Predicting Chronic Absenteeism in High School with Machine Learning

Kevin Monisit
Toms River High School East
Toms River, NJ
monisitkevin@gmail.com

***Abstract*—**Attendance is important for academic success in high school. Therefore, it is imperative to identify students who are likely to become chronically absent in high school before they enter high school. However, identifying students who are likely to be chronically absent is not as simple as extrapolating absences in middle school. Many confounding variables affect the probability of chronic absenteeism, making the decision process more complicated. Machine learning is used to alleviate this issue by learning from a dataset of students and accurately classifying chronic absentees. The dataset used in this paper consists of longitudinal data of Toms River High School East students who graduated in 2020, containing variables known to be associated with chronic absenteeism. Although constrained with a small dataset, an accurate classification model was produced. This paper gives a foundation for future work in chronic absenteeism with its insight in chronic absenteeism and its machine learning applications.

Keywords*—chronic absenteeism, machine learning, classification*

I. INTRODUCTION

Chronic absenteeism can have adverse effects on students. Being chronically absent in school is associated with reduced rates of post-secondary enrollment and higher chances of dropping out of high school [1, 10]. A study of public schools in Utah found that a single year of chronic absenteeism between 8th and 12th grade was "associated with a seven-fold increase in the likelihood of dropping out," making it a better predictor for dropping out of high school than test scores [1]. Students who drop out of high school are more likely to have adverse outcomes in life, such as poverty, incarceration, employment issues, diminished health, etc. [3, 12]. Moreover, being chronically absent in one year led to a thirteen-fold increase in the chance that the student will be chronically absent the next year [2]. In other words, chronic absenteeism is a slippery slope that can become progressively worse and lead to unfavorable outcomes.

Toms River High School East (TRHSE) is one of three high schools in the Toms River Regional School District, located in Ocean County, New Jersey. In a 2018-19 report, TRHSE's chronic absenteeism rate was 18.4%, which was 4.2% higher than the 14.2% state average [5]. Across the grades, chronic absenteeism rises from 9th grade to 12th grade in TRHSE, but this trend is not a special case in TRHSE. Across the nation and for every race, ethnicity, disability, etc., chronic absenteeism "spikes" in high school [1]. With this knowledge, a question may arise: how can TRHSE, or schools in general, effectively help students before they become chronically absent? Schools can help at-risk students with early intervention. However, a second question may arise from the first: how do we know which student needs early intervention?

Educational data mining techniques have already been used to predict student performance and whether a student will drop out of high school [6, 7, 8]. By analyzing certain attributes of students, like socioeconomic background or grade retention, data mining techniques (machine learning) can accurately predict if a student will not graduate on time or drop out [11]. It is important to note the large number of factors that may contribute to the need of intervention that can make the process of prediction difficult for a human. According to L. Wood et al., after analyzing the Education Longitudinal 2002 Study conducted by the U.S. Department of Education (DoED), dropout rates "vary significantly" depending on individual and environmental factors like socioeconomic status and school size [8]. Other factors, such as self-perception and academic achievement and engagement, contribute to the chances that a student will drop out. With this knowledge, it can be concluded that the prediction of adverse outcomes for students requires a lot of insight into certain attributes and their individual contributions to a prediction; with human analysis, identifying students can be time-consuming and may be inaccurate.

Therefore, it is important to not only predict if a student needs intervention with machine learning, but also rank students based on their risk. In other words, curating a list of the top N students who are most at-risk of chronic absenteeism would be helpful in saving time and resources. In fact, school districts have only limited resources and may only be able to assist N students, so a list of top N students would be crucial [7]. This paper focuses on the prediction of chronic absenteeism in high school before entering high school and does not address the process of predicting whether a student will drop out or experience other negative outcomes, but builds upon the same processes that other studies have done in addressing such issues.

II. RELATED WORK

Prior work in the applications of machine learning in data-driven student intervention has been the basis of this paper's research. Mainly, research using student data mining techniques have focused on the prediction of either student performance or possibility of drop out; this is called educational data mining [6]. By using the most important attributes that contribute to what a model attempts to predict, researchers can better predict adverse outcomes or future academic performance. In a systematic literature review, GPA was identified as an attribute that most researchers used in this setting [6]. This makes sense because it is a numerical value that measures academic performance---a strong indicator for drop out. In [14], Quadril et al. used student attributes, such as parent income and irregularity of absences, in order to predict whether a student will drop out using decision trees (a machine learning algorithm). In another paper, [13] used attributes like the number of times a student raised their hands and the parents' level of satisfaction with the school to predict student performance. However, these attributes may be unrealistic to record on a larger scale, and they also require additional feature engineering to make them useful.

