. Specifically, the k-means clustering method showed a clear increase or decrease in the honors program, ethnicity, and FGSC inputs. It is important to note that although FGNC and FGSC are related, FGNC did not show a similar pattern to FGSC in this case. The RPMM model displayed clear patterns for two inputs, gender and major status, and the SS-RPMM model displayed clear patterns for two inputs as well, gender and honors program. See Tables 10 to 13 in the appendix for details.

Table 5: Inputs showing an increase or decrease in average values or percentages between the most at-risk group and the least at-risk group. See Tables 10 to 13 in the appendix for more details.

Input	LCF	k-means	RPMM	SS-RPMM
Gender			X	X
Honors	X	X		X
EOP	X			
Dorm	X			
Student Level	X			
Major Status			X	
Ethnicity	X	X		
FGNC	X			
FGSC	X	X		
Compact	X			
Total GPA	X			
SAT	X			
HS GPA	X			
Total:	11	3	2	2

In addition to identifying practical differences among inputs in the three clusters, a model that has different DF grade rates across the three groups is more informative than a model with similar DF grade rates. These percentages are shown in Table 6 by risk level for each method. The fail percentages using the LCF model were the most differentiated at 28%, 12%, and 6%. This helps interpret the risk level across the three groups and categorize the three groups as "high," "mid," and "low" risk. The risk levels for the k-means clustering, RPMM, and SS-RPMM are less differentiated.

Table 6: Percent fails by risk level for the four methods.

		Risk Level		
		High	Mid	Low
Method	LCF	28%	12%	6%
	k-means	15%	11%	8%
	RPMM	16%	15%	14%
	SS-RPMM	18%	13%	11%

Although the LCF method is the most computationally expensive, it provided the best balance in the cluster sizes, displayed the most input patterns across risk groups, and best distributed the students into high/medium/low-risk groups. This allowed for more meaningful interpretations that could be used to guide academic, personal, and social interventions for PSY 101 as well as increase passing rates for students at SDSU.

In comparison, the k-means clustering had clusters with fewer than 100 students, preventing us from drawing any practical inferences on these groups. Furthermore, the k-means clustering, RPMM model, and SS-RPMM model did not show any differentiated input patterns across the three clusters that would help us identify important demographics or academic information. Also, the SS-RPMM model uses the response in the algorithm building procedure by selecting the most influential inputs to be used in the splitting procedure. In this case, the same inputs were used in over 98% of the splits across the 500 trees, reducing the influence of other inputs and essentially creating 500 identical trees.

Finally, we looked at the Davies-Bouldin, Dunn, and Calinski-Harabasz indices to help identify the model that best fits the data, shown in Table 7. All of these indices aim to find the best model by balancing the intr- and inter-cluster dispersion. Particularly, the goal is to minimize the variance within each cluster but maximize the distances between clusters (Pakhira et al., 2004; Pal and Biswas, 1997; Ray and Turi, 1999; Saitta et al., 2007; Xu, 2011).

Table 7: Davies-Bouldin, Dunn, and Calinski-Harabasz for model comparison.

Method	LCF	k-means	RPMM	SS-RPMM	Goal
Davies-Bouldin	1.56	2.04	10.75	6.40	Minimize
Dunn $\times 10^3$	53.24	0.68	0.65	0.65	Maximize
Calinski-Harabasz	451.39	324.27	107.89	324.27	Maximize

Small Davies-Bouldin, large Dunn, and large Calinski-Harabasz indices suggest models that better fit the data. All three measures suggest that the LCF model better fits the data. In this case, all three quantitative measures suggest that the LCF generally tends to outperform k-means, RPMM, and SS-RPMM when comparing intra- and inter-cluster dispersion.

6. Supplemental Instruction

Students who attend Supplemental Instruction sessions are generally more likely to pass the class when compared to their counterparts who did not attend SI in PSY 101. As previously mentioned, about 15% of the students in the PSY 101 data set earned a DFW grade. However, only 9% of those who attended SI earned a DFW, whereas 20% of those who did not attend SI earned a DFW grade. The rest of this section will explore subgroups that benefit the most from SI, and how attending SI potentially affects the letter grade distribution.

6.1. Group Analysis

Table 8 shows the count and percentage of students who earned DFW grades in each LCF group by SI attendance. Here, students who attended SI at least once are compared to students who did not attend SI in PSY 101. Group 1, the highest risk group, had a decrease in DFW rates of about 12 percentage points. About 28% of the students in this group earned a DFW grade. However, 31% of those students who did not attend SI earned a DFW grade, whereas only 19% of those who did attend earned a DFW grade. This group contains fewer students who lived in the residence halls on campus, about 54%, students who were identified as URM, and students who had low SAT and HS GPAs (see Table 3).

