Using Data Mining to Evaluate Colorado Public Schools Performance

Aaron Holt
University of Colorado Boulder
2900 E College Ave Ap#8
Boulder, CO 80303
+1 (719) 201-0277
aaron.holt@colorado.edu

Anas Salamah
University of Colorado Boulder
4670 White Rock Circle
Boulder, CO 80301
+1 (720) 240-6170
anas.salamah@gmail.com

Hui Soon Kim
University of Colorado Boulder
1855 Athens St Apt. 304
Boulder, CO 80301
+1 (303) 960-6480
huki2996@colorado.edu

ABSTRACT

In this paper, we describe the process we used to apply data mining techniques on Colorado public education data. We want to investigate the performance of public schools in Colorado based on the overall given grade by analyzing the data to find frequent item sets and finding the relationships between these items, as well as, validating the performance based on other outer sources such as US census data which includes financial data of the different cities in Colorado such as family income. We also want to build a classifier that predicts the grade of a school based on its associated information. Finally, we want to validate our classifier by holding out some data as test data and measuring the accuracy of our classifier.

Keywords

Educational Data Mining, frequent pattern analysis, Association rules, Classification.

1. Motivation

Since the introduction of the No Child Left Behind (NCLB) in 2001, data driven decision-making has become an increasingly important topic for schools nationwide. In order for schools to prove they have met the Adequate Yearly Progress requirements defined in NLCB, schools must collect a significant amount of data pertaining to their student demographics, graduation rate, grades, and state assessment scores [1]. The collected data is used to show whether or not a school has achieved its yearly progress goals and to identify areas of improvement.

One of the many challenges schools face is finding valuable ways to use the large amounts of data given. As is often the case with big data, schools find themselves being "data rich but information poor" [2]. Principal Dr. Gregory Decker of Lead Mine Elementary succinctly stated the problem saying "Receiving test data in July is like driving a school bus looking out the rearview mirror. I can see where my students have been but I cannot see where we are going."[2] Thus schools have been attempting to look at information over time to track student achievement. While there has been progress in this area, data from external sources such as US census data is often excluded. In order to improve upon the current methodologies, patterns from Colorado school collected data and external sources, both statically and over time should be considered.

The overarching goal of this project is to find meaningful patterns that highlight successes and failures in schools and develop a prediction model that evaluates the school performance by combining available public school data of Colorado and other external data such as US census. Most schools measure success based on student grades, test scores, and graduation rates. Thus finding correlations between these attributes and those from outside sources is a priority.

The dataset currently proposed comes from several sources. The majority of the data ware collected by the Colorado Department of Education (CDE) and R-Squared Research. The CDE data contains information about students' grades, test scores, ethnicity, and progress at every Colorado public school. The next source comes from the US census, which includes financial data in different cities, counties and the state as a whole. The combination of these sources will hopefully yield new and useful correlations that will help improve the performance of Colorado schools.

2. Literature Survey

2.1 Data Mining Technique

This section reviews the different data mining techniques that may be used for finding significant attributes to affect the school performance and predict the overall grade.

Frequent item sets play an essential role in many data mining tasks that try to find interesting patterns from databases, such as association and correlation rules. In order to find correlations between independent variables and dependent variables, frequent pattern analysis has to be done at first. The Apriori and FP-growth algorithms [3] can generate all frequent item sets with a specific support value. And then finding the association and correlation rules for given frequent item sets, which results in finding the significant attributes that affect a school performance.

Classifications may be used for predictions of school performance. The techniques that are reviewed are Naïve Bayes [4], KNN [4] and ID3 [5]. Bayesian classifier is the supervised data mining technique used to take decisions under uncertain conditions. The concept of probability is used to classify the new entities by looking into past data to predict the performance of newly entered schools. The value of a given class is conditionally independent of the values of other attributes. In Bayesian analysis, combining both the prior and the likelihood to form the posterior probability using the so-called Bayes produces the final classification. K nearest neighbor algorithm (KNN) is known as a lazy learner that makes predictions based on KNN labels assigned to training samples. K nearest neighbor is determined by measuring the distance between the new entered query and previously known samples. The bulk of K nearest neighbors is taken for the prediction of the entered query once the K nearest neighbor is assembled. ID3 is one of the popular DT algorithms that deal with the nominal data sets. ID3 is a classical version of the decision tree induction. ID3 mainly works on the selections of attributes at all the levels of the decision tree that are based on entropy. Basically, a top-down greedy search approach is followed to construct the tree, where each attribute is tested on every tree node.

