MACHINE LEARNING AND DATA ANALYTICS FOR MULTILAYER DATA IN POLICY PLANNING

BY

MUMTAZ HANNA BEE VAUHKONEN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Policy Studies
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Associate Professor Pradeep Dhillon, Chair and Director of Research
Professor Bill Cope
Associate Professor Feng Liang
Assistant Professor Sanmi Koyejo

ABSTRACT

"Does bigger data lead to better decisions?" has been a frequent question for discussion among many decision makers—data scientists as well as organizational leaders and managers. Educational institutions, finance, and the retail industry have had big data for several decades that did not significantly alter decision making as it is doing today, as most of the decisions made were driven more by either small-scale studies or a desire to support a group's belief or interests. However, the advent of computational mechanisms changed the notion of what big data analysis is (Evgeniou, Gaba, & Niessing, 2013). Linking these methods to diverse related data is proving to be a game changer for big data analysis; for example, connecting sales figures to customer behavior is leading to better business decisions. By extending this research into the education domain, data sources at multiple levels in education can be combined to connect the big picture by analyzing and connecting data at higher macro levels to lower levels. The main purpose of this study is to present a framework for analyzing macro data with multiple data types in education for effective policy planning by linking data analysis at multiple macro levels (district and school levels) and extend it demonstrate a proof of concept of connecting to micro levels. Achieving this task required the following steps:

1. Use an unsupervised approach in big data techniques to analyze data at school level and district levels.
2. Develop a supervised classification system to use clusters from step 1 as classes so that changes in school data are constantly reassessed and assigned to new classes.
3. Identify frequent patterns and association rules of various attributes at the school level.

4. Perform a regression analysis on results from the macro level to investigate deeper with additional variables to identify the differences and verify the frequent patterns and association rules.
5. Develop a simulated analysis of student-level performance data to identify similarity and dissimilarity patterns using collaborative models that gain insights into collective intelligence in a recommender system format, which teachers can use to find optimal measures to improve a student's learning.

These steps, put together into an analytical system, can handle large volumes of data and give insights for developing effective macro and micro policies. The results indicate that applying machine learning and data mining models enable extracting more insights at a macro policy planning level.

Dedicated to my beloved husband, Ari Vauhkonen,

and children, Sigriður and Odin Vauhkonen

ACKNOWLEDGMENTS

I extend my hearty gratitude to my advisor, Prof. Pradeep Dhillon, whose patience, support, and dedication to her students have brought me this far. Without her I would not have been able to navigate this doctoral process. She has stood by me through tumultuous times of personal tragedy as well as following times of happiness. Her support has been not only academic but also a strong moral backbone for me to keep taking small steps at all times. I have been extremely fortunate to have her as my advisor.

My immense gratitude to Prof. James D. Anderson, who always had faith and belief to support me in all ways that culminated in completing this large-scale project and multiple domain courses required for completion of this thesis. His flexibility to allow me to take courses from Stanford and UIC, so that I can meet the requirements on schedule, is exemplary of his broad vision to enable students to accomplish their goals.

My gratitude and thanks to Prof. Bill Cope for always being available and giving timely feedback and being flexible with meetings and reviews. His insights have enabled me to frame my research questions to be more precise.

My sincere thanks to Prof. Emeritus Michael Heath, Computational Science and Engineering Department, for his vision and guidance for me to pursue a Machine Learning and Data Mining specialization and enabling me to obtain a joint Ph.D. with the department. My sincere thanks to Prof. Narayana Aluru for enabling my completion of the CSE joint Ph.D. certification.

My sincere thanks for Stephanie Rayl, who has assisted me with numerable requests over the years, from course registrations to scheduling my defense.

Most importantly, no amount of thanks would be enough for my husband, Ari Vauhkonen, for all the love, all those supporting elements in life such as countless nights of washing the dishes so I could carve a bit of extra time for my research. This research would have been enormously harder with little children without his commitment and hard work toward my goals and often counseling on various decisions. And my hearty hugs and thanks for my two adorable children, Sigriður and Odin, for keeping me entertained with humor and their innumerable questions on my research. I am fortunate to have my parents Dr. I. Khasim and Rasool, who have given me the lifetime of love, inspiration, and support to pursue my goals, and thanks are not adequate to express my gratitude. My father's own journey toward his Ph.D. has served as an inspiration for me. I also thank my siblings, Khasim and Khadar, who have constantly supported my decisions to pursue my interests. I also feel immensely fortunate to have highly supportive parents-in-law, Leo and Liisa Vauhkonen. It has been humbling to receive such warm support and constant well wishes from them and from my sister-in-law Anu and her family to complete my research.

I also thank my near and dear friends who have constantly stepped in to help me with children and logistics in the busiest of times with my exams and assignment submissions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The main motivation for this research has been to understand the underlying connections and dynamics between macro attributes for policy interactions and patterns at macro level in the education domain using large-scale data analysis. Policy planning and decision making require tremendous amounts of data gathering and analysis for effective allocation of resources and successful outcomes, whether in education, health, business, or some other domain. Use of statistical models has been a mainstay for policy decision making in the areas of education; however, the emergence of big data analytics and machine learning models have extended the capabilities for extracting numerous insights in education data. Similar to preventive health care, where big data analysis and machine learning models have made tremendous strides, these models have come into increasing in use in the domain of learning analytics in general for both personal computing–based and web-based learning systems. However, macro-level implementation for policy planning is extremely sparse. Policy planning at macro levels is still very time consuming and heavily influenced by varied groups' agendas versus the reality on the ground. The research presented here takes, as an example, California statewide high school macro-level education data from multiple years and shows how data mining and machine learning models can bring additional insights and add value to macro policy planning; in addition, as a future step, it provides a use case as a strong foundation for linking to micro-level analysis based on simulated data.

1.1 Research Questions

For the present research, we define the lowest granularity of macro levels of data to be at the level of each school, with school districts as the next higher level. The type of macro data that exists for California education system is at the individual school level on several attributes, ranging from student performance by school on math and English, teacher/student ratio, number of years of teaching experience, teacher salaries, meal programs, student ethnicity, school finances at district level, and so on. Overall data falls into the categories listed in Table 1, and each dataset in turn has multiple attributes. Full lists of these attributes are provided in Appendices 1, 2, and 3.

Table 1: Datasets at school and district levels for years 2012–2013, 2013–2014 ("2014"), and 2014–2015 ("2015")

Student performance data
Student demographics
Teacher experience and salaries
District-level finance and demographics and expense per student
Online text reviews of parents/guardians/students on select number of high schools in California

By analyzing this data using data mining and machine learning models, the research tries to create a technical framework that can address the following questions.

1. Can we streamline large amounts of education data with numerous attributes and various data types together at a macro level and group schools based on multiple attributes to discover underlying common factors between schools?
2. Can new schools be automatically classified to a category? Can the category of the school be reassigned based on parameter changes from year to year?

3. In this sea of big data, how can we discover and identify influential macro-level patterns that have crucial impact on the success of students in an educational system at school level?
4. Can text mining of the social data input from public stakeholders correlate with the metrics collected by the schooling system?

In addition to these questions, the research also presents a possible foundation for future extension of the analysis to micro levels, so as to link macro and micro levels in order to address questions such as the following:

- Can we establish a technical framework that can bring together the collective knowledge of the teachers under a system of seamless use with efficiency to enhance the learning experience of students?
- Can we create a feedback loop mechanism between patterns that emerge out of micro data and the macro level so that economic, human, and infrastructural resources can be channeled to the appropriate groups of students?

Answers to these questions are vital in transforming education that can cater to individual learning. This study presents the results from a prototype model of an analytics framework system that attempts to provide answers to the questions just raised for high school data from the state of California.

1.2 The Problem Context

The advent of big data has permeated many domains in analyzing patterns, finding associations, and predicting future outcomes. It is especially so in consumer-driven businesses, where it is heavily used to formulate corporate policies. It is being adapted to health care

contexts, both in finding the optimal care for a patient at the micro level and, from a macro perspective, in predicting health outcomes, diagnostics, recovery, and financial aspects for patient cohorts. Health care modeling closely relates to modeling in educational data mining. In the context of education, data mining and machine learning models have been successfully adapted for understanding learning behavior at the micro level. However, such models have been quite slow in making inroads into policy planning at macro levels. Since defining a policy for meeting every individual difference is not a viable method, using mining techniques to find broader patterns and associations would enable formulation of policies that optimize the existing economic and service resources to achieve the set goals. Aside from knowing how best a certain category of students can learn, macro policy planning also involves understanding the broader trends of technological, economic, and social aspects and opinions of different stakeholders, which need to be taken into account in planning curriculum changes for adapting to the future. For example, aside from core theoretical foundations, current technological progress demands a significant amount of change in curriculum to involve heavy scientific, technical, engineering, and mathematical (STEM) emphasis, so that the next generation's human resources are available in required numbers to maintain the innovation and technological progress needed to solve the problem we will encounter in our efforts toward sustainable growth. Information aspects related to these kinds of broader trends often exist in the form of combinations of textual and numeric data. By integrating the qualitative and quantitative results, a clearer picture of trends emerges, thereby enabling a reliable and effective way of informing policy decision making.

Educational data mining (EDM) has emerged as a modeling and design paradigm for formulating algorithms to identify patterns, find associations, and make predictions at micro and macro levels. At a micro level EDM can classify and characterize learners' identifiable

capabilities and performance; at a macro level it can assist in gaining insights to domain knowledge content, assessents, educational functionalities, and applications (Luan, 2002). Education policy consists of setting principles and rules for achieving certain goals in the sphere of education. Policies are formulated at several levels (school, district, state, and country). In order to achieve the goals defined by these policies, certain protocols are established that, in turn, define the procedures for accomplishing the specific tasks set forth to achieve the goals. Some of these tasks have to be performed by individuals (for example, learners) and some by other stakeholder organizations (schools, districts, etc.). Traditionally these policies have been made by committees consisting of a team of researchers and representatives of governing boards and political entities, based to some extent on standard statistical analysis. This often leads to policies based on the assumptions and personal biases of members and political parties involved in those committees. However, with the increasing pace of technological progress, democratic processes involved in decision making, and globalization forces impacting day-to-day lives, this traditional method of policy planning falls short of encompassing the changes needed and the direction the societies are taking. With the advent of big data a clear perspective can be obtained on many aspects and levels by interconnecting information from all agents involved in the process (Moreno-Jimenez et al., 2014). Students, parents, teachers, industry leaders, government officials, and nation builders all have their say and stake in building and guiding society through knowledge. Thus, bringing the opinions, aspirations, and expectations of all these stakeholders together is imperative, and in the current age the possibilities of achieving this are very realistic. Although it is a gargantuan task to provide a solution that is all encompassing, this proposal puts forth a method that, although it tackles only a part of the greater problem, forms a critical component.

This study explores macro data analysis with the following steps and lays a possible foundation to link to micro-level analysis for future research expansion.

- **Macro analysis with clustering**: A large dataset of high schools from the state of California is clustered on their performance indicators, primarily in mathematics, but also on a large set of economic and social attributes of these schools. At a macro level, data is analyzed at a district level as well as school level for the years 2013, 2014, and 2015. Three types of clustering models are applied to determine the clustering algorithm that best suits data in the educational context with combination of quantitative and qualitative variables.
- **Classification model**: A classification model is developed based on the clustering approach described in the previous item. The clusters formed are treated as classes, and a classification model is tested using two approaches: random forests and linear discriminant analysis for multiclass classification. The resulting classification model can automatically classify a new school or reassign a school with changed data into a different class.
- **Frequent patterns and associations at school Level:** Frequent-pattern analysis approaches data from a different direction than clustering does. The main aspect that is analyzed is which attributes co-occur most frequently at a certain level and what associations exist between these attributes. An example of an association rule is: ”If the number of teachers with less than 2 years of experience is lower in a school, the graduation rate is higher for students.” Such insights can be gained from finding frequent patterns and establishing association rules. In this research, frequent-pattern analysis is applied at school-level and district-level data to find association rules.

- **Regression analysis**: Regression analysis is performed in two contexts. In the first, the output from the frequent-pattern analysis is subjected to further regression analysis to establish the nature of the association between the frequent pattern attributes to reveal additional angles to the impact between the attributes that form the topmost association rules. In the second context, regression is run to identify whether it can better explain the variation in the variable of interest with additional attributes.

The following are possible future research steps:

- **Recommendation systems for harnessing collective knowledge:** A sample of students is taken from each of the clusters formed, and individual performance data is simulated based on real performance metrics of school population, along with teacher feedback for each student and their growth reports. A collaborative system is constructed using the recommender system collaborative filtering algorithm. This system will enable teachers to quickly find similar student profiles to the student of their interest and study other teachers' approaches in enhancing the learning method.
- **Identifying Macro Patterns from micro data**: The collaborative system described in the previous item becomes very efficient over time as data increases, and it identifies larger common patterns among similar students' learning success with teacher recommendations. These larger patterns are then propagated to the macro decision makers, who can easily observe what kind of methods are working best for a certain group of students, so that the appropriate resources can be effectively targeted to similar set of students.

The following sequence of machine/statistical learning and data mining methods is proposed to identify the complex collective information that can emerge from large amounts of data. Figure 1 illustrates the flow.

1.3 Key Contributions

In this research I have brought forward the following contributions:

1. Demonstrated a technical framework to analyze large amounts of macro-level education data using machine learning and data mining techniques for drawing crucial insights for multi layered data in policy planning contexts.
2. Identified models that can handle multiple data types for unsupervised analysis. Compared three clustering techniques with California statewide high school data to identify factors for math performance indicators.
3. Created a technical framework that can that can link various levels of macro and micro data for deriving insights using multiple techniques. The school-level data analyzed is linked to district-level finance information to observe the impact of various funding sources on performance indicators. It is further linked to a simulated level of micro data. This large-scale study has not been performed before on a macro scale for high school data and in addition link it to micro levels for a recommendation system for collaborative use of teacher's insights on similar students.
4. I propose a way for identifying class labels for categorizing schools effectively into groups by applying classification models to clustering outputs. Identifying labels manually is an expensive and enormous task. The resulting model of applying classification serves two purposes: (1) to verify that clusters are genuine clusters by

creating a training and testing set of the data and (2) to classify new schools or reclassify existing schools quickly based on the parameter changes that take place during the academic years.

5. Implemented a topic modeling method for online reviews written by students, teachers and community stakeholders on high schools to draw insights that are difficult to gather from official education educational data collection methodologies. This acts are a initial medium to launch more official surveys on concerned topics.
6. Demonstrated a recommendation system using collaborative filtering methods that can be used by teachers to utilize the collective knowledge of other teachers state wide in identifying similar students for recommending a learning approach that works the best for a particular student.

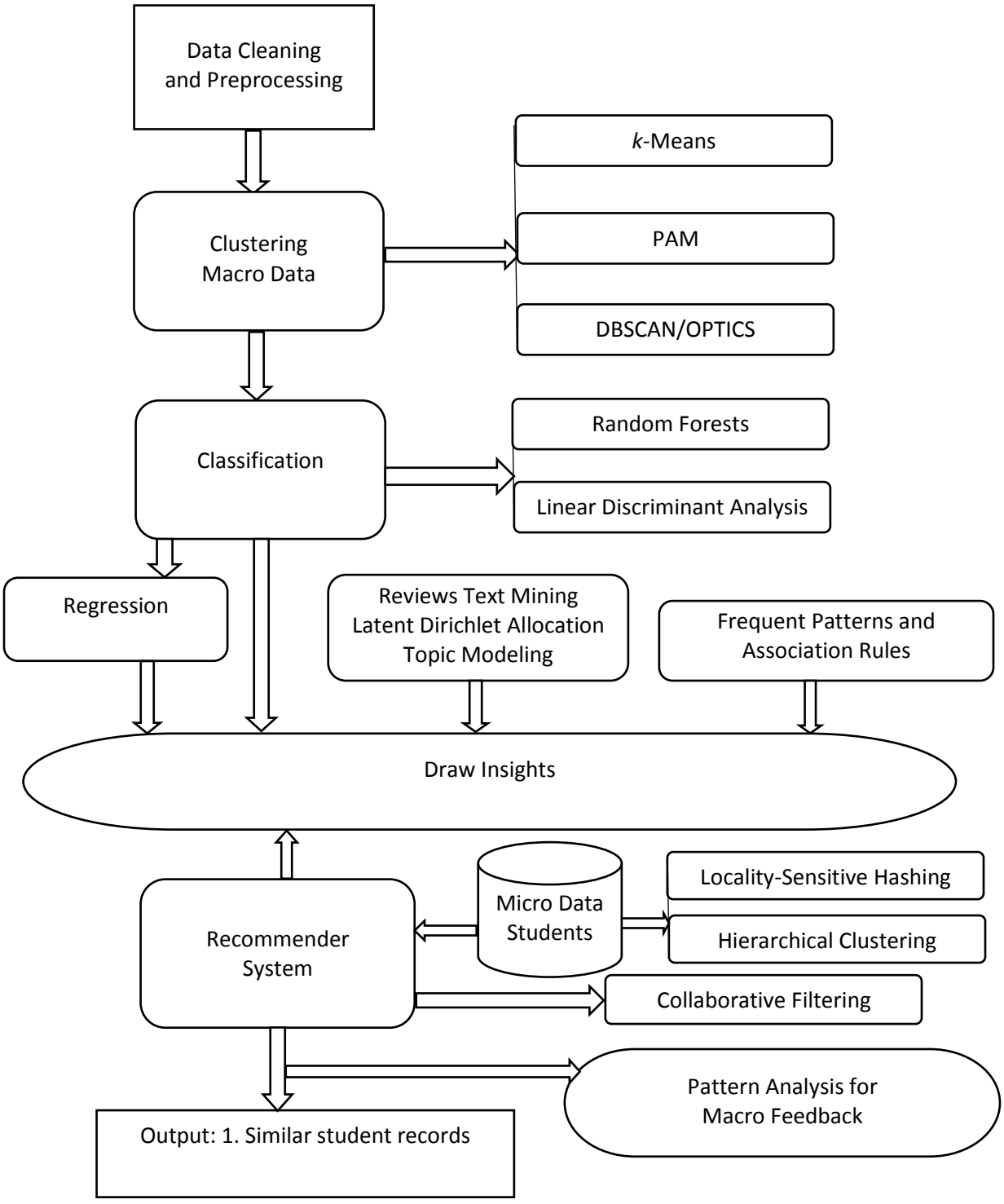


Figure 1: Proposed set of methods and data flows.

7. I present a method to merge text and numeric attributes for analysis and show that by doing so the analysis models can provide more insights.

1.4 Dissertation Structure

Physically, the description of the research is divided into following chapters:

Chapter 2 states the research motivations and problem along with an in-depth discussion of the role of machine/statistical learning and data mining models in the domain of education policy planning. It also explores previous research conducted in this domain.

Chapter 3 describes the datasets used and the machine learning methodologies applied in analysis of the data. A total of 1,636 high schools from state of California were analyzed with multiple attributes belonging to both quantitative and qualitative types.

Chapter 4 elaborates on the computational methods used in the analysis and the algorithms assessed for each part of the framwork.

Chapter 5 presents the results and analysis of the data using various methods in the system framework.

Chapter 6 discusses the results in the light of effective policy planning.

Chapter 7 draws conclusions and plans for future research.

CHAPTER 2

RESEARCH MOTIVATIONS

The fundamental concept that this research explores is that large-scale analysis of macro data using machine learning and data mining models can extract more insights for effective policy planning by linking varying levels of macro data with multiple data types. By applying the appropriate analytic methods, insights and knowledge can be geared toward precisely identifying problem areas, assessing possible solutions, and allocating required resources at the macro level. This chapter explores how big data analysis can add value to education policy decisions, but before delving deeper, a brief overview of data mining in the field of education is explored.

2.1 Data Mining in Education

Data mining (DM) is the use of a computer-based information system (CBIS) (Vlahos, Ferratt, & Knoepfle, 2004) to scan huge data repositories, generate information, and discover knowledge. The meaning of the traditional term "mining" lends itself favorably to data mining, as the goal is to discover and extract valuable knowledge hidden in an ocean of data. Data mining is a part of the knowledge discovery and data mining (KDD) domain. Data mining approaches lead to identifying data patterns, organizing information of hidden relationships, structure association rules, estimate unknown items' values to classify objects, compose clusters of homogenous objects, and unveil many kinds of findings that are not easily produced by a classic CBIS (Peña-Ayala, 2014). The analysis output of data mining represents a valuable support for decision making at all levels. In the context of education, it is a novel approach that can be developed for knowledge discovery, decision making, and recommendation (Vialardi-Sacin, Bravo-Agapito,

Shafti, & Ortigosa, 2009). In the current age, the use of data mining in the education arena has proved inescapable and has given rise to the domain of educational data mining (EDM) research field (Anjewierden, Kollöffel, & Hulshof, 2007).

The data mining field is an amalgamation of disciplines such as probability (Karegar, Isazadeh, Fartash, Saderi, & Navin, 2008), machine learning (Witten, Frank, & Hall, 2011), statistics (Hill & Lewicki, 2006), soft computing (Mitra & Acharya, 2003), artificial intelligence (Bhattacharyya & Hazarika, 2006), and natural language processing (McCarthy & Boonthum-Denecke, 2011). Given the existing research, a majority of data mining approaches, probability, machine learning, and statistics constitute 88 percent of EDM approaches (Peña-Ayala, 2013).

2.2 Rise of Education Data Mining

EDM is an emerging field with the primary goal of applying data mining techniques and tools to education data (Baker & Yacef, 2009). Researchers in EDM apply this domain to tackle problems from institutional effectiveness to enhancing student learning by adopting several standard methods of analysis developed in the data mining domain. An example of this development process is the Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Leventhal, 2010). This process has been applied in numerous EDM studies (Luan, 2002; Vialardi et al., 2011; Wang & Liao, 2011) and is a standard practice for many data mining methods.

The rise of EDM can be traced back to the year 2000, as 98 percent of EDM research has been published starting in 2000. As a specialized field, EDM is still in a relatively juvenile state. As it has grown, EDM has shifted from isolated papers published in conferences and journals to

dedicated workshops, an international conference on educational data mining, a specialized journal of EDM, a handbook (Romero, Ventura, Pechenizkiy, & Baker, 2011), and a growing community of experts.

A major advantage of EDM is that it can be discussed in theory only to a small extent, and the majority of it has to be implemented. As previously mentioned, it has been applied to several domains in education such as student assessment, online learning systems, individualized learning models, and institutional effectiveness. The majority of the implemented systems in EDM fall under two perspectives: *predictive* and *descriptive*. Predictive models forecast future outcomes based on existing data, and descriptive models explain patterns and correlations in existing data. Table 2 (Peña-Ayala, 2014) shows a sampling mix of predictive and descriptive models used in EDM research.

As the ability to collect more data increases, more complex models with a combination of descriptive and predictive approaches are required. The degree of complexity in education policy planning certainly calls for such a combination of approaches. The following section discusses EDM research and implementations so far.

2.3 Data Mining in Education and Policy Planning

The aspect of increasing institutional effectiveness in delivering enhanced outcomes for the students constitutes the main component of education policy planning. Although EDM has made many strides in research in the area of individualized student learning, it is extremely rare to come across analytical systems at macro level. At a macro level, what is available is statewide databases that contain information on school and student performances, statistics on teachers and

staff, financial statements, etc. While the data is available, the analytical method and the approach that is needed to analyze and interpret the data is lacking.

Table 2: Example of predictive and descriptive models used in EDM research

Author of the EDM approach	Discipline	Model	Task	*M*: method *T*: technique	*A*: algorithm *E*: equation *F*: frame
Maull, Saldivar, and Sumner (2010a)	Machine learning, probability	Descriptive	Clustering	*M*: IBL,Bayes theorem	*A*: k-means, EM
Vuong et al. (2011)	Statistic	Descriptive	Functional dependency	*T*: dependency structure	*E*: overall graduation rate
Brunskill (2011)	Probability	Descriptive	Clustering	*M*: Bayes theorem	*A*: EM
Durand, LaPlante, and Kop (2011)	DP	Predictive	Classification	*M*: HMM	*A*: Markov decision processes
Su et al. (2011)	Machine learning	Descriptive	Clustering	*T*: frequencies, variability	*E*: statistical
Hershkovitz et al. (2011a)	Machine learning, statistic	Descriptive	Sequential pattern	*T*: hierarchical clustering	*A*: two-step clustering *E*: descriptive statistic
Chi et al. (2011a)	DP	Predictive	Classification	*M*: HMM	*A*: Markov decision process, four reinforcement learning based feature-selection
Wauters, Desmet, and van den Noortgate (2011a)	Probability	Predictive	Classification	*M*: latent class analysis	*A*: latent class analysis modeling, clustering with latent class analysis
Xu and Recker (2011)	Statistic, probability	Descriptive	Clustering	*M*: decision tree, ensamble, Bayes theorem, distance-based learning	*A*: NaiveBayes, C4.5, KNN, boosting
Barracosa and Antunes (2011a)	Machine learning	Descriptive, predictive	Classification, sequential pattern	*M*: context-free language	*A*: pushdown automata *F*: meta-patterns
Cai et al. (2011)	Statistic	Predictive	Classification	*T*: descriptive statistic	*E*: statistical
Gaudioso et al. (2012)	Machine learning, probability	Predictive	Classification	*M*: decision tree, Bayes theorem, rules induction	*A*: naiveBayes, J48, Jrip, PART
Scheihing, Aros, and Guerra (2012)	Machine learning	Descriptive	Clustering	*M*: latent class analysis, IBL	*A*: k-means

Reproduced from Peña-Ayala (2014).

Policy planning in general is very dynamic, requiring constant assessment of conditions at all levels of education. Policymakers and administrators today demand more data and insights for better decision making. However, a lack of detailed analytic frameworks in this domain hinders making informed policy decisions that can be cost effective from a financial and human resource perspective under current economic and political conditions, which lead to budgetary constraints. Table 3 presents the number of EDM tools implemented in each category of education, and it can be seen that there are none under the category where macro policy analysis and institutional-level decisions can be made. Most of the implementations exist at the level of student records or online learning and computer-based learning systems.

Table 3: EDM implementations by domain in education

Education Systems	**Number Implemented**
Component-based development systems (CBDS)	
Intelligent tutoring system	88
Learning management system	20
Conventional education	20
Computer-based educational system	115
Student-based	
Student behavior modeling	48
Student performance modeling	46
Assessment	45
Student modeling	43
Student support and feedback	21
Curriculum, domain knowledge, sequencing, teacher support	19
Institution-based	
Data mining–based	None
Education databases with no advanced analytics	15

Multiple sources from *Journal of Educational Data Mining*.

The available research is largely on web-based learning systems and some classroom-based observational analytics. Increase in educational software products and dedicated data collection by state agencies have led to large repositories of data that reflects how students learn and perform in a given setting (Koedinger, Cunningham, Skogsholm, & Leber, 2008). Simultaneously, the popularity of Internet-based education has developed into a new dimension of e-learning with content access, instruction, and self-paced learning methods (Castro, Vellido, Nebot, & Mugica, 2007). Studies in web-based learning show that monitoring student learning is an important input to enhancing the quality of teaching (Mazza & Dimitrova, 2004). Online learning systems often lack direct contact between students and teachers, and this lack leads to deficiencies in understanding where students need input, where to make changes to teaching, and so on. The only way to derive information is from data collected from the online interaction

process from students, survey inputs, and system log files. Hershkovitz and Nachmias (2009a) researched online student learning behavior with respect to their pace in online learning and online information organization using analytics techniques. The research shows in general that students' pace of learning is not consistent; however, the analysis does not support informing macro policy extensively. Although the research provides information on improving web-based education, further analysis and exploration can inform how well this web-based education can be integrated in meeting macro goals at large. A study led by Cohen and Nachmias (2011) shows that in web learning environments, micro measures for instructors and macro measures for institutional policy makers can be gathered. Measures such as content usage, student–student and student–teacher online interpersonal interactions, variance in courses, assessment measures, and several other micro measures were analyzed to improve student learning experience and teacher improvement. At the macro level, aspects related to three measures (cost efficiency, instructional quality, and university prestige) were collected. Although the study was targeted at understanding micro and macro issues for a predetermined set of measured attributes from a parametric perspective, it did not apply data mining models to detect underlying patterns and new insights that would connect micro and macro issues without narrowing them down to just three aspects.

Currently existing data mining tools such as DBMiner (Han et al., 1997), SPSS Clementine (now IBM SPSS Modeler; IBM, n.d.), Weka (Machine Learning Group at the University of Waikato, n.d.), RapidMiner (RapidMiner, Inc., 2017), and distributed computing platforms such as Hadoop and Spark are very powerful in dealing with unstructured data analysis. However, they require complex implementations and customizations, as they are not

oriented toward analyzing educational data. The skills required for implementing on these platforms are often very technical and different from the domain of an educator.

Efforts have been made to develop tools geared toward educational analysis. TADA-Ed (Tool for Advanced Data Analysis in Education; Merceron & Yacef, 2005a), enables teachers to visualize and mine students' online homework and exercises to derive patterns that can be useful in enhancing students' learning. EPRules (Romero, Ventura, & De Bra, 2004), was developed based on association rule mining to improve online courseware by employing the GBGP (Grammar Based Genetic Programming) method, widely used in genetic programming, where each individual is represented by a tree derived from a grammar defined by the user that represents all the possible prediction rules. By choosing the top rules and letting the tool analyze further, a teacher can predict student performance and difficulties in a particular course or subject. This tool, however, does not enable looking at groups of students at a large scale, and it also requires technical knowledge for implementation and customization.

Later research by Romero, Ventura, Zafra, and De Bra (2010) implements an association rules tool based on the collaborative filtering method. Collaborative filtering is a method that is used in recommendation systems based on large datasets. It involves filtering for information or patterns based on a larger set of similar users or agents or items to recommend for new cases. The tool developed by Romero and coworkers derives rules on e-learning courses where each course is tagged with certain attributes by domain experts. Once the course is in the system, all activity by a student in relation to the course is derived and stored. Attributes including courses taken, exercises, forums, quizzes, chats, and questions are all recorded. By applying the association rules and collaborative filtering, some rules are derived. Based on the rules, some recommendations are given for each instructor to improve her or his course as well as share the

results with other instructors. Although this research is extremely useful for understanding the performance of each student and each course, and it does try to derive some rules for teachers to gain insights, it does not attempt to address macro issues at a larger scale.

Many tools in EDM currently focus on micro aspects of analyzing student learning. Table 4 shows a list of existing EDM tools.

Table 4: List of EDM tools available

Tool	Objective	Reference
WUM tool	To extract patterns useful for evaluating on-line courses.	(Zaïane and Luo, 2001)
MultiStar	To aid in the assessment of distance learning.	(Silva et al., 2002)
Data Analysis Center	To analyze students' patterns and organize web-based contents efficiently.	Shen, Yang, and Han (2002)
Assistance tool	To provide a tool for students to look for the materials they need.	Shen, Han, Yang, Yang, and Huang (2003)
EPRules	To discover prediction rules to provide feedback for courseware authors.	(Romero et al., 2004)
KAON	To find and organize the resources available on the web in a decentralized way.	(Tane et al., 2004)
GISMO/CourseVis	To visualize what is happening in distance learning classes.	(Mazza and Milani, 2004)
TADA-ED	To help teachers to discover relevant patterns in students' on-line exercises.	(Merceron and Yacef, 2005)
O3R	To retrieve and interpret sequential navigation patterns.	Becker, Vanzin, and Ruiz (2005)
Synergo/ColAT	To analyze and produce interpretative views of learning activities.	(Avouris et al., 2005)
LISTEN tool	To explore huge student–tutor interaction logs.	(Mostow et al., 2005)
TAFPA	To provide helpful analysis of the cognitive process.	Damez, Marsala, Dang, and Bouchon-meunier (2005)
iPDF_Analyzer	To help predict interactive properties in the multimedia presentations produced by students.	(Bari and Benzater, 2005)
Classroom Sentinel	To detect patterns and deliver alerts to the teacher.	(Singley and Lam, 2005)
Teacher ADVisor	To generate advice for course instructors.	Kosba, Dimitrova, and Boyle (2005)
Teacher Tool	To analyze and visualize usage-tracking data.	(Zinn and Scheuer, 2006)
CoSyLMSAnalytics	To evaluate the learner's progress and produce evaluation reports.	Retalis, Papasalouros, Psaromilogkos, Siscos, and Kargidis (2006)
Monitoring tool	To trace deficiencies in student comprehension back to individual concepts.	(Yoo et al., 2006)
MINEL	To analyze the navigational behavior and the performance of the learner.	(Bellaachia and Vommina, 2006)
Simulog	To validate the evaluation of adaptive systems by user profile simulation.	(Bravo and Ortigosa, 2006)
LOCO-Analyst	To provide teachers with feedback on the learning process.	Jovanovic, Gasevic, Brooks, Devedzi, and Hatala (2007)
LogAnalyzer tool	To estimate user characteristics from user logs through semantics.	Andrejko, Barla, Bielikova, and Tvarozek (2007)
Learning log Explorer	To diagnose student learning behaviors.	Jong, Chan, and Wu (2007)
MotSaRT	To help the on-line teacher with student motivation.	(Hurley and Weibelzahl, 2007)
Measuring tool	To measure the motivation of on-line learners.	(Hershkovitz and Nachmias, 2008)
DataShop	To store and analyze click-stream data, fine-grained longitudinal data generated by educational systems.	(Koedinger et al., 2008)
Visualizing trails	To mine and visualize the trails visited in web-based educational systems.	Romero, Ventura, and Salcines (2008)
Measures Tool	To analyze rule evaluation measures.	Ventura, Romero, and Hervas (2008)
Solution Trace Graph	To visualize and analyze student interactions.	(Ben-Naim et al., 2008)
Decisional tool	To discover factors contributing to students' success and failure rates.	(Selmoune and Alimazighi, 2008)
Concept map generation tool	To automatically construct concept maps for e-learning.	Lau, Chung, Song, and Huang (2007)
Author Assistant	To assist in the evaluation of adaptive courses.	(Vialardi et al., 2008)
Meta-Analyzer	To analyze student learning behavior in the use of search engines.	Hwang, Tsai, Tsai, and Tseng (2008)
CIECoF	To make recommendations to courseware authors about how to improve courses.	(Garcia et al., 2009b)
SAMOS	Student activity monitoring using overview spreadsheets.	Juan, Daradoumis, Faulin, and Xhafa (2009)
PDinamet	To support teachers in collaborative student modeling.	Gaudioso, Montero, Talavera, and Hernandez-del-Olmo (2009)
E-learning data analysis	To analyze learner behavior in learning management systems.	Psaromiligkos, Orfanidou, Kytagias, and Zafiri (2009)
Refinement Suggestion Tool	To confirm or reject hypotheses concerning the best way to use adaptive tutorials.	Ben-Naim, Bain, and Marcus (2009)
Edu-mining	To recommend books to pupils.	Nagata, Takeda, Suda, Kakegawa, and Morihiro (2009)
AHA! Mining Tool	To recommend the best links for a student to visit next.	Romero, Ventura, Zafra, and de bra (2010)

Reproduced from Romero et al. (2010).

