Summer 8-10-2019

PREDICTING ABSENTEEISM OF FEMALE STUDENTS IN ALABAMA

Funmilola Okelana
faokelana@my.harrisburgu.edu

BY

FUNMILOLA OKELANA

BSc, University of Ibadan, 2002

MBA, Hult International Business School, 2013

MSc, Information Systems Engineering and Management, Harrisburg University of Science and Technology

MASTERS THESIS

Submitted in partial fulfillment of the requirements for

the degree of Master of Science

in the Graduate School of

Harrisburg University of Science and Technology

2019

Accepted in partial fulfillment of the requirements for

the degree of Masters of Science

in the Graduate School of

Harrisburg University

2019

August 10, 2019

Dr. Siamak Aram, Chair and Faculty Advisor

Analytics Department, Harrisburg University

Abstract

Students are chronically absent when they miss at least 15 days of the school year. Past researchers have identified income and environment as factors that affect school absenteeism. Alabama is a poor state with a high crime rate. The hypothesis for this research is that the absenteeism of female students in Alabama is high. Do we reject or fail to reject this hypothesis. If we fail to reject this hypothesis, then what other factors can affect absenteeism in schools? How can we best predict the absenteeism of female students in Alabama? What is the effect of bad data on predictive models? This research aims to answer the above questions.

Machine learning has proven to be one of the best methods in making good predictions for better decision-making. Different machine learning models are used for making predictions, but the outstanding question is how to identify the best model to predict dependent features. Are features very essential when considering the type of model to be used for prediction?

This research aims to analyze and compare the percentage of prediction and accuracy of prediction using supervised machine learning models while considering features. Based on findings, the recommendation for the best model to predict female students' absenteeism in Alabama school districts was made. Also, the hypothesis that the absenteeism of female students in Alabama is high was rejected. This research is limited to only supervised machine learning models. Information on male students in Alabama is not included in this research

Keywords: Absenteeism, Machine Learning, Predictive Model, Naïve, Random Forest, Boruta

Table of Contents

List of Figures

List of Tables

Relationship to Curricular Practical Training (CPT)

I work in an insurance company as a Business Analyst- Data management intern for my CPT. Masters in Analytics program had helped me in performing my roles better. I can now analyze data better using R. I have acquired good skills to quickly analyze and generate reports to meet management requirements.

Introduction

Research Overview

Education as a leveler can only be achieved when parents, guardians, teachers, and government endeavor to ascertain that students are in school daily and receive the supports they need to study and succeed. (US Department of Education, 2019). The problem of absenteeism stands as a detriment to the purpose of education. Therefore, it is very critical that the issue of absenteeism in schools is well addressed. Previous research showed that school absenteeism rates vary by student health, family income, and environment.

This research investigated the absenteeism of female students in the state of Alabama. This research is expected to reject or fail to reject the hypothesis that income and environment affect school absenteeism while paying attention to just female students in Alabama. While some researchers have used income, health, and environment as significant factors influencing school absenteeism, this research used data that included other features like school enrollment based on gender, suspension from school, gifts and talents, disabilities, etc., to predict absenteeism of female students in Alabama.

Table 1: Alabama State Economy Ranking (Suneson, America's Richest and Poorest States, 2019)

Economy	**Ranking**
Poorest State Ranking	5th Poorest State in the USA
Median Household Income	$48,123
Population	4,874,747 (24th Highest)
Unemployment rate	4.4 percent (22nd highest)
Poverty Rate	16.9 percent (6th highest)

All economy indices stated in table 1 above proved that Alabama could be ranked as one of the poorest states in the USA and, as such, have high tendencies to lack resources to provide adequate educational resources to students

Alabama is also ranked as the 41st safest state in the USA. The state's violent crime rate is nearly two times higher than the national average. (Karksen, 2019). Table 2 below shows some of the indices that classified Alabama as one of the top violent crime states in the USA.

Table 2: Violent Crime in Alabama (Gore, 2019)

Crime (per 100,000)	**Crime Rate/Ranking**
Violent Crime	532
Total Murder	407 (17th most)
Imprisonment Rate	790 (4th highest)
Poverty Rate	17.1% (7th highest)

Table 3 below shows Alabama have a poor public-school ranking

Table 3: Alabama Public School Ranking (Samuel Stebbins and Thomas C. Frohlich, 2018)

School System	**Ranking**
High school graduation rate	87.1% (16th highest)
Public School Spending	$10,142 (14th lowest)
8th grade NAEP proficiency	17.2%(math) 25.6% (reading)
Adults with at least a bachelor's degree	24.7% (7th lowest),

Adults 25-64 with incomes at or above the national median	46.3% (13th lowest)
Pre-k enrollment	42.9%

Records in table 1- 3 above aid in iterating that the hypothesis that absenteeism of female students in Alabama state is expected to be high.

As stated above, the environment of an individual plays a significant role in the individual's behavior. Can we then expect to see high school absenteeism in female students in Alabama?

In this research, supervised machine learning models were used to predict the absenteeism of female students in Alabama. Analysis of how the features used in this research agree with other researchers' findings regarding how income and environment affect school absenteeism. R statistical tool was used to carry out this analysis.

Problem Statement and Justification

Problem and Purpose Statement

The problem is that many machine learning models can be used for prediction, but which model is the best? Are features important when choosing a machine learning model? How do we identify essential features that can help to predict absenteeism of female students in Alabama?

This research aims to analyze and compare the percentage of prediction and accuracy of prediction using supervised machine learning models while considering features. Based on findings, a recommendation for the best model to predict absenteeism of female students in Alabama school districts was made.

This research helped prove or disproof the hypothesis that income and environment are essential factors that affect school absenteeism. It also helped predict the percentage of absenteeism of female students in Alabama. Government agencies, teachers, and guardians can use this information to research further on how to mitigate absenteeism by paying attention to the features used for this prediction.

Research Questions and Objectives

1. Does income affect school absenteeism?
2. Does the environment, including crime rate, affect school absenteeism?
3. What are the features that can best predict absenteeism of female students in Alabama?
4. What machine learning model can best be used for this prediction?

Alabama is a poor state with a high crime rate; therefore, the hypothesis of this research is that school absenteeism is expected to be high. Also, since data used for this research is a continuous variable with a high range, a random forest machine learning model is expected best to predict absenteeism of female students in Alabama state.

This research is quantitative analysis research. Measures of the percentage of predictions and accuracy of models were carried out, and the result was used to draw conclusions in this research. The deliverable for this research is a recommendation of the best model that predicts absenteeism of female students in Alabama based on the accuracy of the model and the features that best predict absenteeism. This research is limited to Supervised Machine Learning models

Literature Review

This review shows samples of research that pointed to the fact that health, family, and environment play a huge role in school absenteeism. This review also highlights some of the methods and theories used in carrying out these researches. Research limitations and areas of research improvement are also discussed. Lastly, this review states how past research is related to this research and intends to improve existing research.

Absenteeism is a plague that had been in existence for decades in schools in the USA. Many researchers have concluded that school absenteeism is primarily associated with students' health, family, and environment. Human beings are largely shaped or influenced by their environmental background. Interactions at distal levels directly impact an individual's behavior. Therefore, it was established that children with health challenges, family conflicts, and poor homes have difficulty attending school regularly. Furthermore, a deep look at community influence on students proved that communities with poor school climate ratings and high crime rates have challenges with school attendance. (Hilary Stempel, Mandy A. Allison, Michael Bronsert, Matthew Cox-Martin, & L. Miriam Dickinson, 2017).