2.2 External Attributes to affect School Performance

Some external attributes such as family income, parent's education, racial distribution and a single parent home Etc. need to be considered because they can affect children's school performance. For example, According to the survey, "Current family income has significant effects on a child's math and reading test scores"[6], "Socioeconomic variables do influence parenting beliefs and behaviors and that these parenting variables influence subsequent change in child's achievement" [7], "Although scores have increased for both Black students and White students, on average Black students do not perform as well as their White peers."[8], "Lower high school graduation rates, lower GPAs, and greater risk for drug abuse are some of the negative outcomes associated with growing up in a single-parent home"[9]. As a result various attributes from other data sources that may affect the school performance will be combined in this project.

3. Proposed Work

3.1 Data Collection and Preprocessing

The initial work involves collecting and processing data. The dataset used in this project consists of Colorado Department of Education (CDE), the R-Squared Research data, and the US census in different formats each with varying attribute types. Data preprocessing is an important step in which missing values are filled, redundancies are removed and filtering is performed so that the database will be ready for use. In order to analyze these data in combined manner, the data integration is performed using common attributes between data sets. According to the attribute types such as normal, numerical, binary and ordinal, a proper data preprocessing technique such as normalization, discretization and linearization is needed.

3.2 Frequent Pattern Analysis

Before moving on to trend analysis over time, it will be useful to establish static trends. There are several genres of trends to search for. The first set involves searching for simple location based items such as where the best and worst schools are located.

Next, correlations between various attributes and student achievement can be explored. For example, one could explore how family incomes of a county affect grades and graduation rates of schools in that county. Results could establish a correlation between wealth and performance.

The discovered static correlations will be used to guide the search for trends over time. Trends over time are likely more valuable as they indicate whether a given strategy is working or not. Similar to the static analysis, trends correlating attributes to student achievement will be explored first.

3.3 Modeling a Classifier

After deciding the significant attributes to influence the school grade, we can model a classifier to predict school performance using those significant attributes

4. How to Evaluate

There are several core evaluation metrics for this project. For both the static and over time analysis, this will primarily be how do various attributes correlate to student achievement. Student achievement for static analysis will explore how an attribute affects overall grades and graduation rate. In the data, the overall grade is an average of reading, writing, math and science. Each school along with its type (Elementary, Middle, or High school) is assigned a grade. Some schools combine more than one type and thus are given a grade for each type. Individual subject grades will also be considered in trend analysis. The trend over time analysis will be similar except for overall grade is replaced with overall growth where growth is defined as the change in individual subject scores over time. Finally, the classifier will be evaluated based on the percentage of time it can accurately predict school grades and graduation rate within a given margin. In order to evaluate our grade classifier, we need to split our data to training and testing data.

5. Milestones

Project Deliverable	**Expected Completion Date**
Data collection and preprocessing	3-6-2015
Static frequent pattern analysis	3-16-2015
Pattern analysis for trends over time	3-30-2015
Classification task	4-10-2015

6. Accomplishment so Far

6.1 Data Preprocessing

We've analyzed two years of school data (2011 and 2012) for school grades of Colorado public schools. The school dataset consists of a total of 17 separate CSV files. Some of which are final school grade, grade change, student number of free meal, enrollment number, school address, and school GPS information. And the US census dataset consists of 8 separate files some of which are education level, family income, number of single family home and race distribution. So we have a total of 25 separate files in this project.