It is important to understand that this paper does not focus on predicting drop out. Rather, this paper focuses on predicting chronic absenteeism. The reason why both drop out and chronic absenteeism is discussed is due to their similarities: both problems consist of predicting a binary label (one value or the other), and both problems are in the same setting. However, predicting chronic absenteeism would require slightly different uses of student attributes. For example, males are more likely than females to drop out of high school. However, the difference in chronic absenteeism rates between males and females are not significantly different [1, 15].

Unlike other related research, I am tasked with using a small dataset due to limited time and resources, such difficulties are described in section III-A. Metrics like AUC and mean empirical risk, found in related research, are used to ensure the reliability of the results of this paper [7, 13].

III. Dataset description and collection

Toms River High School East (TRHSE) is one of three high schools in the Toms River Regional School district, located in Ocean County, New Jersey. The dataset is based on seventy-eight 2020 seniors in TRHSE, containing the following attributes:

- Number of absences from 6th to 12th grade.
- Number of tardies from 6th to 12th grade.
- Is this student on an IEP (Individualized Educational Program)
- Does this student have a free or reduced lunch?
- Is this student on a 504 plan? (Disability)

Borrowing from nomenclature in [4] and report cards from Toms River Regional Schools, the term "tardies" will refer to the number of instances that a student was late to school. Additionally, using the same threshold defined in [4], chronic absenteeism is defined as being absent for more than 10% of a school year; in this case, the threshold of chronic absenteeism is set to be 18 days in a school year for a student who has been enrolled for the whole school year. This paper defines a chronically absent student to be one that was chronically absent in any year in high school.

Additionally, the attributes described above were chosen because they are associated with higher chances of chronic absenteeism [1, 5]. Regarding TRHSE, the student groups that were the most chronically absent in the 2018-19 New Jersey School Performance Report were, but not limited to, disabled students, economically disadvantaged students, minorities, etc.

A. *Dataset Collection Difficulties*

During the time of data collection, the means of curating students based on certain attributes was not possible. In other words, it was not possible to obtain a list of students that were chronically absent, had a 504 plan, etc. Of course, curation of student data would have been possible, but it would require administrative access to sensitive data. With the help of a supervising teacher, each data point went through two steps and people: the supervising teacher would have access to the administrative computer containing student data and a student researcher would record only the data that the supervising teacher read. The student researcher would not look at the screen and would only hear attributes that were requested. Because of this time-consuming procedure and the limited amounts of time, only a small sample of 78 unidentifiable seniors was obtained. Of course, a dataset containing thousands of students along with many attributes to consider would be most beneficial to machine learning models in order to produce accurate results. Nonetheless, steps, which are discussed in section V, were taken to ensure results were reliable.

B. *Exploratory Data Analysis*

Before using machine learning on the dataset, it can be useful to analyze the dataset beforehand. The dataset acquired was analyzed and graphed using python libraries called pandas and matplotlib. An important fact about the dataset is that it contains data from the school year of 2019-20. During this time, the COVID-19 pandemic caused widespread educational changes. Toms River Regional schools transitioned into virtual instruction. The effects of this transition can be seen in figure 1: chronic absenteeism rates steadily rise from 9th to 11th grade and drop significantly during the year of the COVID-19 pandemic. The New Jersey Department of Education stated that because virtual days are counted as days "in membership" of a school, school districts must record whether a student was present or absent. Further, the state received a federal waiver to not record chronic absenteeism rates in their reports [4]. As seen in figure 1, chronic absenteeism rates would not be realistic during the 2019-20 virtual school year. Perhaps students who would be absent were not due to the ease of being present virtually. Another probable reason that chronic absenteeism rates were low was because there was not an established protocol in truly counting present students. Either way, the effects are made apparent in figure 1. If the 2019-20 year was not affected by COVID-19, there would have been more chronically absent students in the dataset.

Two other student attributes that were collected were tardies and absences throughout middle school and high school. With these attributes, it is possible to sum the number of absences and tardies in middle school and high school; when this is done, correlation can be compared in order to determine whether an attribute is a good candidate to be used in the prediction models.