Resolving absenteeism in schools had been a long battle that is yet to be fully mitigated. A group of researchers (Knollmann et al., 2019) investigated a comprehensive assessment of school absenteeism. These researchers found out that multiple risk factors affect school attendance. Students suffering from depression and various types of anxiety have a high tendency to have poor school attendants. Students who dislike their teachers, subject(s), school, or their peers have

challenges with school attendants. Students with family problems have also resulted in missing school. (Knollmann, Reissner, & Hebebrand, 2019)

Another set of researchers looked at the problem of absenteeism from another angle. They looked at the effects of illness and income on school absenteeism. Authors including (Berendes et al., 2019) argued that children's school absenteeism could be mitigated if communicable diseases, including respiratory and diarrheal illness, are effectively controlled. These researchers used survey data from the National Health Interview Survey (NHIS) to analyze associations among income, gastrointestinal disease, and absenteeism**.** The result of the research showed that as income decreased, gastrointestinal and respiratory illnesses increased. The study's outcome iterates the conclusion of other researchers that income is a critical factor that affects absenteeism. If parents of sick children have substantial income, they will afford comprehensive medical treatment for their sick children and reduce school absenteeism. (Berendes, Andujar, Barrios, & Hill, 2019)

(Stempel et al., 2017) conducted research “to examine the association between chronic school absenteeism and adverse childhood experiences (ACEs) among school-age children” (p.837). They used logistic regression, summed ACE score, and latent class analysis to examine the relationship between adverse childhood experience and school absenteeism. The result showed that 4,1% of sample students who experienced chronic school absenteeism were at some point exposed to neighborhood violence, and students’ exposure to neighborhood violence was the only significant point for their ACE score. It was therefore concluded that chronic school absenteeism of school-age children is closely associated with ACE exposure. Therefore, the researchers recommended that collective efforts of parents, pediatricians, public health partners, school

teachers, and administrators are needed to combat ACE exposure and thus reduce school absenteeism. (Hilary Stempel, Mandy A. Allison, Michael Bronsert, Matthew Cox-Martin, & L. Miriam Dickinson, 2017)

Data is key to any research analysis. Therefore most researchers begin their research with data collection. The authenticity of your data determines the accuracy of the result of your research. The study of school absenteeism is such that it requires researchers to use secondary data. This is because it will be almost impossible for the researcher to gather all relevant information needed for their research analysis by themselves. Different researchers have used various secondary data based on their research needs. These data were collected via surveys, and relevant government agencies carried out the surveys. These types of data are actual data with a high level of integrity. Analysis results from these types of data can therefore be trusted. Examples of government agencies are National Center for Health Statistics and The National Center for Children in Poverty.

Once relevant data is collected, the next step is to analyze the data. The type of analysis to be carried out is highly based on the type of data collected. Researchers have used both qualitative analysis and quantitative analysis in carrying out their research on school absenteeism. Others have used a combination of both. For instance, Stempel et al. (2017) used statistical analysis (quantitative analysis) to research school absenteeism. Complex statistical analyses like descriptive analysis and multivariable logistic regression were used to carry out the research. These researchers used the above-named statistical analysis to build three models to analyze adverse childhood experiences (ACEs). They created an individual ACE model, summed ACE score model, and ACEs classes model using latent class analysis. The latent class analysis uses a missing at random system to account for missing values when respondents fail to answer all the ACE

questions. This method also helps to remove outliers from data, thereby providing better analysis results. Mplus statistical program was used to carry out the latent class analysis. (Hilary Stempel, Mandy A. Allison, Michael Bronsert, Matthew Cox-Martin, & L. Miriam Dickinson, 2017)

(Berendes et al., 2019) when comparing school absenteeism with illness and income used linear and logistic regression models, P-values <0.05 was considered statistically significant for adjusted and unadjusted age (Berendes, Andujar, Barrios, & Hill, 2019)

Although various researchers have used CRDC data to carry out different analyses, including descriptive analysis on school absenteeism, there is no research on predicting school absenteeism at various US states using the CRDC data. Therefore, research on the best predictive model to predict school absenteeism using CRDC has yet been performed. This research is vital to the government, parents, pediatricians, and school teachers because predicting school absenteeism at Alabama state can help identify essential factors affecting school absenteeism and enhance the need to address school attendance at Alabama state. This research can also be used as a research model for predicting school absenteeism in other US states.

For this research, predictive models were built to predict absenteeism of female students in Alabama state using CRDC data. This research failed to include information on male students in Alabama state.

Research Design and Methodology

Proposed Solution Overview

Machine learning models are one of the best models for making a prediction. However, the accuracy of prediction varies with models. The choice of model for making predictions depends mainly on the type of data and the features used for your model. The proposed solution for this research is a recommendation of the best-supervised machine learning model that best predicts absenteeism of female students in Alabama based on the features and accuracy of the model.

Research Design Overview

This research is quantitative analysis research. Quantitative analysis involves a systematic approach to resolving an inquiry through exploration, quantification, and confirmation. Therefore, this research involves exploring data and using data to build predictive supervised machine learning models and measure the accuracy of the models. These supervised learning Machine learning models are the Decision Tree Model, Support Vector Model, Naïve Bayes Model, and Random Forest Models.

A model with the highest accuracy of prediction was recommended as the best model to predict the absenteeism of female students in Alabama state. Also, this research aimed to reject or fail to reject the hypothesis if the prediction of absenteeism of female students in Alabama is high.

Secondary data retrieved online from 20115-2016 Civil Rights Data Collection (CRDC) was used for this research. This data was collected from a space of all public local educational agencies (LEAs) and schools, including charter schools, long-term sheltered children justice facilities, alternative schools, and schools for students with disabilities. (CRDC, 2018). Data used for this research was limited to relevant and available information on Alabama school districts.

The research question and objective to be addressed in this research paper is, what is the best machine learning model that best predicts absenteeism of female students in Alabama? How does this prediction confirm the hypothesis that income and environment affect school absenteeism?

Research Design Setting

The setting of this research design includes the following steps:

1 Download data from the CRDC website.

2 Understood the data - After downloading data, it was imperative to observe the data visually. By doing this, data type, size, and type of variable were identified. After visually observing the data, R statistical tool was used to understand the data further. Data was loaded into R, and code "str" was used to understand the data type. Data has 96,360 observations and 1,836 variables. This is extensive data because it contains data from all 50 states of the USA. Because this research is restricted to only Alabama state and only relevant features are needed, research data was reduced to 1,400 observations and 73 variables. It is worth noting here that variables were selected based on variables that have fewer missing data and by guessing.

3 Data Preprocessing- Performed data reprocessing by removing inconsistent, incomplete, or noisy data from source data. Data reprocessing was done by using excel, and R. CDRC data was reprocessed by carrying out the following procedures:

- Identified columns with missing values or inconsistent data. This was done by visually observing the data and using R. Code "str" provided information on

features that have missing data. The research data is made up of numbers. Target variable is an integer and a continuous variable. Likewise, code "sum (is. na (CRDC_Alabama)" was used to identify the number of missing data in the research data. One thousand two hundred sixty-four data were missing in the original data set. From the source data dictionary, the rate of missing data for this research data is only about 5%. Missing data in this source data is represented by -7 (when data element that was supposed to be reported were not reported), -5 (when LEA is unable to report required data element but have an action plan to report such data in the future), -9 (when the data element is not applicable). (CRDC, 2018).

- Some of the data columns have about 60% missing data. Those columns were removed in excel before uploading the file into R. Only columns with very few missing data were used for this research. Some missing data were replaced by 0 (if a lot of the data in the column is zero), while the mean of the column replaced others using the data binning method. The replacements were done using excel, and before loading the file into R. After loading data into R, sum (is.na (CRDC_Alabama)) was used to identify the total sum of missing data in the uploaded file. Seventy-two missing data were found. colSums (is.na (CRDC_Alabama)) was used to identify columns with missing data. All the columns were found to have missing data was found in all co. These missing data were spaces after each data on the excel data. CRDC_Alabama.C<-na.omit(CRDC_Alabama) code was used to clean the missing data. After carrying out the above data preprocessing steps, the data for this research was considered clean for analysis.

4 Data Exploration- Performed data exploration to graphically represent data and produce a statistical summary of the data. Data exploration was done using R. Statistical summary shows that target variable has Min data= 0, Ist Quart= 13, Median = 28.50, mean =46.65, 3rd Quart = 51, Max is 838. The summary also indicated that the target variable has a wide range of data between 0 and 838 and is a continuous variable.