Table 8: DFW grades by cluster groups and SI participation.

Group			Did Not Attend SI			Attended SI			Percent	p-value
Number	Risk	Size	n	DFW's	%	n	DFW's	%	Difference	
1	High	677	499	153	31%	178	34	19%	-12%	0.003
2	Mid	816	657	87	13%	159	12	8%	-5%	0.057
3	Low	695	433	30	7%	262	3	3%	-4%	0.061

Note: The difference in marginal percentages for each group as well as the p-values from a Fisher's exact test comparing the SI and non-SI students in each cluster group is also reported.

Group 2 had about a five percentage point decrease in DFW grades between those who attended SI sessions and those who did not, the largest difference across the risk groups. About 12% of the students in this group earned DFW grades. About 13% of those who did not attend SI earned a DFW grade, whereas 8% of those who attended at least one SI session earned a DFW grade.

Finally, about 6% of the students in Group 3 earned a DFW grade. However, about 7% of those who did not attend SI earned a DFW grade, whereas 3% of the students who did attend earned a DFW grade. This group has an overrepresentation of non-URM students, students who are not first-generation, students with low SAT scores and GPA, and students who live on campus (see Table 3).

6.2. Grade Distribution

In the entire PSY 101 data set, we can see that of those students who attended SI there were fewer students earning D and F grades and more students earning passing grades; see Figure 1. This finding is also true for students in Groups 1, 2, and 3. Not only did students who attended SI move out of the DFW grade range, but those in Groups 2 and 3 also moved out of the C grade range to earn A and B grades. In fact, students who attended SI in these two groups had more A and B grades when compared to those who did not attend SI in the respective group.

A deeper analysis would be required to measure the efficacy of attending SI (Guarcello, 2015). Students have to voluntarily opt into attending SI since it is not mandatory. We thus have no control over the assignment of this treatment. As part of our current research program, we are considering matching methods within cluster groups to measure the impact of SI in this observational study.

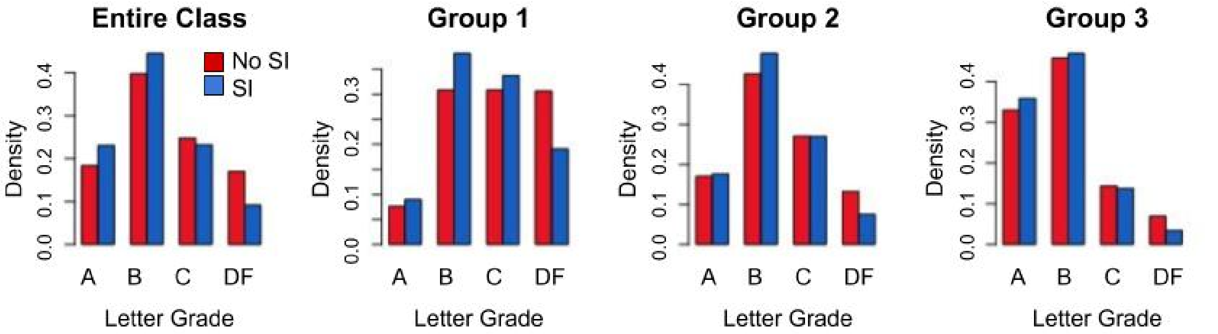

Figure 1: Grade Distribution: Histograms of the final letter grades in PSY 101 by SI attendance for the entire class and Groups 1 to 3.

7. Discussion: Implications for SI and Future Work

We identified three subgroups of students, based on the distance matrix for LCF, that represent different DFW risk levels in PSY 101. The students who were identified in the lowest risk groups are primarily students who live on campus and identify themselves as white. In both of the low-risk groups explored in this paper, there were few students whose parents had no higher education (FGNC or FGSC). These groups also had above average SAT scores and high school GPA. However, the two highest risk groups represent students who come from historically underrepresented populations in higher education. These groups contain an overrepresentation of first-generation college students, commuters, students with below average SAT scores and high school GPAs, and Mexican-American students. In fact, about 54% of the students who were in the high-risk group were identified as URM, compared to 18% of those in the low-risk group and 35% in the PSY 101 data set as a whole. Traditionally, these are populations that are often of interest for higher education institutions due to their underrepresentation and the historical context behind equity and access to opportunities. Specifically at SDSU, a designated Hispanic Serving Institution, these findings help promote discussions around lower-division, high enrollment, general education course support systems provided for students who identify themselves as Mexican-American.