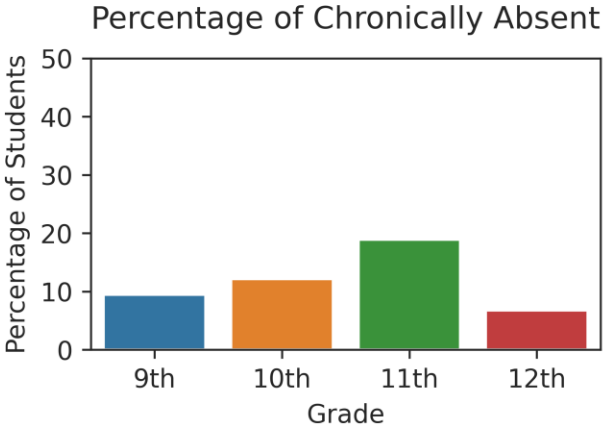

Fig. 1. Percentage of Chronic absenteeism throughout high school. Each bar represents the same cohort used in this paper progressing through the grade years.

Using redundant or needless features can make a model overcomplicated, leading to diminishing performance [13].

High school absences are related to chronic absenteeism; thus, a correlation can show a relationship between attributes and the dependent variable (student was chronically absent). Figure 2 shows a stronger correlation between absences in middle school and absences in high school, r (72) = 0.34, p < 0.01, than between tardies in middle school and absences in high school, r (72) = 0.22, p < 0.01. Four outliers were removed that unrealistically increased these correlations.

IV. Experimental Setup

Problem: This paper aims to predict if a student will be chronically absent in high school using only attributes before high school. The outcome of a student is referred to as **chronically_absent_in_HS**, which is either true (1) or false (0). This variable is computed by checking a student's high school absence record and looking for a year with more than 17 absences. Therefore, the problem itself is a binary classification problem.

Along with predicting student chronic absenteeism, it is also important to know which students need the most attention. School districts may only have a limited number of resources and time to address N students. The classification models used in this paper compute a probability that a student should be classified as one label or the other. These probabilities can be used to rank students by most at-risk. When the students are ranked, school administrators can choose to address the top N students.

Setup: The student dataset is split into a training set and a test set. The test set size is a random sample of the dataset and has 0.25 * N data points, where N is the size of the entire dataset. The training set comprises the remaining data points. When training the models on the training set, k-Fold cross validation was used. As seen in figure 3, the training set is divided into k equal subsections, or folds, where k - 1 folds are used to train the model and one fold is chosen as the validation set. This procedure repeats until all folds are used as a validation set at least once.

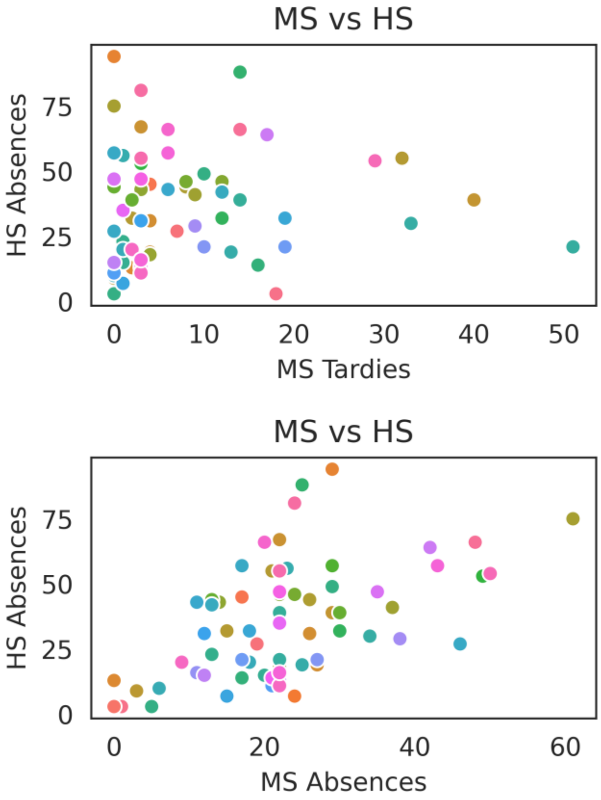

Fig. 2. MS and HS stand for middle school and high school, respectively.