Data visualization was carried out for analysis of correlation or relationship of independent variables with the target variable. The following steps were used to carry out data exploration:

I. Treated factors as factors using R.

II. The five-number summary (Minimum, 1st Quartile, Median, 3rd Quartile, and Maximum) used a boxplot to display the data distribution. An important reason for using this plot was to identify the outliers. When boxplot was carried out on most variables, it was discovered that most of the variables have outliers, including the target variable

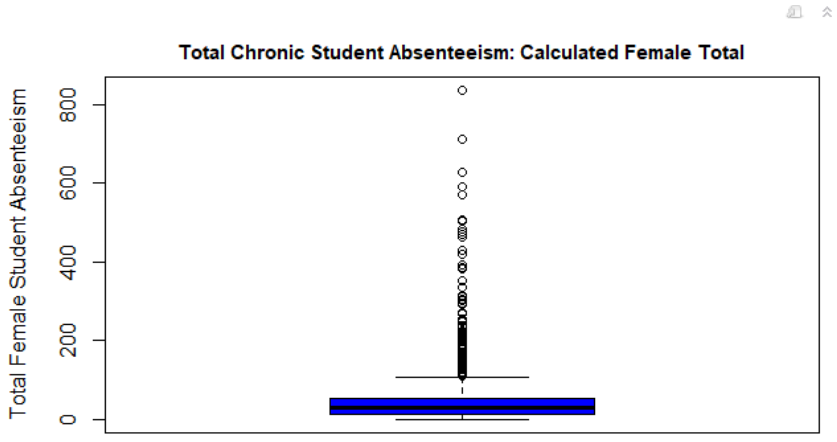


Figure 1: Box plot of Total Chronic Student Absenteeism: Calculated Female Total

Figure 1 shows a box plot with the target variable having many outliers.

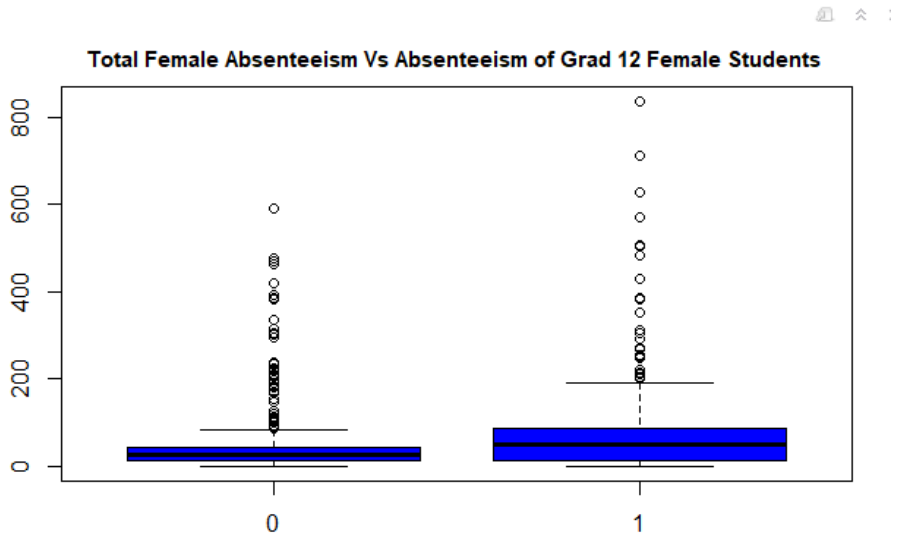


Figure 2: Box plot of Total Female Absenteeism Vs. Absenteeism of Grad 12 Female Students

Comparing absenteeism of total female students with grad 1-12 female students, Figure 2 shows that distribution is very similar with both having a large number of outliers.

Box plot of all the other variables was plotted using Boruta, as shown in figure 14 below (see figure 14)

III. Used Density plot to show normalization of the target variable (Total Chronic Student Absenteeism: Calculated Female Total).

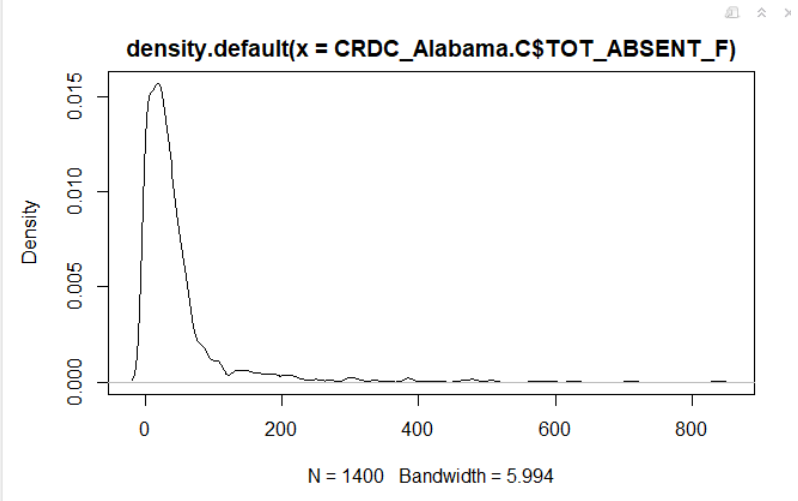


Figure 3:Density Plot

The density plot also confirmed that the data is not normally distributed. It is skewed to the right with many outliers.

IV. Used Histogram to also see the distribution of the data

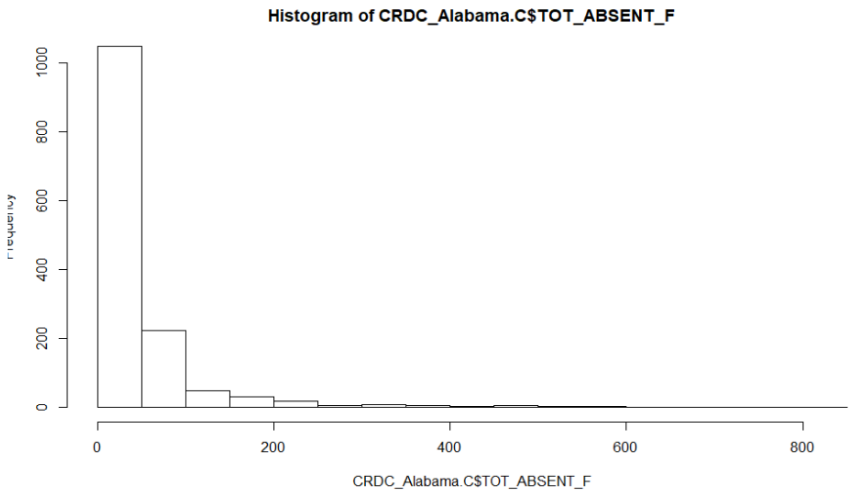


Figure 4: Histogram of distribution of target variable

Figure 4 above clearly shows the skewness of the data to the right. Logarithms were used to correct this skewness.

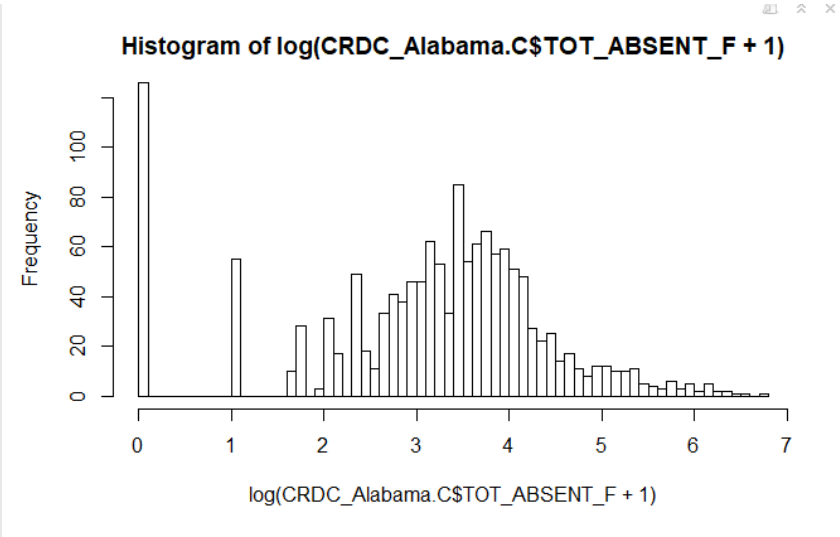


Figure 5: Log of Histogram Distribution

Figure 5 shows that there are some spikes in the distribution of the data. To better understand the data, a frequency distribution of the data using R was carried out. This distribution shows that total female students with 0 number of absenteeism have the highest frequency of 573(40.9%), two times absenteeism have a frequency of 399 (28.5%), four times absenteeism have a frequency of 148(10.57%) 5 times absenteeism have a frequency of 94 (6.71%). With this understanding, it implied that very few students are absent in class.