6.1.1 Data Integration

The main data of this project is the final school grade file because it has information regarding all public schools final grade and each subject grade such as math, writing, etc. The attribute type of each grade is a discrete ordinal number that ranges from 1 to 13. The other attributes of the school dataset consist of a lot of numeric (e.g. % of free meal) and categorical (e.g. School name) values. Since all of the school data files have the unique attribute school name, we used a method of entity identification in data integration, i.e., we have combined data from multiple school files using the attribute school name and produced one integrated school data file.

Moreover, we integrated the above combined school file with US census data. We used the entity identification in data integration again using the attribute ZIP code because all of the census data we have obtained have ZIP code information.

After the above integration process, we could produce one reasonably large integrated file for each year, which consists of 49 attributes as the following table illustrates.

Year	Number of objects	Attributes (total number : 49)
2011	2,061	School Name, School Grade, Student's Grade, Race Distribution, Free Meal, Annual Income, Single Family Home, Education Level, etc.
2012	1,962	

Of course in this process, a lot of irrelevant attributes (e.g. School code) in the school grade were discarded and all of this process was done using Python and Excel.

6.1.2 Data Cleaning

Even if we have integrated all of the data files, there were some missing values in the data table. It is impossible to analyze the relationship between school grade and other attributes accurately without having a value for the school grade so we discarded all of the objects without school grade. Furthermore, the objects without ZIP code cannot be mapped with external attributes from US census data. So those objects have also been discarded. Moreover, some missing values have been filled in automatically using each attribute's mean. The following table shows final total number of data objects we have got after the data cleaning process.

Year	Total object number used (After Data Cleaning)
2011	1,814
2012	1,812

6.1.3 Data Reduction and Discretization

Using a scatter plot, we have found that student's grade of each subject and school grade have strong positive correlation illustrated in the following figure.

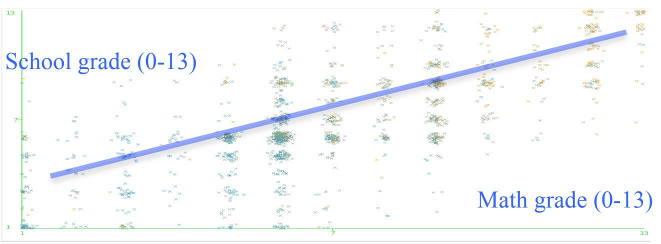

Figure 1: Colorado school grade with corresponding math grades.

[Correlation Coefficients between School and Subject grades]

This positive correlation verifies that student's grades are a direct factor for estimating the school grade. So we focused on analyzing the relationship between school grades and external attributes.

In order to apply the Apriori algorithm to our dataset for the frequent pattern mining, all of the attributes have to be discretized as categorical value. We first took a subset of the data by reducing the number of attributes. This subset consists of School grade, Enrollment, Race distribution, the percentage of free and reduced lunch, the percentage of single family homes, the percentage of families with annual income that is more than $60,000 and the percentage of population with education higher than or equal to bachelor's degree. This subset data then was discretized using 4 equal frequency bins. And then the Apriori algorithm was applied to this dataset for frequent pattern mining.

6.2 Result of Frequent Pattern Mining

6.2.1 Basic Information

The initial mining process resulted in compiling some basic information about Colorado Schools. We can see from the table located in Appendix A that schools with overall grades 10-13 account to 15% of the total number of schools. On the other hand schools with overall school grades 6-9 account for 60% of the total number of schools, which is to be expected because that is the average overall school grade. Thus schools with overall school grades 1-5 account for 15%. In order to better visualize this effect the distribution of schools was plotted onto a Colorado Map, seen in Figure 2.