Afterwards, the fitted model is tested on another test set that it never saw during training. The prediction model's results are determined by its performance on unseen data (the test set). All models are trained and tested on the same dataset splits of students. Models that use randomness (non-deterministic) are replicated by using the same random seed for both the train test split and model fitting.

Models: The models that will be used are as follows: Decision Trees (DT), Random Forests (RF), Adaboost (AB), Support Vector Machines (SVM), and Logistic Regression (LR). These models are implemented with scikit-learn.

V. Dataset Considerations

Small datasets are conducive to overfitting. Overfitting occurs when a learning model fits/corresponds too well to a particular dataset and, ultimately, performs poorly on unseen data. If the problems of overfitting are ignored, the results of a predictive model may be false (over-optimistic or poor performance). When dealing with a small dataset, such as the one used in this paper, a question arises: are the results reliable or over optimistic?

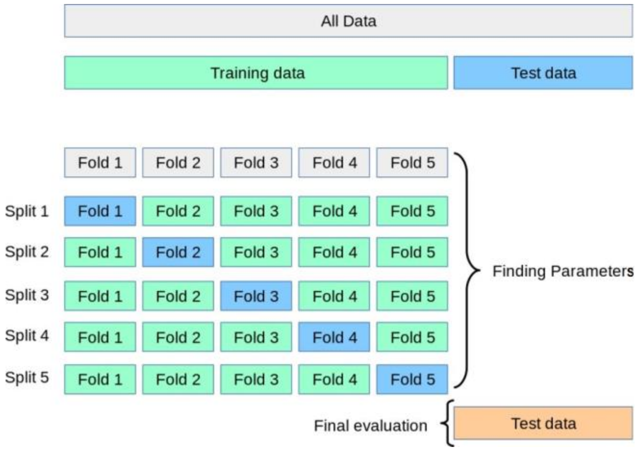

Fig. 3. A diagram of k-fold cross validation with a holdout test set that acts as a model's final evaluation.
Source: Adapted from [16]

A. *Ensuring Reliability*

A random sample of the TRHSE 2020 senior sampling frame, especially a small one, will be homogeneous. In this case, a small sample of high school seniors in TRHSE would yield a sample with very few students that have attributes that contribute to chronic absenteeism. A proportional random student sample would lead to an unbalanced label distribution (non-chronic absentee and chronic absentee). Stratification is the act of dividing a sample into strata, or a defining characteristic, and sampling each stratum. However, the data, at the time of collection, could not easily be curated. Therefore, in order to ensure a model can predict chronic absenteeism based on specific attributes, students of specific underrepresented attributes were sought out in the list in an informal manner. Therefore, generalizations and statistics of the dataset are not reliable. Nonetheless, this does not stop the applications of machine learning for this dataset. See section III-A for more details.

As discussed in section IV, k-Fold cross validation was utilized along with another unseen validation/test set. These procedures were used in order to ensure that the machine learning model would not overfit on the training set. If the model had overfitted to the training set despite k-Fold cross validation, the model will most likely perform poorly on the last held back validation set. Additionally, the held back validation/test set with unseen data was stratified; nearly half of the test set were chronic absentees, and the other half were not. More specifically, when test set size was 0.25 * N, there were 11 non-chronic absentees and 9 chronic absentees in the test set.

B. *Reliability of the Test Set*

As discussed in section 5.1, the test set had a near equal label distribution. Regarding the importance of this fact, a question may arise:

- Why does a near equal label distribution ensure reliability?

In order to answer this question, it is important to discuss the definition of accuracy and the caveats of relying on it. Accuracy is defined to be the number of correct predictions divided by the number of total predictions. However, accuracy itself can be misleading. If a test set has 100 students and all but one student is a non-chronic absentee, a prediction model that blindly labels any student a non-chronic absentee will have an accuracy of 99%. This prediction model cannot be considered reliable because the model has a 0% accuracy for chronically absent students (there is one chronic absentee in the sample and the model predicted wrongly). However, because the test set is not homogeneous and represents the two labels almost equally, the scenario that was described cannot occur; thus, the reliability of the test set as the final model performance evaluation is ensured.