2.3.1 Macro Research

The IBM study on Mobile County, Alabama, public schools (IBM, 2011a) took a large initiative of implementing a data warehouse system that integrated academic and administrative information from 95 schools in the county with capabilities of business intelligence and

performance of students. The results enabled the teachers to understand the pattern of at-risk students.

The Hamilton County, Tennessee, implementation of a data analysis system by IBM (2011b) on 78 schools to track performance metrics and model them to understand the low graduation and high dropout rates resulted in 8% increase in graduation rates. With built-in pattern recognition and predictive algorithms, the system gave teachers and administrators the capability to understand the key predictors that indicated a student failing or dropping out in the future. By identifying these predictors well ahead and providing the right intervention, positive results were obtained. This system also integrates data from many schools to an extent to find and categorize common patterns among at risk students.

Another advanced system that IBM (2013) is developing for schools in Gwinnett County, Georgia, is a project known as Personalized Education Through Analytics on Learning Systems (PETALS), which extensively uses machine learning, predictive modeling, deep content analytics, and advanced case management to identify learning needs of students and provide personalized recommendations for learning. By analyzing longitudinal student data from multiple sources, the goal of the system is to identify similar students, provide the right learning content automatically, and recommend teaching techniques for specific needs.

Though these IBM systems implement some level of macro data analysis, they do not connect the macro and micro levels as an automatic feedback system nor use collective knowledge and recommender systems to enhance student learning.

A micro and macro data mining analysis on the study of Barcelona ports used several methods like association rules, clustering and time series analysis (Nettleton, Fandiño, Witty, & Vilajosana, 2000). The results indicate that by analyzing micro levels of data, association rules

for macro policies can be derived, which enables optimal allocation of economic, infrastructure, and human resources.

Ubels, van Klinken, and Visser (2010) have explored the micro and macro planning gaps in the context of local groups and interests and larger national goals. The research indicates that streamlining goals from local aspirations to national level avoids conflicts.

2.3.2 The Micro–Macro Context for Future Research

From the studies just discussed, it can be seen that big data has made deep inroads at micro levels. The main challenge remaining is how to use the big data techniques to inform macro policy analysis from micro data in a seamless feedback mechanism. Since data mining and machine learning models have an immense capability, an analytic framework that can connect these two aspects is presented.

In order to have a macro-level system for analysis, a data warehouse is essential. Guan, Nunez, and Welsh (2002) propose a model for a data warehouse for storing education information, but they do not suggest the appropriate methods that can be utilized for analyzing the data in the data warehouse. The analytical framework system proposed here can connect with macro and micro levels; to bring the intended change at micro levels effectively takes enormous amount of analysis.

The following chapter covers the background research in technical methods chosen for developing the system.

CHAPTER 3

DATA COLLECTION AND PREPROCESSING

Data for education analysis exists in many forms. Traditionally, quantitative and qualitative data have been most often used for analyzing and coming to conclusions. With the dawn of data analytics and statistical models that have evolved into more complex machine learning and data mining algorithms, different types of data can be brought under the purview of analysis and merged together to gain deeper insights. The data for this research contains a mix of quantitative, qualitative, and text data. The advent of enormous data storage capacities enabled collection of large quantities of observational data. This provides an opportunity to detect actual patterns that exist.

Although the large amount of observational data is a boon for applying the machine learning and data mining algorithms, it also has complex problems of data being not "clean" and readily available to analysis. The data will have many messy elements such as missing values, incorrect values, and invalid characters. This chapter presents, in detail, the types of data utilized for research presented here, the problems encountered with the data, the data cleaning process applied, and the transformations applied to make the data analyzable with the algorithms.

3.1 Data Description

The data used for the analysis falls into the following categories:

I. Large-scale California-wide high school data on Adequate Yearly Progress (AYP) performance metrics for three years: 2013, 2014, and 2015 (attributes tabulated in Appendix 1)

II. California-wide high school data on pupil demographics, second language learners, free meal program participation, and teacher qualifications (attributes tabulated in Appendix 2)

III. California-wide district-level finance and demographic information, including attributes on various sources of revenues (attributes tabulated in Appendix 3)

For a future research step, individual student yearly report data, including grades on various subjects, text-based trimester comments from teachers, and recommendations for progress, was simulated based on the mean and standard deviation performance measures of each school selected as a sample set from the clusters.

The datasets are described in more detail as follows.

I. **High school performance** data at school level from the state of California: A total of 1,636 high schools were used for the macro-level analysis for identifying school clusters. This includes only traditional high schools. Schools for differently abled students and vocational high schools have been excluded. The data consists of multiple variables indicating school-level performance metrics for the years 2013, 2014, and 2015 from the Adequate Yearly Progress Report. This data has been obtained from the California Department of Education. Table 5 lists some of the multiple attributes used for the analysis. The data was originally provided as an Excel file, which was converted to a comma-separated values (CSV) file for analysis in the R language. This raw data had to be further cleaned before it could be analyzed, as it had some problems to overcome. The problems and the cleaning methods are discussed in detail in Section 3.2.

Table 5: Sample of attributes used for clustering from all California high schools

Field Name	Type	Width	Description
cds	Character	14	County/district/school code
rtype	Character	1	Record type: D=district, S=school, X=state
type	Character	1	Type: 1=unified, 2=elementary district, 3=9–12 high district, 4=7–12 high district, E=elementary school, M=middle school, H=high school
sname	Character	50	School name
dname	Character	50	District name
cname	Character	50	County name
Crit1	Character	2	Number of AYP criteria met, based only on participation rate and additional indicators
Crit2	Character	2	Number of AYP criteria possible
m_enr	Character	7	Schoolwide or Local Education Agency (LEA)-wide math enrollment
m_tst	Character	7	Schoolwide or LEA-wide math, number of students tested
m_prate	Character	5	Schoolwide or LEA-wide math participation rate
m_val	Character	7	Schoolwide or LEA-wide math valid scores
m_prof	Character	7	Schoolwide math number of students scoring proficient or above
m_pprof	Character	5	Schoolwide math percent of students scoring proficient or above
mp_aa			

Adequate Yearly Progress (AYP) data from California school system. Sample of important attributes are listed here. Please refer to Appendix 1 for a full list of attributes.

II. **Second dataset and data spread in multiple tables**: The second step of analysis using regression methods was performed to observe whether some other variables correlate with the clustering, such as number of students in the free lunch program, number of migrants, and teacher qualifications. The information for these attributes had to be fetched from other multiple tables provided by the California Department of Education. Since certain attributes were not available for all 1,636 schools, for some steps, only a

sample set of schools from each cluster derived in the initial analysis was taken. Please refer to Appendix 2 for a full attribute list.

III. **District-Level Finance Data**: There are altogether 87 high school and 330 unified school districts in the state of California. For the research implemented here, a total of 337 districts are taken into account, including high schools. The dataset consists of 51 attributes used in the analysis. Appendix 3 details the list of attributes.

Next steps: simulated data with merged numeric and text data: The three datasets just described are macro datasets that do not contain individual, micro-level data. Micro-level data for individual student performance is not legally available from the school system. To establish a technical flow for testing algorithms in future research, micro-level data at the individual level was simulated based on the mean and standard deviation available for the schools taken as a random sample from each of the clusters created from the analysis in step 1. The attributes selected are a reflection of the California system's high school grade reports and annual student reports, containing data on student performance on the courses studied, teacher comments for each trimester, and recommendations for improving performance or college plans.

One of the main goals of the research is also to demonstrate a combined analysis of text data and numeric data. The trimester comments and recommendations of the teachers are text data. After text mining was done, the results were used in the form of numeric variables for merging with the rest of the student report variables for analyzing micro-level data. The following section elaborates further on the text data cleaning and transforming it into a dataset.

3.2 Data Cleaning

Significant efforts are needed to transform this raw data into "clean" or "tidy" data that algorithms can process. It is said that 60% to 80% of the effort in data analytics goes into cleaning and transforming data into sets that can be analyzed (Dasu & Johnson, 2003; Famili, Shen, Weber, & Simoudis, 1997).

When handling data for analysis, it is not out of norm for the initial, raw dataset to be "messy" and "noisy." *Messy* data contains invalid values and characters or missing values or lacks a proper structure to analyze (Wicham, 2014). *Noisy* data implies a random error or variance in a measured variable (Han, Kamber, & Pei, 2011). A "cleaned" or "tidy" dataset is a necessity for any machine learning or data mining algorithm to find meaningful insights.

Tidy or clean data has been mapped into a structure with the following three main characteristics (Wicham, 2014):

1. Each variable forms a column.
1. Each row is an observation.
2. Each type of observational unit forms a table.

By establishing a structure to the meaning of data, it lends itself for efficient processing using statistical methods or computer algorithms. The same rules are applied to unstructured text data such as web pages. In the simplest form, each document (a text document or web page) forms a row and each word a variable (column). The frequency of the words in the document is tallied in the cell where the row and column intersect. Discussion of several specific issues encountered in cleaning up this data follows.

Mix of school-level, district-level, and state-level data: The original data file consisted of a mix of district- and state-level data and school data for all levels (elementary, middle, high,

special education, vocational schools, charter schools). Since the focus of this research is on high school performance, data had to be separated. A total of 2,476 rows of high school data were present, out of which missing values and invalid data were major problems in some observations, as can be seen in Table 6.

Table 6: Sample values of a few attributes

m_val	m_prof	m_pprof	m_ppm
3188508	1065490	33.4	–
17	0	0	–
419	95	22.7	–
85	7	8.2	–
58	12	20.7	–
77	19	24.7	–
96	84	87.5	–
77	21	27.3	–
1	–	–	–
14	0	0	–
276	38	13.8	–

Missing values: Values that are absent where there should be data are classified as missing values. Missing data, in turn, is classified into "missing at random" and "not missing at random." Data is missing at random if the reason it is missing is not related to the actual missing values. This type of missing data does not pose significant problems if removed. Data is not missing at random, on the other hand, when the very cause of it being missing is related to the missing values.

Several methods have been proposed to deal with missing values, such as the following:

1. Removing records with missing values and analyzing the remaining data. This can pose problems where the datasets are small.
2. Imputing values based on neighborhood values and treating them as observed values.

3. Using statistical approaches such as regression or expectation–maximization (EM) algorithms, to estimate the missing values.

Cleaning method selected: For the purpose of this research only those observations that have full data available have been considered, as omitting the records with missing values did not impact the main goal of research. Out of 2,427 high schools in California, data from 1,665 high schools that contained data for all observations has been used for analysis. Missing values were not imputed, because it was essential to have a true picture of analysis in this context, and adequate data was available for analysis.

Invalid characters: There were many invalid characters as well as missing values. Values that contained these characters were treated as missing values. An example is shown in the sample data snippet in Table 7.

Table 7: Sample of invalid and zero characters

m_val	m_prof	m_pprof	m_ppm
17	0	0	–
419	95	22.7	–
85	7	8.2	–
58	12	20.7	–
77	19	24.7	–
96	84	87.5	–
77	21	27.3	–
1	–	–	–
14	0	0	–
276	38	13.8	–

Zero values were treated as actual values after they were researched and confirmed with the California department of education.

Apart from treating the missing values similarly as just discussed, the new data was merged with cluster analysis data, based on the county/district/school code. This ensured that available data was complete from the combination of multiple datasets.

Cleaning and parsing the text for analysis: Cleaning of text data—in fact, cleaning of all raw data—consumes more time than the analysis itself but is a necessary step in feature extraction and preprocessing the lexicon. These extracted features should be able to capture the content of documents in such a way that documents with similar content with different terminology would have similar features. This involves the following steps.

a. *Removing stop words:* Stop words are words such as "is," "the," or "of" that are very frequent and do not contribute significantly to the meaning of the sentence. Figure 2 has an example taken from the Slater school debate.

increas number children area (born move new folk take advantag newli creat job mv) prudent rush school open. gather point handi community, increas sens communiti & communiti spirit, make get school easier & faster without car drive gridlock town

Figure 2: Example of removal of stop words.

b. *Stemming* or *lemmatization:* Stemming is the process of removing suffixes and prefixes, leaving the root or stem of the word. For example, "walk," "walking," and "walked" all have the same base word, or lemma: "walk." Stemming reduces the number of unique words (i.e., it reduces the dimensionality of the dataset) and improves system performance. However, stemming without discernment can create problems for clustering or classification methods (Solka, 2008). For example, stemming "relativity" to "relate" or "probate" to "probe" can cause problems for certain methods of analysis because the meanings of the words in these pairs are very distant from each other. Stemming is an important part of data cleaning and,

when applied in a systematic way, results in saving computational resources and reducing redundancy. The Porter stemming algorithm (Porter, 1980) is a popularly used stemming procedure, and many tools based on variations of this algorithm exist.

Figure 3 shows an example of how a teacher's comment gets transformed after removal of stop words and stemming.

Positive in class. Participates in discussions. Needs improvement in writing. Shows lot of interest in Math. Demonstrates respect and kindness towards peers and teachers

[1] "posit class particip discuss need improv write show lot interest math demonstr respect kind toward peer teacher \n\n"

Figure 3: Result of stop word removal and stemming on teacher's comments.

Section 3.3 describes in more detail the transformations applied to text data and the data types just discussed.

3.3 Data Types and Preprocessing and Validation

Once the dataset is cleaned and structured, the next step is to assess and explore whether the variable values are directly usable or need some transformations to enable them to be analyzed. Attributes of continuous value, categorical attributes, scaled data, and text descriptions exist in the dataset. Certain numeric attributes can be used directly as they are or can be subjected to some form of normalization. Nominal, binary, scaled, and text data, however, each require their different methods to be preprocessed. This section explains the steps taken for transforming certain attributes.

Continuous variables: Attributes with continuous values constitute a significant portion of a dataset. These attributes can be transformed using several techniques such as *z*-score normalization and log transformation. For the current research question, most of the continuous variables are scores of students, number of students, quantities of materials, and so forth. A *z*-score normalization would be suitable for many of these, and it also works as an easy method to detect outliers. Continuous variables can also be subjected to measures of central tendency and many calculations to measure similarity and dissimilarity.

Categorical variables: Categorical data is data that specifies a category or group to which an observation belongs. It can originate from continuous values as well as from qualitative forms. It can be a dichotomous attribute, which can have only two values such as "yes" or "no"; it can also be polytomous, or nominal, with multiple classes, such as grades A, B, C, or D or states in the United States. For most machine learning and data mining algorithms, these variables have to be converted into numeric codes.

A numeric ranking methodology to convert categorical variables to numeric labels has been followed. For example, the categorical attribute MetAttendTarg has three levels ("yes," "no," and "NA"). To enable effective clustering, these values have been converted to 1, 2, and 3, respectively. The nature of the data type has not been altered. Similarly, scaled attributes such as grades have been coded using similar numeric denominations.

Text document transformations: Applying computational methods to text for analyzing them requires transformations that have to be in numeric form. Since most of the similarity and dissimilarity measures require some form of numeric input, text data also need to be converted into a form where the words are represented in the form of numbers. This converted structure is called a term document matrix, described next.

Term document matrix: A term is a word. Keeping a count of how many times a word appears in a document or a corpus is an important part of text mining and is called *term frequency*. The terms that appear the most frequently are not necessarily the most important ones. Because very common words such as "student" or "class" do not contribute much to analysis of the comments given by the teacher, such extremely frequent terms are eliminated from term weighting. The importance of a word comes with its context in the document or corpus. This encoding of text and attributing some degree of importance is called *term weighting*. The method that is most adopted is called inverse document frequency, and the overall method is called *term frequency/inverse document frequency.* The formulation for it is given as:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

where $w_{i,j}$ is the weight of the term i in the document j, $tf_{i,j}$ = number of occurrences of term i in document j, N is the total number of documents, and df_i is the number of documents containing term i.

The following small example details how the transformation is implemented. Given the following three documents a, b, c:

a. Student is excellent in math and always completes homework
b. Student is very good in math and helpful to fellow class mates
c. Student is very motivated in sciences. Has to improve in English

When these are to be transformed, the initial step is to do the cleaning process of removing stop words and stemming. Then the frequency of each word in the documents is tabulated. Figure 4 shows the term frequencies for the three documents.

Terms																		
Docs	alway	and	class	complet	english	excel	fellow	good	hav	help	homework	improv	mate	math	motiv	science	student	veri
3	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	3	2

Figure 4: Term document matrix shows the document frequency of words from three documents.

Multicollinearities: When an attribute can be predicted from another attribute, there is a high degree of correlation between them. Often in datasets, the same information is represented in multiple forms, or measured through more than one attribute. When this happens, there will be multicollinearities in the data. This problem causes regression-based models to fail to identify the real relationships between the variables accurately. Thus, after assessment variables' importance using statistical methods, some variables are removed. For numeric attributes, principal component analysis is applied, and for categorical attributes, a chi-square correlational analysis is performed.

CHAPTER 4

COMPUTATIONAL METHODS

There are several approaches and methods in machine and statistical learning and data mining that can be applied for educational data. Choosing the right method for a given problem is one of the most important steps toward finding meaningful information from massive amounts of data. To identify which method or algorithm best fits the need, it is necessary to choose a few algorithms based on theoretical knowledge and research implementations that may share some similarities with the nature of the problem to be solved. The available methods largely fall into three categories:

1. Supervised learning
2. Unsupervised learning
3. Reinforcement or semi-supervised learning

Supervised Methods: Supervised learning methods solve problems such as classification using a two-step process: a learning step and a predicting step. These methods rely on a training dataset that has labels that indicate what the correct output for each observation is. Once the algorithm is trained on the known labels to create a model (a process called the learning step), it can predict the output variable on new, unseen data. Often a test dataset is used to validate the model. The larger the training dataset, the more accurate the model is in predicting new data.

For example, if we have many schools and they are categorized into few groups based on some indicators, the training dataset can have as a label variable the column that indicates which group a particular school belongs to. Once the model is trained, it can predict the group of a new school that is not yet categorized.

This method is highly efficient in classifying large datasets rapidly using computational models with multiple variables at the same time. Some examples of supervised algorithms are the support vector machine (SVM), logistic regression, discriminant analysis, naïve Bayes, and neural networks. However, there are situations where the labels for the data are not available, so using supervised models is not a viable option, and unsupervised may be the optimal choice.

Unsupervised learning methods: Unsupervised learning methods rely on learning by observation instead of using labeled examples (Han et al., 2011). The main attempt in unsupervised learning is identifying what is similar and what is dissimilar. The unknown patterns or groups in data are identified by comparing one observation to another and measuring their similarity. Based on the nature of data, the similarity or dissimilarity measures are selected. The most common unsupervised learning method is clustering.

Since the dataset for the research problem is large and does not have a defined set of groups, the unsupervised approach has been chosen for the first step of analysis. Section 4.1 details the clustering method and elaborates on the clustering methods applied for analyzing the data.

4.1 Clustering

Clustering is the process of partitioning a set of data observations into related subsets or groups called clusters (Han et al., 2011). The result of clustering often leads to discovering groups that were previously hidden. Each clustering algorithm can give different clustering outcomes on the same data based on the similarity and dissimilarity measures used for analysis. Thus, it is essential to evaluate which is the most optimal clustering algorithm that can be used for the given nature of data. While some algorithms work best for numeric data, some other

models are required for a mixture of data types, or specifically for graphs, images, and network types of data. Fundamentally of clustering methods can be categorized into the following (see Figure 5):

1. Partitioning methods
2. Hierarchical methods
3. Density-based methods

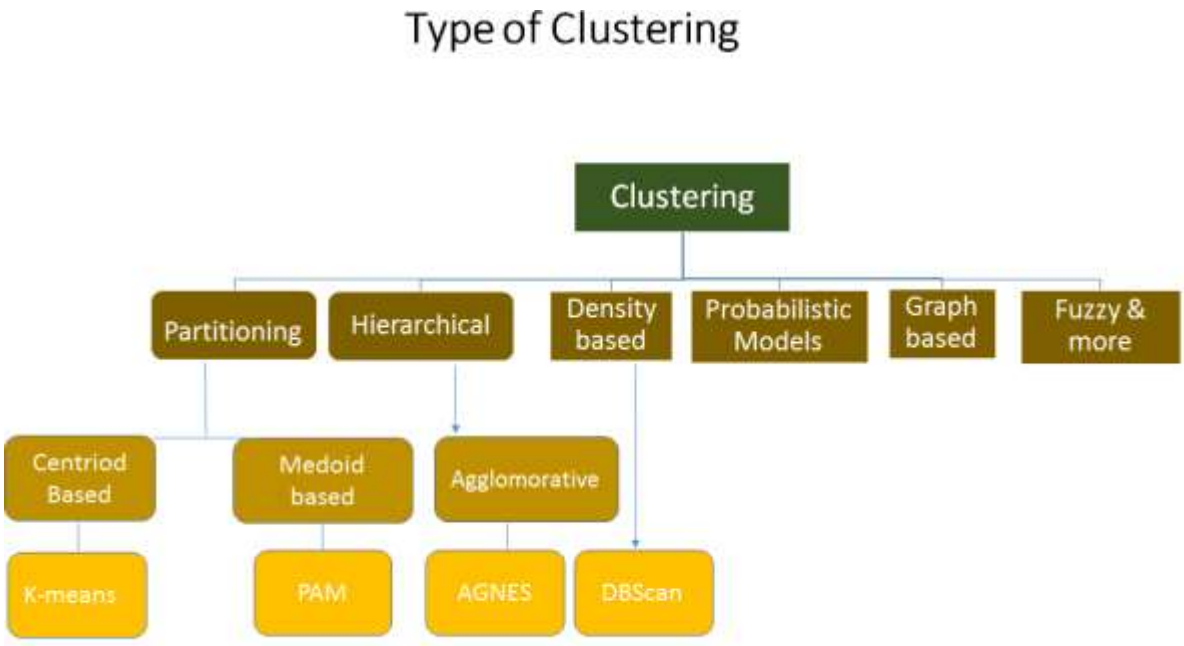


Figure 5: The yellow boxes at the end indicate the clustering algorithms that were used in the following research.

Partitioning methods: Many of the partitioning methods are distance-based methods. Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$. The algorithms use an iterative approach, where each observation is rechecked in every iteration to see whether it fits better in a certain cluster as more and more of the data set is partitioned into clusters. Since finding a global optimum with a large dataset is computationally expensive, most algorithms, such as k-means and k-medoids, rely on a

greedy approach. A greedy approach tries to find a local optimum by using the least number of iterations. A majority of partitioning methods separate the data into exclusive clusters, and each observation can belong to only one cluster. There are fuzzy clustering techniques that give a degree of belongingness of each observation to each of multiple clusters.

An image of a partition-based clustering is shown in Figure 6. *k*-means is a very popular centroid-based algorithm, and partitioning around medoids (PAM) is a medoid-based algorithm that overcomes the drawbacks of mean-based algorithms.

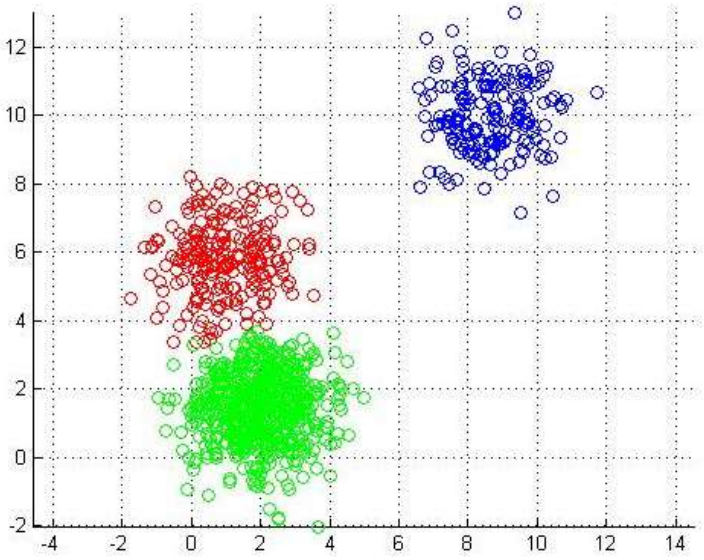


Figure 6: k-*Means clusters. This method largely identifies clusters of spherical shape.*

Hierarchical methods: Hierarchical methods group the data observations in a hierarchical order either from the bottom up (agglomerative) or from the top down (divisive). An *agglomerative* approach starts with each observation being in its own cluster. As the method iterates, it merges the observations in groups till it reaches the top, where all are linked under one cluster. Validation tests indicate how many number clusters are optimal choice. A *divisive* method takes the opposite approach. It starts with all observations under one cluster and

iteratively divides the observations into multiple clusters till all the observations are in coherent, related clusters. The two types of approach are compared in Figure 7. In either case, the result is a tree structure that represents lower levels of observations as subtrees, shown in Figure 8.

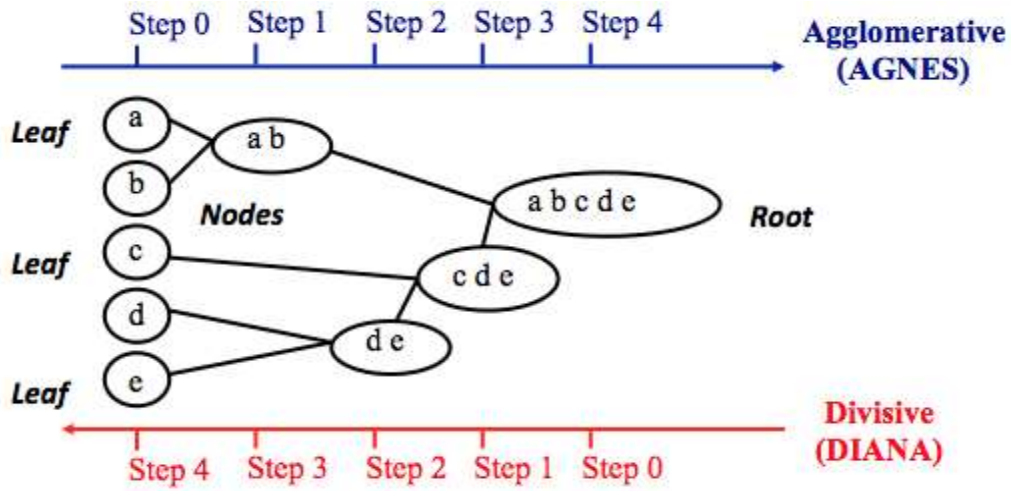


Figure 7 Agglomerative and divisive approaches in hierarchical methods. (Reproduced from Han et al., 2011.)

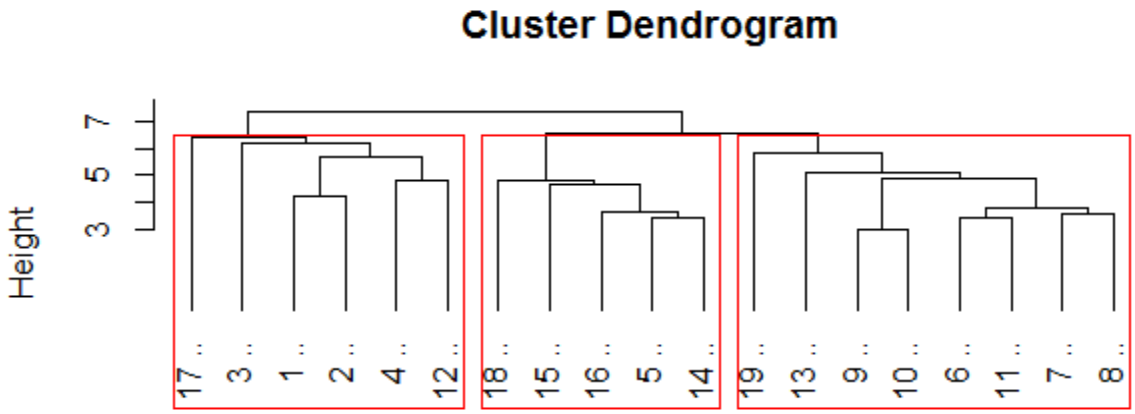


Figure 8: A tree structure is normally used to present hierarchical clustering visually from the top down.

Density-based methods: Density-based clusters are unique in terms of identifying clusters with arbitrary shape. While partitioning and hierarchical methods identify largely oval, spherical kind of clusters, density-based methods can find many shapes of clusters, for example, one with an S shape. By modeling dense versus sparse spaces in data, the algorithms can identify various shapes of clusters.

Clustering analysis has been in increasing use in many domains such as customer profile analysis, product reviews, cancer research, web search, business intelligence, biology, and, increasingly, education.

The present research makes use of an algorithm from each of the three clustering approaches discussed above and compares them to see which method suits the data best at a macro level and at micro level. Descriptions of the algorithms follow.

4.1.1 The *k*-Means Algorithm

k-means is a centroid-based clustering algorithm. It is the most commonly used and simplest of the algorithms that can give a quick overview of data clustering and has thus almost become a benchmark clustering algorithm. Its strength lies in its ability to handle a large dataset and identify spherical datasets easily in fewer iterations than other algorithms. How *k*-means works is detailed below.

Given a set of data points, *D*, initially the number of clusters desired, *k*, has to be specified.

Algorithm: *k-Means*

Input:

D, a dataset
k, number of clusters desired
Output: k clusters (data partitioned into k clusters), $C_1, \ldots, C_k$.

Method:

1. Randomly choose *k* data objects from *D* as the initial cluster centers
2. Iterate
3. With each data point,
    a. Compare to the initial *k* points
    b. Assign to the cluster that is most similar based on the mean value of the objects in the cluster
4. Update the cluster means by calculating the mean value of the objects for each cluster
5. Iterate until no more change in cluster assignments happens

In this algorithm, initially a random pick of *k* points, the specified number of clusters, is chosen as the centroids. For example, if 100 observations of data exist and an input of three clusters is given, initially three data objects are randomly drawn from the 100 data objects and made the initial centroids of three clusters. Next, each data object from the remaining 97 is picked and compared to the three centroids and the distance measured between the data object and the centroid. The data object is assigned to whichever centroid is closest or most similar to it. With each data point added to the centroid group, the mean of the group is recalculated and the centroid readjusted. This process is repeated until all data observations are assigned to the clusters and no further changes happen to data assignment.

The centroid $\mathbf{c}_i$ of the cluster C_i is the center point of the cluster, and it represents the cluster. Data points assigned to the cluster are measured for similarity using the Euclidean distance. For a given data object $p \in$ and cluster C_i, the distance between it and the cluster's centroid $\mathbf{c}_i$ is measured by dist(p, $\mathbf{c}_i$) with the Euclidean distance formula for n-dimensional space is

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}$$

The resulting cluster quality is also checked for intracluster variance; that is, how close or far apart the point is to the center of the cluster.

There are several pros and cons for the algorithm. It can handle large datasets and can compute clusters relatively quickly compared to some other algorithms. It is efficient at identifying spherical-shaped clusters, and the runtime complexity is $O(nkt)$, where n is the total number of data observations or objects, k is the number of clusters, and t is the number of iterations. Generally, $k < n$ and $t < n$. This makes the method highly scalable in processing extremely large datasets.

However, the *k*-means algorithm can handle only data types where the mean can be computed. Datasets having multiple data types have to be excluded. Any data set with hidden, arbitrarily shaped clusters does not scale well with the *k*-means algorithm. Also, *k*-means clustering results can be highly impacted by outliers.

k-means was tested with the research data at hand, as it is a benchmarking algorithm for clustering and gives a quick overview of what is missing in analysis. It worked more as a steppingstone to build up to the next level of analysis with more complex algorithms that can handle categorical data.

4.1.2 The PAM, CLARA, and AGNES Algorithms

Partitioning Around Medoids (PAM) is an algorithm that, instead of utilizing the mean of the data, it takes advantage of being median based. Like *k*-means, it initially chooses a specified number of *k* clusters arbitrarily from data objects—in this case, as medoids—and follows an iterative approach. Where PAM starts to differ from *k*-means is that, when each object is assigned to an initial medoid, in the subsequent iteration each data object that is added to each cluster is replaced as a medoid and checked whether it forms better clustering by calculating the distance from the new object *p* to every other object. For example, if an object *p* is assigned to

cluster C_1 with medoid *m*, in the next iteration, when another object is added p is replaced as the medoid, and assessed whether it produces a better cluster quality by comparing its distance to every other point assigned to the clusters. This replacing and reassessing of cluster quality introduces a cost function, which calculates the difference in absolute error value if a current medoid is replaced by a new object. If the outcome of this cost evaluation is negative, then the new data object takes the medoid's place, if it is positive, the swapping does not takes place, but rather the data object is merely added as a cluster member. The main goal of the algorithm is to reduce the average dissimilarity of data objects to their closest selected medoid center.