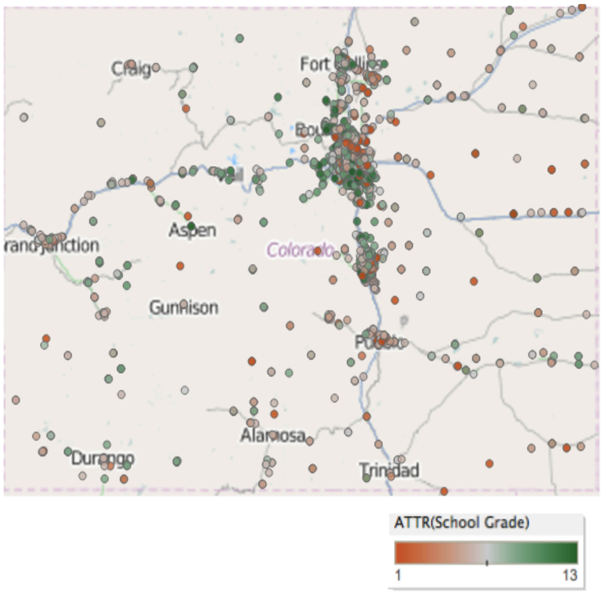

Figure 2: Colorado school locations with corresponding grades.

As you can see there is a high density of schools between the Denver metropolitan area and Fort Collins and the overall school grade in that area is good. However, to the west and south, the overall grade is between bad to average.

Following our timeline, we started our mining process by finding the static frequent patterns on the 2012 school grades dataset, which represents the prior school year. As mentioned earlier, the school grade attribute is an indicator of how well a school is doing overall. It takes into consideration student test scores, student growth from previous years, and graduation rate if the school is a high school. Thus searching for patterns and correlations affecting school grades is a good starting point for understanding and describing the data. Initially we selected the attributes 'Race', 'Income', and 'Education Attainment' to mine in order to verify previous research stating that these attributes had the most impact on the performance of schools. Related attributes such as 'percentage of students on free and reduced lunch' and 'percentage of single families' were also selected. We further discretized these attributes into four bins enabling us to run the Apriori Algorithm as we mentioned in 6.1.3. The attribute breakdown is shown in Appendix B.

6.2.2 Apriori Results

The Apriori algorithm was used to help determine which attributes affected school grade the most. The algorithm was run using the following values:

Initial Apriori Specifications	
Metric	Lift
Minimum Lift	2.0
Minimum Support	0.075

The lift and support values were initially 1.1 and 0.075, but with a low lift the number of rules generated were too high to manage. Thus we have increased to the lift value to 2 to make the output more manageable, and to find only strong correlations.

With these values every single attribute was a big enough to be a single item set. On following passes of the algorithm the item sets decreased until there was only one 4-itemset relating to the school grade. A 4-itemset, which has a lift value of 4.13, is shown below.

School Grade = 0-5.5	PCT Hispanic = 0.45+	PCT White = 0-0.38	PCT Free Meal = 0.65+

This rule effectively says that a 'bad school', or a school with a low overall grade, will likely have a high percentage of Hispanic students, a lower percentage of White students, and a high percentage of students on free and reduced lunch. Other indicators were also found in smaller rules such as a low percentage of the population with family income of $60,000+(0-0.4), and a low percentage of Pacific Islanders (0-0.0002).

Using these values no rules were found that correlated attributes to good schools, thus the support was lowered to 0.075 in hopes of finding rules from good schools. The biggest frequent item set pertaining to new schools was a 3-itemset with a lift value of 3.2.

School Grade = 9.5+	PCT Hispanic = 0-0.110	PCT Free Meal = 0-0.2

This finding is interesting because it shows attributes 'PCT Hispanic' and 'PCT Free Meal' are good indicators for finding good and bad schools. A good school grade is negatively correlated with 'PCT Hispanic' and 'PCT Free Meal' whereas a bad school is positively correlated with 'PCT Hispanic' and 'PCT Free Meal'. Another indicator (with a lift greater than 2) for a good school was the percentage of the population having a bachelors degree or higher.

There were no rules with a lift greater than 2 for schools with a grade range of (5.5-6.5). The largest frequent item set for schools with scores (5.5-9.5) was a 3-itemset with a lift of 2.77.

School Grade = 6.5-9.5	PCT $60,000+ = 0.7+	PCT Free Meal = 0-0.2

This rule verifies that families with an income greater than $60,000 typically do not meet the requirements for free lunch.