In a similar study about SI, Rath et al. (2007) mention that URM students may not perform as well as their non-URM peers due to a variety of systematic, personal, social, and academic barriers. Particularly related to SI, students from URM backgrounds may not have access to college resources, are less likely to be prepared for college-level courses, lack the self confidence to perform well as a result of institutionalized stereotypes, and may feel a lack of social belonging in higher education environments when compared to their non-URM peers (Fischer, 2007; Massey et al., 2011; Rath et al., 2007; Steele and Aronson, 2005; Telles and Ortiz, 2008). However, this does not suggest that SI is not as effective with URM students. A deeper analysis within each subgroup would be needed to measure these effects.

We also see the intersection of these characteristics in the concentration of students who participate in EOP or the Compact for Success programs. Both of these programs recruit students from first-generation and URM backgrounds. It is important to note that this analysis made no conclusions about about the efficacy of these programs. Further studies are required to accurately measure the impact that the EOP and the Compact for Success programs have on these populations. For example, we may use a matching algorithm to compare students who were in EOP to those who were not, to minimize bias due to non-randomized "treatment" assignment, and measure the difference in both populations.

Those students in the most at-risk group who attended SI had about a 12% decrease in DFW rates when compared to students who did not attend SI. Similarly, those students in the average risk group who attended SI had about a 5% decrease in DFW rates when compared to students who did not attend SI. Since this resource is free and available to all students, it may behoove members of EOP and Compact for Success as well as the Commuter Resource Center, Center for Intercultural Relations, and other organizations that target these at-risk populations to play active roles in the advertisement of SI. In doing so, DFW rates may be reduced in these courses and the achievement gap and graduation rates may be positively impacted.

In terms of evaluating the impact of SI on these subgroups, students who attend SI for PSY 101 at SDSU are generally more likely to pass this class when compared to their counterparts who did not attend SI (Guarcello, 2015). This general pattern translated to what we observed in the three subgroups at different levels, regardless of the level of risk. The groups that benefit the most from SI are Groups 1 and 2. As mentioned above, these groups contain students who predominantly come from historically disadvantaged groups.

For future work, a surrogate split method can be introduced to deal with missing values at each split rather than imputing the missing values before building the model (Feelders, 1999). Ideally, this surrogate split method must not significantly increase the computational complexity of the algorithm. In terms of computational efficiency, the forest requires about 10 minutes to build for 2188 observations and 500 trees on a dual-core, 2.7 *GHz* MacBook Pro with 8 *GB* RAM. The algorithm may be easily put in parallel at the forest level since the latent class trees are grown independently. More importantly, computing the distance matrices in the clustering algorithm requires about two hours. That said, our software for this distance matrix building process may be further optimized, and this task may also be performed through parallel computing.

8. Acknowledgments

This work is supported in part by NSF BIGDATA grant 1633130. We thank the Associate Editor and three reviewers for comments that lead to an improved evaluation and exposition of our methods.

References

BAKER, R. S. AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining 1,* 1, 3–17.

BRAXTON, J. M. 2000. *Reworking the student departure puzzle*. Vanderbilt University Press, 2000.

BREIMAN, L. 2001. Random forests. *Machine Learning 45,* 1, 5–32.

BRUSCO, M. J., SHIREMAN, E., AND STEINLEY, D. 2016. A comparison of latent class, k-means, and k-median methods for clustering dichotomous data. *Psychological methods 22,* 3, 563.

CARUANA, R., KARAMPATZIAKIS, N., AND YESSENALINA, A. 2008. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, 96–103.

CARUANA, R. AND NICULESCU-MIZIL, A. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 161–168.

CASELLA, G. AND BERGER, R. 2002. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

CHAN, J. Y. AND BAUER, C. F. 2014. Identifying at-risk students in general chemistry via cluster analysis of affective characteristics. *Journal of Chemical Education 91,* 9, 1417–1425.

COLLINS, L. M. AND LANZA, S. T. 2010. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Vol. 718. John Wiley & Sons.

CSU. 2017a. The California state university: 2017-2018 CSU undergraduate impacted programs matrix. https://www.calstate.edu/sas/documents/impactedprogramsmatrix.pdf.

CSU. 2017b. The California state university: CSU campus impaction information 2017-2018. http://www.calstate.edu/sas/impaction-campus-info.shtml.

CSU. 2018. Impacted undergraduate majors and campuses, 2018-19. https://www2.calstate.edu/attend/degrees-certificates-credentials/Pages/impacted-degrees.aspx.

DAVIES, R., NYLAND, R., BODILY, R., CHAPMAN, J., JONES, B., AND YOUNG, J. 2017. Designing technology-enabled instruction to utilize learning analytics. *TechTrends 61,* 2, 155–161.

FARSIDES, T. AND WOODFIELD, R. 2003. Individual differences and undergraduate academic success: The roles of personality, intelligence, and application. *Personality and Individual Differences 34,* 7, 1225–1243.

FEELDERS, A. 1999. Handling missing data in trees: surrogate splits or statistical imputation? In *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 329–334.

FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S., AND AMORIM, D. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research 15,* 1, 3133–3181.

FISCHER, E. M. J. 2007. Settling into campus life: Differences by race/ethnicity in college involvement and outcomes. *The Journal of Higher Education 78,* 2, 125–161.

FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics, New York.

GRAY, G., MCGUINNESS, C., AND OWENDE, P. 2014. Non-cognitive factors of learning as predictors of academic performance in tertiary education. In *7th International Conference on Educational Data Mining*. International Educational Data Mining Society, 107–114.

GRAY, G., MCGUINNESS, C., OWENDE, P., AND HOFMANN, M. 2016. Learning factor models of students at risk of failing in the early stage of tertiary education. *Journal of Learning Analytics 3,* 2, 330–372.

GUARCELLO, M. A. 2015. Blended learning and bottlenecks in the California State University system: An empirical look at the importance of demographic and performance analytics. Ph.D. Thesis, University of San Diego.

HARTIGAN, J. A. AND WONG, M. A. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28,* 1, 100–108.

HOUSEMAN, E. A., CHRISTENSEN, B. C., YEH, R.-F., MARSIT, C. J., KARAGAS, M. R., WRENSCH, M., NELSON, H. H., WIEMELS, J., ZHENG, S., WIENCKE, J. K., ET AL. 2008. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics 9,* 1, 365–380.

JAMES, G., WITTEN, D., HASTIE, T., AND TIBSHIRANI, R. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

JAYAPRAKASH, S. M., MOODY, E. W., LAURÍA, E. J., REGAN, J. R., AND BARON, J. D. 2014. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics 1,* 1, 6–47.

KOESTLER, D. C., MARSIT, C. J., CHRISTENSEN, B. C., KARAGAS, M. R., BUENO, R., SUGARBAKER, D. J., KELSEY, K. T., AND HOUSEMAN, E. A. 2010. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics 26,* 20, 2578–2585.

LIAW, A., WIENER, M., ET AL. 2002. Classification and regression by random forest. *R News 2,* 3, 18–22.

MASSEY, D. S., CHARLES, C. Z., LUNDY, G., AND FISCHER, M. J. 2011. *The source of the river: The social origins of freshmen at America's selective colleges and universities*. Vol. 61. Princeton University Press.

MAULL, K. E., SALDIVAR, M. G., AND SUMNER, T. 2010. Online curriculum planning behavior of teachers. In *Proceedings of the Third International Conference on Educational Data Mining*.

PAKHIRA, M. K., BANDYOPADHYAY, S., AND MAULIK, U. 2004. Validity index for crisp and fuzzy clusters. *Pattern Recognition 37,* 3, 487–501.

PAL, N. R. AND BISWAS, J. 1997. Cluster validation using graph theoretic concepts. *Pattern Recognition 30,* 6, 847–857.

PAPAMITSIOU, Z. AND ECONOMIDES, A. A. 2014. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Journal of Educational Technology & Society 17,* 4, 49–64.

RATH, K. A., PETERFREUND, A. R., XENOS, S. P., BAYLISS, F., AND CARNAL, N. 2007. Supplemental instruction in introductory biology I: Enhancing the performance and retention of underrepresented minority students. *CBE-Life Sciences Education 6,* 3, 203–216.

RAY, S. AND TURI, R. H. 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*. Calcutta, India, 137–143.

ROMERO, C. AND VENTURA, S. 2010. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40,* 6 (Nov.), 601–618.

SAITTA, S., RAPHAEL, B., AND SMITH, I. F. 2007. A bounded index for cluster validity. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 174–187.

SCHREIBER, J. B. AND PEKARIK, A. J. 2014. Technical note: Using latent class analysis versus k-means or hierarchical clustering to understand museum visitors. *Curator: The Museum Journal 57,* 1, 45–59.

SDSU. 2018a. Compact for success. http://compactforsuccess.sdsu.edu/.

SDSU. 2018b. Educational opportunity programs and ethnic affairs. http://studentaffairs.sdsu.edu/EOP/.

SHALABH. 2009. Statistical learning from a regression perspective. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172,* 4, 935–935.

SMITH, E. AND HANLEY, G. 2013. Reducing bottlenecks and improving student success. http://www.calstate.edu/bot/agendas/sep13/Agenda.pdf.

STEELE, C. M. AND ARONSON, J. 2005. Stereotypes and the fragility of academic competence, motivation, and self-concept. *Handbook of Competence and Motivation*, 436–455.

TALAVERA, L. AND GAUDIOSO, E. 2004. Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on Artificial Intelligence in CSCL. 16th European Conference on Artificial Intelligence*. 17–23.