The following is the algorithm for PAM:

Algorithm: *PAM*

Input:

k: the number of clusters
D: a data set containing *n* objects
Output: k clusters $C_1, \ldots, C_k$

Method:

1. Choose *k* data objects arbitrarily from *D* as initial medoids
2. Iterate
    a. Assign each remaining object to the cluster with the nearest medoid
    b. Randomly select another nonrepresentative data object O_r
    c. Compute the total cost *S* of swapping representative object medoid, to the new random picked data object
    d. If $S < 0$, then swap the medoid with the new object O_r to form the new set of *k* medoids
3. Repeat until no more change in clusters happens

The objective of the method is

$$F(x) = \text{minimize} \sum_{\text{i}=1}^{\text{n}} \sum_{j=1}^{n} d(i, j) z_{ij}$$

(Kaufman and Rosseeuw 1987),

Since the data for research contains several types of attributes, PAM must be used with a distance measure that can compute distances using nominal and categorical attributes as well as numeric ones. The *Gower General Similarity Coefficient* (Gower, 1971) is a distance measure that has been applied to objects that have both numeric and categorical attributes. It is the most popular measure for mixed-data-type datasets and clustering. The Gower distance ranges from 0 to 1.

Although the PAM algorithm is very efficient for small to midsize datasets, the costs involved in computation significantly increases with large datasets that can have millions of observations. To handle such scales, the CLARA (Clustering LARge Applications) algorithm has been developed (Kaufman & Rousseeuw, 1990). CLARA takes a random sample from the large dataset, computes the clusters for the sample using PAM, and then assigns each object in the rest of the dataset to the nearest cluster. After taking multiple random samples, it returns the best clustering outcome for the larger dataset.

AGNES (AGglomerative NESting; Kaufman & Rousseeuw, 1990) is an agglomerative hierarchical clustering method. This method initially places each object into its own cluster. The clusters are then merged one step at a time based on some similarity measure. For example, suppose C_1, C_2, C_3, C_4, and C_5 are objects. Initially they are all clusters of one. In the next iteration, suppose C_1 and C_3 have the smallest distance between them. Then they can be combined into one cluster. This process repeats till all the objects are brought under one cluster. This approach is a single-linkage approach, where each cluster is represented by all the objects in that cluster. Validation methods indicate the number of clusters that represent the best outcome for the data. The measure can be Euclidean or some other distance measure. In the context of this

research we have applied Ward's method, which minimizes the sum of squared Euclidean distances (Ward, 1963). This results in minimizing of total within-cluster variance.

AGNES clustering is used in this research for text mining analysis on the comments part of student data that the teacher inputs for each student for each trimester. Here the comments are clustered based on similarity. Each group of cluster is given a numeric nominal value for the dataset, to be processed further by a recommendation system.

4.1.3 The DBSCAN Algorithm

Most partitioning approaches identify clusters that are more or less spherical in shape. Density-based methods such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) are geared toward finding arbitrarily shaped clusters. DBSCAN identifies dense regions followed by sparse regions. These dense regions are defined as clusters, and they can be in any shape (Ester, Kriegel, Sander, & Xu, 1996).

DBSCAN quantifies the density of regions using two parameters: the radius, Eps (ε), of a core object, and a user-specified metric, MinPts. A data object is considered a core or center object if there are at least MinPts of data objects in the radius ϵ ε neighborhood of the core object. A dense region is defined by two notions of connectivity between the data points around a core object.

Directly density reachable: A data object **o** is said to be directly density reachable to a core object **c** if **o** is within the ε radius of **c**. With this criterion, a core object creates a dense region based on its ε neighborhood.

Density reachable: A data object **o** is density reachable from **c** if there is a chain of objects $\mathbf{o}_1, \ldots, \mathbf{o}_n$ such that with $\mathbf{o}_1 = \mathbf{c}$ and $\mathbf{o}_n = \mathbf{o}$, where each $\mathbf{o}_{i+1}$ is directly density reachable

from $\mathbf{o}_i$ with respect to ϵ ε and MinPts (all the objects on the path must be core objects, with the possible exception of **c**). Figure 9 illustrates this concept. Density regions are defined by core objects with the data points at least equal to MinPts in the neighborhood of radius. Multiple core objects are connected to become dense regions.

Density connected: To connect all the core objects as well as the neighboring data points in the dense regions, DBSCAN uses the notion of density connectedness (Han et al., 2011). Two objects $\mathbf{o}_1$, $\mathbf{o}_2$ $\in$ dataset *D* are density connected with respect to ε and MinPts if there is an object **c** $\in$ *D* such that both $\mathbf{o}_1$ and $\mathbf{o}_2$ are density reachable from**c** with respect to ε and MinPts. If $\mathbf{o}_1$ and $\mathbf{o}_2$ are density connected, and $\mathbf{o}_2$ and $\mathbf{o}_3$ are density connected, then so are $\mathbf{o}_1$ and $\mathbf{o}_3$.

Following is the algorithm for DBSCAN:

Algorithm: DBSCAN

Input:

Dataset D, radiusϵ eps, MinPts
Output: Dense Regions as clusters

Method:

1. *C* = 0
2. For each point *P* in dataset *D*
    a. If *P* is visited continue to next point,
    b. Mark *P* as visited,
    c. NeighborPts = regionQuery(*P*, eps)
    d. If sizeof(NeighborPts) < MinPts
            Mark *P* as NOISE
        Else
            C = next cluster
          expandCluster using *P*, NeighborPts, *C*, eps, MinPts

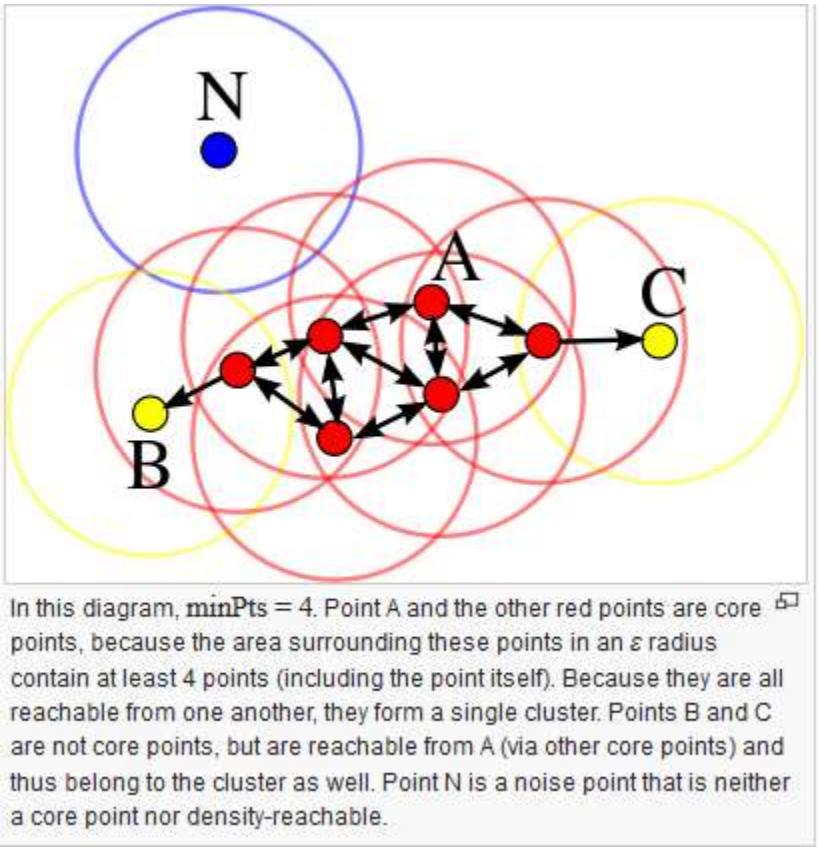


Figure 9: Dense region formation in DBSCAN. (Source: Wikimedia Commons, author Chire, retrieved from https://en.wikipedia.org/wiki/File:DBSCAN-Illustration.svg)

The DBSCAN algorithm used with the correct dataset and knowledge has many advantages. The number of clusters does not need to be specified. If a domain expert can understand the nature of data and specify the MinPts and the radius effectively, the outcome of cluster quality is high. Any shape clusters can be detected. This is especially useful in data that is continuous and changes frequently. Noisy data and outliers are effectively handled by this algorithm.

However, there are drawbacks to this algorithm as well. The choice of MinPts and ε have to be well made; they have a great impact on the quality of clusters. The data has to be largely numeric for this algorithm. The output is also heavily dependent on the distance function chosen.

If there is a high variation in density of the data, choosing the correct input parameters will be different.

4.2 Classification

Although clustering algorithms solve the problem of identifying the categories that exist in a set of data with several unknowns, classification predicts qualitative response for a given data set with known qualitative or categorical groups. Predicting a qualitative response for a given data set is called classifying, since it involves assigning an observation to a category or class. Most classification models develop a probability prediction for each class for a certain observation and then assign it to the class that has the highest probability.

Classification problems are very common in several domains in the real world; for example:

- Categorizing students based on math performance
- Categorizing patients on rate of recovery so they can identify associated traits for each group
- In banking, deciding whether a transaction is fraudulent or not

Classification methods can be applied to multiple data formats, such as numeric data observations, mixed data types, text data, images, audio, and video. Thus, classification can be an extremely useful method in handling multiple scenarios. In the research undertaken here, classification is applied in putting schools into the categories that the clustering method identified. If a school's data is updated, a new school is added, or schools complete data that is currently missing, they need to be assigned a class. Instead of performing clustering again, a

classification model, trained with the existing complete data of schools used for forming clusters, can be used to assign a class to these new data observations automatically.

Classification is divided into two types:

1. **Binary classification:** Binary classification consists of categorizing the observations into two classes. Methods try to predict the best probability of an observation belonging to one of the two classes. For a binary model, responses are typically coded as a 0 and 1.
2. **Multiclass classification**, where there are more than two classes and the observation can belong to any one of them. Certain methods can provide a clear category that an observation belongs to, and other approaches give an output that contains the degree of probability that an observation belonging to each category.

Two classification methods, random forest and linear discriminant analysis, are next compared to see which approach is suitable for the nature of data for the research. Both methods have very effective approaches for multiclass classification. Random forest is based on a tree ensemble classification model, while linear discriminant analysis is a probability-based approach that predicts the class based on a given set of predictors.

4.2.1 Random Forests

A random forest (RF) is an ensemble of a large collection of decision trees (Breiman, 2001). A decision tree is a method where, starting at the top of a tree model, an attribute splits on a decision condition. Figure 10 illustrates the decision splits. Initially the decision is made on a single predictor, which is chosen based on a strength of predictors. Then subsequent splits down the nodes of the tree are performed on other attributes.

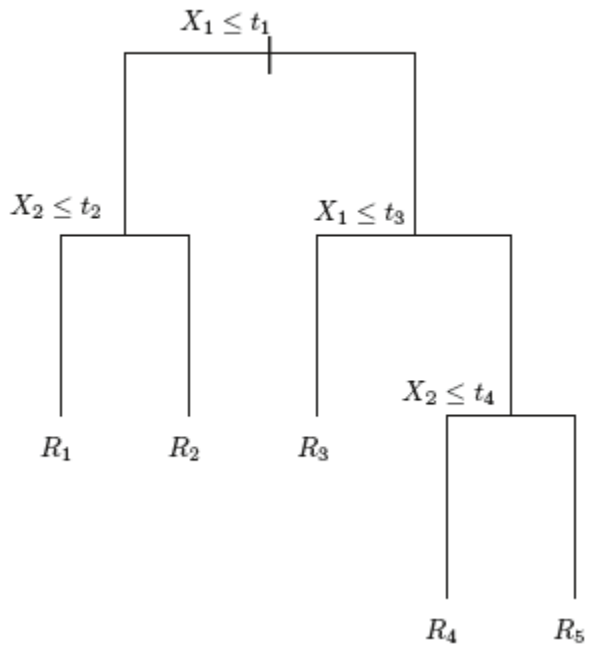


Figure 10: Decision tree that splits on four attributes.

In the Random Forests algorithm (James, Witten, Hastie, & Tibshirani, 2013, pp. 320–321), a large number of decision trees are constructed with splits made on different attributes. Each of these individual decision trees is constructed using a randomly selected sample of rows (with replacement) of the same size as the original dataset and a sampling of predictors on each tree. This results in decorrelating the trees and overcoming the problem of a single strong predictor dominating the splits. By choosing a random sampling set of *m* predictors from the set *K* of all available predictors, a random tree ensures that all the trees are not dominated by one or two strong predictors in the decision splits. At each split a fresh sample of *m* predictors is taken. On average, $m \approx \sqrt{p}$; that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors. Thus $(p - m)/p$ predictors will not even consider the strongest predictor and will create more chances for other predictors to be

considered. Since Random Forests averages many trees, the problem of averaging many correlated trees will be avoided, because only uncorrelated trees can reduce the variance. Section 5.4.1 presents the Random Forests analysis for the classification problem discussed for the present research.

4.3 Text Mining

The main goal of using text mining in this research context is to merge it with numeric data for analysis. Such domains as health care have started to experiment with integrating patient record data with doctor comments (Foster, Liberman, & Stine, 2013). Although numeric attributes give measures of several indicators of student performance and profile, text descriptions provided by the teacher on each student carry additional information that can be valuable from a qualitative perspective. Beyond performance scores and ethnic/gender similarities, text descriptions of students can be used to derive similarities in student personalities and abilities. For example, all high-scoring students may not be similar, certain learning methods may have enabled certain sets of students to improve rather than other sets of students. Mere scores will not indicate such detailed information. By accessing the comments and student inputs from teachers, a more detailed profile can be obtained and be compared to other students for grouping.

This section briefly describes the nature of text, followed by the algorithms used in processing the text data for grouping the teacher descriptions of students into similar or dissimilar groups.

4.3.1 Structured versus Unstructured Data

Data that is stored in databases such as a relational database or Access, with a certain pattern and order with querying capabilities, is called structured data. Data that includes text and images, on the other hand, is unstructured data, which does not have a defined form of rows and columns for storage. Textual data sources are very diversified and can exist in free text form or semi-formatted form such as HTML or XML. They need to be transformed for analysis. Text mining is the process of transforming and analyzing unstructured data that is in text form.

Components in text: The basic unit of analysis in text mining is a document. It is a sequence of words connected and controlled by grammatical rules. A document consists of words as the basic entities, then phrases made out of words, followed by sentences. A group of sentences makes a paragraph, and it forms the fundamental unit of a series of related ideas, actions, or meanings (Strunk, 2007). A document can contain just one sentence, groups of paragraphs, or just a phrase; it can be by a single author or multiple authors. A document in the context of text mining can be a user review, an email, a research article, a letter, or many other types. Figure 11 shows the hierarchy of a document. The document in turn belongs to a set of documents for analysis, called a corpus. A lexicon, which is a set of all unique words in a corpus, is created out of a corpus.

Figure 11: Hierarchy of a document corpus.

Each text mining algorithm performs best with a particular set of input formats of text. Text input can be cleaned into a *unigram* model, where each word forms a unit; a *bigram* model, where two words form a unit; or an *n-gram* model, in which three or more words form a unit. There are other models that are based on a part-of-speech (POS, or POST for part-of-speech tagging) approach. The main focus of text analysis in this research is to identify document similarity, so that each student record that has text input can be measured for how similar it is to another student's record. Once the similarity and dissimilarity is determined, students can be given categorical representation for similar groups. These categories are used as qualitative attributes along with numeric attributes.

To enable the similarity/dissimilarity measurements, a vector-space model (Salton, 1989) is used. In this model, each document *d* is converted into a vector in the term-space model. The term frequencies are weighted by the *tf-idf* (term frequency/inverse document frequency) term weighting model (described in Section 3.3), in which each document can be represented as

$$(tf_1 \log(n/df_1), tf_2 \log(n/df_2), \ldots, tf_i \log(n/df_i))$$

where tf_i is the frequency of the ith term in the document and df_i is the number of documents that contain the ith term. Since the documents are of different lengths, the length of each document vector is normalized so that it is of unit length ($||dtfidf|| = 1$). In other words, each document is a vector on the unit hypersphere. Given a set A of documents and their corresponding vector representations, the ***composite*** vector D_A is defined to be $D_A = \sum_{d \in A} d$, and the ***centroid*** vector C_A to be $C_A = D_A/|A|$.

Cosine Similarity: To compute the similarity between two documents d_i and d_j in the vector-space model, the cosine similarity is the most commonly used measure, which is defined to be $\cos(d_i, d_j) = d_i^t \, dj \, /(||d_i|| \, ||d_j||)$. The cosine formula can be simplified to $\cos(d_i, d_j) = d_i \, td_j$, when the document vectors are of unit length. This measure equals 1 if the documents are identical and 0 if they are very dissimilar—in other words, if the vectors are orthogonal to each other (Zhao, Karypis, & Fayyad, 2005).

4.3.2 Hierarchical Clustering of Documents

Three different algorithms are mainly tested, apart from *k*-means, as a benchmarking algorithm to find document similarity. The first is hierarchical clustering of documents.

As discussed in Section 4.1, hierarchical clustering of documents can be accomplished in two ways: by partitioning methods and by agglomerative methods. Partitioning algorithms take a top-down approach, where all the documents begin under a single cluster that is repeatedly bisected till each document is in its own cluster, and methods are applied to identify the right groupings of clusters. Agglomerative approaches, on the other hand, build the hierarchical

solution by initially assigning each document to its own cluster and then repeatedly selecting and merging pairs of clusters, from the bottom up, to obtain a single large cluster that encompasses all documents at the top. Thus, agglomerative algorithms build the tree from the bottom (i.e., its leaves) toward the top (i.e., root). For the document clustering used in the present research, the agglomerative approach proves better in terms of document similarity, based on earlier research (Duda, Hart, & Stork, 2001).

Agglomerative clustering can be accomplished using different linkage methods: single-linkage, complete-linkage, average (UPGMA), and centroid (WPGMC).

The single-link scheme measures the similarity of two clusters in terms of maximum similarity between the documents from each cluster (Sneath & Sokal, 1973). The similarity between two clusters S_r and S_t is given by

$$\text{sim}_{\text{single–link}}(S_r, S_t) = \max_{d_i \in S_r, d_j \in S_t} \{\cos(d_i, d_j)\}$$

On the other hand, complete-linkage computes the minimum similarity between a pair of documents to measure their similarity (King, 1967). The similarity is expressed as

$$\text{sim}_{\text{complete–link}}(S_r, S_t) = \min_{d_i \in S_r, d_j \in S_t} \{\cos(d_i, d_j)\}$$

However, both the single- and the complete-link methods often do not give the expected results because they compute similarity on a limited amount of information (single-link) or assume that all the documents in the cluster are very similar to each other (complete-link approach). The UPGMA scheme (Unweighted Pair Group Method using arithmetic Averages; Jain & Dubes, 1988, p. 80), also known as group average, overcomes these problems by measuring the similarity of two clusters as the average of the pairwise similarity of the documents from each cluster:

$$\text{sim}_{\text{UPGMA}}(S_r, S_t) = \frac{1}{n_i n_j} \sum_{d_i \in S_r, d_j \in S_t} \cos(d_i, d_j) = \frac{D_i^t D_j}{n_i n_j}$$

Section 5.6 elaborates on the results from the complete linkage versus the average linkage method.

4.3.3 Topic Modeling

Topic modeling in text mining is a statistical and probabilistic model that relates to identifying the major areas of discussion in a given document or set of documents. In this research context, topic modeling using the latent Dirichlet allocation (LDA) method is applied to identify the major topics of discussion in the reviews submitted online by students, parents, and community stakeholders for each school. The premise is that by recognizing the topics discussed, the text mining process will be more likely to uncover insights on the true nature of concerns expressed by stakeholders.

LDA is a generative model that identifies topics in a set of documents. Instead of clustering the documents, it tries to identify the topics in the document. It assumes that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. It was first presented as a graphical model by Blei, Ng, and Jordan (2003). In LDA the assumption that each document is a mixture of topics come with a Dirichlet prior. (The Dirichlet distribution is a family of continuous multivariate probability distributions parameterized by a vector α $\boldsymbol{\alpha}$ of positive reals and is used as a prior distribution in LDA, which is a Bayesian model.)

The LDA process happens in three steps. First, the number of words used in a document is determined by sampling with Poisson distribution. In the second stage, a distribution over

topics for a document is elicited from the Dirichlet distribution. In the third step, based on the document specification distribution, topics are generated, and then the words for each topic are classified. LDA identifies the topics and indicates the top topics. It pools all the words related to each other closely into one topic. For example, "bark," "bone," and "leash" will be pooled into one topic, which can be identified as "dog." This is done by creating word probabilities, and it does not suffer from the problem of polysemy (in which one word has multiple meanings) and synonymy (in which different words have the same meaning). It also considers words that have a combination of an adjective and a noun, such as "good student." This ability allows the method to identify the topics much more effectively when presented with a group of words. LDA has been successfully applied in many domains of research and applications, such as topic detection, emotion detection, and word sense disambiguation (Blei et al., 2003).

LDA does a fairly optimal job of identifying the topics, and an experienced person knowledgeable in the domain can accurately identify the topic.

4.3.4 Locality-Sensitive Hashing

The locality-sensitive hashing (LSH) algorithm is a method for performing near-neighbor search and identifying similar items in high-dimensional spaces. Text documents are usually high-dimensional in nature once they are converted to a term document matrix. Each unique word becomes a dimension, thus making each document a high-dimensional vector. Since comparing the text input given by teachers between students and identifying similar student profiles is the main goal, LSH can be an effective method in identifying pairs of student comments that are similar to each other. LSH also has the advantage of handling very large numbers of documents.

LSH analysis can be accomplished in the following steps.

1. *Shingling:* Shingling involves creating a set of shingles out of a document. A shingle can be words or characters in the document that are considered as tokens in this context. Technically, a *k*-shingle or a *k*-gram sequence of tokens is created out of a document. For example, suppose that document *d* = {abcadb} and the tokens are the characters here. This document yields a set of 2-shingles (*k* = 2): $S(D_1)$ = {ab, bc, ca, ad, db}. The number of characters in each shingle, *k* (here it is 2), can be determined based on the length of the documents. Thus document D_1 becomes a set of *k*-shingles, and each unique shingle is a dimension, thereby creating a very sparse vector space. The main assumption is that documents that are similar have many common shingles. This leads to the issue of picking the right size of *k* (the number of tokens in a shingle). If *k* is too small, then most documents have all the shingles; if too large, then there is no similarity. Thus, it is essential to explore the data and experiment to obtain the correct shingle size (Rajaraman, Leskovic, & Ullman, 2014).

Typically the distance measure that is used for LSH is Jaccard similarity (Rajaraman et al., 2014),

$$\text{sim}(D_1, D_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}$$

though other measures such as Hamming distance can be applied too based on the nature of the data at hand.

Given that shingles are created, if most pairs of similar documents are to be searched, it will take enormous amount of time. For example, if *N* = 1 million documents, computing the pairwise similarities would take about 5 days at 10^5 seconds/day and 10^6 comparisons/second, and if *N* = 10 million, it takes 1 year (Rajaraman et al., 2014). The next step, min-hash, is implemented to overcome this problem.

2. *Min-hash:* To perform min-hash, the document vectors have to be converted into bit vectors and encoded using bit vectors of (0/1), where rows = shingles and columns = documents, as shown in the following example (Figure 12), described by Rajaraman et al. (2014).

1. Each document is a column: Example: sim(C_1, C_2) = ? Size of intersection = 3; size of union = 6, Jaccard similarity (not distance) = 3/6
2. $d(C_1, C_2)$ = 1 – (Jaccard similarity) = 3/6

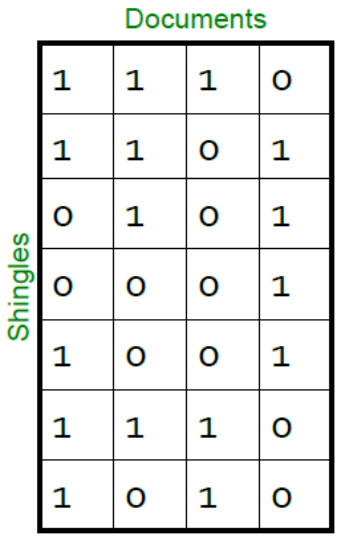


Figure 12: Bit vectors for min-hash.

The next step after creating the 0/1 matrix is finding the similarity of columns by computing the small signatures (small summaries) of each column (each column is a document). Similarity of columns = similarity of signatures. Since each column signature will be small, creating a hash function for it that can hold in RAM becomes easier. A hash function is a method that can assign small codes or keys to large values so they can be stored effectively and securely. This hashing function is called min-hash function that can handle Jaccard similarity.

Given C_1, column 1 and C_2, column 2 (columns are documents), the goal is to find a hash function $h(\cdot)$ such that:

If sim(C_1, C_2) is high, then with high probability $h(C_1) = h(C_2)$

If sim(C_1, C_2) is low, then with high probability $h(C_1) \neq h(C_2)$

4.4 Frequent Patterns and Association Rules

Frequent patterns are patterns that occur frequently in a data set and establish a recurring relationship between the elements present in a data set. Finding frequent patterns plays an important role in mining association rules, identifying correlations, and underlying relationships that otherwise may not be noticeable. Frequent-pattern mining was proposed by Agrawal, Imieliński, and Swami (1993) in the context of analyzing market data of customer transactions. For example, in a large database of customer transactions by using frequent pattern analysis one can find that if a person is buying bread and butter they also tend to buy milk. These items become frequent item sets. These frequent patterns can in turn be used to derive association rules; for example, "90% of transactions that purchase bread and butter also purchase milk." The antecedent of this rule consists of "bread and butter" and the consequent consists of "milk" alone. The number 90% is the confidence factor of the rule (Agrawal et al., 1993). This information can be used to decide how much milk to stock, based on the purchases of bread and butter. Similarly, one can check rules that have customers buying memory storage as antecedent, which will show other products that might be impacted by removing the memory storage from the shop shelves.

This method can be transformed and applied in education contexts. In this research we treat each school as a transaction and each attribute as an item with multiple levels. For example, in the data set, one of the patterns and association rules that emerges is

$$\{\text{FreeReducedMealsPerC=4,TeachersFTEC=2}\} \Rightarrow \{\text{CohortGraduatesPerC=4}\}$$

The frequent pattern–derived association rule indicates that in schools where FreeReducedMealsPer is above 75% and the number of Teachers Full time or Equivalent is above 50 in a school, Cohort Graduates percent is higher than 75%. Analysis using regression model cannot identify this relationship.

This section provides a brief technical overview of this method and discusses how it can be used for the data for this research and in education.

A set of attributes that co-occur under certain conditions creates a pattern, and these sets of attributes that occur together are referred to as itemsets in frequent pattern analysis. An itemset that contains *K* items is called a *K*-itemset. For example, {memory sticks, laptops} is a 2-itemset. Frequent patterns are analyzed based on the concepts of support, confidence, and lift.

In a database, let D be the set of data transactions T that are no null transactions and $T \subseteq I$ with a transaction TID identifying each transaction, where $I = \{I_1, I_2, I_3, I_4, \ldots, I_n\}$ be an itemset. If A is a set of items, a given transaction T contains A if $A \subseteq T$. A frequent pattern occurs if repeatedly A items occur together. An association rule is a connection that indicates $A \Rightarrow B$ where $A \Rightarrow I$ and $B \Rightarrow I$ and both are not null and $A \cap B$ is not null. The **support** s of $A \Rightarrow B$ indicates the percentage of transactions in the set D that contain $A \cup B$, and the **confidence** c is the percentage of transactions containing A that also contain B (Han et al., 2011).

This is represented in probability terms

$Support(A \Rightarrow B) = P(A \cup B)$

$$Confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)}$$

Algorithms Apriori, Eclat, and FP-Tree are the most popular algorithms often used to compute frequent patterns and association rules. For this research, the Apriori and Eclat algorithms were used for analysis, since the dataset is not exceedingly large.

Frequent patterns have been used in education data analysis for identifying student career choice patterns (Campagni, Merlini, & Sprugnoli, 2012). A sequential mining algorithm to analyze learning behaviors to discover frequent sequential patterns was proposed by Huang, Chen, and Cheng (2007) to suggest recommendations for students in selecting learning content.

However, frequent pattern and association rule mining at the macro level for policy decision making is lacking. The research presented here presents the results in Section 5.5 and highlights the importance to policy planning.

4.5 Recommender Systems: Collaborative Filtering

Information about student learning and measures to improve their learning on a daily basis is available to the teachers within their school system. Although this information does give an opportunity to discuss with other teachers on what type of methods work for improving learning on different types of students, enlarging that access to a large amount of data with collective intelligence from many teachers and students with various backgrounds will provide an enhanced tool box to identify remedial measures or teaching methods that are effective for improving the learning experiences for students with multiple backgrounds.

Although each student is an individual and differs from others, at the same time, there are similarities within groups of individuals. Those similarities can be identified to an extent, given a large amount of data. Collaborative filtering systems and recommender systems in general aim to achieve that.

Recommender systems are widely used in product sales online. When a customer (say, John) buys a certain book, the system identifies other customers with profiles similar to John's and recommends items that other customers have purchased and liked. Applying the same model

to students in a school system, a teacher who wants to improve learning experience for a student in her class can easily access a system where she can identify similar students from across larger populations and obtain more details on what teaching methods or recommendations the other teachers have used or are using to bring improvements in the child and how successful they have been. In a sense, the teacher has access to the collective knowledge and intelligence of teachers in many different locations, rather than being restricted to a localized source of knowledge such as other teachers at the same school. This is not to dismiss the importance of localized knowledge; rather the emphasis is on opening up access to more channels of input and information that was not formerly possible.

Education data lends itself very easily to recommendations models because so much of the data is quantified and ranked. However, to connect behavioral data of students that exists in the form of text remains a challenge. A recommender system approach will be an attempt to solve that problem. An advantage with individual data in education is that the information is not too sparse.

CHAPTER 5

DATA ANALYSIS AND RESULTS

The final goal of data analysis is to make optimal decisions. This is feasible only when one is able to interpret the results correctly and is comfortable in understanding them. This section attempts to achieve that goal by presenting the analysis in depth.

Analysis was performed using the steps shown in Figure 1 (repeated here as Figure 13).

5.1 Analysis Environment: R and RStudio

The thesis analysis was performed using the R language and in the RStudio development environment (Rstudio, 2016). R is a mature language for data analysis that is most popular in statistics and integrates well into computer science notions of software application programming. It has been integrated into Microsoft Azure and can be run inside Hadoop and Hive. It has been tested under various domains, and proved to be most suitable for developing a vertical demonstration of a system that can establish a link between macro and micro data analysis. R also makes it easy to integrate the current code into a larger system if developed in the future.

5.2 Data Exploration

The initial data set is a large data set containing all the traditional high schools. After data cleaning and preprocessing a total of 1,636 high schools had data eligible for analysis. To establish the fact that macro data and micro data can be linked together for policy planning, an example of math performance for high school students is analyzed. The highest macro level started is at school level, since this gives an optimal level to have enough difference and

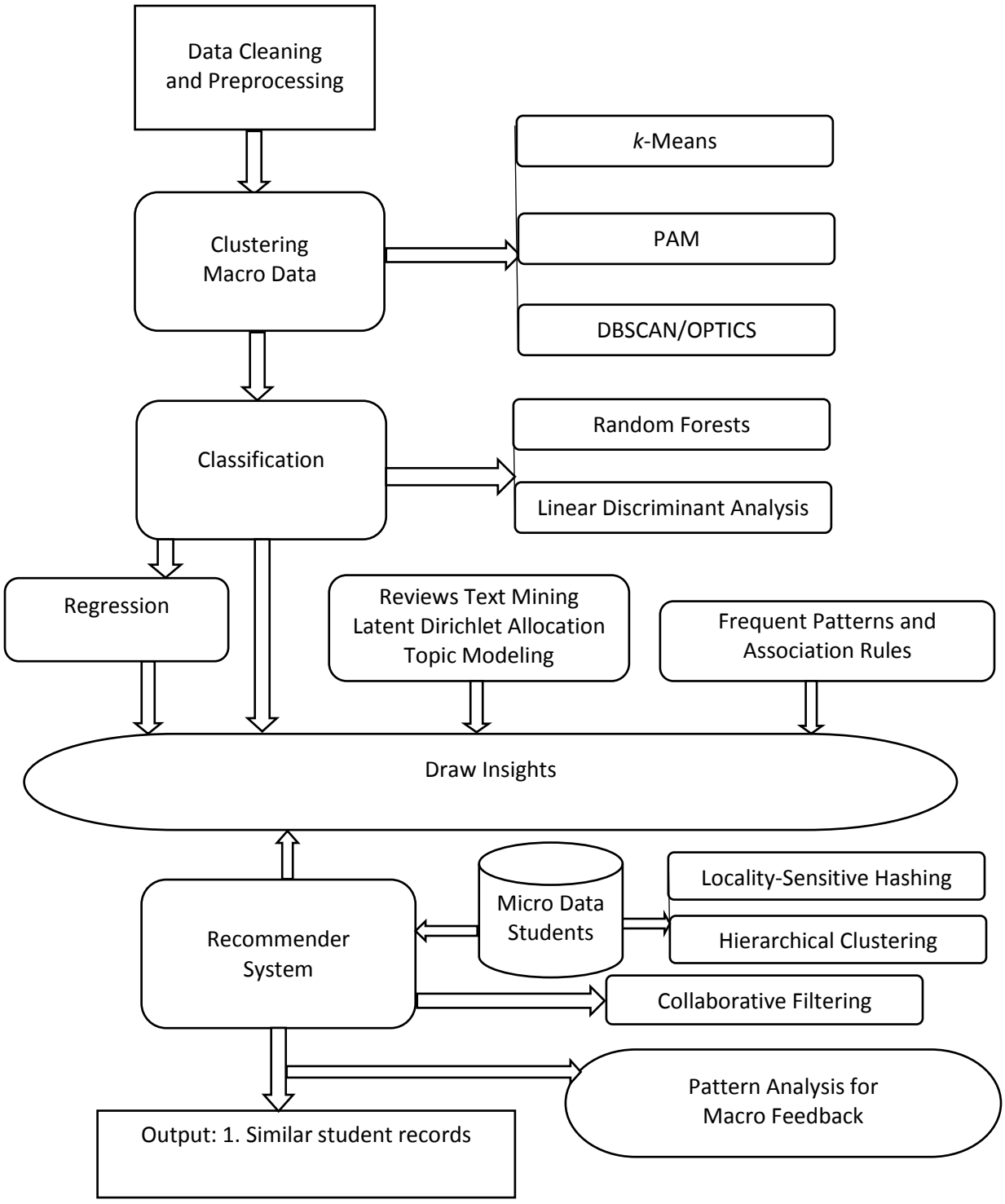


Figure 13: Steps in the analysis.

commonalities. Starting at district level would seem highly generic and may hide more detailed information in the aggregation process.