6.3 Relating External Attributes to School Grades

One of the attributes in the US census data that have been mined to see if there exists a relationship between it and the school grade is the total number of enrolled students in a school. We have also compared the percentage of different ethnicities in schools with the school grade attribute. In addition, we explored the relationship between the percentage of students in a single family home and students on free or reduced lunch with the school grade.

6.3.1 Visual Verification

After preprocessing, integration, and discretization of the data, we wanted a way to check that the results produced made sense. Thus, in order to verify some of the results, graphs of various attributes were made.

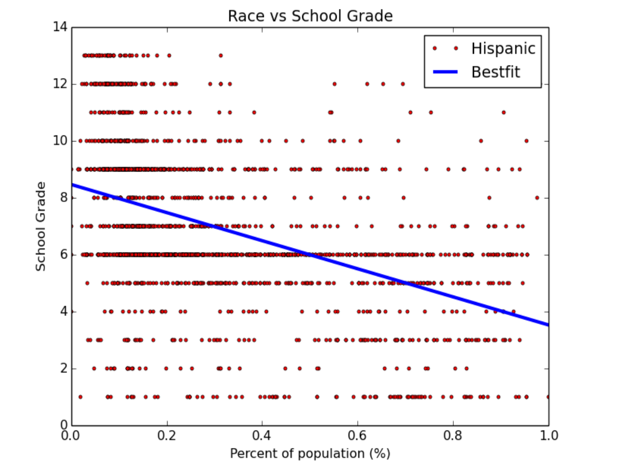

Figure 3: Percent Hispanic versus school grade.

In the section 6.22 it was found that a high percentage of Hispanic population meant a bad school and a low percentage meant a good school. This is seen in Figure 3 above as a linear bestfit line shows a negative correlation. A similar trend can be also seen with the percentage of Black students compared to the school grade in Figure 4.

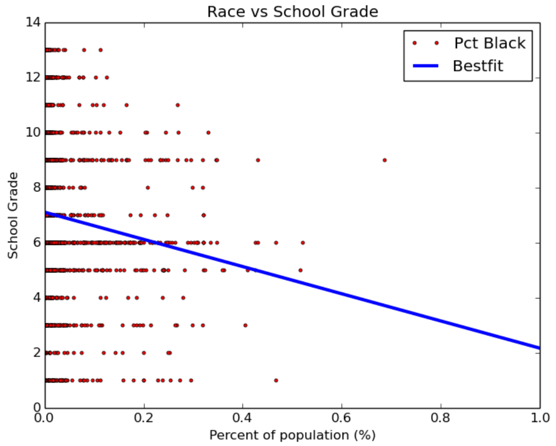

Figure 6*: Percentage Black students versus school grade.

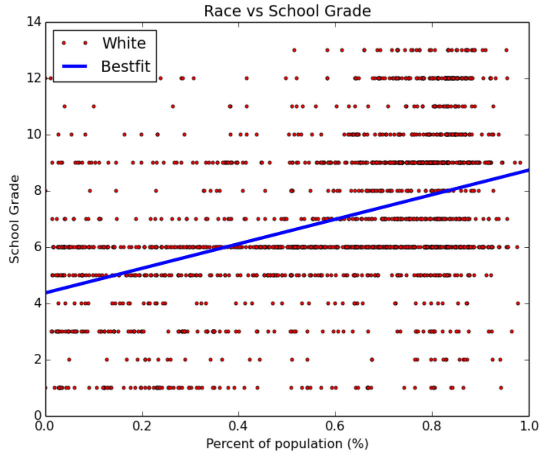

Figure 5: Percent White versus school grade.

It was also discovered that a low percentage of white population meant a bad school. This is seen in positive correlation of the linear bestfit line in Figure 5 above. One item not found using the Apriori algorithm that can be seen here is that a high percentage of Whites means a better school.