VI. Results

Before a student enters high school, it is important to predict whether a student will be chronically absent or not. Similarly, it is equally important to know the most at-risk students by ranking students by the probabilities that a student will be chronically absent. The term "risk" will convey the probability that a student will be chronically absent in high school. These risks are calculated when a classification model makes a prediction. This section will evaluate the predictions of the classification models with traditional metrics and ensure the reliability of probability rankings using a metric called mean empirical risk.

A. *Evaluation of Classification Predictions*

The classification of to-be chronically absent students can be evaluated by traditional metrics like ROC, AUC, accuracy, recall, and so on. As discussed in section V-B, other evaluation metrics should be used. In a binary classification problem, there are true/false positives and true/false negatives. Furthermore, there is a threshold for which a data point can be considered positive or negative based on risks calculated by classification models. This information can be represented by an ROC curve (Receiver Operating Characteristic Curve) and the corresponding AUC (Area Under the ROC Curve).

The effects of a small sample size are seen in the rigidness of the ROC curve in figure 4 where test set size is N * 0.25. In a larger test set, the curves would be less rigid and the distinctions between model performance would be clearer. However, the curves in the graph overlap each other in certain areas. When the test set is 0.3 * N, the number of positive students and negative students in the test set is 14 and 10, respectively. Consequently, the ROC curves are more distinct, but the accuracies of the models decrease. This means that when there are more chronically absent students in the test set, there are less chronically absent students in the training set, which leads to a lower AUC and overall accuracy. These effects are more significant in a small sample that has a low number of chronically absent students.

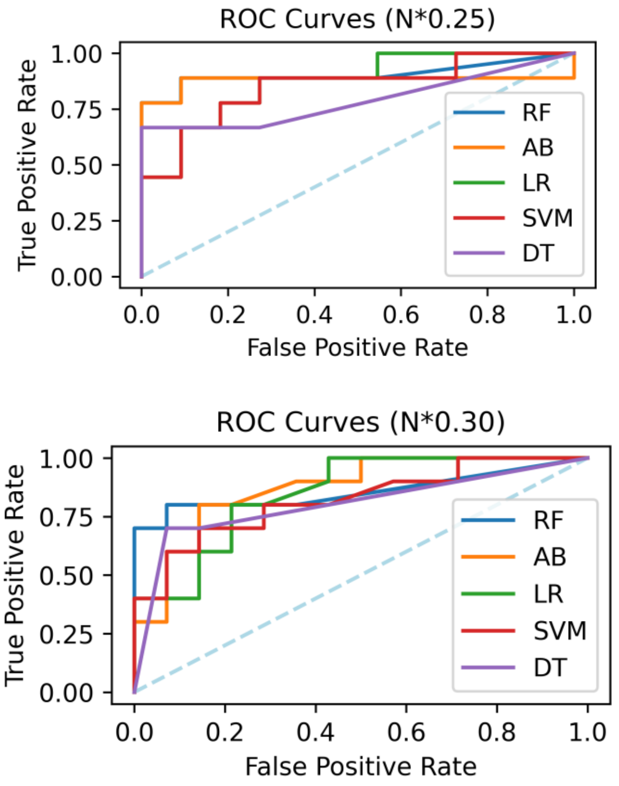

Fig. 4. Graphs show ROC curves. N is the number of data points. When the test set increases from 25% to 30% of N, ROC curves are more distinct.

The issue of overlapping and distinct ROC curves can be alleviated by considering a model's AUC instead. AUC is the area under the ROC curve of any individual model and can be used as a numerical performance metric. In table I, the AUC scores show that Random Forest Classifier outperforms the rest of the models. AdaBoost and Logistic Regression have nearly similar performance, and SVM does not perform as well as the others.

While AUC can be used to determine the better classification model, another metric can be used that better fits the context of the classification problem. Schools have limited time and resources to address every student who is at-risk of chronic absenteeism. Thus, school districts are concerned about the top N students who are most likely to be chronically absent. Mean empirical risk is an alternative metric that evaluates the goodness of student rankings while being interpretable: are the students whose risk scores are the highest truly the ones that are at-risk? If a classification model assigns a 90% risk score to a considerable number of students that are ultimately false positives, then validity of rankings cannot be ensured.