TELLES, E. M. AND ORTIZ, V. 2008. *Generations of exclusion: Mexican-Americans, assimilation, and race*. Russell Sage Foundation.

TSAI, C.-F., TSAI, C.-T., HUNG, C.-S., AND HWANG, P.-S. 2011. Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. *Australasian Journal of Educational Technology 27,* 3, 481–498.

UMKC. 2017. Supplemental instruction. http://info.umkc.edu/si/.

VAN BUUREN, S. 2013. *mice: Multivariate imputation by chained equations*. R package version 2.30.

VAN DEN BERGH, M., SCHMITTMANN, V. D., AND VERMUNT, J. K. 2017. Building latent class trees, with an application to a study of social capital. *Methodology 13,* S1, 13–22.

XU, B. 2011. Clustering educational digital library usage data: Comparisons of latent class analysis and k-means algorithms. Ph.D. Thesis, Utah State University.

9. APPENDIX

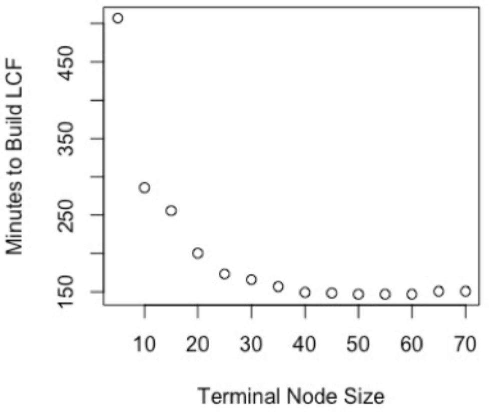

Figure 2: Time it took to build an LCF by the required terminal node size. Run time does not decrease significantly beyond the required terminal node size larger than 40.

Table 9: Inputs showing increase or decrease in average values or percentages between the most at-risk group to the least at-risk group depending on the required terminal node size.

Input	Required Terminal Node Size										Total
	5	10	15	20	25	30	35	40	45	50	
Gender	X										1
Honors			X		X	X	X	X	X	X	7
EOP			X	X	X	X	X	X	X	X	8
Dorm X	X	X	X		X	X	X	X	X	X	9
Student Level				X		X	X	X	X	X	6
Major Status				X							1
Ethnicity			X			X	X	X	X	X	6
FGNC			X		X	X	X	X	X	X	7
FGSC			X		X	X	X	X	X	X	7
Compact				X	X	X	X	X	X	X	7
Total GPA	X					X	X	X		X	5
Age				X							1
SAT		X	X		X	X	X	X	X	X	8
HS GPA		X	X		X	X	X	X	X	X	8
HS Grad Year				X		X	X	X	X	X	6
Total:											

Table 10: Descriptive statistics for the three cluster groups and class - LCF

Input	Description	PSY 101 Data Set		Group 1		Group 2		Group 3	
Size	Group Size	2188		677		816		695	
Risk				High		Mid		Low	
Response - Pass/Fail Counts									
Grade	Fail	325	(15%)	187	(28%)	99	(12%)	39	(6%)
	Pass	1863	(85%)	490	(72%)	717	(88%)	656	(94%)
Categorical Variables									
Gender	F	1391	(64%)	429	(63%)	462	(57%)	500	(72%)
	M	797	(36%)	248	(37%)	354	(43%)	195	(28%)
Honors	No	2105	(96%)	677	(100%)	795	(97%)	633	(91%)
	Yes	83	(4%)	0	(0%)	21	(3%)	62	(9%)
Disabled	No	2156	(99%)	671	(99%)	802	(98%)	683	(98%)
	Yes	32	(1%)	6	(1%)	14	(2%)	12	(2%)
EOP	No	1978	(90%)	532	(79%)	769	(94%)	677	(97%)
	Yes	210	(10%)	145	(21%)	47	(6%)	18	(3%)
Dorm	No	1200	(55%)	548	(81%)	501	(61%)	151	(22%)
	Yes	988	(45%)	129	(19%)	315	(39%)	544	(78%)
Student Level	First Year	1194	(55%)	351	(52%)	398	(49%)	445	(64%)
	Sophomore	730	(33%)	168	(25%)	332	(41%)	230	(33%)
	Junior	169	(8%)	78	(12%)	71	(9%)	20	(3%)
	Senior	95	(4%)	80	(12%)	15	(2%)	0	(0%)
Major Status	Pre Major	476	(22%)	197	(29%)	151	(19%)	128	(18%)
	Major	1712	(78%)	480	(71%)	665	(81%)	567	(82%)
Ethnicity	AfrAm	92	(4%)	41	(6%)	32	(4%)	19	(3%)
	Asian	100	(5%)	18	(3%)	31	(4%)	51	(7%)
	Fil	199	(9%)	78	(12%)	86	(11%)	35	(5%)
	Intr	96	(4%)	63	(9%)	27	(3%)	6	(1%)
	MexAm	520	(24%)	285	(42%)	177	(22%)	58	(8%)
	Mult	166	(8%)	34	(5%)	80	(10%)	52	(7%)
	NatAm	17	(1%)	6	(1%)	10	(1%)	1	(<1%)
	OthHisp	115	(5%)	31	(5%)	41	(5%)	43	(6%)
	Other	56	(3%)	7	(1%)	22	(3%)	27	(4%)
	PI, NatHW	9	(<1%)	4	(1%)	4	(<1%)	1	(<1%)
	SE Asian	58	(3%)	26	(4%)	17	(2%)	15	(2%)
	White	760	(35%)	84	(12%)	289	(35%)	387	(56%)
FGNC	No	1848	(84%)	458	(68%)	709	(87%)	681	(98%)
	Yes	340	(16%)	219	(32%)	107	(13%)	14	(2%)
FGSC	No	1421	(65%)	280	(41%)	528	(65%)	613	(88%)
	Yes	767	(35%)	397	(59%)	288	(35%)	82	(12%)
Compact	No	1899	(87%)	469	(69%)	737	(90%)	693	(100%)
	Yes	289	(13%)	208	(31%)	79	(10%)	2	(<1%)
Numeric Variables									
Age		19	(1.76)	20	(2.64)	19.25	(1.29)	18.78	(0.52)
Total GPA		2.09	(1.45)	2.01	(1.29)	2.08	(1.42)	2.17	(1.63)
SAT		1111	(154)	962	(142)	1130	(92)	1234	(89)
HS GPA		3.62	(0.63)	3.33	(0.96)	3.69	(0.37)	3.82	(0.26)
HS Graduation Year		2015	(1.72)	2014	(2.54)	2015	(1.32)	2015	(0.55)