Data exploration and variable reduction: Since there are numerous variables with many measurements, it is extremely important to identify the multicollinearities that can impact the outcome of analysis. To an extent, clustering algorithms can handle the multicollinearities; however, the validation methods used for the analysis are not completely immune to them. Thus it is essential to remove such attributes that can skew a model. For example, the graph in Figure 14 indicates the primary multicollinearities that exist among some main variables in the data. The variable e_enr—schoolwide English Language Arts (ELA) enrollment—is highly collinear with m_enr and m_tst. If all the variables that are highly correlated are included, it is likely that the correlated variable might mask the impact of other variables that can give some insights (James et al., 2013). Thus, when the focus is on math performance, and English language enrollment predicts enrollment in math, it is essential to choose math enrollment rate as an attribute to be included in analysis.

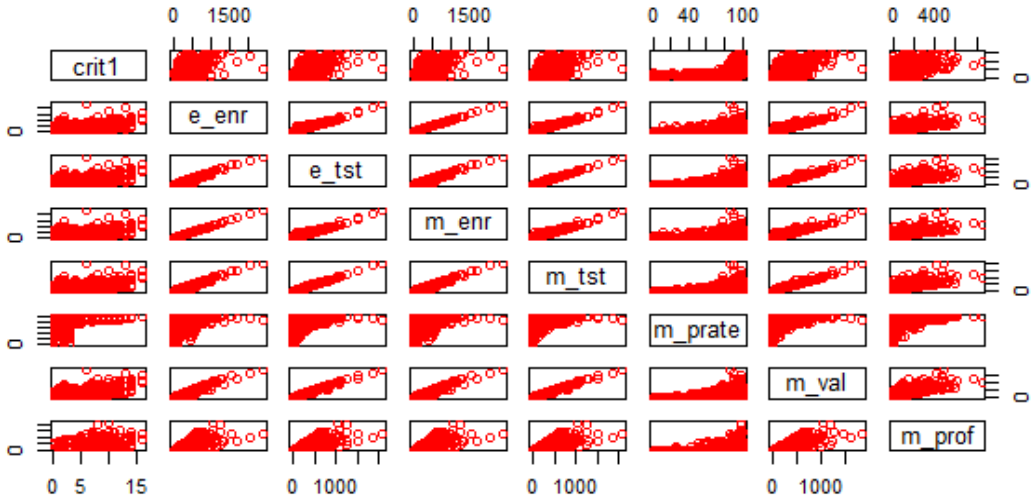


Figure 14 Identifying multicollinearities.

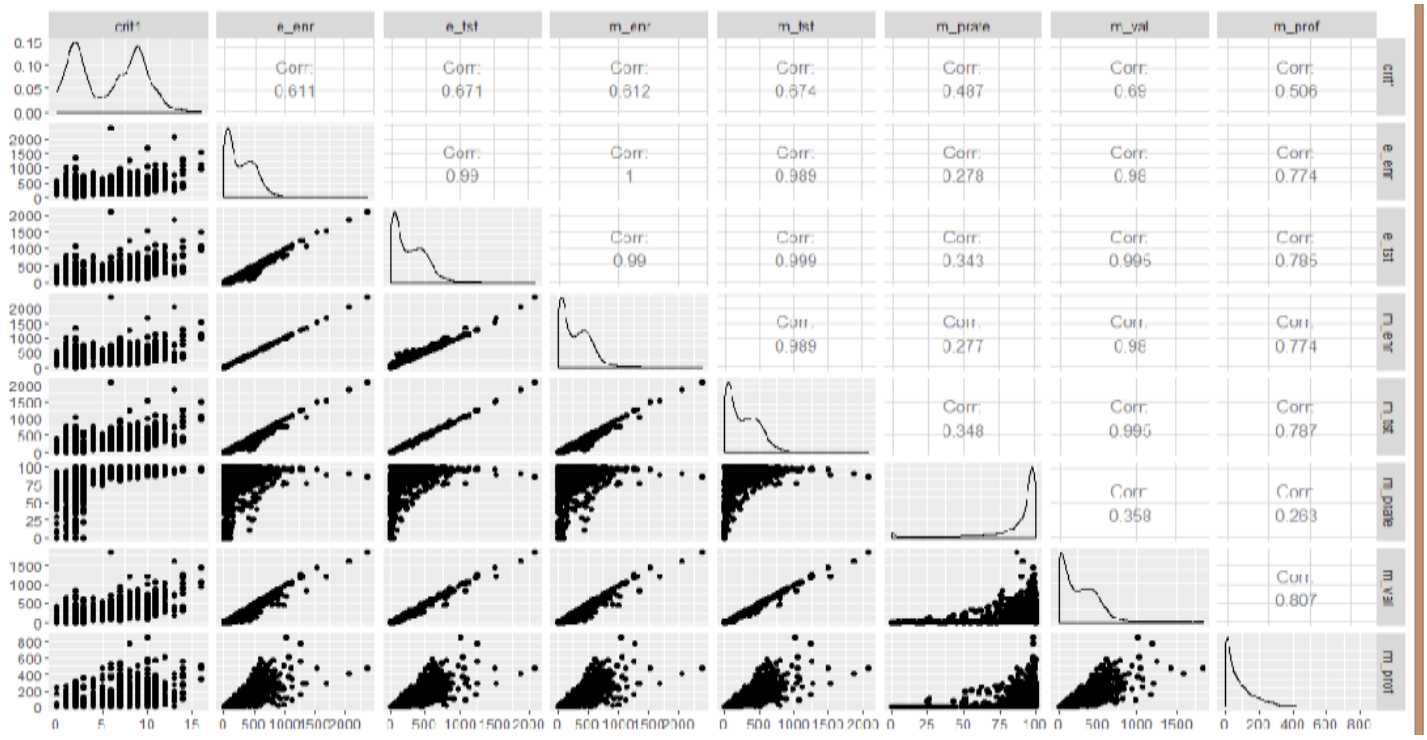


Figure 15 Identifying correlations.

5.3 Macro Clustering Analysis

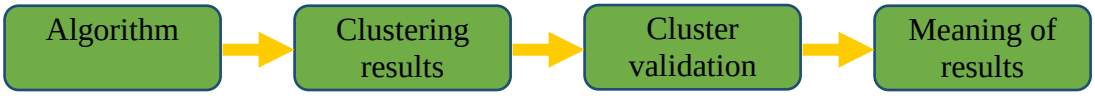
For analyzing this data, three clustering methods have been tested. Two approaches in partitioning methods, *k*-means and PAM, as well as the density-based clustering algorithm DBSCAN/OPTICS, are compared to assess which clustering technique is most suitable for the data in obtaining best clusters. Since no current available research exists in clustering these data sets at a macro level, a comparison is required.

Step 1: The Clustering of school data is performed at multiple levels

1. California high schoolwide data. This analysis uses the schoolwide data attributes and groups the schools using the algorithms. The attributes used are focused on student performance in math and other available indicators.
2. Data by major groups: African American, Asian, Hispanic, and Caucasian.

5.3.1 Schoolwide Macro Clustering

The clustering info is presented in the following steps:



***k*-Means:** As described in Section 4.1.1, *k*-means is a partitioning method based on Euclidean distance and can handle only numeric attributes. Thus, only numeric attributes on math performance, enrollment criteria, and graduation results have been included for *k*-means for analysis. The algorithm divided the 1,636 schools into the three clusters shown in Figure 16.

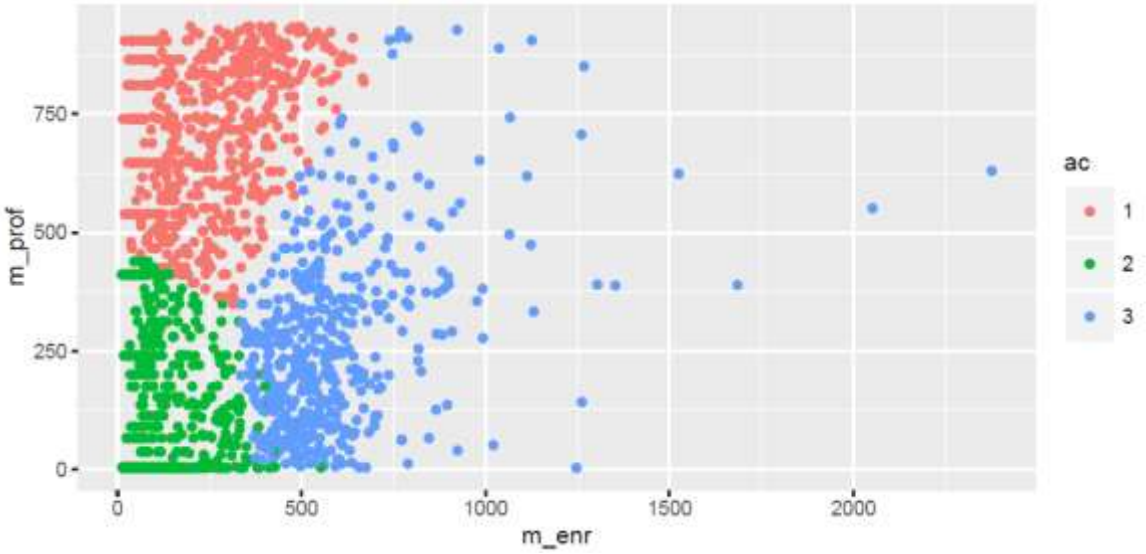


Figure 16 Plot of a three-cluster outcome using k-means.

***k*-means cluster solution validation:** The sum of squared errors (SSE) was computed for the actual data as well as for 250 runs of the clustering method on random data for validating the method and determining the optimal number of clusters for *k*-means. The plot in Figure 17 demonstrates that the clustering of the actual data is significant and that a three-cluster solution was optimal for schoolwide data for the year 2015. The optimal number of clusters is determined where the elbow bends in the graph for actual data and starts to flatten out.

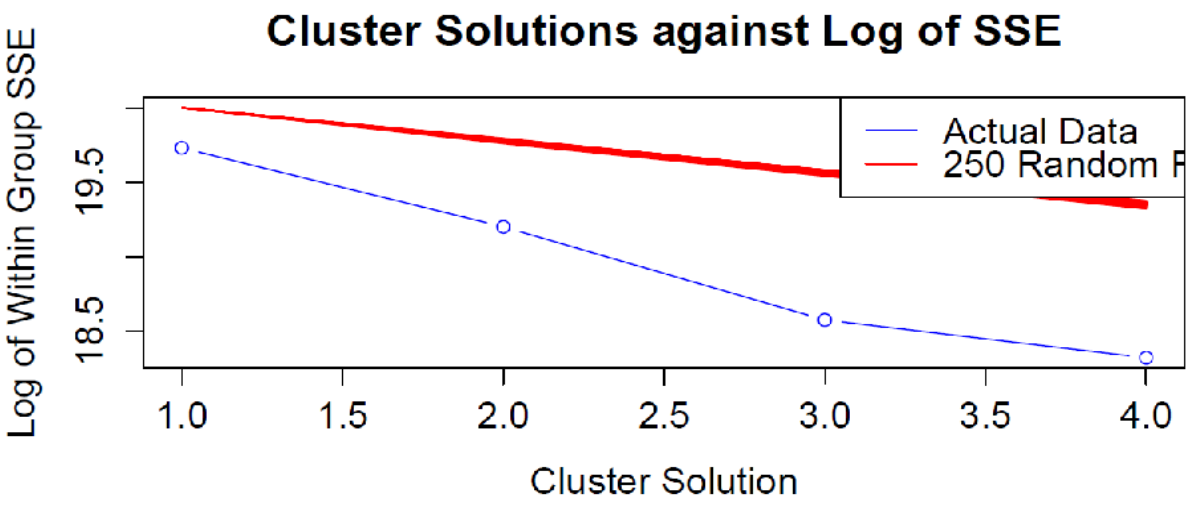


Figure 17 Sum of squared errors (SSE) validation for number of clusters.

Meaning of the *k*-means results: The results indicate that based on multiple attribute partitioning into clusters, the 1,636 schools analyzed fall into three significant groups. The differences between them are statistically significant, based on the SSE evaluation. The differences are not based on a single variable but multiple ones. Looking at the cluster means for the variables, we see difference in the value means of some of the key numeric attributes, as listed in Table 8.

Table 8: Schoolwide results with k-means

Attribute	**Cluster1**	**Cluster2**	**Cluster3**
crit1	6.360845	3.040917	8.339286
m_enr	242.9347	87.4419	580.244
m_tst	224.6603	72.18167	543.7976
m_prate	92.50096	76.79705	93.82341
m_val	211.094	62.58101	519.0615
m_prof	701.8464	125.2766	302.4444

crit1: Number of AYP criteria met, based only on participation rate and additional indicators
m_enr: Schoolwide math enrollment
m_tst: Schoolwide or LEA-wide math number of students tested
m_prate: Schoolwide or LEA-wide math participation rate
m_val: Schoolwide or LEA-wide math valid scores
m_prof: Schoolwide math number of students scoring proficient or above

The following are sample schools from the different clusters:

Cluster 1: Saratoga High

Cluster 2: Riverdale High

Cluster 3: Thousand Oaks High

A random sampling of the schools from each cluster indicates that Cluster 1 consists of schools with high performance numbers, located in urban areas. Further examination shows that the schools have more diversity and a high degree of multiethnic population. Cluster 2 school sampling indicates that most of them are located in interior rural and semiurban areas or schools with less diversity. Cluster 3 schools show a profile of wealthy neighborhood schools in urban and semiurban settings. Appendix 4 shows the demographic profile of the example school from each cluster.

Although *k*-means as a benchmarking algorithm gives decent results, it does not withstand the effects of outliers and cannot combine categorical and other data types. The next algorithm, PAM, addresses these issues.

PAM: As described in Section 4.1.2, PAM (Partitioning Around Medoids) is a partition-based algorithm like *k*-means, but instead of using mean values of variables, it centers its clusters on actual data objects called medoids, so it is not restricted to numeric variables that can be averaged, and it can use distance metrics other than Euclidean distance. In this study the Gower distance is applied so that variables that are categorical can be included in the analysis. A few categorical variables, such as sw_gr_met (schoolwide graduate rate met: yes, no, na) and MetAttendTarg (attendance target met: yes, no, na), are included. The levels are converted to numeric codes, but the attribute type stays as categorical. By applying Gower distance, a dissimilarity measure, the resulting clusters are more succinct and robust to outliers. PAM was tested for a three and four clusters. As the visualizations in Figures 18 and 19 and the sum of squared error validation demonstrates, a four-cluster approach is most suitable with PAM while including the categorical attributes.

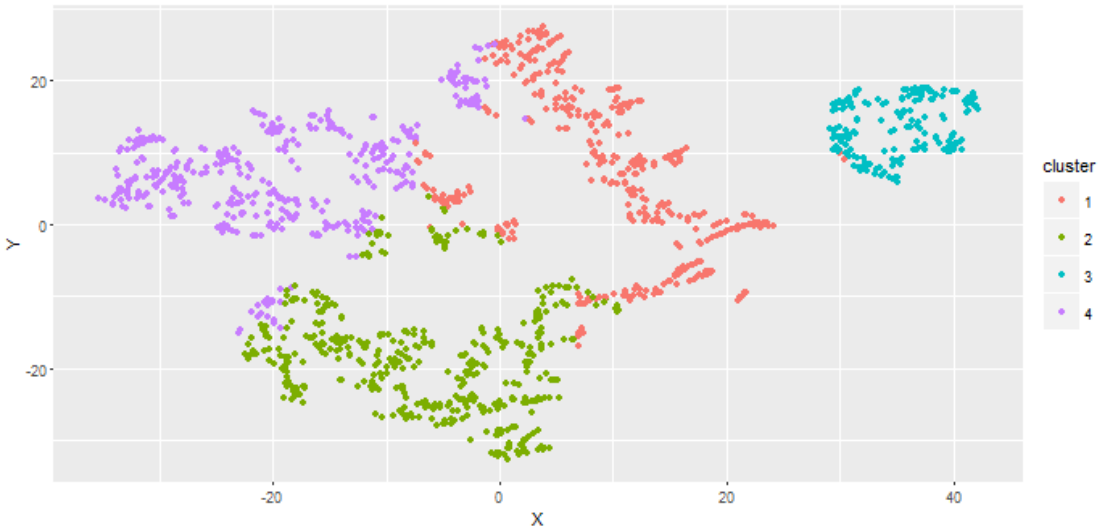


Figure 18: Four clusters obtained with PAM based on the Gower distance.

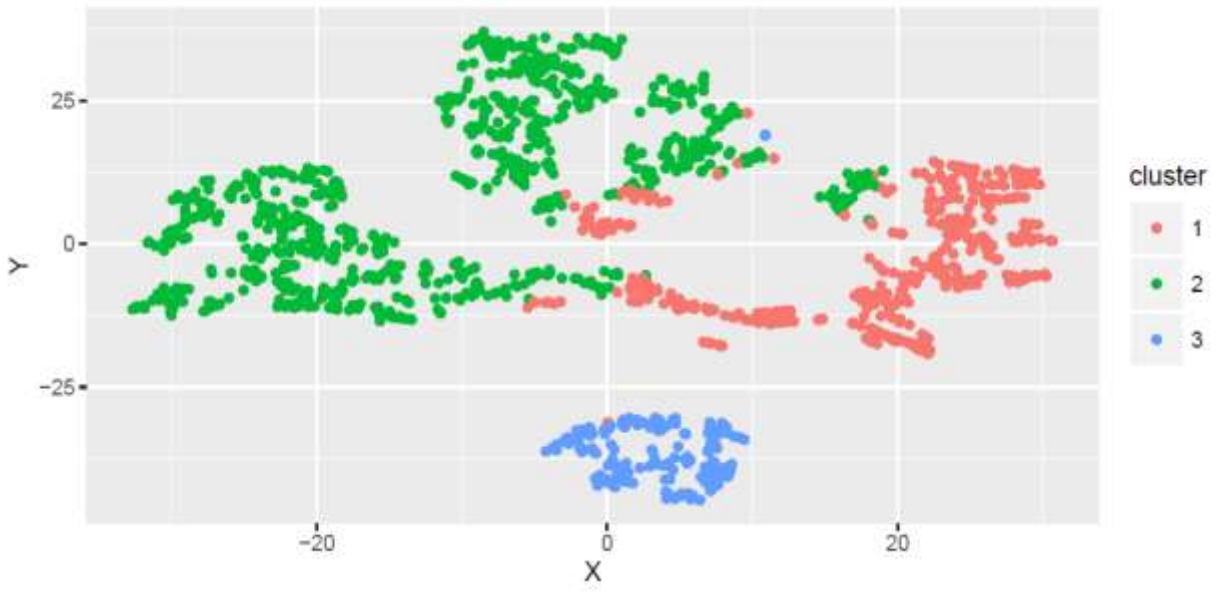


Figure 19: Three clusters obtained with PAM.

PAM cluster validation: Computing SSE for the clusters found by PAM resulted in an optimal number of four clusters. This indicates that adding categorical variables resulted in additional information for forming more clusters. The Gap statistic and the silhouette method were also used for assessing the optimal number of clusters, but neither resulted in significant numbers, in terms of giving more weight to those clusters; thus, the SSE measure is used. A visual comparison between the four-cluster and three-cluster plots in Figures 18 and 19 also indicate that a four-cluster partitioning has more distinct clustering quality. Figure 20 illustrates the optimal number clusters as measured by SSE.

Figure 20: SSE for different number of clusters found by PAM.

Meaning of PAM cluster results: Cluster validation showed that at a macro level, the high school data is most effectively grouped into four clusters. Table 9 shows the same key math performance attributes in the dataset that were examined in Table 8 but for the medoids of the four clusters found with PAM. This table shows that each of the clusters is distinct. The only math-related attribute that is almost equal across clusters is m_prate, which is the schoolwide math participation rate, and this is a good indicator that all schools had more or less equal participation rates in math, so the clustering was not biased. Cluster 2 is the only cluster where m_prate is slightly lower (87 versus 96 across all clusters), and it is also a cluster that has grouped the worst-performing schools.

Cluster 1 in the PAM results consists of schools average-performing schools, showing a standard normal distribution in terms of performance. Cluster 2 consists largely of schools that are disadvantaged in terms of student backgrounds and their performance. Cluster 3 indicates the high-profile schools in urban settings with large enrollments and student groups that are high

performing. Cluster 4 shares a similar profile to Cluster 3; however, it also has some elite schools as well as some underperforming schools. Further geographic and economic investigation into these schools would highlight more differentiating factors. The regression section takes a random sample of schools from the list and attempts to identify other varying attributes that can explain these differences in performance.

The results from PAM clustering do prove that it is a very effective algorithm to use for the nature of data in this research that has multiple data types. Since *k*-means cannot handle nonnumeric data types, PAM is a more suitable model.

Table 9. Sample attribute representative values for each cluster for schoolwide data for PAM algorithm

Attribute	**Cluster1**	**Cluster2**	**Cluster3**	**Cluster4**
crit1	3	2	9	9
m_enr	120	49	526	364
m_tst	114	34	500	337
m_prate	96	87	96	96
m_val	107	21	480	318.5
m_prof	29	3	211	68
m_pprof	23.15	10.5	37.3	21.3

crit1: Number of AYP criteria met, based only on participation rate and additional indicators
m_enr: Schoolwide Math Enrollment
m_tst: Schoolwide Math Number of Students Tested
m_prate: Schoolwide Math Participation Rate
m_val: Schoolwide Math Valid scores
m_prof: Schoolwide Math number of students scoring Proficient or Above
m_pprof: Schoolwide Math Percent of students scoring Proficient or Above

The histograms in Figure 21 for the variable m_pprof (schoolwide percent proficient or above in math) illustrates the cluster differences.

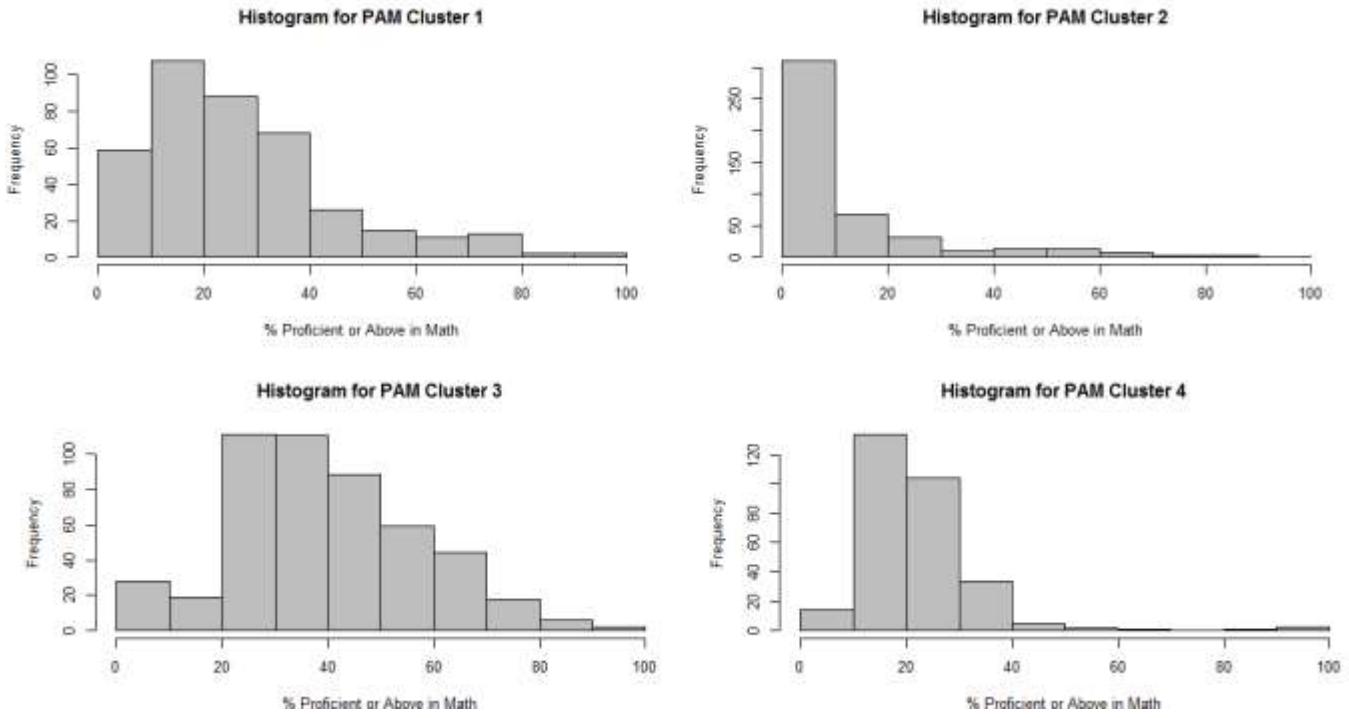


Figure 21: Histograms for variable m_pprof for the four PAM clusters.

DBSCAN/OPTICS: DBSCAN is a density-based method as described in Section 4.1.3. The large-scale data version of OPTICS (ordering points to identify the clustering structure) was used for analyzing the data. Applying it to the data for the research problem gave very unsatisfactory results. Figure 22 is a plot of the results from DBSCAN where it created multiple sparse clusters and treated more than half the observations as noise points. The independent discreate observations could have posed a density problem for DBSCAN. Figure 22 illustrates the density of the clusters.

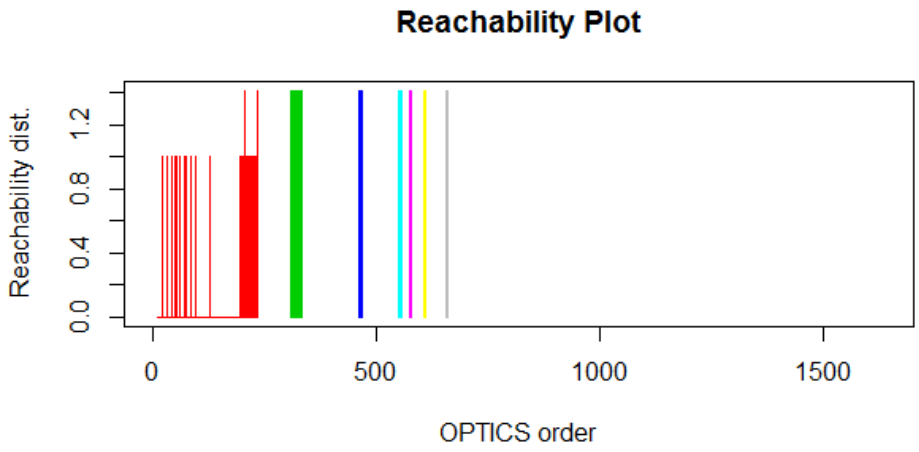
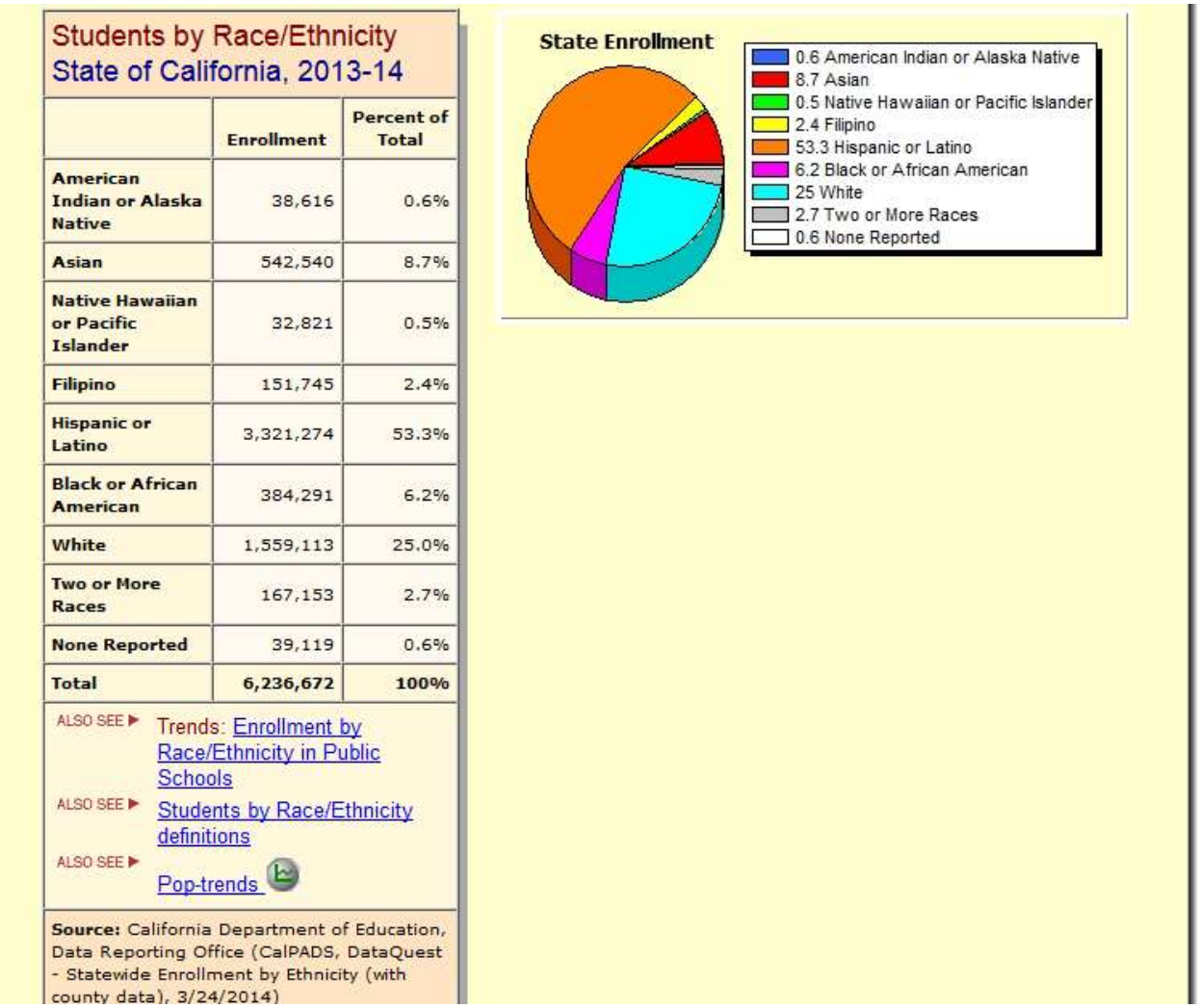


Figure 22: Clustering results from DBSCAN, showing multiple sparse clusters with most observations treated as noise.

5.3.2 PAM Clustering by Student Ethnicity in 1,636 High Schools

California is an ethnically diverse state, and statistics are stored on all major groups. In order to make sure that majority of the population can be productive and contribute to the economy, imparting education successfully for various student groups is critical. A macro analysis that can connect to the micro level of student learning methods is imperative to ensure optimal use of resources. Figure 23 gives a statistical overview of the overall ethnicity profile of student enrollment in California high schools. Figure 24 compares the ethnic distribution of students in California and in the United States.



Students by Race/Ethnicity
State of California, 2013-14

	Enrollment	Percent of Total
American Indian or Alaska Native	38,616	0.6%
Asian	542,540	8.7%
Native Hawaiian or Pacific Islander	32,821	0.5%
Filipino	151,745	2.4%
Hispanic or Latino	3,321,274	53.3%
Black or African American	384,291	6.2%
White	1,559,113	25.0%
Two or More Races	167,153	2.7%
None Reported	39,119	0.6%
Total	**6,236,672**	**100%**

ALSO SEE ▶ Trends: Enrollment by Race/Ethnicity in Public Schools

ALSO SEE ▶ Students by Race/Ethnicity definitions

ALSO SEE ▶ Pop-trends

Source: California Department of Education, Data Reporting Office (CalPADS, DataQuest - Statewide Enrollment by Ethnicity (with county data), 3/24/2014)

Figure 23: Ethnic distribution of California high school students.

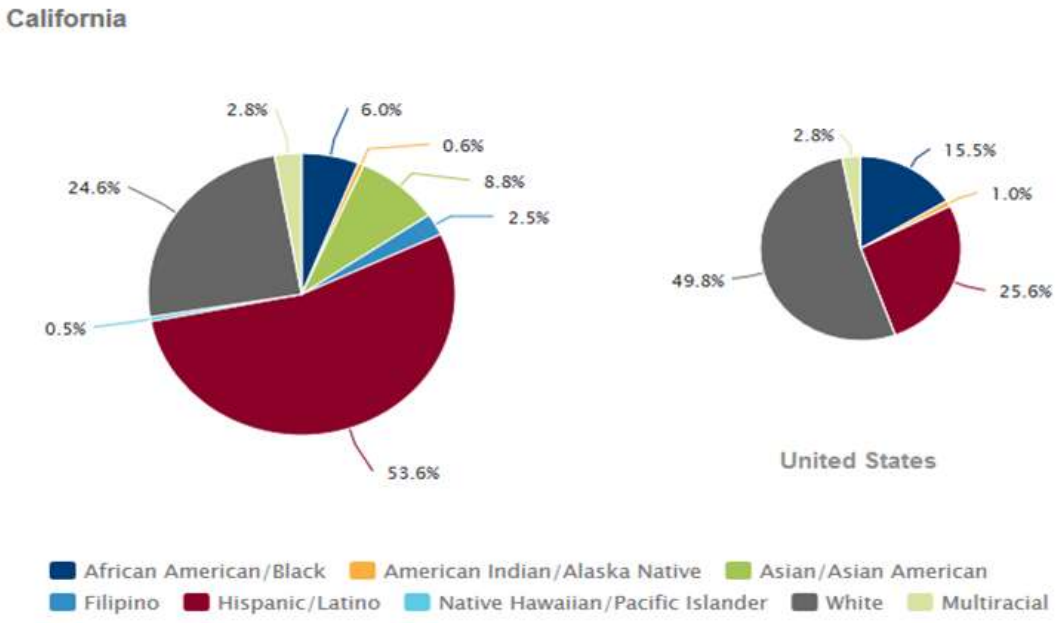


Figure 24: Comparison of ethnic distribution of students in the State of California and the United States. (Source: California State Department of Education.)