V. Plotted scattered plots to show relationships between a few of the variables in the dataset target variables. This plot was used to determine if there are any relationships between the variables, the type of relationships, and how strong the relationships are. The scattered plots below showed that there are no linear relationships between most of the variables.

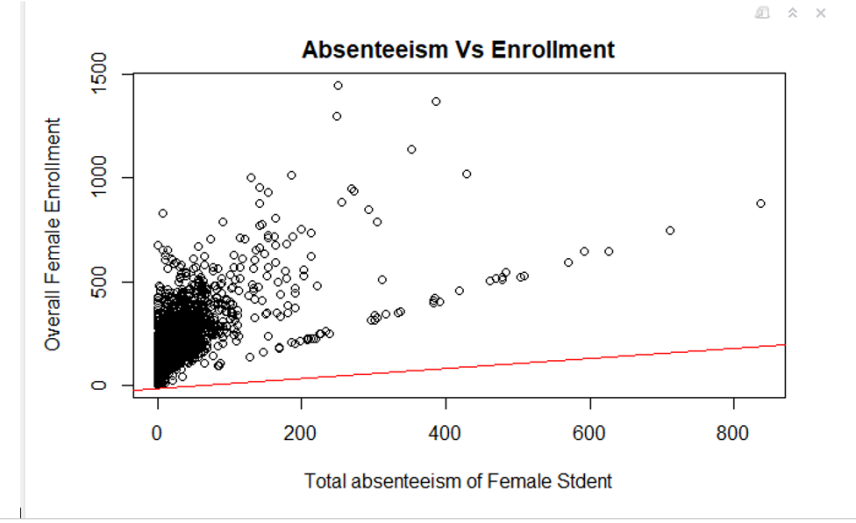


Figure 6: Scattered Plots- Absenteeism Vs. Enrolment

A plot of Overall Female Enrollment with Total absenteeism of Female students showed a non-linear relationship between the two variables. The relationship is also very weak.

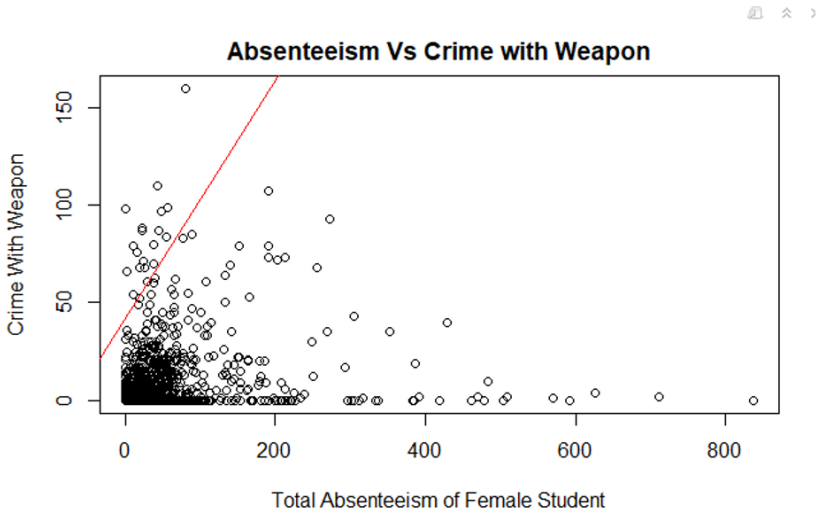


Figure 7: Scattered plot: Absenteeism Vs. Crime with Weapon

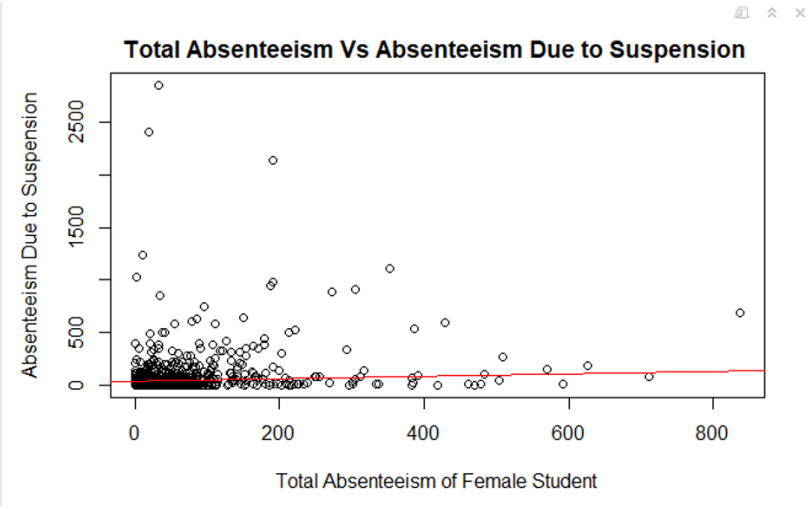


Figure 8: Scattered Plot: Absenteeism Vs. Absenteeism due to suspension

Figures 7 and 8 also showed that the variables do not have linear relationships with each other. It can be inferred from the scattered plots above that most of the variables have weak relationships. Because of these findings, a correlation plot was plotted to show which variables have weak or no relationship with each other.

VI. Plotted Correlation plots

For this research, the initial selection of variables for analysis and building model was based on guesses and columns with minimum missing data. The correlation metric was then carried out on the initially selected variables. The correlation metric was plotted in groups to prevent overplotting. The Features with solid and perfect correlation were dropped. Features with slight to high correlation were thus used for building prediction models. Features with a correlation between +/- 0.5 to +/-0.8 were utilized for creating prediction models. Features with a correlation greater than +/- 9.0 were dropped to prevent multicollinearity. Spearman Correlation Coefficient was used because the variables are all continuous variables with non-linear relationships.