Note: Inputs that are differentiated across the three groups are highlighted.

Table 11: Descriptive statistics for the three cluster groups and class - k-means

Input	Description	PSY 101 Data Set		Group 1		Group 2		Group 3	
Size	Group Size	2188		2047		85		50	
Risk				High		Mid		Low	
Response - Pass/Fail Counts									
Grade	Fail	325	(15%)	312	(15%)	9	(11%)	4	(8%)
	Pass	1863	(85%)	1741	(85%)	76	(89%)	46	(92%)
Categorical Variables									
Gender	F	1391	(64%)	1304	(64%)	36	(62%)	51	(61%)
	M	797	(36%)	743	(36%)	22	(38%)	32	(39%)
Honors	No	2105	(96%)	2047	(100%)	58	(100%)	0	(0%)
	Yes	83	(4%)	0	(0%)	0	(0%)	83	(100%)
Disabled	No	2156	(99%)	2019	(99%)	56	(97%)	81	(98%)
	Yes	32	(1%)	28	(1%)	2	(3%)	2	(2%)
EOP	No	1978	(90%)	1847	(90%)	52	(90%)	79	(95%)
	Yes	210	(10%)	200	(10%)	6	(10%)	4	(5%)
Dorm	No	1200	(55%)	1123	(55%)	47	(81%)	30	(36%)
	Yes	988	(45%)	924	(45%)	11	(19%)	53	(64%)
Student Level	First Year	1194	(55%)	1167	(57%)	2	(3%)	25	(30%)
	Sophomore	730	(33%)	688	(34%)	1	(2%)	41	(49%)
	Junior	169	(8%)	137	(7%)	17	(29%)	15	(18%)
	Senior	95	(4%)	55	(3%)	38	(66%)	2	(2%)
Major Status	Pre Major	476	(22%)	418	(20%)	46	(79%)	12	(14%)
	Major	1712	(78%)	1629	(80%)	12	(21%)	71	(86%)
Ethnicity	AfrAm	92	(4%)	88	(4%)	0	(0%)	4	(5%)
	Asian	100	(5%)	97	(5%)	1	(2%)	2	(2%)
	Fil	199	(9%)	196	(10%)	2	(3%)	1	(1%)
	Intr	96	(4%)	69	(3%)	27	(47%)	0	(0%)
	MexAm	520	(24%)	500	(24%)	9	(16%)	11	(13%)
	Mult	166	(8%)	157	(8%)	2	(3%)	7	(8%)
	NatAm	17	(1%)	16	(1%)	1	(2%)	0	(0%)
	OthHisp	115	(5%)	108	(5%)	2	(3%)	5	(6%)
	Other	56	(3%)	50	(2%)	2	(3%)	4	(5%)
	PI, NatHW	9	(<1%)	9	(<1%)	0	(0%)	0	(0%)
	SE Asian	58	(3%)	57	(3%)	1	(2%)	0	(0%)
	White	760	(35%)	700	(34%)	11	(19%)	49	(59%)
FGNC	No	1848	(84%)	1723	(84%)	46	(79%)	79	(95%)
	Yes	340	(16%)	324	(16%)	12	(21%)	4	(5%)
FGSC	No	1421	(65%)	1310	(64%)	41	(71%)	70	(84%)
	Yes	767	(35%)	737	(36%)	17	(29%)	13	(16%)
Compact	No	1899	(87%)	1760	(86%)	58	(100%)	81	(98%)
	Yes	289	(13%)	287	(14%)	0	(0%)	2	(2%)
Numeric Variables									
Age		19	(1.76)	19	(0.99)	25	(6.58)	19	(0.77)
Total GPA		2.09	(1.45)	2.07	(1.44)	1.76	(1.63)	2.77	(1.42)
SAT		1111	(154)	1112	(138)	846	(320)	1273	(93)
HS GPA		3.62	(0.63)	3.70	(0.32)	0.52	(1.17)	3.95	(0.22)
HS Graduation Year		2015	(1.72)	2015	(1.00)	2009	(6.38)	2015	(0.75)