B. Evaluation of Student Rankings

Classification models used in this paper compute a probability that a data point is a specific label. In our case, this label is binary: a student will or will not be chronically absent in high school. One proposed method of determining the goodness of risk scores is called mean empirical risk [7]. According to Lakkaraju et al., the procedure for determining mean empirical risk is as follows:

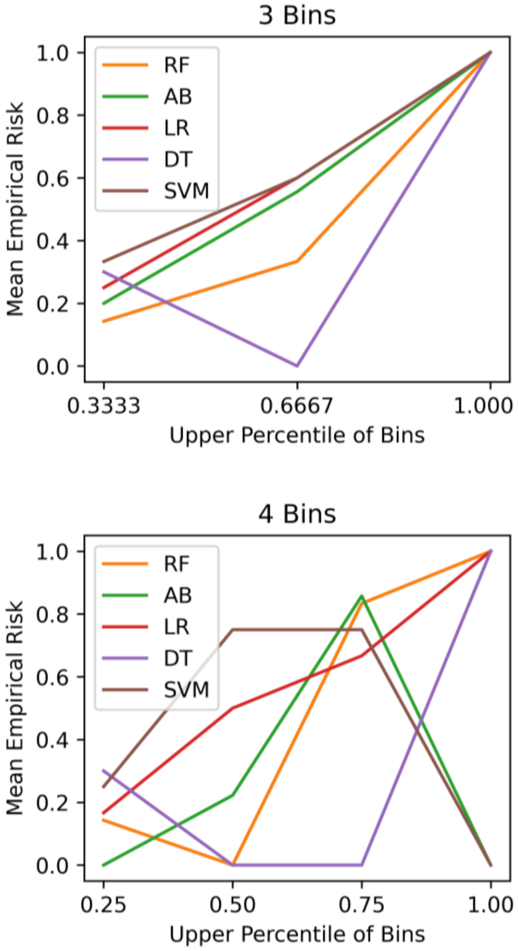

Fig. 5. The instances where the curves have a value of 0 is when a bin has 0 or 1 student instance. This deviance from monotonicity does not devaluate a model's ranking. However, it is an indicator that more data points would be useful.

1. Sort students in order by their risk
2. Create N bins where each bin encompasses an equally divided range of probabilities (for example, the first bin of ten bins would encompass probabilities from 0 to 0.1)
3. Calculate the mean empirical risk for each bin by computing the fraction of actual positives divided by the total number of students in the bin.
4. Prediction models are determined good if and only if the "empirical risk curve is monotonically non-decreasing" [7].

In this paper, the number of bins for the creation of the mean empirical risk curve is 4. Due to the small test set for this paper, the mean empirical risk curve was not monotonically non-decreasing for bins greater than 3. When computing the mean empirical risk curve for a graph of 10 bins, there were bins that

TABLE I

Results when test set size is N * 0.25

Classifier	AUC	Accuracy
Random Forest Classifier	0.904	90%
AdaBoost Classifier	0.876	85%
Logistic Regression	0.869	80%
SVM	0.848	80%

had no students. When a bin has no students, the mean empirical risk for that bin is 0 which breaks the monotonicity of the curve. This breakage implies a false impression that a model incorrectly ranks students. Instead, the false impression occurs because there are not enough students with probabilities that are encompassed by each bin (assuming the number of bins is greater than 4).

As seen in figure 5, many models have a monotonic non-decreasing curve. However, the details of the curves, if the models were tested on a larger test set, are lost due to the low number of bins. Whether monotonicity would be broken if the number of bins and data points is increased cannot be known. These graphs only show the general shape and direction. However, the empirical risk curves show promising insight into the goodness of the models' risk rankings.

VII. DISCUSSION

By analyzing a sample of Toms River High School East seniors that graduated in 2020, machine learning models were able to accurately predict whether a student may or may not be chronically absent in high school before they enter high school. However, it must be made clear that the dataset used contained students that went through extensive educational changes due to COVID-19. This means that those who would be chronically absent were not because of virtual schooling. Data analysis confirms there was an unusually low number of chronic absentees in 2019-20. How the pandemic affected both the prediction results and data were not addressed.

Additionally, the dataset used in this paper is very limited in scope. Not only is the dataset relatively small, but it only contains students from one year from one school in one region. Therefore, no conclusive statements can be made on the accuracy that machine learning may or may not have in predicting chronic absenteeism elsewhere. Other student attributes not included in this paper may better contribute to the prediction of chronic absenteeism. Only with many data points, resources, and extensive data analysis can conclusive and generalizable statements be made.