PAM clustering of math performance by Hispanic/Latino high school students:

Hispanic/Latino students constitute 53.6% of the enrolled student population in the state of California. They are the single largest ethnic group in the state. However, their academic performance, especially in math, is not on par with white or Asian groups. In math, 69% of Asian students achieved the state targets, compared to 49% of whites, 21% of Latinos, and 16% of blacks (California Department of Education, 2015).

Enhancing the learning and performance of students who presently suffer from academic deficits has lasting economic, social, and political implications for the success of the state as a whole. The question arises how performance trends are distributed relative to ethnic composition among California schools. To answer this question, PAM clustering algorithm was used after

validating with schoolwide data set. The cluster validation using SSE indicated a three-cluster outcome as optimal. Figure 25 shows the cluster plot for three clusters.

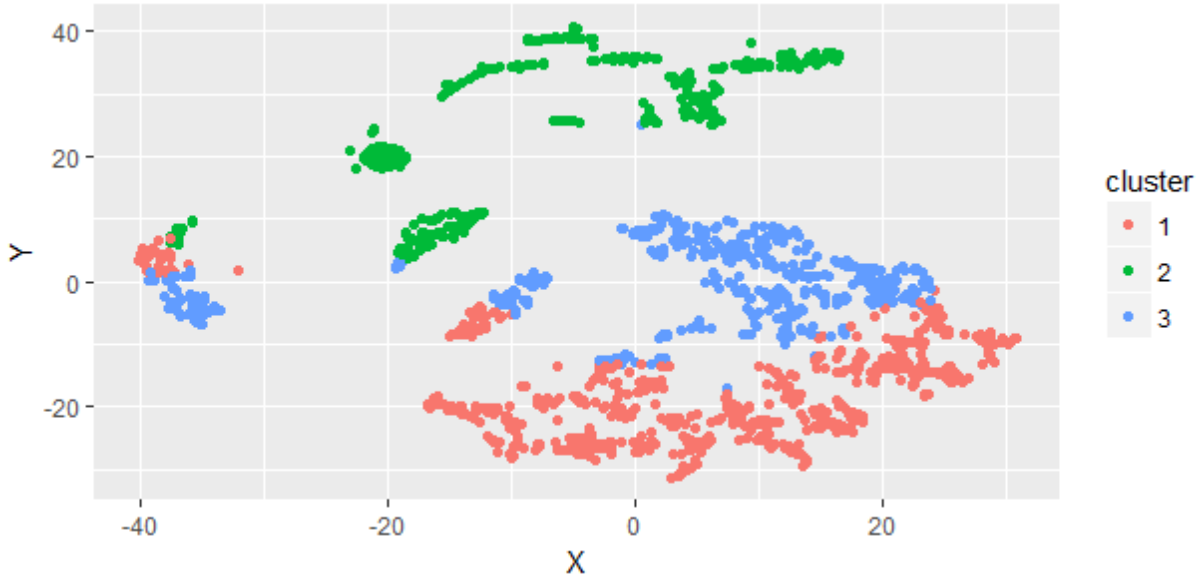


Figure 25: Plot of 3 cluster solution using PAM for Hispanic/Latino data.

Results for PAM clustering, Hispanic: The results, shown in Table 10, indicate that the three clusters are distinct. The first difference that is highlighted is the Hispanic student enrollment in each cluster. Cluster 1 has medium enrollment indicated by the number of students participating in math test, Cluster 2 has small enrollment, and Cluster 3 has large enrollments. The results in the clusters indicate that the performance gap between Cluster 1 and Cluster 3 is much less than with Cluster 2, which contains schools that are set in remote locales as well as state-run academies for bilingual programs.

Table 10: Representative values of key math performance attributes for Hispanic/Latino high school students in California.

Attribute	Cluster 1	Cluster 2	Cluster 3
mt_hi	72	10	280
mv_hi	67	6	272
mpp_hi	18.1	6	20
enp_hi	32	8	131

mt_hi: Math tested, Hispanic or Latino
mv_hi: Math valid scores, Hispanic or Latino
mpp_hi: Math students scoring proficient or above, Hispanic or Latino
enp_hi: ELA number students scoring proficient or above, Hispanic or Latino

The clustering does validate the CDE result of an average of 21% state wide Math proficiency for Hispanic group; however, by separating groups within that profile, it is easier to target remedial programs to the students who need them most. Figure 26 consists of histograms illustrating math proficiency rates by cluster.

Clustering also helps to identify the percentage cohorts and delve deeper into those for identifying similarities in specific groups. Identifying the attributes of students who perform below the norm and above the norm in each cluster and the variables at play can give deep insights for policy. For example, Los Altos High School in Cluster 1 is a wealthy school district that also admits 28% of Hispanic/Latino students from nonwealthy neighborhoods; its Hispanic math proficiency rate is at 31%. Los Osos school in Cluster 3, on the other hand, with a high ethnicity mix and not as wealthy as Los Altos, has a Hispanic proficiency rate at 44% with a graduation rate of 90%, some of the highest in the state. Thus, developing a system that can analyze the attributes at macro and micro levels can assist teachers to use collective knowledge to enhance student learning.

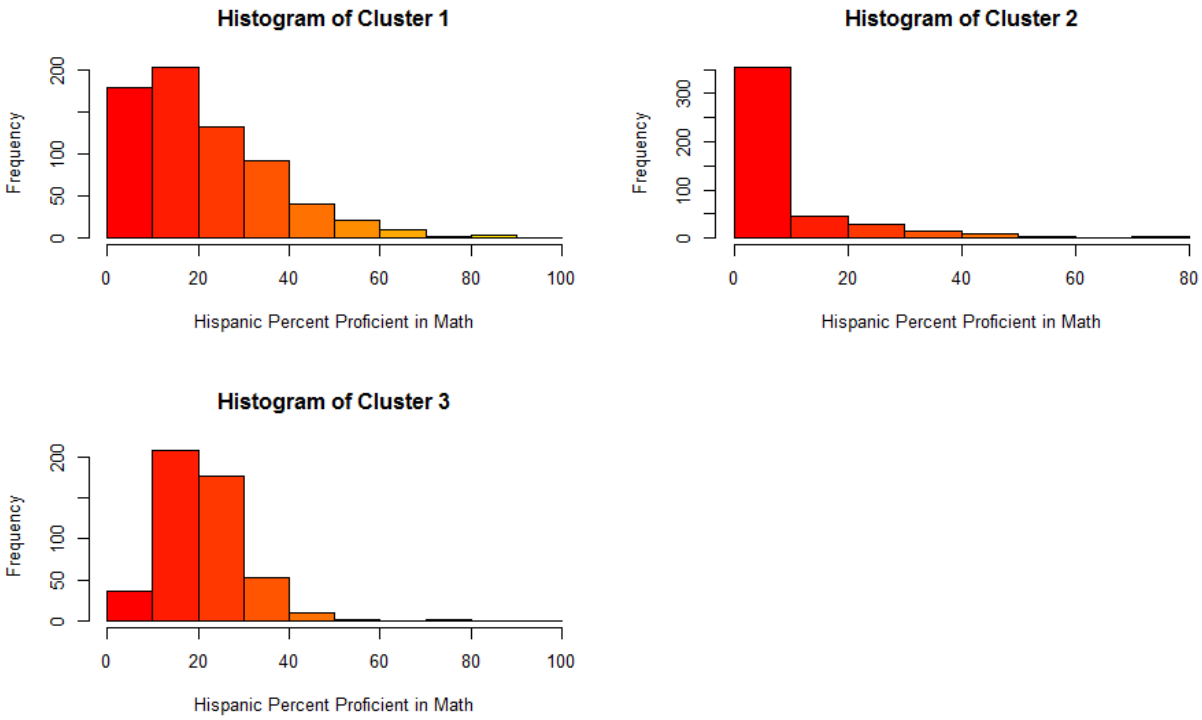


Figure 26: Hispanic math proficiency rates across state of California in three clusters.

PAM clustering for white student group: Figure 27 depicts the clustering output from PAM for clustering by white students only. SSE analysis (not shown) favored a three-cluster outcome for the white student set. Table 11 shows the corresponding representative values, and Figure 28 shows the histograms.

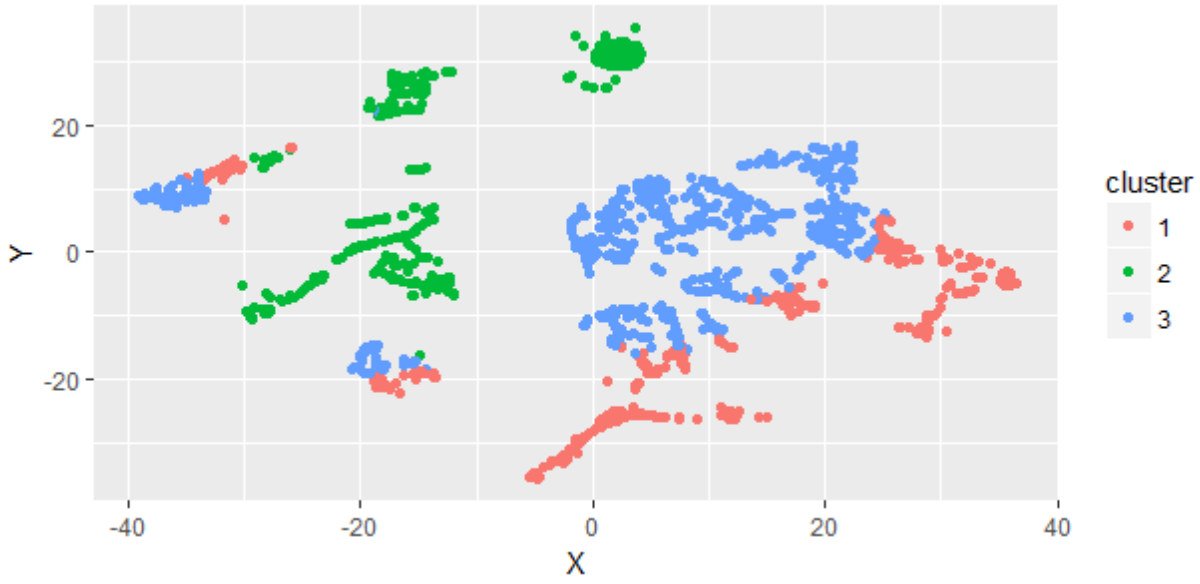


Figure 27: Three-cluster results of using PAM algorithm for white student data.

Table 11: Representative math attributes for white student group

Attributes	**Cluster 1**	**Cluster 2**	**Cluster 3**
mp_wh	93	96	95
mv_wh	7*	88	232
mpp_wh	12	39.5	49

mp_wh: Math participation, white
mv_wh: Math valid scores, whitempp_wh: Math students scoring proficient or above, white

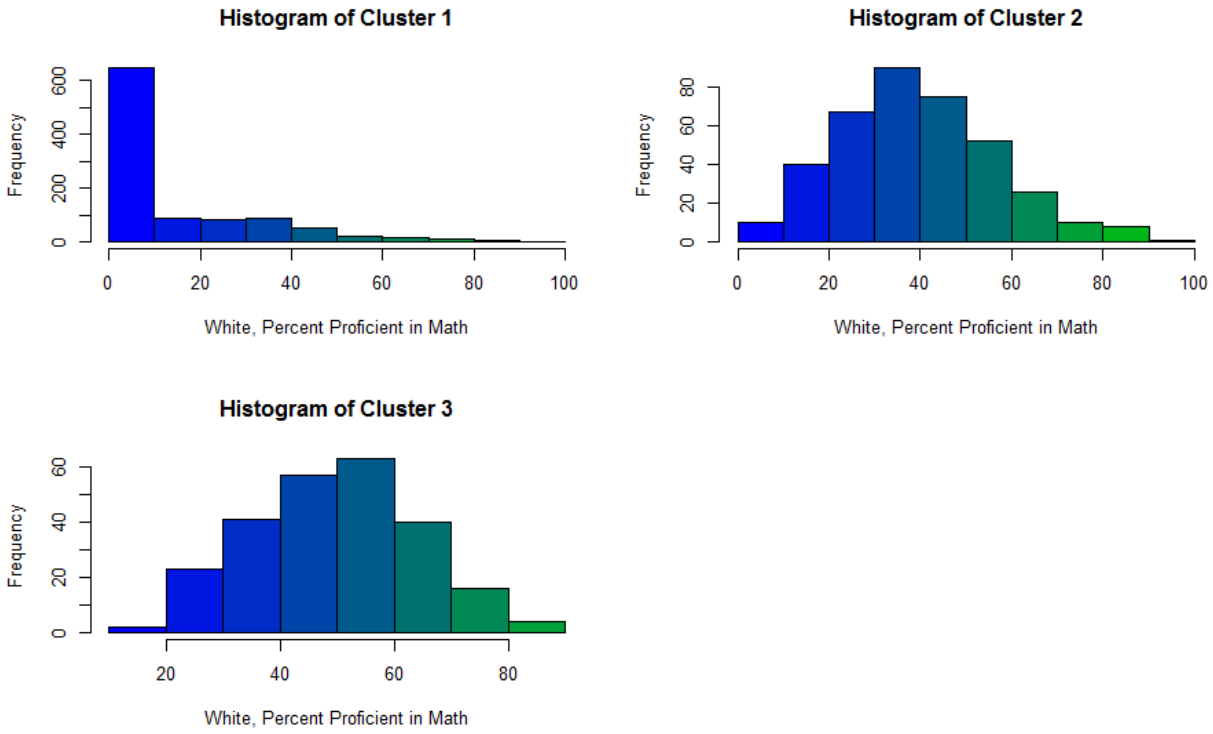


Figure 28: White math proficiency rates across state of California in three clusters.

There is a high variation of values in several attributes, even though the math participation rate is very similar across the three clusters: 93%, 96%, and 95%, respectively. For example, the math percent proficient or above (mpp_wh) varies significantly, as indicated in Table 10 and the histograms. Cluster 1 represents the low-performing schools with a representative value of 12%, while Cluster 2 shows a medium representative value of 39.5% and Cluster 3 a higher representative value of 49%. This pattern is noticed across multiple variables, and it does show that the PAM algorithm is extremely effective in detecting real clusters with large datasets with and high dimensionality.

PAM Clustering for African Americans: The clustering results from PAM for the African American student group resulted in a two-cluster model, shown in Figure 29. The

histograms for the two groups are shown in Figure 30, and representative values are tabulated in Table 12.

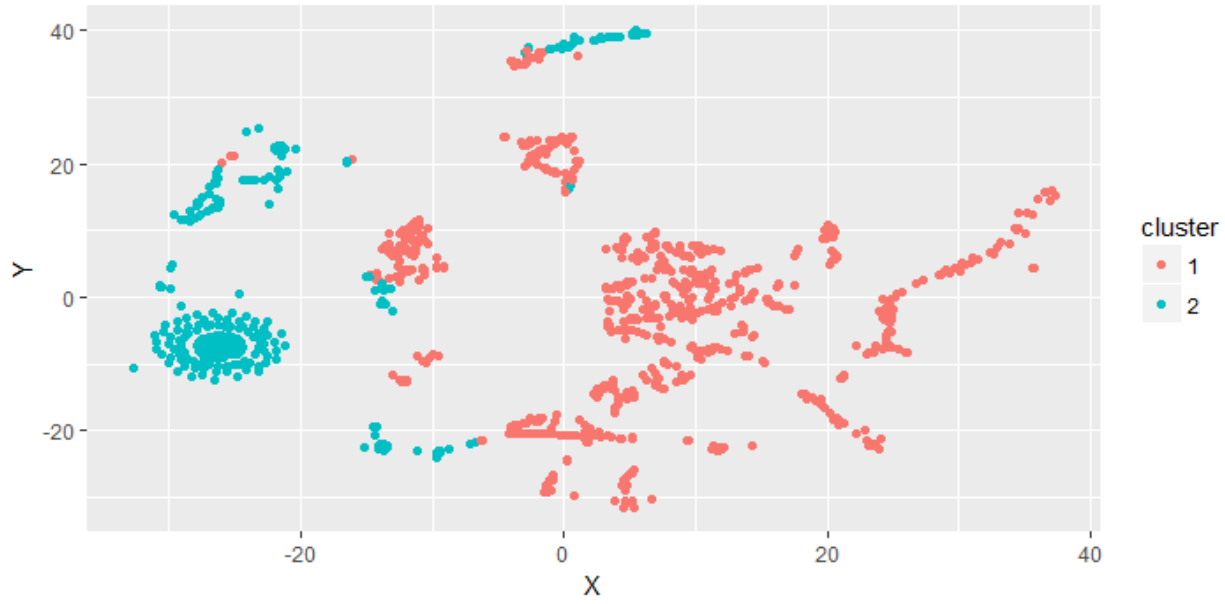


Figure 29: Clustering by African American student group.

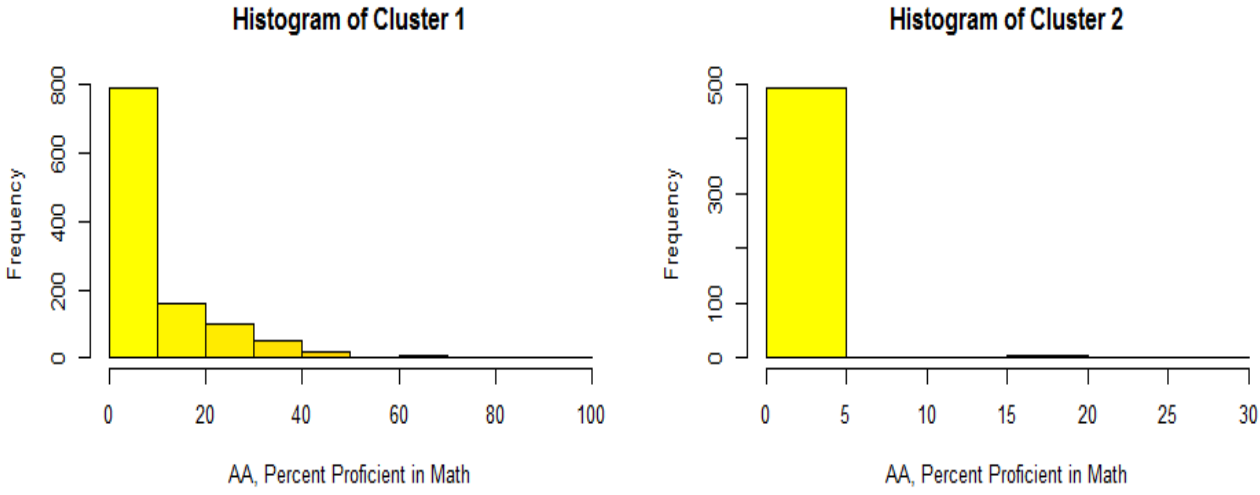


Figure 30: Histograms of African American students for the attribute "percent proficient in math" for the two clusters.

Table 12: Representative values of math performance attributes for African American students

Attribute	Cluster 1	Cluster 2
mpp_aa	7.907686	0.428373
mv_aa	20.09541	2.321429
mnp_aa	2.818905	0.1130952

mpp_aa: math percent of students scoring proficient or above, African American
mv_aa: math valid scores, African American
mnp_aa: math number of students scoring proficient or above, African American

The resulting two clusters have wide differences. The number of students enrolled in math for cluster 1 is 27,807 versus 2,348 for cluster 2. Cluster 1 came from schools where the enrollment is visible, and cluster 2 from schools where very few African American students are enrolled. Another prominent aspect is that students whose math scores are on the higher side of the graph also come from schools that have high-performing students. The highest math percent proficient or above for the African American group in California is 90.5%, and it is from the Whitney (Gretchen) High School in Cerritos, where about 23 African American students are enrolled. The ethnicity of the school is predominantly Asian (and Filipino), with a significant minority Hispanic students, as shown in Figure 31.

Figure 31: Enrollment at Whitney (Gretchen) High School, Cerritos, California, where the math percent proficient or above is 90% for African Americans, one of the highest rates in the nation.

On the other hand, Cabrillo High (Figure 32) has math percent proficient or above percent at 3% for African Americans. The school profile is among the lowest in the nation. Further comparisons based on enrollments and math performance indicate that schools that are predominantly white or black seem to be the schools where African American students perform the worst, and the more diversity a school has and the higher the overall performance of the school, the higher the African American students' will be also.

Figure 32: Cabrillo High School in Santa Barbara County, where math percent proficient of African American students is among the worst in the United States.

Clustering for Asians: The Asian community includes students with ethnic backgrounds in East Asia, South Asia, or South East Asia. PAM clustering for Asians resulted in three clusters (Figures 33 and 34). The first cluster represented a set of schools where the Asian student enrollment is extremely minimal but existing and the performance is along the line of the average for all populations. The second cluster reflected a set of schools where little or no Asian enrollment was present and, if any, their performance was poor. The third cluster reflected schools where Asian enrollment population is considerable. Table 13 indicates that Cluster 3 results reflect an above-average performance for Asians in schools that have significant enrollment, as indicated by the attribute me_as. Cluster 1, with little enrollment of Asian students, saw poor performance, as did cluster 2, where the enrollment is almost negligible.

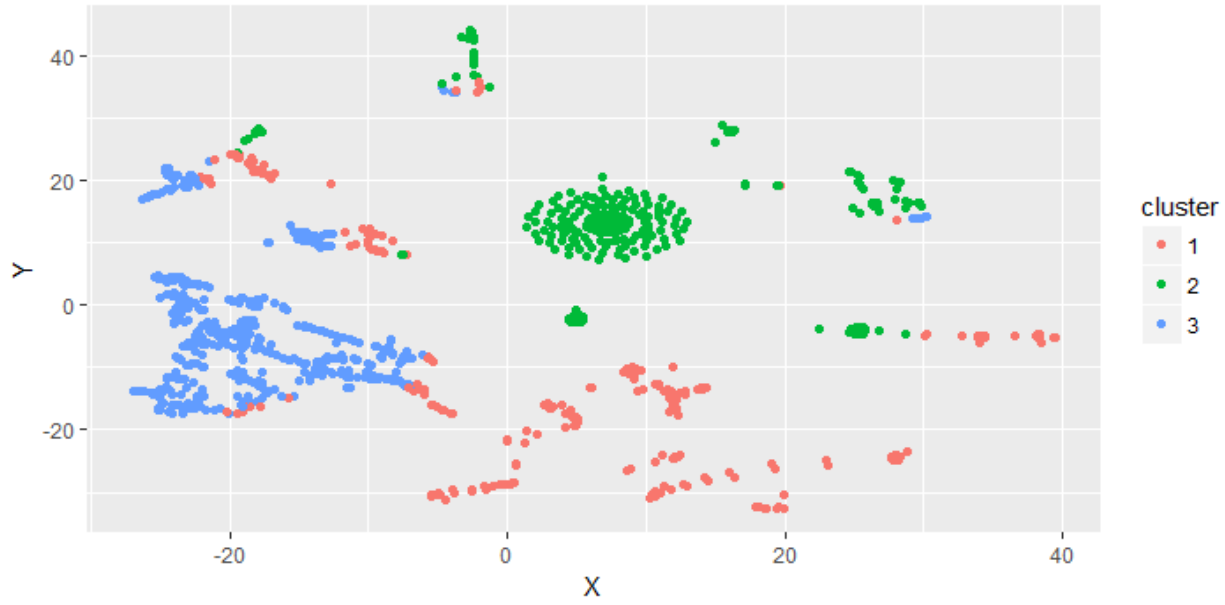


Figure 33: Clustering results for Asians (including South Asia, South East Asia, and East Asia).

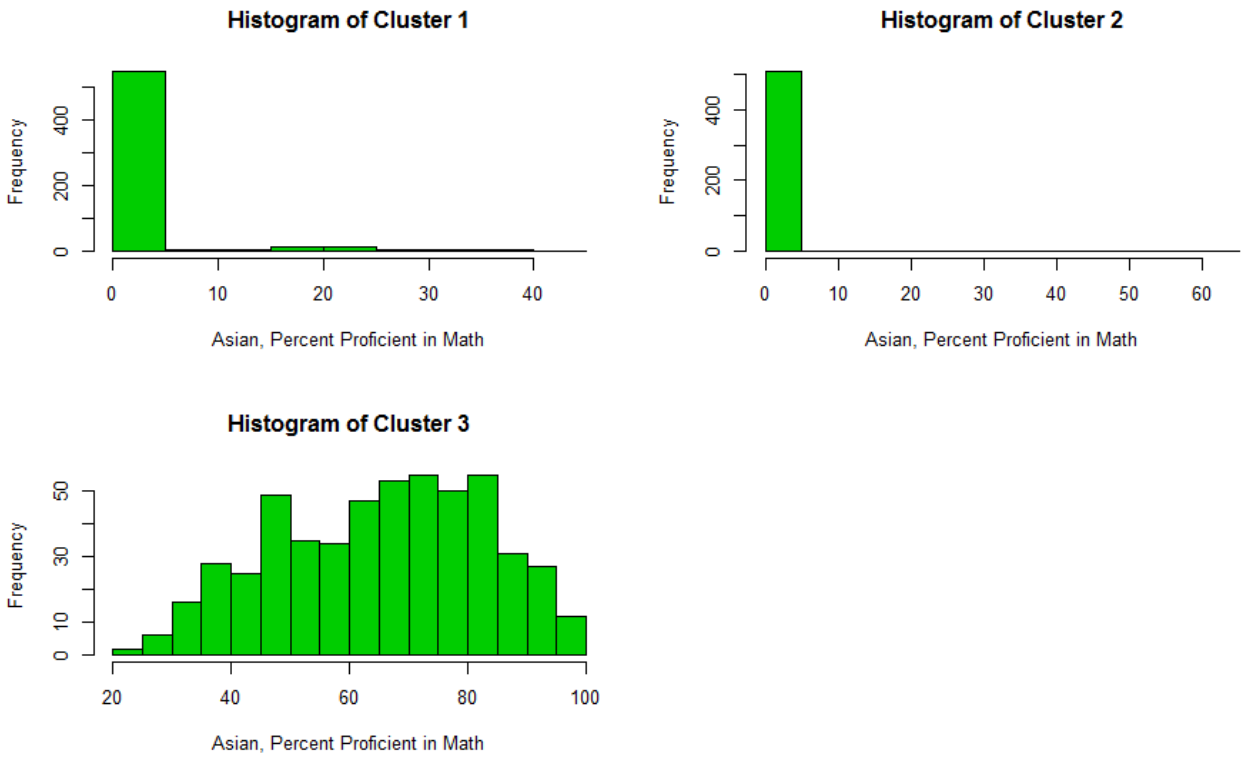


Figure 34: Histogram of percent proficient or above in math for Asians (including South Asian, South East Asian, East Asian).

Table 13: Asian math performance indicators for three clusters

Attribute	Cluster 1	Cluster 2	Cluster 3
mpp_as	1.834057	0.3558594	65.59086
me_as	7.100167	1.029297	73.77143
mv_as	6.085142	0.5761719	68.37333

mpp_as: math percent proficient or above for Asians
me_as: math enrollment for Asians
mv_as: total valid math scores, Asians

5.3.3 Clustering of District-Level Data

District data is analyzed to see whether differences in performance outcomes can be captured at the district level and which attributes indicate that. The analysis can give insight to identify whether the policies or indicators at district level impact or highlight the differences at school level.

The district-level data consist of demographics, performance indicators, and finance information. The table in Appendix 3 lists all 51 attributes used for analysis. Finance data is not published at the school level, and it is up to individual schools to release the information. Clustering and regression are performed on district data for the following purposes:

1. To observe link between the school clustering results and district clustering results and draw insights from analysis
2. To identify multiple factors at the district level that associate with student performance

The PAM clustering approach was implemented for this dataset because, like the schoolwide data, it also contains a combination of categorical and numeric attributes. Total SSE for identifying the number of clusters yielded four clusters for 2015 data and three clusters for 2014 data, as shown in the graphs in Figures 35 and 36 respectively.

Figure 35: Total SSE for 2015 district finance data for clustering.

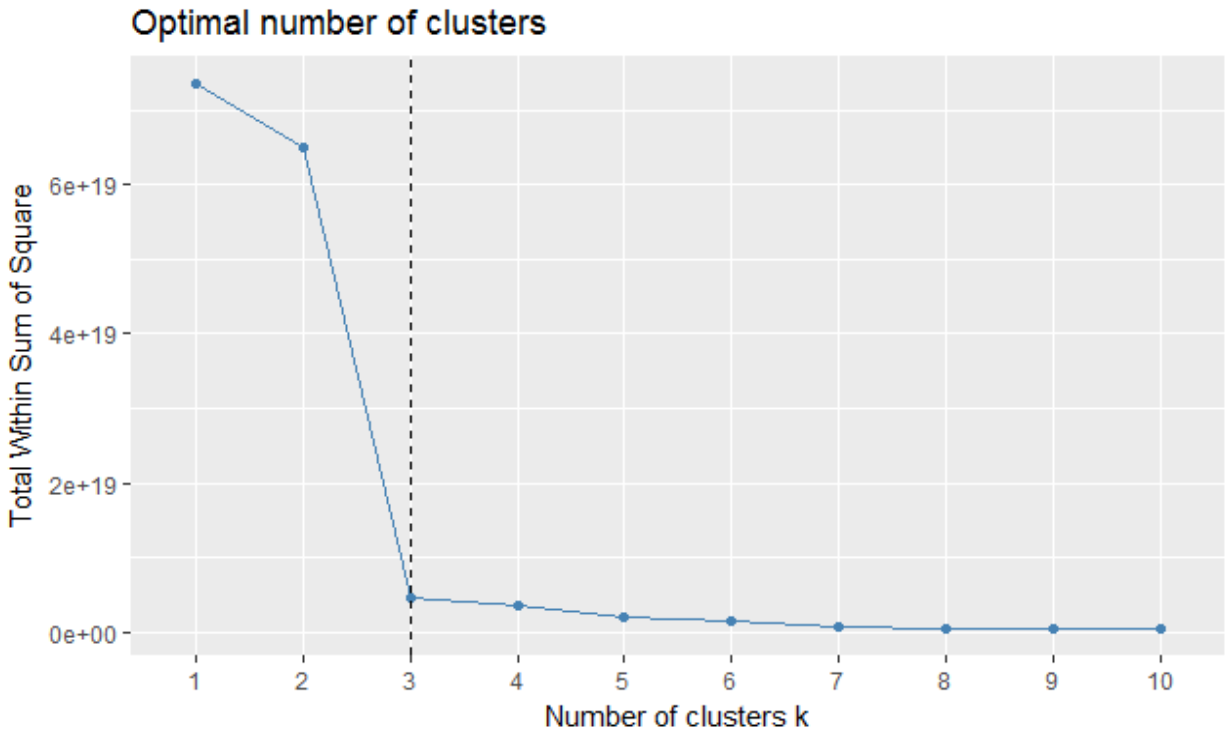


Figure 36: Total SSE for 2014 district-level finance data for clustering.

District Data Clustering Analysis Results: As just seen, clustering results for 2015 and 2014 differ in terms of clustering outcomes: while 2015 has a four-cluster outcome (Figure 37), 2014 has a three-cluster outcome (Figure 38). Changes between the years indicate data reporting on foster youth and SAT writing score attributes in 2015. The demographic diversity reporting on Native American population was more detailed in 2015 than in previous years. Whereas macro analysis at the school level brought out differences in performance indicators, district data analysis captured the differences from the following perspectives:

1. Economic
2. Demographic
3. Teachers

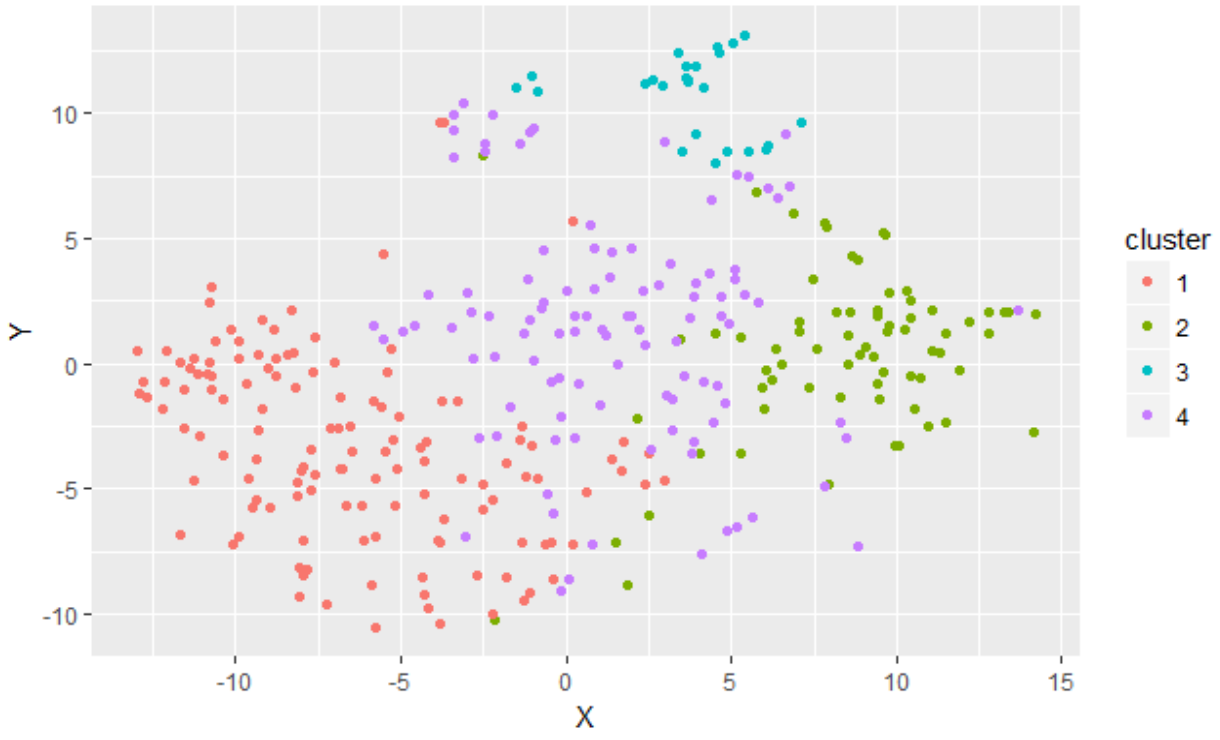


Figure 37: 2015 data, four-cluster solution of district-level data analysis.