Note: Inputs that are differentiated across the three groups are highlighted. Inputs were standardized and centered around 0. Traditional k-means was used to create the clusters.

Table 12: Descriptive statistics for the three cluster groups and class - RPMM

Input	Description	PSY 101 Data Set		Group 1		Group 2		Group 3	
Size	Group Size	2188		690		997		501	
Risk				High		Mid		Low	
Response - Pass/Fail Counts									
Grade	Fail	325	(15%)	79	(16%)	100	(14%)	146	(15%)
	Pass	1863	(85%)	422	(84%)	590	(86%)	851	(85%)
Categorical Variables									
Gender	F	1391	(64%)	301	(60%)	483	(70%)	607	(61%)
	M	797	(36%)	200	(40%)	207	(30%)	390	(39%)
Honors	No	2105	(96%)	481	(96%)	669	(97%)	955	(96%)
	Yes	83	(4%)	20	(4%)	21	(3%)	42	(4%)
Disabled	No	2156	(99%)	494	(99%)	682	(99%)	980	(98%)
	Yes	32	(1%)	7	(1%)	8	(1%)	17	(2%)
EOP	No	1978	(90%)	468	(93%)	627	(91%)	883	(89%)
	Yes	210	(10%)	33	(7%)	63	(9%)	114	(11%)
Dorm	No	1200	(55%)	239	(48%)	373	(54%)	588	(59%)
	Yes	988	(45%)	262	(52%)	317	(46%)	409	(41%)
Student Level	First Year	1194	(55%)	278	(55%)	389	(56%)	527	(53%)
	Sophomore	730	(33%)	166	(33%)	196	(28%)	368	(37%)
	Junior	169	(8%)	38	(8%)	58	(8%)	73	(7%)
	Senior	95	(4%)	19	(4%)	47	(7%)	29	(3%)
Major Status	Pre Major	476	(22%)	75	(15%)	232	(34%)	169	(17%)
	Major	1712	(78%)	426	(85%)	458	(66%)	828	(83%)
Ethnicity	AfrAm	92	(4%)	13	(3%)	23	(3%)	56	(6%)
	Asian	100	(5%)	34	(7%)	27	(4%)	39	(4%)
	Fil	199	(9%)	53	(11%)	62	(9%)	84	(8%)
	Intr	96	(4%)	15	(3%)	39	(6%)	42	(4%)
	MexAm	520	(24%)	112	(22%)	168	(24%)	240	(24%)
	Mult	166	(8%)	35	(7%)	46	(7%)	85	(9%)
	NatAm	17	(1%)	3	(1%)	7	(1%)	7	(1%)
	OthHisp	115	(5%)	24	(5%)	35	(5%)	56	(6%)
	Other	56	(3%)	15	(3%)	15	(2%)	26	(3%)
	PI, NatHW	9	(<1%)	1	(<1%)	1	(<1%)	7	(1%)
	SE Asian	58	(3%)	15	(3%)	18	(3%)	25	(3%)
	White	760	(35%)	181	(36%)	249	(36%)	330	(33%)
FGNC	No	1848	(84%)	431	(86%)	588	(85%)	829	(83%)
	Yes	340	(16%)	70	(14%)	102	(15%)	168	(17%)
FGSC	No	1421	(65%)	333	(66%)	469	(68%)	619	(62%)
	Yes	767	(35%)	168	(34%)	221	(32%)	378	(38%)
Compact	No	1899	(87%)	444	(89%)	599	(87%)	856	(86%)
	Yes	289	(13%)	57	(11%)	91	(13%)	141	(14%)
Numeric Variables									
Age		19	(1.76)	19	(2.30)	19	(1.7)	19	(1.47)
Total GPA		2.09	(1.45)	1.98	(1.48)	2.03	(1.47)	2.19	(1.42)
SAT		1111	(154)	1116	(163)	1106	(157)	1113	(146)
HS GPA		3.62	(0.63)	3.65	(0.55)	3.56	(0.77)	3.65	(0.56)
HS Graduation Year		2015	(1.72)	2015	(2.26)	2015	(1.71)	2015	(1.39)