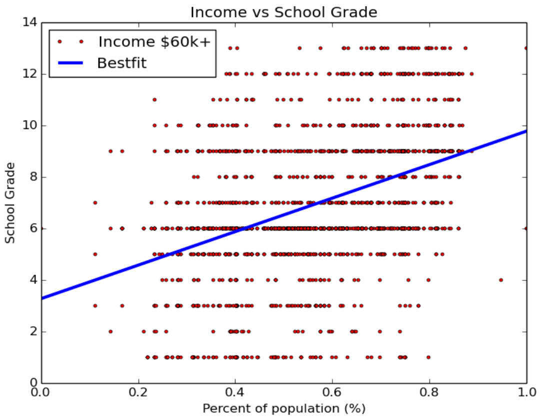

Figure 6: Percentage of families with income greater than $60,000 versus school grade.

In section 6.22 the Apriori algorithm found a rule showing that a higher income was correlated with a school grade of 6.5-9.5. The linear bestfit line in Figure 6 shows that a higher percentage of families above the $60k mark is positively correlated with school grade.

We also tested the Free and Reduced Lunch attribute and Figure 7 shows the negative correlation that we mentioned before between the Free and Reduced Lunch attribute and school grades.

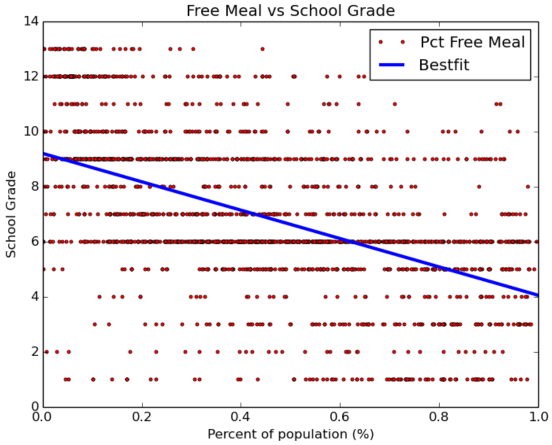

Figure 7: Percentage of students on free or reduced lunch versus school grade.

A strong correlation can also be seen in the attribute that describes the percentage of families with a single parent in figure 8.

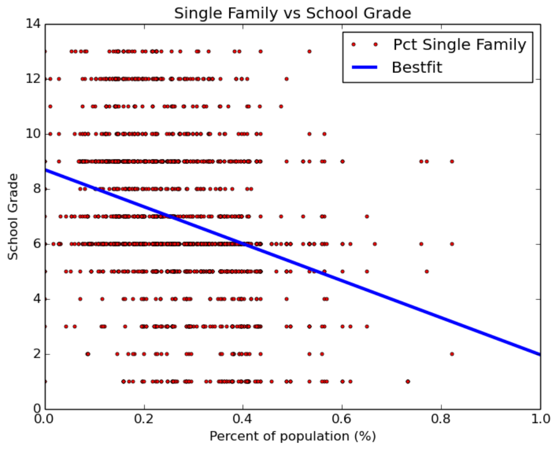

Figure 8: Percentage of students with single parents versus school grade.

We have also produced graphs that represent the relationship between the other background and the school grade in Figures 8 and 9. However, due the low percentage of those different backgrounds, the bestfit is not a true indicator of the relationship between them and the school grade.

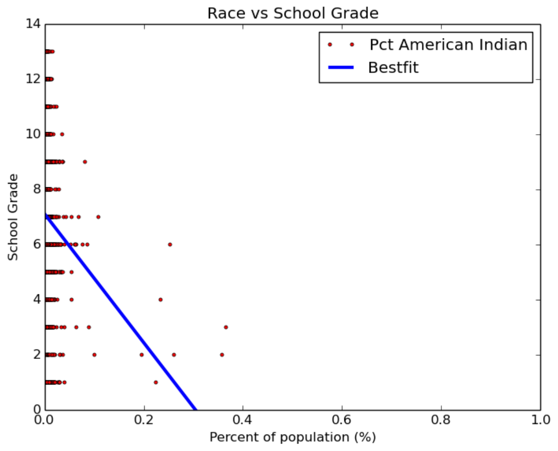

Figure 8: Percentage of American Indian students versus school grade.