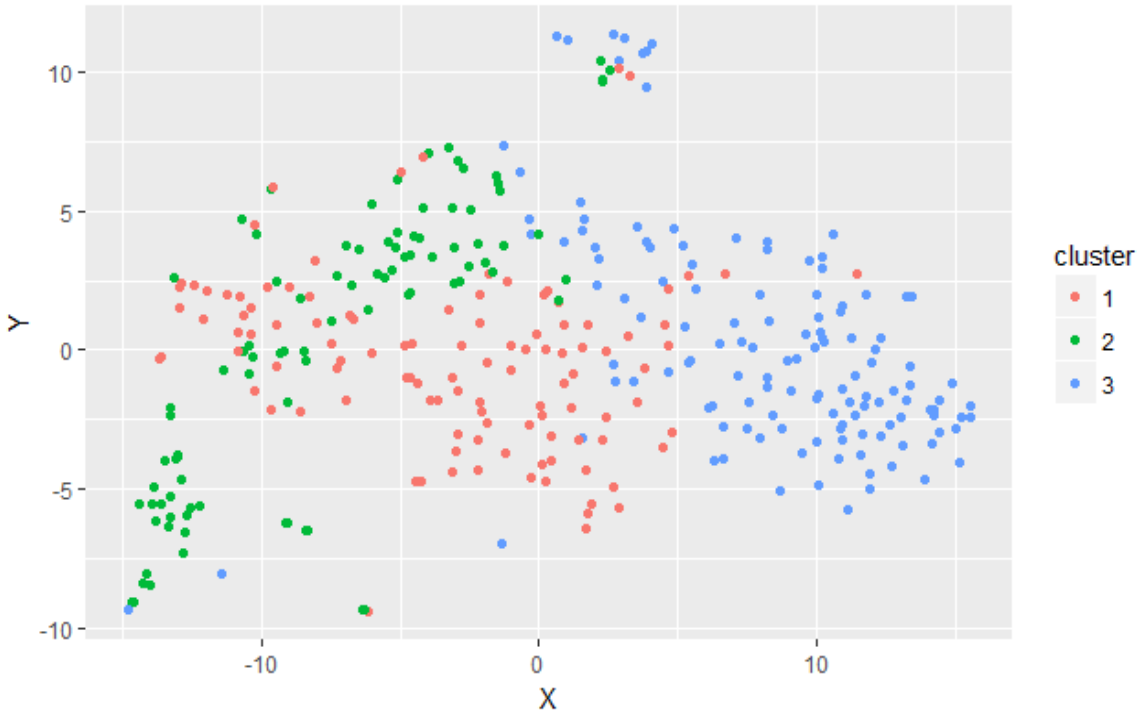


Figure 38: 2014 data, three-cluster solution of district-level clustering.

The results from the clustering analysis will be discussed under the following categories in the context of student performance.

Economic aspects: The economic attributes available for the clustering are listed in Table 14. Tables 15 and 16 show that there are differences in student outcomes, and those are not necessarily driven by financial inputs driven by the districts. Districts where TotalGenFundRevenuesPerStudent is high and federal and local revenue per student is high do not automatically have higher graduation outcomes nor higher preparedness for meeting the University of California/California State University entrance requirements. While graduation rates are seemingly close to each other for all the clusters, the preparedness for university admissions varies drastically.

Table 14: Economic attributes available for district-level clustering

TotalGenFundExpendituresPerStudent
TotalGenFundExpenditures
GenFundExpbyActivityPer
ActivityInstrucRelatedSvcsExpPer
TotalGenFundRevenues
TotalGenFundRevenuesPerStudent
PupilServicesExpPer
InstrucrelatedSvcsPerStudentNum
CertificatedSalariesPerStudent
ClassifiedSalariesPerStudent
PupilServicesPerStudentNum
FederalRevenuePerStudent
StateRevenuePerStudent

Table 15: 2015 district clustering results on key economic attributes in context of student graduation outcomes

Cluster	**Enrollment (in 1,000s)**	**Total-GenFund-Revenues $**	**Total-GenFund-Revenues PerStudent $**	**Federal-Revenue-Per-Student $**	**Local-Revenue-Per-Student $**	**Grads-Mtg-UCCSU Percent***	**Cohort Graduates%**
1	13,469	120,620,393	10,453	701	598	34.2	87.3
2	10,223	90,375,754	9,299	354	767	53.4	93.7
3	339	4,930,741	15,900	1,196	1,308	18	84.1
4	3,130	29,433,880	10,008	530	715	33.4	90.5

*Graduates meeting University of California/California State University (UC/CSU) entrance requirements

Table 16: 2014 district clustering results on key economic attributes in context of student graduation outcomes

Cluster	**Enrollment (in 1,000s)**	**Total-GenFund-Revenues $**	**Total-GenFund-Revenues-PerStudent $**	**Federal-Revenue-Per-Student $**	**Local-Revenue-Per-Student $**	**Grads-Mtg-UCCSU %**	**Cohort Graduates%**
1	14,575	122,197,243	8,762	483	574	44	89.6
2	3,357	24,859,252	9,163	467	831	36.3	90.2
3	4,480	43,109,034	9,583	693	568	31.9	86.3

Demographic aspects: The demographic attributes consist of percentage representation of each major ethnic group of students in the state. Tables 17 and 18 show the main attributes based on clustering and year in context of the three student outcomes in district performance indicators. The clustering for years 2015 and 2014 indicates that district student populations that are very diverse (such as Cluster 2 in 2015 and Cluster 1 in 2014) tend to perform better compared to skewed demographics and student populations that has mostly two large ethnic groups.

Linking this aspect to the analysis at the school level, where performance numbers in mathematics and language based on student demographic profile are used in clusters, it confirms the same outcome: diversity in a school setting enhances student performance outcomes.

Teachers: Observing the teacher-related attributes and based on the regression model (results shown at the end of this section) run on all 51 attributes, the clustering data (Tables 19 and 20) shows that teacher-related parameters do not produce any significant differences between the clusters, as the data across clusters does not vary drastically and seems to show a statewide uniformity in terms of teacher attributes. The only attribute that had a slight significance level was the teacher lowest salary offered.

Table 17: Student demographics by cluster for 2015 district data

Cluster	American-Indian-Alaska-NativePer	Asian-Per	African-AmericanPer	Filipino-Per	Hispanic-Per	Hawaiian-PacIslanderPer	Two-More-Races-Per	None-Reported-Per	White-Per	GradsMtg-UCCSUPer
1	0.3	1.9	2.7	0.9	75.9	0.3	1	0.2	9.7	34.2
2	0.3	7.3	2.6	2.6	26.6	0.4	5.3	0.2	45.2	53.4
3	3.6	0.5	0.9	0	20	0	1.9	0	59.9	18
4	0.6	1.4	1.2	0.7	49.6	0.2	2	0.2	38.5	33.4

Table 18: Student demographics by cluster for 2014 district data

Cluster	American-Indian-Alaska-NativePer	Asian-Per	African-AmericanPer	Filipino-Per	Hispanic-Per	Hawaiian-PacIslanderPer	Two-More-Races-Per	None-Reported-Per	White-Per	GradsMtg-UCCSUPer
1	0.4	8.1	3.9	2.6	44.3	0.5	3.4	0.2	29.1	44
2	0.9	1.8	1.2	0.6	27.3	0.3	3.1	0.2	56	36.3
3	0.3	1.2	1.1	0.4	78.3	0.2	0.7	0.1	10.4	31.9

Table 19: 2015 teacher data at district level and student outcomes by cluster

Cluster	PerPupil-RatioTo-Teacher	Teaching-Days	Salary-Change	2Year-Teachers	Teacher-Highest-Salary-Offered-District	Teacher-Lowest-Salary-Offered	Avg-Years-Teaching	TeachersNum	Grads-Mtg-UCCSU-Per	SAT-AvgResults-Mathematic
1	22.4	180	4	27	90,039	44,070	12	601	34.2	451
2	22.6	180	3.5	20	90,881	44,105	11	467	53.4	545
3	14.5	180	3	1	74,075	38,651	10	24	18	0
4	21.6	180	3	7	84,032	41,714	11	151	33.4	498

Table 20: 2014 teacher data at district level and student outcomes by cluster

Cluster	PerPupil-RatioTo-Teacher	Teaching-Days	Salary-Change	2Year-Teachers	Teacher-Highest-Salary-Offered-District	Teacher-Lowest-Salary-Offered	Avg-Years-Teaching	TeachersNum	Grads-Mtg-UCCSU-Per	SAT-AvgResults-Mathematic
1	23.6	180	2.5	25	88,386	43,133	12	692	44	513
2	22.1	180	1.1	6	79,503	38,779	12	153	36.3	517
3	22.1	180	2.9	9	81,338	41,380	11	207	31.9	447

Other attributes: Investigating other attributes (for example, English language learner, foster care, free meal program) shows that the FreeReducedMeals and English Language Learners show statistically significant difference in graduate and student performance (Tables 21 and 22), based on regression model using clusters as groups at district level. However, similar analysis at the school level does not show FreeReducedMeals as an effective attribute. This difference is captured in district data and not at the school level.

Table 21: 2015 District data analysis meal programs and foster care and ELA

Clus-ter	Enroll-ment	English-Learners-Per	APExam-Graduating-Class-TestTakers	Free-Reduced-MealsPer	Average-Daily-Attendance	FRPM-ELFoster-UnduplD
1	13,469	26.7	690	77	12,099	82.2
2	10,223	10.4	767	28.8	9,400	32.6
3	339	5.7	0	65.5	280	70.4
4	3,130	18	145	56.3	2,825	59.3

Table 22: 2013–14 District data clustering results on meal program, foster care and ELA

Clus-ter	Enroll-ment	English-Learners-Per	APExam-GraduatingClass-TestTakers	Free-Reduced-MealsPer	Average-Daily-Attendance	FRPM-ELFoster-UnduplDistrict
1	14,575	15.2	913	50.6	13,796	56.1
2	3,357	9.3	161	43.6	2,557	46
3	4,480	30.6	165	79.9	4,184	82.9

The clustering section discussed analysis for school data and district data using the PAM algorithm. The following section takes the school clustering model a step further to develop a classification model, and the purpose and results are discussed.

5.4 Multiclass Classification Models

The cluster analysis can be further developed into a classification model. The number of clusters can be classes of schools. When the metrics change in the schools year by year, a classification model using the cluster number as the labels can be used to classify schools into

different clusters automatically. This will assist in automatically tracking the changes happening in the school profiles.

Two methods of classification are implemented and compared to see which methods give the optimal results.

5.4.1 Random Forests

The random forests (RF) method, described in Section 4.2.1, with multiclass classification model, was implemented for the school data. The existing data was divided into a training set consisting of 1,200 schools and a testing set of 436 schools. The results, shown in Table 23, show that as a classification model, RF performed extremely well, with an out-of-bag (OOB) error rate of 2.08%, which means that the model has an accuracy of 97.92%. The term "out-of-bag" comes from "bagging" (itself short for "*b*ootstrap *agg*regation"), a predecessor to the RF method. Each tree in bagging uses roughly two-thirds of random observations from the dataset, and the remaining one-third are "out of the bag" and used for prediction, from which the error rate is calculated (James et al., 2013, pp. 317–318). RF is more efficient than bagging, because bagging uses all p predictors rather than $m \approx \sqrt{p}$ as in RF (James et al., 2013, p. 320), but the error rate is still called the OOB error rate. The RF for our data constructed 500 trees, with each tree using at least three variables for splits.

Table 23 Classification results from Random Forests, and prediction error (OOB) rate estimate

```
Confusion matrix:
    1   2   3  4 class.error
1 315   0  10  0  0.03076923
2   1 280   0  2  0.01060071
3   8   0 513  1  0.01724138
4   2   0   1 67  0.04285714

  OOB estimate of  error rate: 2.08%


               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3
```

Variable importance to each class: Table 24 displays the degree of importance that each variable had in determining the class of all observations.

Table 24: Global variable importance based on an ensemble of 500 trees used for classification

	1	2	3	4
crit1	0.08171577	0.292916430	0.111727156	0.040636457
m_prof	0.06051559	0.036337580	0.099152131	0.013095791
m_pprof	0.01805971	0.005167715	0.019244789	0.009188163
MetAttendTarg	0.01372140	0.053312883	0.002793011	0.539580917
m_prate	0.01935891	0.002813436	0.031228918	0.018331714
m_enr	0.04781987	0.183087512	0.082133818	0.029845733
m_tst	0.08244825	0.150313281	0.123536738	0.032843118
m_val	0.10019747	0.175375658	0.168065251	0.035799922
grad_app	0.22950181	0.491398457	0.150826131	0.473344809

If a significant predictor is removed from the model, some observations will be incorrectly classified by the remaining model. The proportion of observations that will be misclassified when a given predictor is removed is known as the mean decrease in accuracy for that predictor. These values are graphed in Figure 39 and tabulated in Table 25. For example, by removing math valid scores (m_val), on average 0.14, or about one in seven, of the observations in the data set will be misclassified.

The Gini index is a measure of how purely the nodes in the tree represent the classes; it is small when most of the observations at each node belong to just one class (James et al., 2013, p.

312). The mean decrease in Gini indicates the degree of purity contributed to a class by a certain variable. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient. These values are also graphed in Figure 39 and tabulated in Table 25.

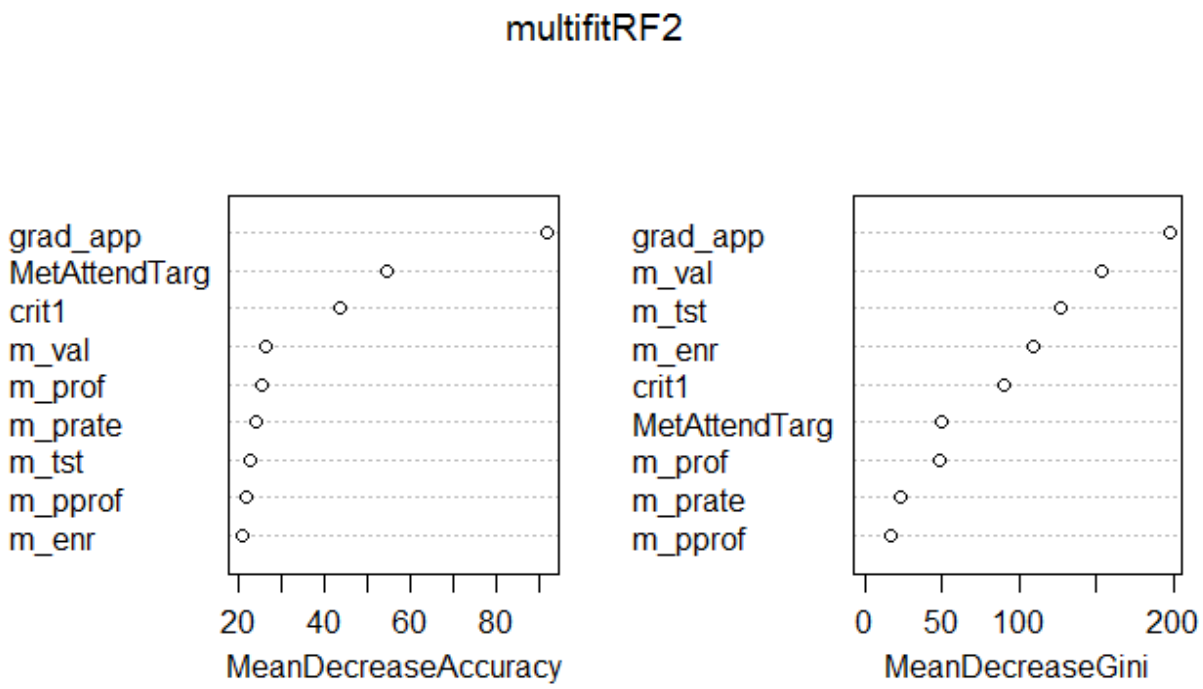


Figure 39: Graphs of mean decrease in accuracy and mean decrease in Gini index for each variable used in the model.

Table 25: Mean decrease in accuracy and mean decrease in Gini index for the variables used in the model

	MeanDecreaseAccuracy	MeanDecreaseGini
crit1	0.14204254	89.45832
m_prof	0.06900397	47.69411
m_pprof	0.01495739	16.81380
MetAttendTarg	0.04856027	49.73028
m_prate	0.02053097	22.16602
m_enr	0.09362631	109.04634
m_tst	0.11306822	126.81207
m_val	0.14367948	153.68037
grad_app	0.27069901	198.01052

Finally, Figure 40 graphs the observations on a two-dimensional plane, with the classes color coded. Most of the observations line up neatly along three axes, and the classes are mostly confined to specific segments of specific axes, indicating strong prediction power.

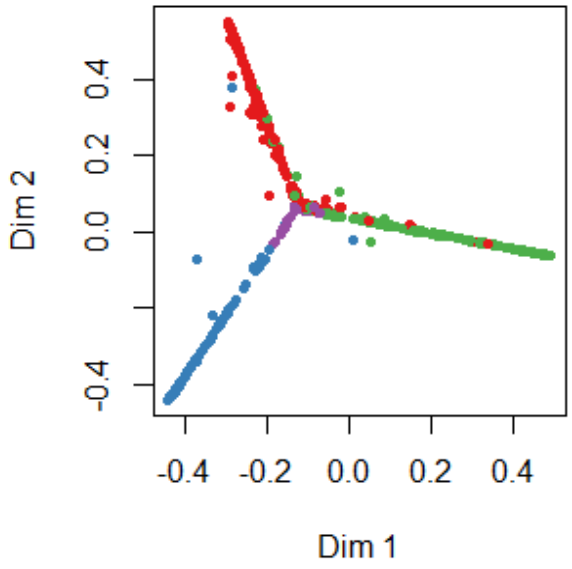


Figure 40: Plot indicates the separation of the four classes on a two-dimensional plane.

Meaning of the results from Random Forests: The results from Random Forests show that, as a classification model, it is highly successful in separating the observations and assigning them to their respective classes with a very low error rate. Thus, if any new school is added to the data, or any update happens to an existing school's parameters, reclustering of data is not needed. The RF-trained model will be able to classify the new or altered observation to the cluster it would now belong to with a 97.92% accuracy.

5.4.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA; not to be confused with the latent Dirichlet allocation described in Section 4.3.3) was also performed on the schools dataset. The results show that as a classifier it is also efficient for the given dataset, though not as strong as RF. Tables 26, 27, and 28 show the confusion matrix, total numbers of observations in each class, and the accuracy (diagonal entries) and misclassification rates (off-diagonal entries) for the 1,200 schools in the training dataset. It shows an average accuracy rate of 94.6%.

Table 26: Confusion matrix for test data set for linear discriminant analysis

	1	2	3	4
1	94	0	7	0
2	1	72	0	0
3	4	0	224	0
4	1	2	1	31

Table 27: Number of observations from the training set in each class

```
> multifitlda$counts
  1   2   3   4
325 283 522  70
```

Table 28: Accuracy and misclassification rates for each class assignment in LDA for test set

	1	2	3	4
1	0.93069307	0.00000000	0.06930693	0.00000000
2	0.01369863	0.98630137	0.00000000	0.00000000
3	0.01754386	0.00000000	0.98245614	0.00000000
4	0.02857143	0.05714286	0.02857143	0.88571429

Since there is a considerable (3.3-point) difference between LDA and Random Forests, this test shows that RF performs better as a classifier for this data. LDA performs best when the

data assumptions are linear and have a Gaussian distribution. RF, on the other hand, is distribution agnostic.

5.5. Frequent Patterns and Association Rules

The dataset used for frequent pattern analysis (described in Section 4.4) is school-level demographics, teacher education and teaching experience, foster care, graduation numbers, and percent meeting university requirements, and it is merged with the math performance attribute. A total of 51 attributes are used for the pattern analysis. The full list of attributes is given in Appendix 3. Data from years 2014 and 2015 has been applied for pattern analysis, and differences are observed to see whether any new patterns exist from each years. Some of the examples of the attributes are given in Table 29.

Table 29: Sample attributes at school level used for frequent pattern analysis

EnrollmentC
EnglishLearnersPerC
AmericanIndianPerC
AsianPerC
AfricaAmericanPerC
HispanicPerC
FluentEnglishProficientPerC
FosterYouthNumC
FreeReducedMealsPerC
CohortGraduatesPerC
PerPupilRatioTeacherC
1stYearTeachersC
2YearTeachersC
AvgYearsTeachingC (average years teaching experience)
TeachersFTEC (number of full-time teachers)
Grad_app
m_pprof (math percent proficient or above)

Data attributes are discretized and binned based on the mean, median, and quantile values. For example, the attribute Enrollment ranges from 2 to 4,814, with a median of 1,366. Thus, the values were binned into 10 categories to form the variable EnrollmentC. Similarly, the attribute AvgYearsTeaching has minimum of 0, maximum of 21, and median of 11. This attribute was binned into four categories as AvgYearsTeachingC.

Table 30 shows the topmost frequent patterns and association rules derived out of them:

Table 30: Most frequent patterns and association rules

Frequent Patterns and Associations
{FluentEnglishProficientPerC=4,PerPupilRatioTeacherC=2} ⇒ {CohortGraduatesPerC=4}
{EnrollmentC=4,PerPupilRatioTeacherC=2} ⇒ {CohortGraduatesPerC=4}
{EnrollmentC=4,X2YearTeachersC=2} ⇒ {CohortGraduatesPerC=4}
{FreeReducedMealsPerC=1} ⇒ {EnglishLearnersPerC=1}
{TeachersFTEC=3} ⇒ {CohortGraduatesPerC=4}
{FreeReducedMealsPerC=3} ⇒ {CohortGraduatesPerC=4}
{EnrollmentC=3,TeachersFTEC=2} ⇒ {CohortGraduatesPerC=4}
{FosterYouthNumC=3,PerPupilRatioTeacherC=2} ⇒ {AvgYearsTeachingC=1}
{EnrollmentC=4,AvgYearsTeachingC=1} ⇒ {TeachersFTEC=2}
{EnrollmentC=4,CohortGraduatesPerC=4} ⇒ {TeachersFTEC=2}
{EnglishLearnersPerC=2,TeachersFTEC=2} ⇒ {AvgYearsTeachingC=1}
{m_pprof=1, MetAttendTarg=3, AmericanIndianPerC=1, WhitePerC=1, FreeReducedMealsPerC=4, AvgYearsTeachingC=1} ⇒ {HispanicPerC=5}

Suppose we wish to learn what conditions lead to a school having the highest value of cohort graduation percentage. In the list of frequent patterns and association rules, we look for the CohortGraduatePerC at the highest level, 4, among the consequents, and we find several patterns in the antecedents. For example, the association rule "{FluentEnglishProficientPerC=4,PerPupilRatioTeacherC=2} ⇒ {CohortGraduatesPerC=4}" indicates that the cohort graduation percentage is at its highest level (above 75%) when the FluentEnglishProficientPercent is above 75% and pupil teacher ratio is around 20.

These patterns uniquely contribute to analysis where regression models fail to identify some interactions and key relationships between variables. The frequent patterns with multiple variables can be projected into a higher-dimensional space to form association rules. These rules give several insights to enhance the decision-making capacity for policy planning. The plot in Figure 40, for example, shows the support and confidence for eight orders of rules.

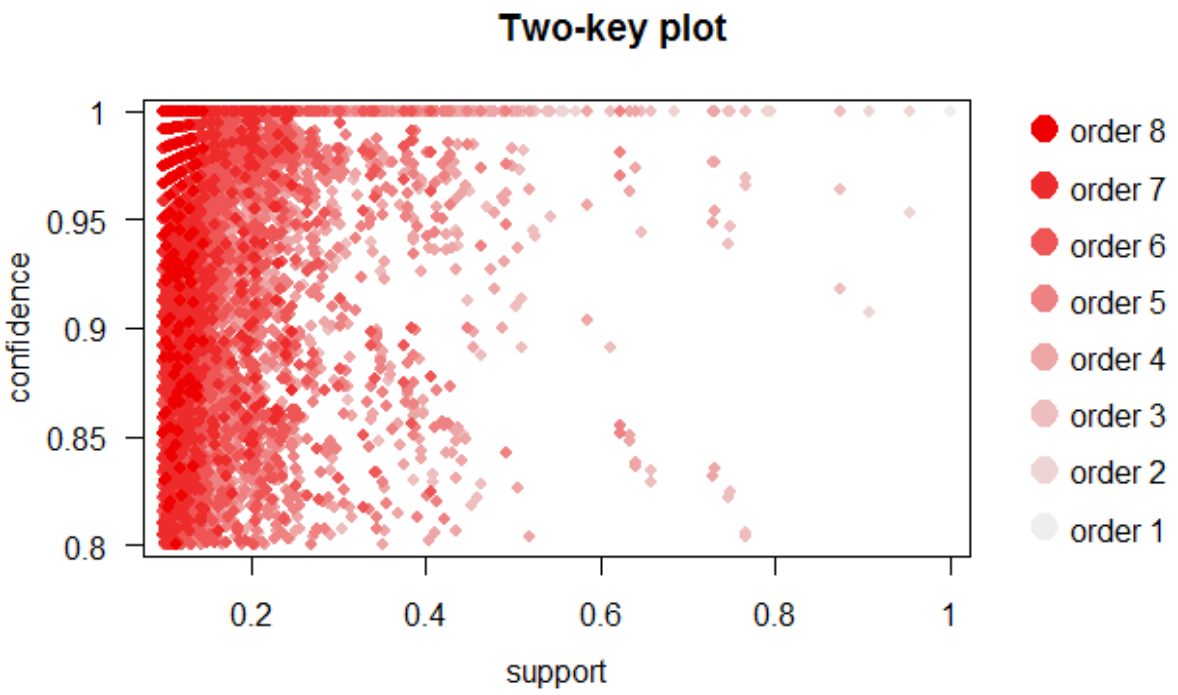


Figure 41: Support and confidence for multiple-order frequent patterns.

5.6 Regression and Results

Although unsupervised learning was able to group the schools clearly into clusters and assign a cluster number to each group, there are other attributes in the data set that can be combined with cluster number as an attribute to observe whether they can explain the variation

in math proficiency. The attributes in Table 31 were added onto the original schoolwide dataset for regression analysis.

Table 31: Additional attributes for regression analysis

Attribute	**Meaning**
FreeReducedMealProgram	Number of students in the free/reduced-price meal program
Foster	Number of students in foster care
Homeless	Number of students who are homeless
DirectCertification	Number of students who are direct certified
EnglishLearnerEL	Number of students who are learning English as a second language

The results given in Table 32 show that, whereas being in the free/reduced-price meal program does not have a statistically significant effect on the math proficiency, being in foster care or homeless or an English language learner or schools having direct certification programs had a direct and significant impact on math proficiency. The cluster number used as categorical attribute also explains the variance in math proficiency, as it is expected to if the clustering is distinct. The results indicate that, increase in each student into Foster care, reduced Math proficiency by -1.2, while each student increase as an English Language Learner lead to reduction of math proficiency in the school by 2.85 points. The impact on math proficiency is worse being homeless vs foster care while both had a negative effect.

Table 32: Significance of results of regression on external variables and cluster number

```
Analysis of Variance Table

Response: m_prof
                       Df  Sum Sq Mean Sq  F value    Pr(>F)
MetAttendTarg           1   28409   28409   4.5673  0.032748 *
FreeReducedMealProgram  1    2484    2484   0.3993  0.527535
Foster                  1  181038  181038  29.1053 7.945e-08 ***
Homeless                1  263721  263721  42.3982 1.010e-10 ***
DirectCertification     1   62683   62683  10.0775  0.001531 **
EnglishLearnerEL        1  199767  199767  32.1163 1.736e-08 ***
clusterNum              3 6778376 2259459 363.2510 < 2.2e-16 ***
Residuals            1513 9411017    6220
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

District Level Regression Analysis:

Regression analysis using generalized linear models (GLM) and Multivariate adaptive regression splines (MARS). Results in Appendix 5 show that certain attributes that do not highlight importance at school level show larger patterns of impact at district level. Free Reduced Meal Program though did not contribute to enhanced math proficiency rates at school level based on above results, at district level it did contribute positively towards increasing the number of cohort graduates. The MARS analysis takes the analysis deeper and captures the non linear relationships between the attributes and identifying the value thresholds where major changes are occurring in relation to the response attributes(s), as the scaled value results indicate below.

Table 33: MARS coefficients with hinge functions

```
  marsf14$coefficients
                                                      CohortGraduatesPer
(Intercept)                                                  -0.1499600
h(SATAvgResultsMathematic--0.651021)                          0.3087407
h(-0.651021-SATAvgResultsMathematic)                         -0.3071306
h(ClassifiedSalariesPerStudent-0.850356)                     -1.4124754
h(TotalGenFundExpendituresPerStudent-1.20877)                 1.0546839
h(2.09633-AfricanAmericanPer)                                 0.2152435
h(1.81244-GradsMtgUCCSUPer)                                  -0.2061950
h(LocalRevenuePerStudent--0.164014)                          -0.1608632
h(FederalRevenuePerStudent-0.666129)                         -0.1308797
h(-0.0783653-PerPupilRatioToTeacher)                         35.0906696
h(APExamGraduatingClassTestTakersNum--0.407696)              -0.1079499
h(-0.407696-APExamGraduatingClassTestTakersNum)             -39.3047838
h(-1.37847-WhitePer)                                        -20.3600720
```

5.7 Foundations for the Future: The Micro Analysis

The macro analysis tested and evaluated methods to group statewide schools into multiple clusters that had inherent differences. The regression step validated that the clusters treated as a categorical attribute, along with a new set of attributes can explain the math proficiency differences among schools. The next logical step is to start evaluating methods that can handle individual student data at the micro level. Since individual data in the schools is held very private, data has been simulated to reflect the real data in the school records that teachers maintain for each student. The micro analysis is conducted in the hopes of developing a collaborative recommender system to use the collective intelligence of teachers across the state in identifying the right method and medium of instruction to enhance a student's learning. Simultaneously, by keeping track of the types of patterns that emerge in tackling different learning issues for various students, a feedback mechanism can be established with the macro policy levels for optimal policy decisions.

The following steps are involved in developing the recommender system model:

1. Secure individual data (here, data is simulated). The data consists of student grade records, scores, and teacher text input for each trimester and a growth plan. Multiple data types exist, including numeric, categorical, and text data.
2. Conduct text mining on the teacher input sections of student records and convert the content into numerically coded categorical vectors that can be merged into the student data set. The text mining component uses *k*-means for benchmarking and hierarchical text clustering for similar documents. It is also compared with locality-sensitive hashing (LSH) for document similarity.
3. Build the similar student profile recommender system using the collaborative filtering model, where a teacher can find profiles of students similar to a target student for whom the teacher is interested in exploring various learning approaches that can suit for recommendation.
4. Keep a count and record of similar student profile groupings and the methods that lead to improvement in their learning, and provide these patterns back to macro levels.

Because these models are based on machine learning, they can easily recalibrate as new data is added to the system and give more frequent feedback.

5.7.1 Text Mining

Text Mining has been applied for this research for two scenarios:

1. Topic Modeling
2. Extract similar teacher descriptions of students to merge text and numeric data

Topic Modeling: The main interest driving topic modeling of parent, student and community stake holder opinions on the schools they belong to is to identify some insights that might be difficult to capture in official data collection by educational agencies. Latent Dirichlet Allocation has been applied to model the topics for over 300 reviews of stake holders for 20 schools randomly selected from all the clusters. Table 34 below shows an example output of analysis for Mountain View High School. The results indicate that 6 trending topics are most distinct and prominent in the reviews. The key words for each topic shows and overview of the topic that is being discussed. Overall the results show the reviews had a positive tone about all topics that trended from the nature of the 'high school programs', the 'clubs' in the school, class environment' and 'graduation' outcomes.

Table 34: Topic Modeling results for Mountain View High School, CA

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
[1,]	"good"	"area"	"ago"	"amaz"	"academ"	"graduat"
[2,]	"program"	"attend"	"area"	"backyard"	"campus"	"offer"
[3,]	"absolut"	"care"	"attent"	"fortun"	"care"	"person"
[4,]	"amaz"	"help"	"club"	"spartan"	"challeng"	"various"
[5,]	"anyth"	"overal"	"delight"	"three"	"class"	"achiev"
[6,]	"appreci"	"public"	"extracurricular"	"thrive"	"environ"	"activ"
[7,]	"appropri"	"solid"	"four"	"abl"	"excel"	"adulthood"

Below Figure 42 represents log likelihood of the model where it indicates that 6 topics are optimal to choose from the list of topics.

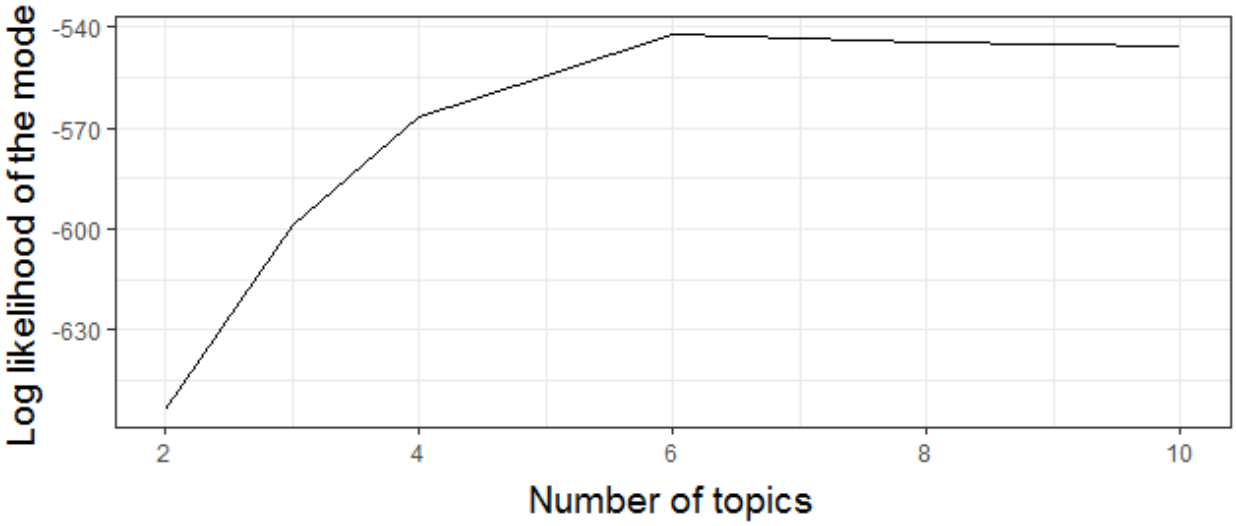


Figure 42: Log Likelihood of the Reviews from Mountain View High schools

Merging text and Numeric Data: One of the goals proposed for this research is to integrate text data with numeric and categorical data for analysis. Some 80% of the information is usually stored in text form, which is unstructured data. Extracting information from text for analysis can give useful insights that can be missed in numeric analysis. The main goal of text mining at the micro level in the present research context is to identify document similarity and dissimilarity of student feedback information provided by teachers for each student. In order to evaluate the most optimal approach for text mining, three approaches are tested.

1. *k*-means (this method is largely tested as a benchmarking algorithm for the other methods)
2. Hierarchical analysis of documents
3. Locality-sensitive hashing

In order to have the results presentable, a random sample of 20 documents is taken to illustrate graphs that are easy to comprehend.

***k*-means:** Since *k*-means is a benchmarking algorithm, no study feels complete without other algorithms compared to *k*-means results. The data in Table 33 shows that *k*-means clustered the data into four groups based on Euclidean distance. The top row in the table is the document number, and the lower row indicates the class assigned to each document. For example, documents 1 and 2 are classified into the first cluster.

Table 35: Grouping of observations by k-means