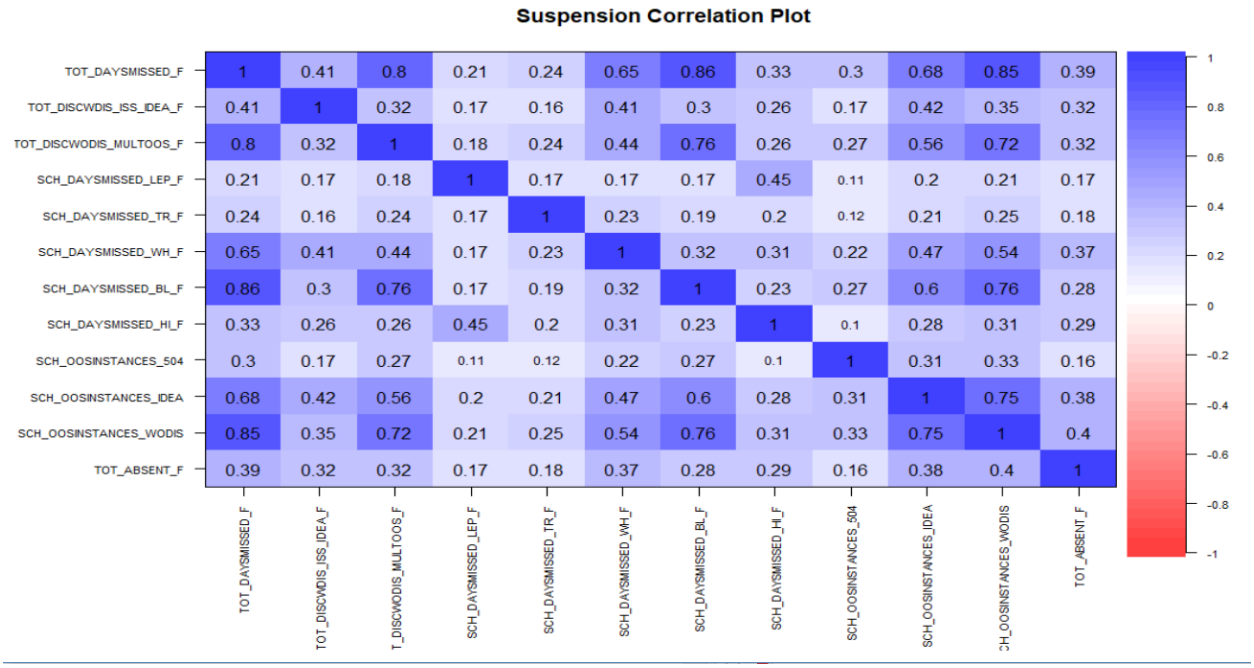


Figure 9: Suspension Correlation Plot

For instance, SCH_OOSINTANCES_504 and SCH_DAYMISSED_TRF have a weak relationship with other variables and so were dropped. TOT_DAY_MISSED_F and SCH_DAYMISSED_BL_F variables have a strong correlation and so were dropped.

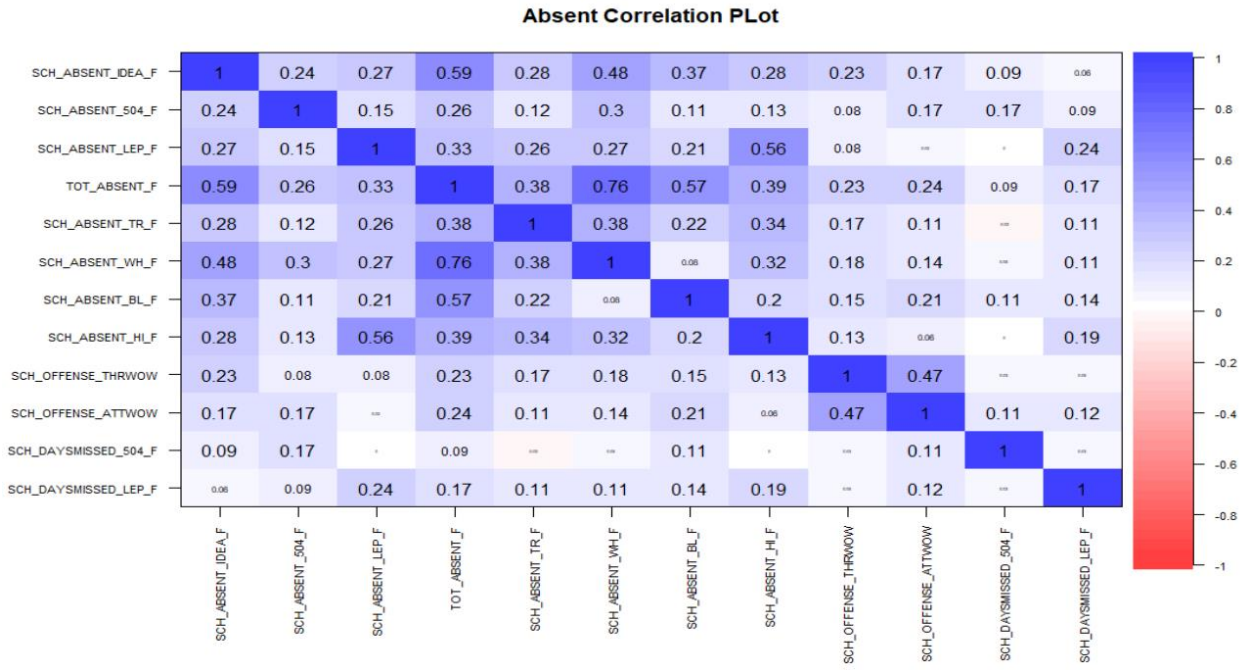


Figure 10: Absent Correlation Plot

Only SCH_ABSENT_BL_F, TOT_ABSENT_F, and SCH_ABSENT_HL_F variables were selected because they have a good relationship with each other.

Other plots that showed a correlation between the variables are as shown in the appendix below. Correlation plots with the Boruta plot were used in selecting the best features for building the prediction models for this research analysis.

5 Trained model on the data – Trained the data by dividing observations into two. 80% is train data and 20% test data. Both train and test made up 100% observation. Trained data was used to build the models.

6 Built Prediction Models

Model 1- Random Forest Model: The target variable for analysis is a continuous variable. Random Forest Model was chosen because it works well with both categorical and continuous variables.

Random Forest also mitigates the effects of multicollinearity and missing data. Thus, features that are not relevant to the models were removed. R was used to carry out this model. Boruta, caret, mlbench, and randomforest libraries in R were used to build this model. Boruta library in R was used to identify features that are important and features that are not important for the Random Forest Model. Figure 14 below shows the important and unimportant features. Green indicates the features are important, while red indicates that the features are not necessary.

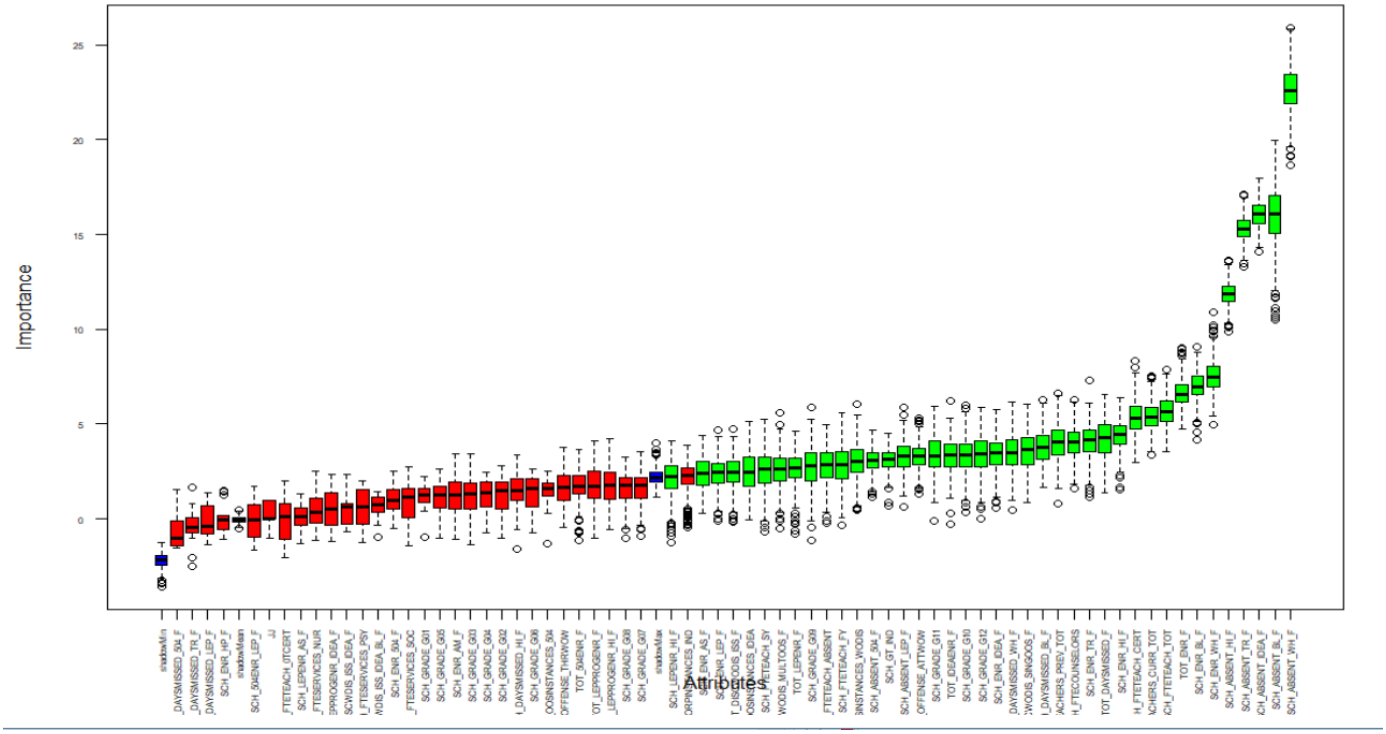


Figure 11: Boruta Plot

Model 2- Naïve Bayes Model. Nine figures were used in building this model. This is because only nine features have fair to significant correlation amongst themselves. 80% training data was used in making the model.