Note: Inputs that are differentiated across the three groups are highlighted. Traditional RPMM model was used. BIC and terminal node size were used as stopping criteria.

Table 13: Descriptive statistics for the three cluster groups and class - SS RPMM

Input	Description	PSY 101 Data Set		Group 1		Group 2		Group 3	
Size	Group Size	2188		848		971		369	
Risk				High		Mid		Low	
Response - Pass/Fail Counts									
Grade	Fail	325	(15%)	154	(18%)	129	(13%)	42	(11%)
	Pass	1863	(85%)	694	(82%)	842	(87%)	327	(89%)
Categorical Variables									
Gender	F	1391	(64%)	51	(6%)	971	(100%)	369	(100%)
	M	797	(36%)	797	(94%)	0	(0%)	0	(0%)
Honors	No	2105	(96%)	765	(90%)	971	(100%)	369	(100%)
	Yes	83	(4%)	83	(10%)	0	(0%)	0	(0%)
Disabled	No	2156	(99%)	842	(99%)	945	(97%)	369	(100%)
	Yes	32	(1%)	6	(1%)	26	(3%)	0	(0%)
EOP	No	1978	(90%)	789	(93%)	843	(87%)	346	(94%)
	Yes	210	(10%)	59	(7%)	128	(13%)	23	(6%)
Dorm	No	1200	(55%)	525	(62%)	306	(32%)	369	(100%)
	Yes	988	(45%)	323	(38%)	665	(68%)	0	(0%)
Student Level	First Year	1194	(55%)	381	(45%)	699	(72%)	114	(31%)
	Sophomore	730	(33%)	335	(40%)	222	(23%)	173	(47%)
	Junior	169	(8%)	82	(10%)	33	(3%)	54	(15%)
	Senior	95	(4%)	50	(6%)	17	(2%)	28	(8%)
Major Status	Pre Major	476	(22%)	188	(22%)	212	(22%)	76	(21%)
	Major	1712	(78%)	660	(78%)	759	(78%)	293	(79%)
Ethnicity	AfrAm	92	(4%)	22	(3%)	52	(5%)	18	(5%)
	Asian	100	(5%)	35	(4%)	52	(5%)	13	(4%)
	Fil	199	(9%)	90	(11%)	63	(6%)	46	(12%)
	Intr	96	(4%)	40	(5%)	28	(3%)	28	(8%)
	MexAm	520	(24%)	191	(23%)	252	(26%)	77	(21%)
	Mult	166	(8%)	67	(8%)	70	(7%)	29	(8%)
	NatAm	17	(1%)	9	(1%)	7	(1%)	1	(<1%)
	OthHisp	115	(5%)	44	(5%)	51	(5%)	20	(5%)
	Other	56	(3%)	28	(3%)	23	(2%)	5	(1%)
	PI, NatHW	9	(<1%)	2	(<1%)	4	(<1%)	3	(1%)
	SE Asian	58	(3%)	21	(2%)	31	(3%)	6	(2%)
	White	760	(35%)	299	(35%)	338	(35%)	123	(33%)
FGNC	No	1848	(84%)	728	(86%)	751	(77%)	369	(100%)
	Yes	340	(16%)	120	(14%)	220	(23%)	0	(0%)
FGSC	No	1421	(65%)	576	(68%)	476	(49%)	369	(100%)
	Yes	767	(35%)	272	(32%)	495	(51%)	0	(0%)
Compact	No	1899	(87%)	726	(86%)	875	(90%)	298	(81%)
	1	289	(13%)	122	(14%)	96	(10%)	71	(19%)
Numeric Variables									
Age		19	(1.76)	19	(1.44)	19	(1.90)	20	(1.91)
Total GPA		2.09	(1.45)	2.23	(1.36)	1.85	(1.55)	2.41	(1.27)
SAT		1111	(154)	1149	(155)	1091	(136)	1077	(176)
HS GPA		3.62	(0.63)	3.58	(0.64)	3.69	(0.51)	3.52	(0.84)
HS Graduation Year		2015	(1.72)	2014	(1.40)	2015	(1.86)	2014	(1.87)

Note: Inputs that are differentiated across the three groups are highlighted.