```
> kmeansResult$cluster
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1  1  3  3  2  4  3  3  4  4  4  2  2  2  2  2  3  3  2  2
```

Hierarchical clustering: Hierarchical clustering results also indicate how closely similar or dissimilar a document is to another based on the tree proximity of the nodes. Once these results are obtained, the group that the observations are assigned to is taken as a transformed qualitative attribute. The hierarchical approach classified the documents into five major groups based on the cut points chosen. Whereas one group dominates in the tree, another group has only two elements that are similar to each other. The clustering shown in Figure 42 is based on the UPGMA method discussed in Section 4.3.2.

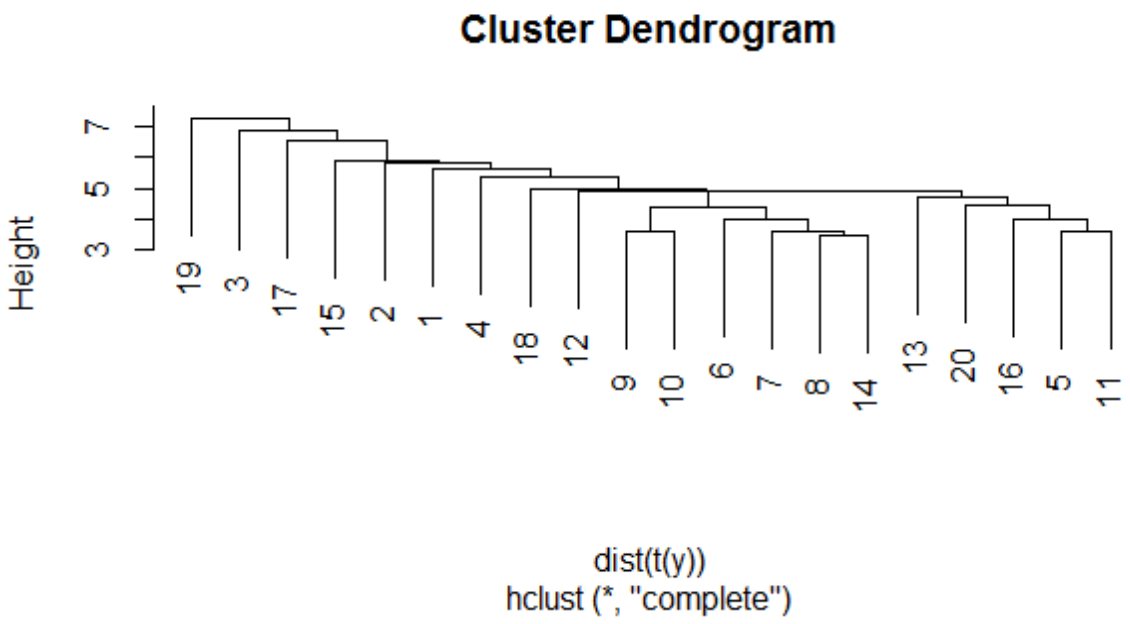


Figure 43: Hierarchical clustering produced by the UPGMA.

Locality-sensitive hashing: The results from locality-sensitive hashing, shown in Table 34, indicate that for document 1, documents 16, 5, and 8 are the closest related documents, and they fall in one category. Documents 14 and 8 also share similarity, but documents 14 and 1 do not have many similar features. Likewise, documents 10 and 9 are similar, but documents 6 and 10 are not, while documents 6 and 9 are. This shows that LSH is more sensitive in aligning the document relations. In Figure 42, hierarchical clustering also highlights the relation in the tree between 9 and 10, but not between 6 and 9; in the end, though, they get assigned to a single cluster. By applying LSH, it is easy to distinguish a step further. If too many documents are returned as one cluster, applying LSH would fine-tune it and precisely pinpoint which aspects they are being grouped under.

Table 36: Pairwise document similarities from locality-sensitive hashing for a set of 20 documents