7 Evaluated Model- Validated the overall performance of the model using test data. Also used correlation of feature to improve the accuracy of our model by randomly using different correlations values until one that best improved the model was found.

8 It provided a recommendation of the best model to predict absenteeism of female students in Alabama based on accuracy.

9 Drew conclusions on my research

Data used for this research is secondary data. Data is sourced by downloading the data from the CRDC website. This data is used because it is actual data collected by CDRC from a two-yearly survey at the request of the US Department of Education, Office for Civil Rights (OCR). This data is a valuable resource for government, educators, school officials, parents, researchers, and anyone who seek data on student fairness and opportunities (CRDC, 2018)

The survey participants for the data collection are educators and parents of students from all public local educational agencies (LEAs) and schools, including long-term secure children fairness amenities, charter schools, alternative schools, and schools working with children with disabilities.

Quantitative analysis was carried out on the accuracy of the models. Percentage of predictions and accuracy of models were measured. The model with the highest percentage of accuracy was chosen as the recommended model for absenteeism of female students in Alabama school districts

Key Outputs and Deliverables

The key output of this research is a recommendation of the best prediction model for predicting absenteeism of female students in Alabama based on features. Also, confirmation to reject or fail to reject the hypothesis that income and environment affect school absenteeism was determined. Personal and sensitive data were excluded from the research data.

Results and Discussion

Data

There are 72 variables and 1400 observations in the research dataset. Statistical summary of target variable- "Total Chronic Student Absenteeism: Calculated Female Total" shows that target variable has Min data= 0, Ist Quart= 13, Median = 28.50, mean =46.65, 3rd Quart = 51, Max is 838. The summary also indicated that the target variable has a wide range of data between 0 and 838 and is a continuous variable. Statistical summary of all other variables showed that most of the variables have a wide range. For instance, the variable indicating out-of-school days due to school suspension for black females has min=0 and max= 2714.

The statistical summary and various plots of these variables indicated that the variables do not have a linear relationship, and the data is not normally distributed. The data have too many missing values and too many outliers. The correlation plot indicated that most of the variables are either not correlated or have weak correlations. Only a handful of variables have any significant correlation. Best features are included in the appendix

Findings from Prediction Models

1. RandomForest Model:

Boruta in RandomForest showed that 40 independent variables were important while 31 independent variables are not necessary. Built model with only important featured, and the result showed: Number of trees:500; No. of variables tried at each split: 23; Mean of squared residuals: 319.5303; % Var explained: 93.21. The model failed to predict the probability of female absenteeism, and thus accuracy was "NA."

2. Naive Bayes Model:

It was discovered that although 40 independent features were essential, only about nine features have fair to significant correlation (0.4- 0.8) amongst themselves. Plotting the Naive Bayes model using only the nine fairs to good correlated variables showed the following probabilities

Table 4: Probability of Female Student Absenteeism in Alabama

Number of times of Absenteeism by group	Probability
0-1	8%
1-2	3%
2-3	1%
3-4	3%
5-6	2%
11-12	3%
17-18	3%
Others (up to 263)	< 1%

This result shows that the probability of high absenteeism of female students in Alabama state is low. Most female students in Alabama do not miss school more than 15 times a year.

See the appendix for a list of the best features.

3. Support Vector Machine and Decision Tree Models

Although an attempt was made to find predictions using these two models, these models also failed to predict the absenteeism of female students in Alabama State using the CRDC data set.

Conclusion, Summary, and Recommended Future Work

From the research result, the hypothesis that the absenteeism of female students in Alabama is high can be rejected. This result can also be partially inferred from the target variable's frequency distribution, which indicated that about 40% of all female students in Alabama have 0 absenteeism.

It can also be proven from this research that feature selection is highly critical to machine learning. The variables for this research do not have a good correlation, resulting in poor feature selection for the model. Again, most data have outliers that are difficult to eliminate because the data range is tremendously high. Therefore, this research proves that data is vital for any analysis, and poor data selection can lead to a poor research result.

CRDC needs to do a better job in data collection, thereby reducing the number of missing data sets. The agency needs to improve on the quality of its survey questions. Further studies can be done on how to clean better the CRDC data for use on Naïve Bayes, Random Forest models, and Support Vector Machine Models. Latent class analysis may be used to mitigate the effect of outliers in future research. Also, further research on identifying predictive models that can tolerate the features in CRDC data should be carried out. Analysis of absenteeism of male students in Alabama should be carried out to see if we can fail to reject the hypothesis that school absenteeism in Alabama is high.

The limitation of this research is that it failed to compare the accuracy of prediction of the Decision tree model, Random Forest model, the Naïve Bayes model, and Support Vector Machine model. This is because data is terrible, so the predictive model failed to make predictions.

Although this research has limitations, Naïve Bayes Model can be recommended to be the best model of the four models ran. This is because only the Naïve Bayes model was able to generate probability based on the features selected.

It should be noted that the results of this research can be flawed because of the bad data used for this analysis.

References

Berendes, D., Andujar, A., Barrios, L., & Hill, V. (2019, March 09). Associations Among School Absenteeism, Gastrointestinal and Respiratory Illness, and Income - United States, 2010-2016. *MMWR. Morbidity And Mortality Weekly Report, 68*(9), pp 209-213.

CRDC. (2018, September). Civil Rights Data Collection. *2015-16 Data Notes*, 1. Washington DC, USA: Department of Eduction.

Gore, L. (2019). *Where Alabama Lands on List of Most Dangerous States*. Retrieved from www.al.com: https://www.al.com/news/2018/01/where_alabama_lands_on_list_of.html

Hilary Stempel, M. M., Mandy A. Allison, M. M., Michael Bronsert, P. M., Matthew Cox-Martin, P., & L. Miriam Dickinson, P. (2017). Chronic School Absenteeism and the Role. *Academic Pediatrics, 17*(8), :837–843.

Karksen, L. (2019, January). *Alabama Ranks in Bottom Ten of Safest States in US*. Retrieved from WHNT NEWS 19: https://whnt.com/2019/01/22/alabama-ranks-in-bottom-ten-of-safest-states-in-u-s/

Knollmann, M., Reissner, V., & Hebebrand, J. (2019, March). Towards a Comprehensive Assessment of School Absenteeism: Development and Initial Validation of the Inventory of School Attendance Problems. *European Child & Adolescent Psychiatry, 28*(3), pp. 399-414. doi:10.1007/s00787-018-1204-2.

Mohammed, F. O. (2019). *Predicting Absenteeism Using Mutiple Machine Learning Models.* Anly 530 final Project.