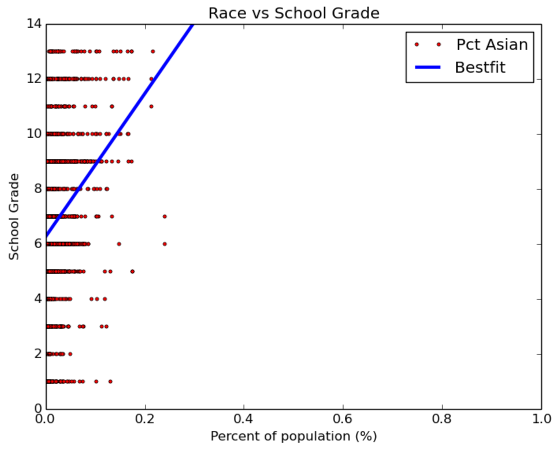

Figure 9: Percentage of students with an Asian background versus school grade.

7. Remaining Tasks

There are two remaining tasks left for the project. The first is to finish mining the 2011 results and compare that to the 2012 results. By comparing the two years changes in correlations and rules can be discovered. Subtasks for the 2011 data include: frequent patterns, rules, and visual verification. After these are completed they can be compared to the 2012 results. The second task is to create a classifier that uses a school location to determine the school grade. Subtasks for the classifier include: Splitting the data into training and testing set, picking one or more classifiers, and feature engineering.

8. References

[1] Mandinach, Ellen, Margaret Honey, and Daniel Light. *A Theoretical Framework for Data-Driven Decision Making*. Rep. San Francisco: AERA, 2006. Print.

[2] Salpeter, Judy. *Data: Mining with a Mission.* Rep. N.p., 15 Mar. 2004. Web. 20 Feb. 2015.

[3] Survey on Frequent Pattern Mining Bart Goethals HIIT Basic Research Unit Department of Computer Science University of Helsinki

[4] M. J. Islam, Q. M. Jonathan Wu, M. Ahmadi and M.A. Sid-Ahmed, “Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers”, 2007 International Conference on Convergence Information Technology, IEEE DOI 10.1109/ICCIT.2007.148

[5] An Implementation of ID3 --- Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of

New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, Australia

[6] The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit- By Gordon B. Dahl and Lance Lochner

[7] How Does Parents' Education Level Influence Parenting and Children's Achievement? - Pamela E. Davis-Kean University of Michigan

[8] Achievement Gaps U.S. Department of Education NCES 2009-455, How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress Statistical Analysis Report

[9] Academic Achievement of Children in Single Parent Homes: A Critical Review-by Mark S. Barajas Western Michigan University

Appendix A: School Grade Distribution

School Grade	Total
1	96
2	40
3	100
4	39
5	185
6	539
7	182
8	91
9	270
10	89
11	34
12	109
13	37

Appendix B: Discretized Data

Attribute Breakdown				
Attribute	Bin 1	Bin 2	Bin 3	Bin 4
School Grade	0-5.5	5.5-6.5	6.5-9.5	9.5+
PCT American Indian	0-0.002	0.002-0.005	0.005-0.0010	0.0010+
PCT Asian	0-0.005	0.005-0.015	0.015-0.035	0.035+
PCT Black	0-0.004	0.004-0.011	0.011-0.031	0.031+
PCT Hispanic	0-0.110	0.110-0.210	0.210-0.450	0.450+
PCT White	0-0.38	0.38-0.66	0.66-0.80	0.80+
PCT Pacific Islander	0-0.0002	0.0002-0.0020	0.0020-0.0040	0.0040+
PCT Free Meal	0-0.2	0.2-.4	0.4-0.65	0.65+
PCT Single Family	0-0.17	0.17-0.25	0.25-0.35	0.35+
PCT Family Income $60000+	0-0.4	0.4-0.55	0.55-0.7	0.7+
Education Bachelors or Higher	0-0.19	0.19-0.33	0.33-0.51	0.51+