```
# A tibble: 12 x 3
         a      b      score
     <chr>  <chr>      <dbl>
1   doc-1 doc-16 0.06060606
2   doc-1  doc-5 0.08823529
3   doc-1  doc-8 0.02941176
4  doc-10  doc-9 0.16666667
5  doc-14 doc-20 0.14814815
6  doc-14  doc-3 0.02173913
7  doc-14  doc-8 0.07692308
8  doc-16  doc-5 0.16666667
9  doc-20  doc-3 0.02127660
10 doc-20  doc-8 0.07407407
11  doc-3  doc-8 0.02272727
12  doc-6  doc-9 0.04347826
```

5.7.2 Collaborative Filtering

Collaborative filtering analysis was performed using two distance approaches: Jaccard distance and Pearson coefficient. Table 35 shows that Jaccard similarity identifies the five students most similar to a given student more closely, and it also indicates the degree of similarity between each pair of students. By identifying the five students that are most similar to each other, it becomes easier for a teacher to see what kind of learning methods followed by a particular student can be useful for other similar students. This enables monitoring what types of students will find a certain method effective. In turn this enables more efficiently assignment of economic and personnel resources while utilizing the collective knowledge of teachers across the state of California.

Table 37: Results of collaborative filtering based on Jaccard distance

	1	2	3	4
2	0.9000000			
3	0.8000000	0.8000000		
4	0.8500000	0.7619048	0.9444444	
5	0.8000000	0.8000000	0.8888889	0.8421053

The results in Table 35 based on Jaccard distance gives a list of five students who are most similar to Student 1, the student for whom we are trying to find similar students to match. Thus, in the results, Student 2 is closely related to Student 1, where the distance value is close to 1 (0.90). Student 3 is equally similar to Students 1 and 2 and much closer to Student 4 (0.94) retrieved in the list, and so on.

Table 36 shows the results from using the Pearson coefficient as the proximity measure; however, the Jaccard distance was able to retrieve students with more proximity.

Table 38: Results of collaborative filtering based on Pearson coefficient

	1	2	3	4
2	0.8761736			
3	0.7339804	0.7881582		
4	0.7347926	0.7313434	0.8353625	
5	0.7426042	0.7870584	0.7532487	0.7245103

Collaborative filtering not only pulls all the students who are similar to the input student for search; it also provides information on how similar each of these students are among themselves, thus giving a choice for the teacher on how many students' learning records can be accessed. A software interface can fetch the student record for the teacher while keeping identifying information private.

As the dataset grows, the same recommendation model can be used to see which learning methods work for which group of students, identify which cluster or clusters they belong to, and

forward the patterns to macro-level decision makers. The decision makers can divert the required resources (financial, social, infrastructural, and human) as needed.

For example, among the five students retrieved, it is shown that the common factor that four of the students retrieved had was enrolling in a math remedial program that has tutor assistance, and they showed learning improvement, so it is highly suggestive that the student they are all similar to also has the likelihood to succeed using the remedial program with a tutor assistance. Thus, for such a group of students, more tutoring resources can be provided rather than individualized computer-based programs. For a different set of students, a self-paced computer-based learning system could prove more useful than using a tutor.

CHAPTER 6

DISCUSSION AND FUTURE RESEARCH

The analytic frame is a complex system that can handle large datasets for analysis, and at each stage it feeds the outcome into the next level of analysis until it reaches the micro levels. Once at micro level, patterns in learning measures and outcomes among similar students are detected and sent as feedback to macro level for policy analysis. After evaluation of multiple algorithms for each step of analysis, the best-performing algorithms were selected. The final system is illustrated in Figure 43.

This leads to discussion on the following aspects.

1. Technical aspects
2. Connecting education policy at macro and micro levels
3. Economic and resources perspective
4. Ethics and privacy perspective

6.1 Technical Aspects

From a technical standpoint, implementing the analytic system over statewide high school data in all of California has been time consuming but successful. The results indicate that the system is able to process the data and give desired outputs successfully at each step. Each state of the framework is evaluated with multiple methods in order to decide which approach is most suitable for the data at hand. In clustering, PAM, along with the Gower distance, proved to perform the best for the dataset with multiple data types. Many clustering algorithms are unable to handle data that can include

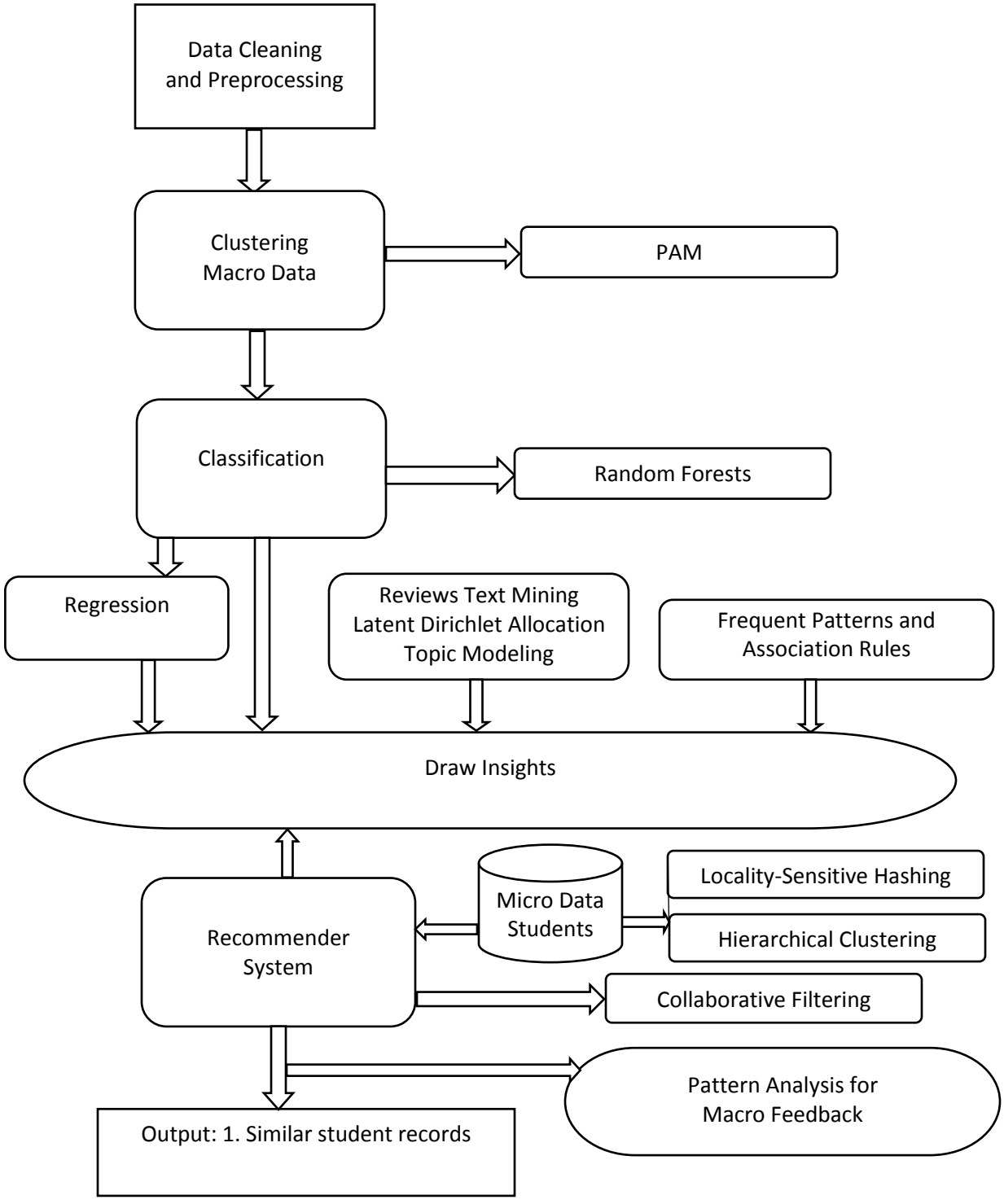


Figure 44: Final configuration of the analytic framework.

multiple data types for qualitative and quantitative items, and PAM is among the few that can perform very effectively.

In the classification stage, Random Forests (RF) proved to be a highly suitable approach compared to linear discriminant analysis, and RF is the choice for the analytic system. In the text mining stage to merge numeric and text data, *k*-means, hierarchical clustering, and locality-sensitive hashing (LSH) are compared. Both hierarchical clustering and LSH show promising results, with LSH being more accurate in computing similarity. Collaborative filtering becomes the highlight at the micro application level, where experience with students with common characteristics can give insights into the best learning methods for them.

The use of recommender systems where collaborative filtering (CF) is implemented is extremely sparse in the education domain, with only two studies mentioned in connection with the use of student responses to questions on an online learning system (Bergner et al., 2012) to predict the missing interactive model a certain student could have, based on similar students' response to the learning modules.

CF models are increasingly adopted in health systems. CF mines aggregated user behavior and information to identify intricate and unforeseen patterns that are difficult to identify by human analysis with a small set of data attributes. Recommendations generated from analyses of these CF-detected patterns have demonstrated significantly greater reliability than those made using more traditional demographic categories (Caplan & Rosenthal, 2013). Caplan and Rosenthal state that the core idea behind applying CF to clinical decision making is to make decisions about a patient based on historical data derived from multiple "similar" patients presenting multiple "similar" cases. Strecher (2007) states tha "collaborative filtering in the health area could match the coping strategies, medical decisions, and preferences of similar

others with specific needs and interests of the user." Education domains precisely reflect the same scenario for students. Instead of medical decisions, decisions are made about learning strategies and which strategy or methods certain students can respond to and certain students cannot. This enables decision makers from the highest levels to teacher levels to observe the correct feedback necessary to enhance each student's learning.

Currently a system of the nature just described for education is nonexistent. Hindering these applications are challenges unique to higher education. First, the domain lacks the high-capacity computational infrastructure, financial budgets, and human resources required for effective collection, cleaning, analysis, and distribution of large datasets. From a technical standpoint, having the required technical human resources makes a significant difference in the implementation of large-scale systems. This remains one of the biggest challenges across the nation; much of the expertise required for policy analysis digitization is not readily available.

6.2 Connecting the Micro and Macro Levels

One of the most important themes of this dissertation is connecting the micro and macro levels. The framework is able to demonstrate that it is possible, while integrating multiple data types. The outcome from classification and regression directly feeds into the micro levels by identifying the cluster of the students. Many learning analytics methods have emerged and formed into a subdomain of learning analytics (Cope & Kalantzis, 2016). However, lack of systems that can integrate the macro levels with the individual learning methods has resulted in isolated development of these methods. The results have shown that by integrating the micro levels with collaborative models and sending feedback to the macro level, untapped modes of

improving student learning can be introduced that can significantly benefit large-scale school systems, both financially and from a human resource perspective.

Although the system studied here was intended to explore the technical mechanisms to tackle the problem of establishing a link between macro and micro contexts, it can also have growing capability of integrating with a larger global student system, where teachers can interact and identify various learning methods. As Dhillon (2015) discusses, as education is geared up from local economies to a larger stage of global economies, big data becomes ever more important in connecting various elements of knowledge, skills, and emerging outcomes of global movement of human talent. This process is already visible in higher education institutions in the United States, where education institutions are adopting a service model to open up to large student cohorts of international students. This affects not only how economic policy decisions are made for higher education institutions but also local students' access to education. Big data can be extremely helpful in analyzing various components of this process so that local resentment can be converted to collaboration and all students can benefit from exposure to global trends.

6.3 Ethics and Privacy

The advent of big data has triggered a radical shift in how research is perceived. Commenting on computational social science, Lazer et al. (2009) argue that it offers "the capacity to collect and analyze data with an unprecedented breadth and depth and scale." Moretti (2007) adds another dimension while referring to analysis of texts as a profound change at the levels of epistemology and ethics. Big data reframes key questions about the constitution of

knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality (boyd & Crawford, 2014).

Similar to may new technological advances, big data also has inspired both utopian and dystopian schools of thought. On one talon, it enables efficient solutions and insights for diverse aspects of life on earth, such as medical research, fighting terrorism, predicting environmental disasters, enhancing the supply chain and delivery of customer goods, and recommending user commodities. On the other talon, however, the emergence of big data is viewed as a menacing manifestation of Big Brother–style invasions of privacy, governmental overwatch, increased corporate control on individuals, and manipulations of individuals by profit-seeking entities. Heavy ethical questions arise in terms of who would have access to a student's data, who is responsible for securing it, and where to draw the line on utilizing it for various solutions. Although the benefits far outweigh the concerns, the issue of privacy and ethics often becomes debatable.

It is inevitable that big data will make inroads into every sphere of education, and often what is essential is to have an official mandate on privacy for student data and mechanisms for not allowing it to be used for permanent profiling of students. Many efforts are in the lead in creating ethical standards in education for big data. The *Journal of Learning Analytics* has dedicated a special issue to "Ethics and Privacy in Learning Analytics" (Gasevic, Dawson, & Jovanovic, 2016).

The failure of InBloom offers numerous insights and tends caution on implementing systems that maintain student databases with detailed information. Nonetheless, student performance data is maintained in the schools, and students are routinely categorized for learning levels in the high schools. InBloom's aggressive approach, in an environment where privacy

laws and policies have not yet emerged, and its top-down model of the system with expectation of rapid adoption in a slow-moving customer environment, sealed its fate. As technology systems are increasingly adopted, models that ensure fruitful outcomes have to be introduced organically in education policy settings as they deal with private data of students that can be misused and excessively profiled. The focus on how the data can benefit in improving learning outcomes of students at a large scale is essential. Offering these solutions in a state like California that shows more openness to adopting newer technologies can demonstrate the pros and cons that can come out of such projects and establish a mechanism to address the concerns of the community stakeholders.

In the research implemented here, to avoid violations of privacy, the individual-level students' identifying information is anonymized, and double-blind methods were followed in this analytic system framework, in which it is not possible for users in the system to identify exactly who the other students are in the state. The teachers have avenues to connect to the teacher of the other students but cannot identify individual students.

CHAPTER 7

CONCLUSION

The research presented here attempts to solve the problem of connecting macro and micro aspects of policy planning using big data. This problem has not been tackled before in education policy planning. The solutions for the framework system presented here have been thoroughly analyzed so that the most effective algorithm is implemented for handling the nature of the data involved. While traditional statistical models excel in finding insights on experimental data, big data analytics brings to the table its efficacy in identifying patterns in enormous amounts of data.

The clustering models have proven effectiveness in grouping a large set of data based on multiple parameters, thereby identifying aspects that were not previously noticeable. Although the PAM method has been used in categorizing learners, it has not been used at a scale for statewide schools on performance indicators combined with categorical attributes. This study demonstrated that it is possible to combine multiple data types and aspects to implement clustering for clearer insights.

The classification model also adds the classic machine learning advantage of automatically realigning the school categories based on changes in any of the indicators used for creating the models, and it consistently keeps the data updated, so that policy makers have constant access to the latest information for effective decision making rather than information that has been gathered in previous years.

The regression model enables policy makers to identify external or new variables that become visible to see their impact on schools and student performance factors. It ties the

traditional analytics approach and the big data analytics approach together, making the best use of both models.

The frequent patterns and association rules identify hidden patterns that exist among the attributes that belong to multiple areas of education (student performance, teaching, graduation, demographics, etc.). The association rules that are formed prove to be effective for explore the aspects at a deeper level to understand the nature of the dynamics between the attributes.

The recommendation system model proposed is a unique contribution and outcome of this research, where it will become possible to utilize the collective knowledge of teachers effectively for enhancing the learning of students across the state of California if the system is implemented. This leads to more exchange of ideas between the teachers and access to multiple points of resource allocation methods for macro policy decision makers, by identifying the patterns.

7.1 Future Steps

This research shows that beyond the potential to enhance student outcomes through just-in-time, diagnostic data that is essential for learning and instruction, the evolution of higher education practice overall could be significantly enhanced through data-intensive research and analysis. A worthy next step would be to improve our capacity to process and understand today's increasingly large, heterogeneous, noisy, and rich datasets rapidly at micro levels and link the results to the macro levels for effective policy planning.

This system is an example that can tackle large-scale data and scale it effortlessly. Implementing elaborate systems at a statewide level would require time, infrastructure, and

human resources. The next step is to approach the State Education department to allow for a prototype implementation of the system for testing purposes.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia (Eds.), *Proceedings of the 1993 ACM-SIGMOD international conference on management of data* (pp. 207–216). New York, NY: ACM.

Anjewierden, A., Kollöffel, B., & Hulshof, C. (2007). Towards educational data mining: using data mining methods for automated chat analysis to understand and support inquiry learning processes. In C. Romero, M. Pechenizkiy, T. Calders, & S. R. Viola (Eds.), *Proceedings of the international workshop on applying data mining in e-learning* (pp. 23–32). Worcester, MA: International Educational Data Mining Society.

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: a review and future vision. *Journal of Educational Data Mining, 1*(1), 1–15.

Berendt, B., & Preibusch, S. (2014). Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence and Law, 22*(2), 175–209.

Bergner, Y., Dröschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-based collaborative filtering analysis of student response data: machine-learning item response theory. In K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the fifth international conference on educational data mining* (pp. 95–102). Worcester, MA: International Educational Data Mining Society.

Bhattacharyya, D. K., &Hazarika, S. M. (2006). *Networks, data mining, and artificial intelligence: trends and future directions.* New Delhi, India: Narosa Publishing House.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

boyd, d., & Crawford, K. (2014). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Campagni, R., Merlini, D., & Sprugnoli, R. (2012). *Sequential patterns analysis in a student database.* Presented at the workshop on Mining and Exploiting Interpretable Local Patterns (I-Pat) at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Bristol, UK, 24–28 September. Retrieved from http://local.disia.unifi.it/merlini/papers/IPat2012.pdf

Caplan, E., & Rosenthal, N. (2013). Collaborative filtering: an interim approach to identifying clinical doppelgängers. *Health Affairs* Blog. Retrieved from http://healthaffairs.org/blog/2013/06/17/collaborative-filtering-an-interim-approach-to-identifying-clinical-doppelgangers/

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In L. C. Jain, R. A. Tedman, & D. K. Tedman (Eds.), *Studies in computational intelligence: Vol. 62. Evolution of teaching and learning paradigms in intelligent environment* (pp. 183–221). Berlin and Heidelberg, Germany: Springer-Verlag.

CDE: California Department of Education (2015). http://www.cde.ca.gov/

Cohen, A., & Nachmias, R. (2011). What can instructors and policy makers learn about Web-supported learning through Web-usage mining. *The Internet and Higher Education 14*(2), 67–76.

Cope, W., & Kalantzis, M. (2016). Big data comes to school: implications for learning, assessment, and research. *AERA Open, 2*(2), 1–19. http://journals.sagepub.com/doi/pdf/10.1177/2332858416641907

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning.* Hoboken, NJ: Wiley-IEEE.

Dhillon, P.A. (2015). *International organizations and education.* Education Policy, Organization, and Leadership. University of Illinois, Urbana-Champaign.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Hoboken, NJ: Wiley.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, U. M. Fayyad (Eds.), *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96)* (pp. 226–231). Palo Alto, CA: AAAI Press.

Evgeniou, T., Gaba, V., & Niessing, J. (2013). Does bigger data lead to better decisions? *Harvard Business Review*. https://hbr.org/2013/10/does-bigger-data-lead-to-better-decisions

Famili, A., Shen, W.-M., Weber, R., & Simoudis, E. (1997). Data preprocessing and intelligent data analysis. *Intelligent Data Analysis, 1*(1), 3–23.

Foster, D. P., Liberman, M., & Stine, R. A. (2013). *Featurizing text: converting text into predictors for regression analysis.* Philadelphia: University of Pennsylvania.

Gasevic, D., Dawson, S., & Jovanovic, J. (2016). Ethics and privacy as enablers of learning analytics. *Journal of Learning Analytics, 3*(1), 1–4.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics 27*, 857–874.

Guan, J., Nunez, W., & Welsh, J. (2002). Institutional strategy and information support: the role of data warehousing in higher education. *Campus-Wide Information Systems, 19*(5), 168–174.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Han, J., Chiang, J. Y., Chee, S., Chen, J., Chen, Q., Cheng, S., et al. (1997). DBMiner: a system for data mining in relational databases and data warehouses.

Hershkovitz, A., & Nachmias, R. (2009a). Consistency of students' pace in online learning. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Educational data mining 2009: 2nd international conference on educational data mining, proceedings. Córdoba, Spain. July 1–3, 2009* (pp. 71–80). Worcester, MA: International Educational Data Mining Society.

Hershkovitz, A., & Nachmias, R. (2009b). Learning about online learning processes and students' motivation through Web usage mining. *Interdisciplinary Journal of Education Learning and Learning Objects, 5*(1), 197–214.

Hill, T., & Lewicki, P. (2006). *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. Tulsa, OK: StatSoft.

Huang, Y.-M., Chen, J.-N., & Cheng, S.-C. (2007). A method of cross-level frequent pattern mining for web-based instruction. *Educational Technology & Society*, *10*(3), 305–319..

IBM. (2011a). *Mobile County Public Schools: Smarter Planet Leadership Series. Analytical insights help keep students on track.* Retrieved from http://www-03.ibm.com/software/businesscasestudies/th/en/corp?synkey=X421855L73231S22

IBM. (2011b). *Hamilton County Department of Education: Deeper student insights leave a deep impact.* Smarter Planet Leadership Series.

IBM. (2013). *IBM and Georgia's largest school system bring personalized learning to life.* Retrieved from https://www-03.ibm.com/press/us/en/pressrelease/42759.wss

IBM. (n.d.). *SPSS Modeler.* Retrieved from http://www-03.ibm.com/software/products/en/spss-modeler

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data.* Englewood Cliffs, NJ: Prentice Hall.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R.* New York, NY: Springer Science+Business Media.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical data analysis based on the L_1 norm and related methods* (1st ed., pp. 405–416). Amsterdam, The Netherlands: North-Holland.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis.* Hoboken, NJ: Wiley.

Karegar, M., Isazadeh, A., Fartash, F., Saderi, T., & Navin, A. H. (2008). Data-mining by probability-based patterns. In V. Luzar-Stiffler, V. H. Dobric, & Z. Bekic (Eds.), *Proceedings of the 30th international conference on information technology interfaces* (pp. 353–360). Zagreb, Croatia: SRCE University Computing Centre.

Keshtkar, F., Morgan, B., & Graesser., A. (2012). Automated detection of mentors and players in an educational game. In Proceedings of the 5th international conference on educational data mining (pp. 212–213).

King, B. 1967. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association, 62*(317), 86–101.

Koedinger, K. R., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. In R. S. J. d. Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining 2008: the first international conference on educational data mining, proceedings. Montréal, Québec, Canada, June 20–21, 2008* (pp. 157–166). Worcester, MA: International Educational Data Mining Society.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., et al. (2009). Social science. Computational social science. *Science, 323*(5915), 721–723

Leventhal, B. (2010). An introduction to data mining and other techniques for advanced analytics. *Journal of Direct, Data and Digital Marketing Practice, 12*(2), 137–153. 1210.1057/dddmp.2010.35

Luan, J. (2002). *Data mining and knowledge management in higher education—potential applications.* Paper presented at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada. http://eric.ed.gov/ERICWebPortal/detail?accno=ED474143

Machine Learning Group at the University of Waikato. (n.d.). *Weka 3: data mining software in Java.* Retrieved from http://www.cs.waikato.ac.nz/~ml/weka/index.html

Mazza, R., & Dimitrova, V. (2004). Visualising student tracking data to support web-based distance education. In *Proceedings of the 13th international World Wide Web conference*

(pp. 154–161). New York, NY: Association for Computing Machinery. Retrieved from http://www.iw3c2.org/WWW2004/docs/2p154.pdf

McCarthy, P. M., & Boonthum-Denecke, C. (2011). *Applied natural language processing: identification, investigation and resolution.* Hershey, PA: International Science Reference.

Merceron, A., & Yacef, K. (2005a). TADA-Ed for educational data mining. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning 7*(1), 1–16.

Merceron, A., & Yacef, K. (2005b). Educational data mining: a case study. In C.-K. Lool, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *International conference on artificial intelligence in education: supporting learning through intelligent and socially informed technology* (pp 467–474). Amsterdam, The Netherlands: IOS Press.

Mitra, S., & Acharya, T. (2003). *Data mining: multimedia, soft computing, and bioinformatics.* Hoboken, NJ: John Wiley & Sons, Inc.

Moreno-Jimenez, J., Cardeñosa, J., Gallardo, C., & Villa-Moreno, M. Á. (2014). A new e-learning tool for cognitive democracies in the Knowledge Society. Computers in Human Behavior, 30, 409-418.

Moretti, F. (2007). *Graphs, maps, trees: abstract models for a literary history.* London, UK: Verso.

Nettleton, D., Fandiño, V., Witty, M., & Vilajosana, E. (2000). Using a data mining work bench for micro and macro economic modelling. In N. Ebecken & C. A. Brebbia (Eds.), *Data Mining II* (pp. 25–34). Southampton, UK: WIT Press/Computational Mechanics.

Peña-Ayala, A. (2013). *Educational data mining: application and trends.* Springer International.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications 41* (4, pt. 1), 1432–1462.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137. doi: 10.1108/eb046814

Rajaraman, A., Leskovec, J., & Ullman, J. D. (2014). *Mining of massive datasets* (2nd Ed.). Cambridge, UK: Cambridge University Press.

RapidMiner, Inc. (2017). *RapidMiner.* Retrieved from http://www.rapidminer.com

Romero, C., Ventura, S., & De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modeling and User-Adapted Interaction 14*(5), 425–464.

Romero, C., Ventura, S., Zafra, A., & De Bra, P. (2010). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers & Education 53*(3), 828–840.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. d. (2011). *Handbook of educational data mining.* Boca Raton, FL: CRC Press.

RStudio. (2016). *RStudio - Open source and enterprise-ready professional software for R.* Retrieved from https://www.rstudio.com

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Boston, MA: Addison-Wesley.

Solka, J. (2008). Text data mining: theory and methods. *Statistics Surveys, 2,* 94–112.

Strecher, V. (2007). Internet methods for delivering behavioral and health-related interventions (eHealth). *Annual Review in Clinical Psychology, 3,* 53–76.

Strunk, W. Jr. (2007). *The elements of style.* Minneapolis, MN: Filiquarian Publishing, LLC.

Ubels, J., van Klinken, R., & Visser, H. (2010). The micro–macro gap. Bridging the micro–macro gap: gaining capacity by connecting levels of development action. In J. Ubels, N.-A. Acquaye-Baddoo, & A. Fowler (Eds.), *Capacity development in practice* (pp. 167–179). London, UK: Earthscan Ltd.

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., et al. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *User Modeling and User-Adapted Interaction, 21*(1–2), 217–248.

Vialardi-Sacin, C., Bravo-Agapito, J. Shafti, L., & Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Educational data mining 2009: 2nd international conference on educational data mining, proceedings. Córdoba, Spain. July 1–3, 2009* (pp. 190–199). Worcester, MA: International Educational Data Mining Society.

Vlahos, G. E., Ferratt, T. W., & Knoepfle, G. (2004). The use of computer-based information systems by German managers to support decision making. *Journal of Information & Management, 41*(6), 763–779.

Wang, Y.-h., & Liao, H.-C. (2011). Data mining for adaptive learning in a TESL-based e-learning system. *Expert Systems with Applications, 38*(6), 6480–6485. DOI:10.1016/j.eswa.2010.11.098

Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236–244.

Wicham, H. (2014). Tidy data. *Journal of Statistical Software, 59*. doi:10.18637/jss.v059.i10

Williams, T., & Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction, 43,* 23-29. DOI: 10.1016/j.autcon.2014.02.014

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Zhao, Y., Karypis, G., & Fayyad, U. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery, 10*(2), 141–168.

APPENDIX A

ADEQUATE YEARLY PROGRESS: TABLE OF ATTRIBUTES

Field #	Field Name	Type	Width	Description
1	cds	Character	14	County/District/School code
2	rtype	Character	1	Record type: D=district, S=school, X=state
3	type	Character	1	Type: 1=unified, 2=elementary district, 3=9–12 high district, 4=7–12 high district, E=elementary school, M=middle school, H=high school
4	sped	Character	1	A=Alternative Schools Accountability Model (ASAM), E=Special Education, C=Combination ASAM and Special Education
5	charter	Character	1	Y=charter, not direct funded, D=direct funded charter, blank=not a charter school
6	sname	Character	50	School name
7	dname	Character	50	District name
8	cname	Character	15	County name
9	met_all	Character	4	Yes = met all 2015 AYP criteria, No = did not meet all 2015 AYP criteria (note: the criteria includes only the participation rate and additional indicators)
10	crit1	Character	2	Number of AYP criteria met based only on participation rate and additional indicators
11	crit2	Character	2	Number of AYP criteria possible
12	e_enr	Character	7	Schoolwide or LEA-wide ELA enrollment
13	e_tst	Character	7	Schoolwide or LEA-wide ELA number of students tested
14	e_prate	Character	5	Schoolwide or LEA-wide ELA participation rate
15	e_pr_met	Character	4	Schoolwide or LEA-wide ELA participation rate met
16	m_enr	Character	7	Schoolwide or LEA-wide math enrollment
17	m_tst	Character	7	Schoolwide or LEA-wide math number of students tested
18	m_prate	Character	5	Schoolwide or LEA-wide math participation rate

Field #	Field Name	Type	Width	Description
19	m_pr_met	Character	4	Schoolwide or LEA-wide math participation rate met
20	ee_aa	Character	7	ELA enrollment, black or African American
21	et_aa	Character	7	ELA tested, black or African American
22	ep_aa	Character	5	ELA participation rate, black or African American
23	epm_aa	Character	4	ELA participation rate, met black or African American
24	me_aa	Character	7	Math enrollment, black or African American
25	mt_aa	Character	7	Math tested, black or African American
26	mp_aa	Character	5	Math participation rate, black or African American
27	mpm_aa	Character	4	Math participation rate met, black or African American
28	ee_ai	Character	7	ELA enrollment, American Indian or Alaska Native
29	et_ai	Character	7	ELA tested, American Indian or Alaska Native
30	ep_ai	Character	5	ELA participation rate, American Indian or Alaska Native
31	epm_ai	Character	4	ELA participation rate met, American Indian or Alaska Native
32	me_ai	Character	7	Math enrollment, American Indian or Alaska Native
33	mt_ai	Character	7	Math tested, American Indian or Alaska Native
34	mp_ai	Character	5	Math participation rate, American Indian or Alaska Native
35	mpm_ai	Character	4	Math participation rate met, American Indian or Alaska Native
36	ee_as	Character	7	ELA enrollment, Asian
37	et_as	Character	7	ELA tested, Asian
38	ep_as	Character	5	ELA participation rate, Asian
39	epm_as	Character	4	ELA participation rate met, Asian
40	me_as	Character	7	Math enrollment, Asian
41	mt_as	Character	7	Math tested, Asian
42	mp_as	Character	5	Math participation rate, Asian
43	mpm_as	Character	4	Math participation rate met, Asian
44	ee_fi	Character	7	ELA enrollment, Filipino
45	et_fi	Character	7	ELA tested, Filipino
46	ep_fi	Character	5	ELA participation rate, Filipino

Field #	Field Name	Type	Width	Description
47	epm_fi	Character	4	ELA participation rate met, Filipino
48	me_fi	Character	7	Math enrollment, Filipino
49	mt_fi	Character	7	Math tested, Filipino
50	mp_fi	Character	5	Math participation rate, Filipino
51	mpm_fi	Character	4	Math participation rate met, Filipino
52	ee_hi	Character	7	ELA enrollment, Hispanic or Latino
53	et_hi	Character	7	ELA tested, Hispanic or Latino
54	ep_hi	Character	5	ELA participation rate, Hispanic or Latino
55	epm_hi	Character	4	ELA participation rate met, Hispanic or Latino
56	me_hi	Character	7	Math enrollment, Hispanic or Latino
57	mt_hi	Character	7	Math tested, Hispanic or Latino
58	mp_hi	Character	5	Math participation rate, Hispanic or Latino
59	mpm_hi	Character	4	Math participation rate met, Hispanic or Latino
60	ee_pi	Character	7	ELA enrollment, Native Hawaiian or Pacific Islander
61	et_pi	Character	7	ELA tested, Native Hawaiian or Pacific Islander
62	ep_pi	Character	5	ELA participation rate, Native Hawaiian or Pacific Islander
63	epm_pi	Character	4	ELA participation rate met, Native Hawaiian or Pacific Islander
64	me_pi	Character	7	Math enrollment, Native Hawaiian or Pacific Islander
65	mt_pi	Character	7	Math tested, Native Hawaiian or Pacific Islander
66	mp_pi	Character	5	Math participation rate, Native Hawaiian or Pacific Islander
67	mpm_pi	Character	4	Math participation rate met, Native Hawaiian or Pacific Islander
68	ee_wh	Character	7	ELA enrollment, white
69	et_wh	Character	7	ELA tested, white
70	ep_wh	Character	5	ELA participation rate, white
71	epm_wh	Character	4	ELA participation rate met, white
72	me_wh	Character	7	Math enrollment, white
73	mt_wh	Character	7	Math tested, white
74	mp_wh	Character	5	Math participation rate, white
75	mpm_wh	Character	4	Math participation rate met, white
76	ee_mr	Character	7	ELA enrollment, two or more races
77	et_mr	Character	7	ELA tested, two or more races

Field #	Field Name	Type	Width	Description
78	ep_mr	Character	5	ELA participation rate, two or more races
79	epm_mr	Character	4	ELA participation rate met, two or more races
80	me_mr	Character	7	Math enrollment, two or more races
81	mt_mr	Character	7	Math tested, two or more races
82	mp_mr	Character	5	Math participation rate two or more races
83	mpm_mr	Character	4	Math participation rate met, two or more races
84	ee_sd	Character	7	ELA enrollment, socioeconomic disadvantaged
85	et_sd	Character	7	ELA tested, socioeconomic disadvantaged
86	ep_sd	Character	5	ELA participation rate, socioeconomic disadvantaged
87	epm_sd	Character	4	ELA participation rate, met socioeconomic disadvantaged
88	me_sd	Character	7	Math enrollment, socioeconomic disadvantaged
89	mt_sd	Character	7	Math tested, socioeconomic disadvantaged
90	mp_sd	Character	5	Math participation rate, socioeconomic disadvantaged
91	mpm_sd	Character	4	Math participation rate met, socioeconomic disadvantaged
92	ee_el	Character	7	ELA enrollment, English learner
93	et_el	Character	7	ELA tested, English learner
94	ep_el	Character	5	ELA participation rate, English learner
95	epm_el	Character	4	ELA participation rate met, English learner
96	me_el	Character	7	Math enrollment, English learner
97	mt_el	Character	7	Math tested, English learner
98	mp_el	Character	5	Math participation rate, English learner
99	mpm_el	Character	4	Math participation rate met, English learner
100	ee_di	Character	7	ELA enrollment, students with disabilities
101	et_di	Character	7	ELA tested, students with disabilities
102	ep_di	Character	5	ELA participation rate, students with disabilities
103	epm_di	Character	4	ELA participation rate met, students with disabilities
104	me_di	Character	7	Math enrollment, students with disabilities
105	mt_di	Character	7	Math tested, students with disabilities

Field #	Field Name	Type	Width	Description
106	mp_di	Character	5	Math participation rate, students with disabilities
107	mpm_di	Character	4	Math participation rate met, students with disabilities
108	e_val	Character	7	Schoolwide or LEA-wide ELA number of valid scores
109	e_prof	Character	7	Schoolwide or LEA-wide ELA number of students scoring proficient or above
110	e_pprof	Character	5	Schoolwide or LEA-wide ELA percent of students scoring proficient or above
111	e_ppm	Character	4	Schoolwide or LEA-wide ELA percent proficient or above met (not applicable for 2015 AYP)
112	m_val	Character	7	Schoolwide or LEA-wide math valid scores
113	m_prof	Character	7	Schoolwide math number of students scoring proficient or above
114	m_pprof	Character	5	Schoolwide math percent of students scoring proficient or above
115	m_ppm	Character	4	Schoolwide math percent proficient or above met (not applicable for 2015 AYP)
116	ev_aa	Character	7	ELA valid scores, black or African American
117	enp_aa	Character	7	ELA number of students scoring proficient or above, black or African American
118	epp_aa	Character	5	ELA percent proficient or above, black or African American
119	eppm_aa	Character	4	ELA percent proficient or above met, black or African American (not applicable for 2015 AYP)
120	mv_aa	Character	7	Math valid scores, black or African American
121	mnp_aa	Character	7	Math students scoring proficient or above, black or African American
122	mpp_aa	Character	5	Math percent of students scoring proficient or above, black or African American
123	mppm_aa	Character	4	Math percent proficient or above met, black or African American (not applicable for 2015 AYP)
124	ev_ai	Character	7	ELA valid scores, American Indian or Alaska Native

Field #	Field Name	Type	Width	Description
125	enp_ai	Character	7	ELA number students scoring proficient or above, American Indian or Alaska Native
126	epp_ai	Character	5	ELA percent of students scoring proficient or above, American Indian or Alaska Native
127	eppm_ai	Character	4	ELA percent proficient or above met, American Indian or Alaska Native (not applicable for 2015 AYP)
128	mv_ai	Character	7	Math valid scores, American Indian or Alaska Native
129	mnp_ai	Character	7	Math students scoring proficient or above, American Indian or Alaska Native
130	mpp_ai	Character	5	Math percent of students scoring proficient or above, American Indian or Alaska Native
131	mppm_ai	Character	4	Math percent proficient or above met, American Indian or Alaska Native (not applicable for 2015 AYP)
132	ev_as	Character	7	ELA valid scores, Asian
133	enp_as	Character	7	ELA number of students scoring proficient or above, Asian
134	epp_as	Character	5	ELA percent of students scoring proficient or above, Asian
135	eppm_as	Character	4	ELA percent proficient or above met, Asian (not applicable for 2015 AYP)
136	mv_as	Character	7	Math valid scores, Asian
137	mnp_as	Character	7	Math students scoring proficient or above, Asian
138	mpp_as	Character	5	Math percent of students scoring proficient or above, Asian
139	mppm_as	Character	4	Math percent proficient or above met, Asian (not applicable for 2015 AYP)
140	ev_fi	Character	7	ELA valid scores, Filipino
141	enp_fi	Character	7	ELA number of students scoring proficient or above, Filipino
142	epp_fi	Character	5	ELA percent of students scoring proficient or above, Filipino
143	eppm_fi	Character	4	ELA percent proficient or above met, Filipino (not applicable for 2015 AYP)
144	mv_fi	Character	7	Math valid scores, Filipino
145	mnp_fi	Character	7	Math percent of students scoring proficient or above, Filipino

Field #	Field Name	Type	Width	Description
146	mpp_fi	Character	5	Math percent of students scoring proficient or above, Filipino
147	mppm_fi	Character	4	Math percent proficient or above met, Filipino (not applicable for 2015 AYP)
148	ev_hi	Character	7	ELA valid scores, Hispanic or Latino
149	enp_hi	Character	7	ELA number students scoring proficient or above, Hispanic or Latino
150	epp_hi	Character	5	ELA percent of students scoring proficient or above, Hispanic or Latino
151	eppm_hi	Character	4	ELA percent proficient or above met, Hispanic or Latino (not applicable for 2015 AYP)
152	mv_hi	Character	7	Math valid scores, Hispanic or Latino
153	mnp_hi	Character	7	Math students scoring proficient or above, Hispanic or Latino
154	mpp_hi	Character	5	Math percent of students scoring proficient or above, Hispanic or Latino
155	mppm_hi	Character	4	Math percent proficient or above met, Hispanic or Latino (not applicable for 2015 AYP)
156	ev_pi	Character	7	ELA valid scores, Native Hawaiian or Pacific Islander
157	enp_pi	Character	7	ELA number of students scoring proficient or above, Native Hawaiian or Pacific Islander
158	epp_pi	Character	5	ELA percent of students scoring proficient or above, Native Hawaiian or Pacific Islander
159	eppm_pi	Character	4	ELA percent proficient or above met, Native Hawaiian or Pacific Islander (not applicable for 2015 AYP)
160	mv_pi	Character	7	Math valid scores, Native Hawaiian or Pacific Islander
161	mnp_pi	Character	7	Math students scoring proficient or above, Native Hawaiian or Pacific Islander
162	mpp_pi	Character	5	Math percent of students scoring proficient or above, Native Hawaiian or Pacific Islander
163	mppm_pi	Character	4	Math percent proficient or above met, Native Hawaiian or Pacific Islander (not applicable for 2015 AYP)
164	ev_wh	Character	7	ELA valid scores, white
165	enp_wh	Character	7	ELA number of students scoring proficient or above, white

Field #	Field Name	Type	Width	Description
166	epp_wh	Character	5	ELA percent of students scoring proficient or above, white
167	eppm_wh	Character	4	ELA percent proficient or above met, white (not applicable for 2015 AYP)
168	mv_wh	Character	7	Math valid scores, white
169	mnp_wh	Character	7	Math students scoring proficient or above, white
170	mpp_wh	Character	5	Math percent of students scoring proficient or above, white
171	mppm_wh	Character	4	Math percent proficient or above met, white (not applicable for 2015 AYP)
172	ev_mr	Character	7	ELA valid scores, two or more races
173	enp_mr	Character	7	ELA number of students scoring proficient or above, two or more races
174	epp_mr	Character	5	ELA percent of students scoring proficient or above, two or more races
175	eppm_mr	Character	4	ELA percent proficient or above met, two or more races (not applicable for 2015 AYP)
176	mv_mr	Character	7	Math valid scores, two or more races
177	mnp_mr	Character	7	Math students scoring proficient or above, two or more races
178	mpp_mr	Character	5	Math percent of students scoring proficient or above, two or more races
179	mppm_mr	Character	4	Math percent proficient or above met, two or more races (not applicable for 2015 AYP)
180	ev_sd	Character	7	ELA valid scores, socioeconomic disadvantaged
181	enp_sd	Character	7	ELA number of students scoring proficient or above, socioeconomic disadvantaged
182	epp_sd	Character	5	ELA percent of students scoring proficient or above, socioeconomic disadvantaged
183	eppm_sd	Character	4	ELA percent proficient or above met, socioeconomic disadvantaged (not applicable for 2015 AYP)
184	mv_sd	Character	7	Math valid scores, socioeconomic disadvantaged
185	mnp_sd	Character	7	Math students scoring proficient or above, socioeconomic disadvantaged

Field #	Field Name	Type	Width	Description
186	mpp_sd	Character	5	Math percent of students scoring proficient or above, socioeconomic disadvantaged
187	mppm_sd	Character	4	Math percent proficient or above met, socioeconomic disadvantaged (not applicable for 2015 AYP)
188	ev_el	Character	7	ELA valid scores, English learner
189	enp_el	Character	7	ELA students scoring proficient or above, English learner
190	epp_el	Character	5	ELA percent of students scoring proficient or above, English learner
191	eppm_el	Character	4	ELA percent proficient or above met, English learner (not applicable for 2015 AYP)
192	mv_el	Character	7	Math valid scores, English learner
193	mnp_el	Character	7	Math students scoring proficient or above, English learner
194	mpp_el	Character	5	Math percent of students scoring proficient or above, English learner
195	mppm_el	Character	4	Math percent proficient or above met, English learner (not applicable for 2015 AYP)
196	ev_di	Character	7	ELA valid scores, students with disabilities
197	enp_di	Character	7	ELA students scoring proficient or above, students with disabilities
198	epp_di	Character	5	ELA percent of students scoring proficient or above, students with disabilities
199	eppm_di	Character	4	ELA percent proficient or above met, students with disabilities (not applicable for 2015 AYP)
200	mv_di	Character	7	Math students with a valid scores, students with disabilities
201	mnp_di	Character	7	Math students scoring proficient or above, students with disabilities
202	mpp_di	Character	5	Math percent of students with disabilities scoring proficient or above
203	mppm_di	Character	4	Math percent proficient or above met, students with disabilities (not applicable for 2015 AYP)
204	grad14	Character	6	Graduation rate for 2014, class of 2012–13

Field #	Field Name	Type	Width	Description
205	grad15	Character	6	Graduation rate for 2015, class of 2013–14
206	tr_15	Character	6	LEA or schoolwide graduation target rate for 2015
207	sw_gr_met	Character	1	Met schoolwide graduation (Y=yes, N=no, blank=not applicable)
208	sg_gr_met	Character	1	Met student group graduation rates (Y=yes, N=no, blank=not applicable)
209	e_pr_app	Character	5	Participation rate ELA (Yes, No, YMA=yes met on appeal, N/A)
210	m_pr_app	Character	5	Participation rate math (Yes, No, YMA=yes met on appeal, N/A)
211	e_pp_app	Character	5	Percent proficient ELA (not applicable for 2015 AYP)
212	m_pp_app	Character	5	Percent proficient math (not applicable for 2015 AYP)
213	grad_app	Character	5	Graduation Rate (Yes, No, YMA=yes met on appeal, N/A)
				Note: To meet this criterion, the graduation rate must be met at the schoolwide or LEA level and for all numerically significant student groups.
214	ela_met	Character	5	ELA met based on the ELA participation rate only (Yes, No)
215	math_met	Character	5	Math met based on the ELA participation rate only (Yes, No)
216	e_targ15	Character	5	ELA 2015 percent proficient target (not applicable for 2015 AYP)
217	m_targ15	Character	5	Math 2015 percent proficient target (not applicable for 2015 AYP)
218	AvgDailyAttend	Character	3	Attendance rate (calculated by using average daily attendance data) (DNS = did not submit data; N/A)
219	SchAttend	Character	3	Attendance rate (calculated by using number of days attended and enrolled) (DNS = did not submit data; N/A)
220	MetAttendTarg	Character	3	Met attendance rate target (Yes, No, N/A)

ELA: English language arts

LEA: Local educational agency: As defined in ESEA, "a public board of education or other public authority legally constituted within a State for either administrative control or direction of, or to perform a service function for, public elementary schools or secondary schools in a city, county, township, school district, or other political subdivision of a State, or for a combination of

school districts or counties that is recognized in a State as an administrative agency for its public elementary schools or secondary schools."

APPENDIX B

SCHOOL DEMOGRAPHICS DATA

Field #	School Demographic Attributes
1	SchoolName
2	CountyName
3	DistrictName
4	SchoolType
5	CDSNumber
6	Zip
7	City
8	Enrollment
9	EnglishLearners
10	EnglishLearnersPer
11	AmericanIndianPer
12	AsianPer
13	AfricanAmericanPer
14	FilipinoPer
15	HispanicPer
16	NativeHawaiianPer
17	TwoRacesPer
18	NoneReportedPer
19	WhitePer
20	FluentEnglishProficientNum
21	FluentEnglishProficientPer
22	FosterYouthNum
23	FreeReducedMealsNum
24	FreeReducedMealsPer
25	FRPMNum
26	CohortGraduatesPer
27	PerPupilRatioTeacher
28	X1stYearTeachers
29	X2YearTeachers
30	AvgYearsTeaching
31	TeachersNum
32	TeachersFTE

APPENDIX C

DISTRICT-LEVEL FINANCIAL DATA

Field #	Attribute Names
1	DistrictName
2	Enrollment
3	FreeReducedMealsNum
4	EnglishLearnersPer
5	EnglishLearnersNum
6	FosterYouthNum
7	FreeReducedMealsPer
8	AverageDailyAttendance
9	AmericanIndianAlaskaNativePer
10	AsianPer
11	AfricanAmericanPer
12	FilipinoPer
13	HispanicPer
14	HawaiianPacIslanderPer
15	TwoMoreRacesPer
16	NoneReportedPer
17	WhitePer
18	FRPMELFosterUnduplD
19	RedesignatedFEP
20	CohortGraduatesPer
21	ELMakingAnnualGrowthTargetPer
22	APExamGraduatingClassTestTakers
23	SATAvgResultsMathematic
24	CohortGraduatesNum
25	GradsMtgUCCSUPer
26	SATAvgResultsWritingNum
27	PerPupilRatioToTeacher
28	TeachingDays
29	SalaryChange
30	X2YearTeachers
31	TeacherHighestSalaryOfferedDistrict.
32	TeacherLowestSalaryOffered
33	AvgYearsTeaching
34	TeachersNum
35	Suspensions

36	CurrentExpoEducperADA
37	TotalGenFundExpendituresPerStudent
38	TotalGenFundExpenditures
39	GenFundExpbyActivityPer
40	GenFundExpbyActivityNum
41	ActivityInstrucRelatedSvcsExpPer
42	TotalGenFundRevenues
43	TotalGenFundRevenuesPerStudent
44	PupilServicesExpPer
45	InstrucrelatedSvcsPerStudentNum
46	CertificatedSalariesPerStudent
47	ClassifiedSalariesPerStudent
48	PupilServicesPerStudentNum
49	FederalRevenuePerStudent
50	StateRevenuePerStudent
51	LocalRevenuePerStudent

APPENDIX D

SAMPLE SCHOOL PROFILES FROM EACH CLUSTER

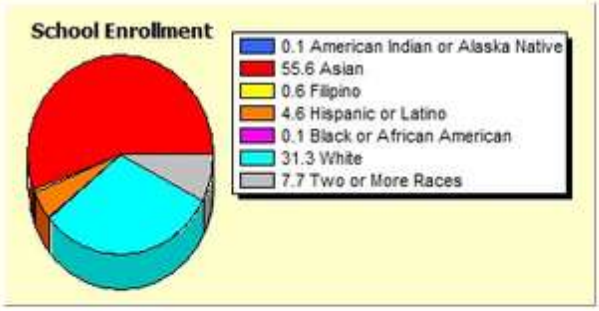
SELECTED BY *K*-MEANS

Cluster 1: Saratoga High

Students by Race/Ethnicity
Saratoga High School,
2013-14

	School		District
	Enrollment	Percent of Total	Percent of Total
American Indian or Alaska Native	2	0.1%	0.1%
Asian	794	55.6%	29.8%
Native Hawaiian or Pacific Islander	0	0.0%	0.3%
Filipino	9	0.6%	0.4%
Hispanic or Latino	65	4.6%	7.1%
Black or African American	1	0.1%	0.3%
White	447	31.3%	53.2%
Two or More Races	110	7.7%	8.6%
None Reported	0	0.0%	0.4%
Total	1,428	100%	100%

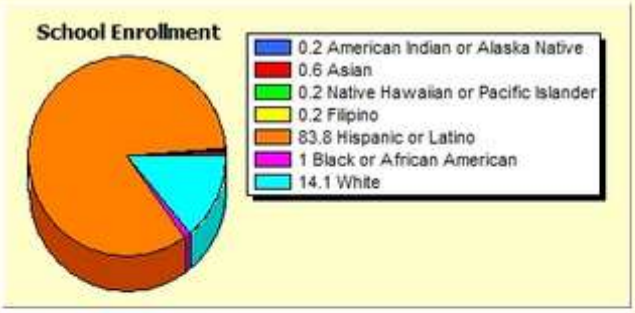
Note: Saratoga High's Ethnic Diversity Index is 42.

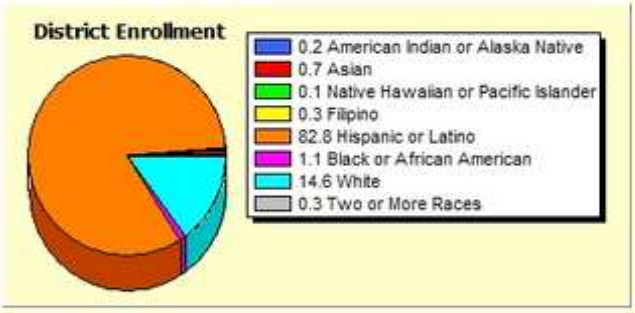


Cluster 2: Riverdale High

Students by Race/Ethnicity
Riverdale High School,
2013-14

	School		District
	Enrollment	Percent of Total	Percent of Total
American Indian or Alaska Native	1	0.2%	0.2%
Asian	3	0.6%	0.7%
Native Hawaiian or Pacific Islander	1	0.2%	0.1%
Filipino	1	0.2%	0.3%
Hispanic or Latino	434	83.8%	82.8%
Black or African American	5	1.0%	1.1%
White	73	14.1%	14.6%
Two or More Races	0	0.0%	0.3%
None Reported	0	0.0%	0.0%
Total	518	100%	100%





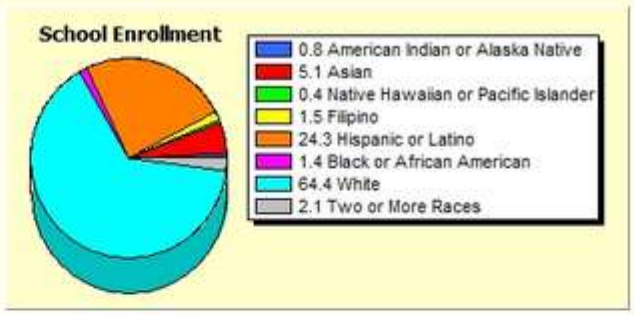
Cluster 3: Thousand Oaks High

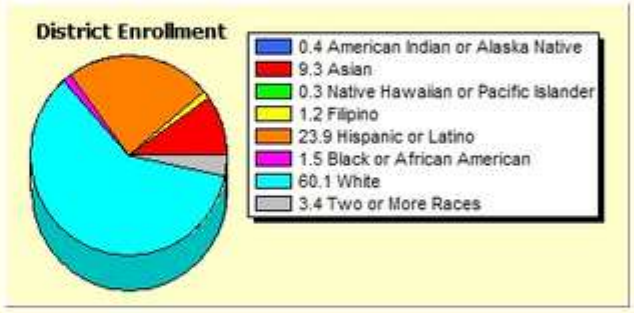
Students by Race/Ethnicity Thousand Oaks High School, 2013-14

	School		District
	Enrollment	Percent of Total	Percent of Total
American Indian or Alaska Native	19	0.8%	0.4%
Asian	118	5.1%	9.3%
Native Hawaiian or Pacific Islander	9	0.4%	0.3%
Filipino	35	1.5%	1.2%
Hispanic or Latino	559	24.3%	23.9%
Black or African American	32	1.4%	1.5%
White	1,483	64.4%	60.1%
Two or More Races	49	2.1%	3.4%
None Reported	0	0.0%	0.0%
Total	2,304	100%	100%

Note: Thousand Oaks High's Ethnic Diversity Index is 37.

ALSO SEE ▶ Students by Race/Ethnicity definitions





APPENDIX E

REGRESSION OUTPUT FROM GLM AND MARS

```
Call:
glm(formula = f14s$CohortGraduatesPer ~ ., data = f14s[, -17])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3628  -0.3017   0.0406   0.3382   2.6693

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -8.157e-16  3.850e-02   0.000 1.000000
Enrollment                       2.597e+00  2.197e+00   1.182 0.238103
FreeReducedMealsNum             -1.935e+00  6.961e-01  -2.780 0.005786 **
EnglishLearnersPer              -2.811e-01  9.549e-02  -2.944 0.003504 **
FreeReducedMealsPer              4.186e-01  3.853e-01   1.086 0.278194
AverageDailyAttendance          -2.623e-01  1.455e+00  -0.180 0.857050
AmericanIndianAlaskaNativePer   -1.290e+00  2.633e+00  -0.490 0.624402
AsianPer                        -2.216e+00  4.707e+00  -0.471 0.638176
AfricanAmericanPer              -1.454e+00  2.892e+00  -0.503 0.615577
FilipinoPer                     -7.121e-01  1.431e+00  -0.498 0.619171
HispanicPer                     -5.232e+00  1.143e+01  -0.458 0.647374
HawaiianPacIslanderPer          -1.586e-01  2.410e-01  -0.658 0.511039
TwoMoreRacesPer                 -5.963e-01  1.272e+00  -0.469 0.639535
NoneReportedPer                 -2.990e-01  6.671e-01  -0.448 0.654296
WhitePer                        -4.537e+00  9.780e+00  -0.464 0.643068
FRPMELFosterUnduplDistrict      -2.036e-01  3.875e-01  -0.525 0.599684
RedesignatedFEP                  3.324e-02  3.618e-01   0.092 0.926858
ELMakingAnnualGrowthTargetPer    3.099e-02  5.691e-02   0.545 0.586461
APExamGraduatingClassTestTakersNum -3.745e-01  1.769e-01  -2.117 0.035087 *
SATAvgResultsMathematic          9.143e-01  3.289e-01   2.780 0.005794 **
GradsMtgUCCSUPer                 2.046e-01  5.719e-02   3.578 0.000406 ***
PerPupilRatioToTeacher          -1.807e-02  4.104e-02  -0.440 0.659943
TeachingDays                     2.363e-01  1.803e-01   1.310 0.191127
SalaryChange                     5.509e-02  4.564e-02   1.207 0.228456
X2YearTeachers                  -1.496e-01  1.501e-01  -0.996 0.319925
```

```
TeacherHighestSalaryOfferedDistrict  2.875e-02  1.811e-01   0.159 0.873948
TeacherLowestSalaryOffered          -3.298e-01  1.490e-01  -2.214 0.027598 *
AvgYearsTeaching                    -2.560e-02  5.194e-02  -0.493 0.622416
TeachersNum                         -1.068e+00  1.579e+00  -0.676 0.499478
Suspensions                         -2.536e-02  4.191e-02  -0.605 0.545680
CurrentExpoEducperADA               -9.699e-01  6.130e-01  -1.582 0.114717
TotalGenFundExpendituresPerStudent   8.840e-01  7.905e-01   1.118 0.264398
TotalGenFundExpenditures            -1.937e+00  2.991e+00  -0.648 0.517804
GenFundExpbyActivityPer             -9.061e-02  1.586e-01  -0.571 0.568320
GenFundExpbyActivityNum              1.101e-01  2.556e-01   0.431 0.666923
ActivityInstrucRelatedSvcsExpPer     3.452e-01  2.991e-01   1.154 0.249389
TotalGenFundRevenues                 2.329e+00  2.435e+00   0.956 0.339702
TotalGenFundRevenuesPerStudent       2.381e-01  3.054e-01   0.780 0.436197
PupilServicesExpPer                  9.869e-02  2.342e-01   0.421 0.673806
InstrucrelatedSvcsPerStudentNum     -3.665e-01  2.176e-01  -1.684 0.093214 .
CertificatedSalariesPerStudent       8.181e-01  2.256e-01   3.626 0.000340 ***
ClassifiedSalariesPerStudent        -7.157e-01  2.285e-01  -3.132 0.001917 **
PupilServicesPerStudentNum          -1.148e-01  2.623e-01  -0.438 0.661979
FederalRevenuePerStudent            -1.497e-01  5.556e-02  -2.694 0.007462 **
StateRevenuePerStudent              -9.532e-02  5.623e-02  -1.695 0.091131 .
LocalRevenuePerStudent              -1.624e-01  5.730e-02  -2.834 0.004916 **
SATAvgResultsWritingNum             -5.355e-01  3.216e-01  -1.665 0.097030 .
EnglishLearnersNum                   6.391e-01  5.103e-01   1.252 0.211415
```

```
> ev
                                nsubsets   gcv    rss
ClassifiedSalariesPerStudent          12 100.0  100.0
GradsMtgUCCSUPer                      10  71.3   73.1
APExamGraduatingClassTestTakersNum    10  64.3   67.0
PerPupilRatioToTeacher                10  64.3   67.0
AfricanAmericanPer                     8  44.3   48.7
WhitePer                               7  33.7   39.6
SATAvgResultsMathematic                6  28.5   34.6
TotalGenFundExpendituresPerStudent     5  25.9   31.4
LocalRevenuePerStudent                 4  22.8   27.7
FederalRevenuePerStudent               1  10.0   12.8
```

```
> marsf14$coefficients
                                                 CohortGraduatesPer
(Intercept)                                           -0.1499600
h(SATAvgResultsMathematic--0.651021)                   0.3087407
h(-0.651021-SATAvgResultsMathematic)                  -0.3071306
h(ClassifiedSalariesPerStudent-0.850356)              -1.4124754
h(TotalGenFundExpendituresPerStudent-1.20877)          1.0546839
h(2.09633-AfricanAmericanPer)                          0.2152435
h(1.81244-GradsMtgUCCSUPer)                           -0.2061950
h(LocalRevenuePerStudent--0.164014)                   -0.1608632
h(FederalRevenuePerStudent-0.666129)                  -0.1308797
h(-0.0783653-PerPupilRatioToTeacher)                  35.0906696
h(APExamGraduatingClassTestTakersNum--0.407696)       -0.1079499
h(-0.407696-APExamGraduatingClassTestTakersNum)      -39.3047838
h(-1.37847-WhitePer)                                 -20.3600720
>
```