Samuel Stebbins and Thomas C. Frohlich, 2. W. (2018, February). *Geographic Disparity: States with the Best (and Worst) Schools*. Retrieved from

https://www.usatoday.com/story/money/economy/2018/02/08/geographic-disparity-states-best-and-worst-schools/1079181001/

Suneson, G. (2018, October 4). *America's Richest and Poorest States*. Retrieved from 24/7 Wall St.: https://247wallst.com/special-report/2018/10/04/americas-richest-and-poorest-states-6/3/

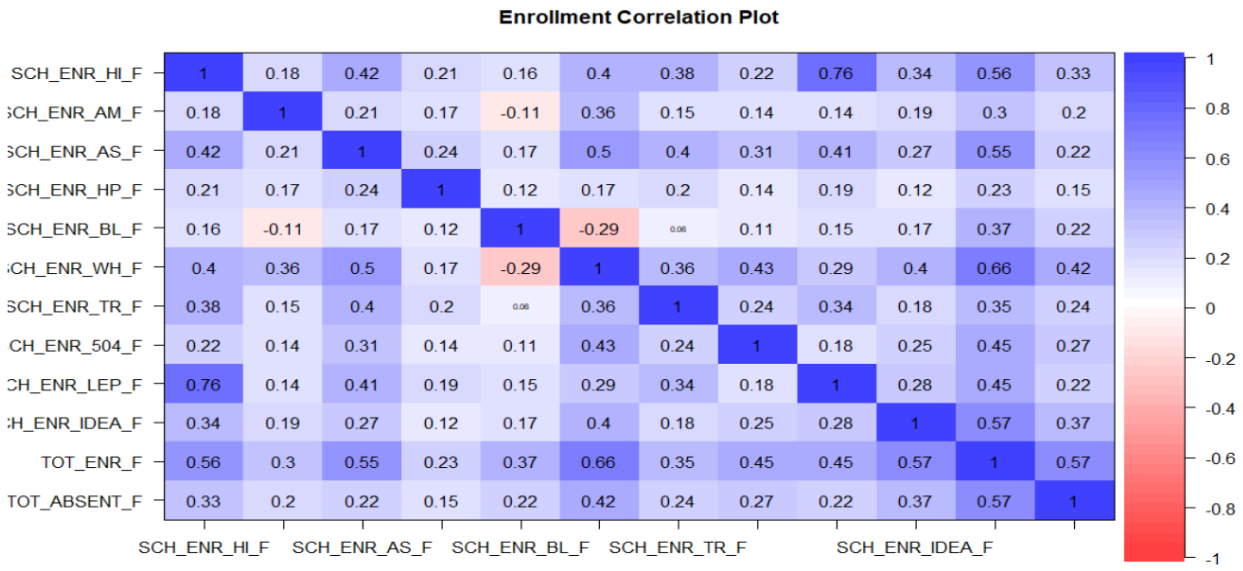
Suneson, G. (2019, October 4). *America's Richest and Poorest States*. Retrieved from 24/7Wall St: https://247wallst.com/special-report/2018/10/04/americas-richest-and-poorest-states-6/3/

US Department of Education. (2019, January). *Chronic Absenteeism in the Nation's Schools*. Retrieved from www2.ed.gov: https://www2.ed.gov/datastory/chronicabsenteeism.html#intro

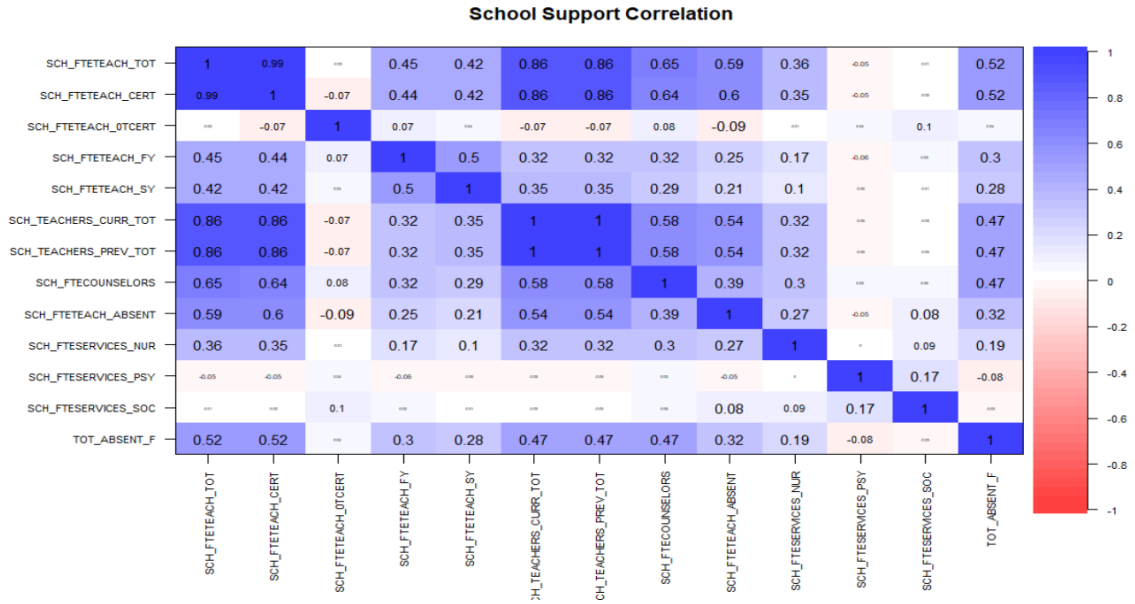
Appendix

Frequency distribution of target variable

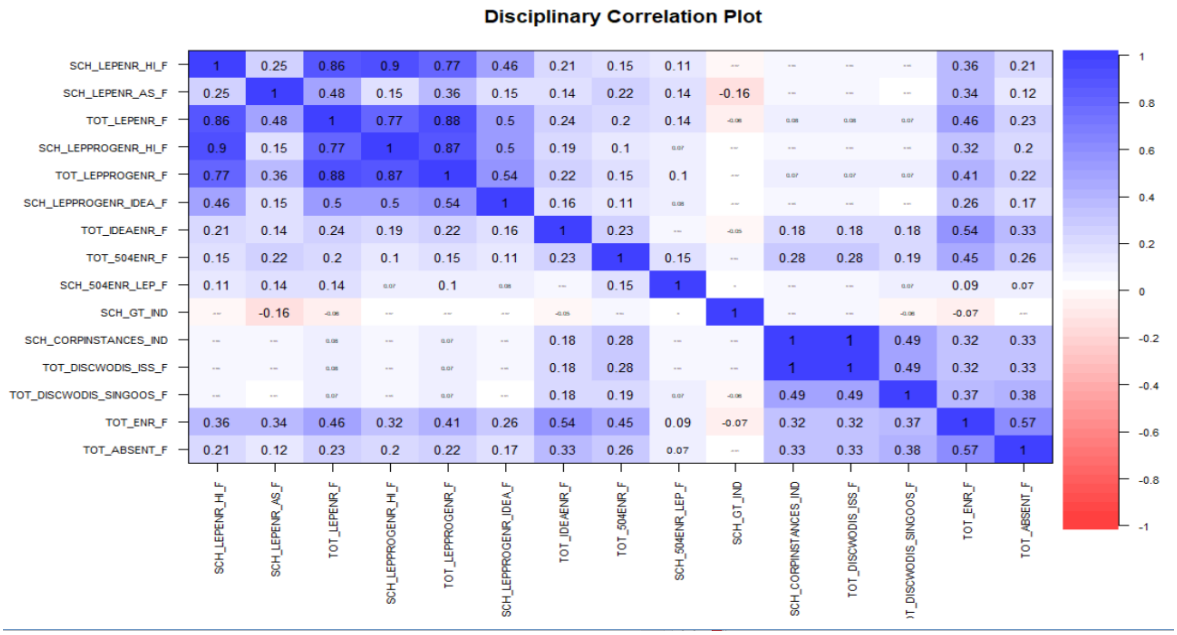
	frequencies	percentage	cumulativepercentage
0	573	40.92857143	40.92857
2	399	28.50000000	69.42857
4	148	10.57142857	80.00000
5	94	6.71428571	86.71429
6	18	1.28571429	88.00000
7	58	4.14285714	92.14286
8	20	1.42857143	93.57143
9	12	0.85714286	94.42857
10	23	1.64285714	96.07143
11	6	0.42857143	96.50000
12	6	0.42857143	96.92857
13	7	0.50000000	97.42857
14	4	0.28571429	97.71429
15	3	0.21428571	97.92857
16	7	0.50000000	98.42857
17	2	0.14285714	98.57143
18	4	0.28571429	98.85714
19	4	0.28571429	99.14286
20	1	0.07142857	99.21429
21	1	0.07142857	99.28571
22	1	0.07142857	99.35714
23	1	0.07142857	99.42857
27	1	0.07142857	99.50000
29	1	0.07142857	99.57143
30	1	0.07142857	99.64286
31	1	0.07142857	99.71429
34	1	0.07142857	99.78571
39	1	0.07142857	99.85714
41	1	0.07142857	99.92857
114	1	0.07142857	100.00000



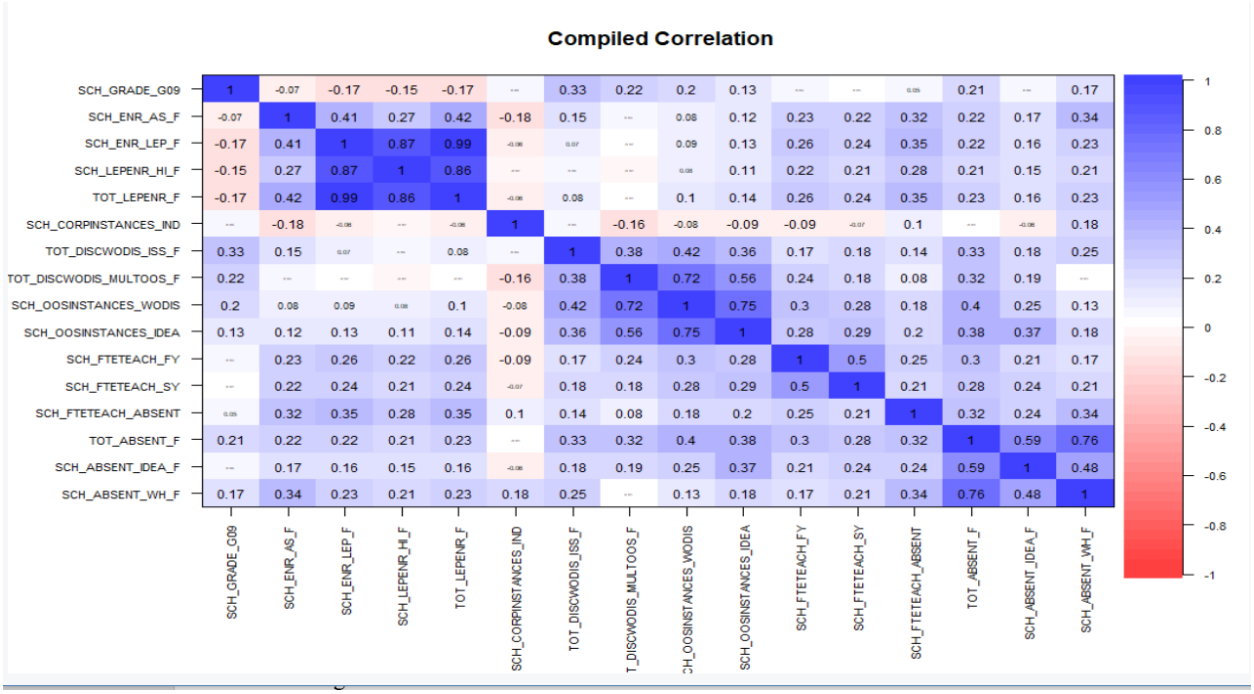
Enrollment Correlation Plot



School Support Correlation plot



Complied Correlation plot



Boruta Decision- Decision on importance of variable

	decision
JJ	Rejected
SCH_GRADE_G01	Rejected
SCH_GRADE_G02	Rejected
SCH_GRADE_G03	Rejected
SCH_GRADE_G04	Rejected
SCH_GRADE_G05	Rejected
SCH_GRADE_G06	Rejected
SCH_GRADE_G07	Rejected
SCH_GRADE_G08	Rejected
SCH_GRADE_G09	Confirmed
SCH_GRADE_G10	Confirmed
SCH_GRADE_G11	Confirmed
SCH_GRADE_G12	Confirmed
SCH_ENR_HI_F	Confirmed
SCH_ENR_AM_F	Rejected
SCH_ENR_AS_F	Confirmed
SCH_ENR_HP_F	Rejected
SCH_ENR_BL_F	Confirmed
SCH_ENR_WH_F	Confirmed
SCH_ENR_TR_F	Confirmed
TOT_ENR_F	Confirmed
SCH_ENR_LEP_F	Confirmed
SCH_ENR_504_F	Rejected
SCH_ENR_IDEA_F	Confirmed
SCH_LEPENR_HI_F	Tentative
SCH_LEPENR_AS_F	Rejected
TOT_LEPENR_F	Confirmed
SCH_LEPPROGENR_HI_F	Rejected
TOT_LEPPROGENR_F	Rejected
SCH_LEPPROGENR_IDEA_F	Rejected
TOT_IDEAENR_F	Confirmed
TOT_504ENR_F	Rejected
SCH_504ENR_LEP_F	Rejected
SCH_GT_IND	Confirmed
SCH_CORPINSTANCES_IND	Tentative
TOT_DISCWODIS_ISS_F	Confirmed
TOT_DISCWODIS_SINGOOS_F	Confirmed
TOT_DISCWODIS_MULTOOS_F	Confirmed
SCH_DISCWDIS_ISS_IDEA_BL_F	Rejected
TOT_DISCWDIS_ISS_IDEA_F	Rejected
SCH_OOSINSTANCES_WODIS	Confirmed
SCH_OOSINSTANCES_IDEA	Confirmed
SCH_OOSINSTANCES_504	Rejected
SCH_DAYSMISSED_HI_F	Rejected
SCH_DAYSMISSED_BL_F	Confirmed
SCH_DAYSMISSED_WH_F	Confirmed
SCH_DAYSMISSED_TR_F	Rejected
TOT_DAYSMISSED_F	Confirmed
SCH_DAYSMISSED_LEP_F	Rejected
SCH_DAYSMISSED_504_F	Rejected
SCH_OFFENSE_ATTWOW	Confirmed
SCH_OFFENSE_THRWOW	Rejected
SCH_ABSENT_HI_F	Confirmed
SCH_ABSENT_BL_F	Confirmed
SCH_ABSENT_WH_F	Confirmed
SCH_ABSENT_TR_F	Confirmed
SCH_ABSENT_LEP_F	Confirmed
SCH_ABSENT_504_F	Confirmed

```
SCH_ABSENT_IDEA_F          Confirmed
SCH_FTETEACH_TOT           Confirmed
SCH_FTETEACH_CERT          Confirmed
SCH_FTETEACH_0TCERT         Rejected
SCH_FTETEACH_FY            Confirmed
SCH_FTETEACH_SY            Confirmed
SCH_TEACHERS_CURR_TOT      Confirmed
SCH_TEACHERS_PREV_TOT      Confirmed
SCH_FTECOUNSELORS          Confirmed
SCH_FTETEACH_ABSENT        Confirmed
SCH_FTESERVICES_NUR         Rejected
SCH_FTESERVICES_PSY         Rejected
SCH_FTESERVICES_SOC         Rejected
```

Best Features

```
'data.frame':  1400 obs. of  8 variables:
 $ SCH_ENR_AS_F       : num  0 0 0 0 2 2 2 2 2 2 ...
 $ SCH_ENR_LEP_F      : num  0 0 0 0 11 5 5 14 5 14 ...
 $ TOT_LEPENR_F       : num  0 0 0 0 25 40 22 69 56 105 ...
 $ SCH_FTETEACH_ABSENT: num  0 0 2 2 11 27 12 27 3 20 ...
 $ SCH_ABSENT_WH_F    : num  0 0 0 0 20 41 8 17 47 29 ...
 $ SCH_ABSENT_IDEA_F  : num  0 0 0 0 2 5 5 5 8 11 ...
 $ TOT_ABSENT_F       : Factor w/ 195 levels "1","2","3","4",..: 1 1 1 1 55
118 50 64 106 95 ...
 $ TOT_ABSENT         : Factor w/ 195 levels "1","2","3","4",..: 1 1 1 1 55
118 50 64 106 95 ...
```