VIII. FUTURE WORK

Future related work should confront this topic with a comprehensive dataset. This dataset should include a diverse set of students from different schools and regions. Researchers should investigate different attributes that affect chronic absenteeism as well. Furthermore, researchers should not investigate predicting chronic absenteeism in high school, but in middle school. A study of Baltimore City Schools concluded that most high school dropouts had a pattern of chronic absenteeism spanning "several years." Not only that, the habits of chronic absenteeism for students entering ninth grade are "extremely difficult to change" [9]. Future researchers should investigate predicting chronic absenteeism in middle school in order to intensely focus more resources on early middle school intervention. Unfortunately, it was not practical to address middle school chronic absenteeism for this paper, but perhaps other researchers with more resources can address this topic.

IX. CONCLUSION

This paper attempts to determine if machine learning models can determine if a student will be chronically absent in high school using only a student's middle school attributes. With only 78 data points to train and test with, actions, such as using a holdout test set with a near equal label distribution, were taken in order ensure that machine learning models did not overfit and did not produce over-optimistic results. Random Forest classifier outperformed all the models used, but more research needs to be done in order to make conclusive statements about the results in this paper.

Nonetheless, it is my hope that this paper provides promising insight into student-focused machine learning applications and encourages school administrators to look for innovative data-driven ways to address problems, like chronic absenteeism, moving forward.

X. ACKNOWLEDGEMENT

I would like to thank my ASR teacher, Dr. Kretz, for helping me collect data and giving me the opportunities and foundation to pursue this project. I would also like to thank Mrs. Anders and the rest of the administration for allowing me to do my work.

XI. REFERENCES

[1] U.S. Dept. of Education. "Chronic Absenteeism in the Nation's Schools." www2.ed.gov. https://www2.ed.gov/datastory/chronicabsenteeism.html [Accessed Sep. 4, 2020].

[2] "Research Brief: Chronic Absenteeism," University of Utah, Salt Lake City, USA, 2012.

[3] David M. Cutler, "Education and Health: Evaluating Theories and Evidence" in the National Poverty Center, Jun. 2006, doi:10.3386/w12352.

[4] NJ Dept. of Educ. "Chronic Absenteeism Guidance," Guidance for Reporting Student Absences and Calculating Chronic Absenteeism, May 2020 [Accessed Sep. 1, 2020]. [Online]. Available: https://www.state.nj.us/education/students/safety/behavior/attendance/ChronicAbsenteeismGuidance.pdf.

[5] NJ Dept. of Educ. "Toms River High School East," NJ School Performance Report, 2018-2019. [Accessed Aug. 22, 2020]. [Online].

Available: https://rc.doe.state.nj.us/report.aspx?type=school&lang=english&county=29&district=5190&school=030&schoolyear=2018-2019#P24995f74c0aa41e3b21249e28c64c140_5_oHit0.

[6] A. Sahiri et al., “A Review on Predicting Student’s Performance using Data Mining Techniques,” in The Third Information Systems International Conference, 2015, doi: 10.1016/j.procs.2015.12.157.

[7] H. Lakkaraju et al., “A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes,” 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1909-1918, doi: 10.1145/2783258.2788620

[8] L. Wood et al., “Predicting Dropout Using Student- and School-Level Factors, School Psychology Quarterly, 201,. doi: 10.1037/spq0000152.

[9] M. Iver, “A Portrait of the 2008-09 Dropouts in the Baltimore City Schools,” Baltimore Educ. Research. Consortium, Baltimore, MD, USA, Aug. 2009.

[10] R. Balfanz and V. Byrnes, “The Importance of Being in School: A Report on Absenteeism in the Nation’s Public Schools,” John Hopkins University Center for Social Organization of Schools, Baltimore, MD, 2012.

[11] A. Bowers et al., “Do We Know Who Will Drop Out? A review of the Predictors of Dropping out of High School: Precision, Sensitivity, and Specificity,” in The High School Journal, vol. 96, 2013, pp. 77-100, doi:10.7916/d86w9n4x.

[12] A. Sum et al. “The Consequences of Dropping Out of High School,” Northeastern University, Boston, MA, 2009.

[13] M. Pojon, “Using Machine Learning to Predict Student Performance,” M. Sc. thesis, Dept. Natural Sciences, Univ. Tampere, Tempere, Finland, Jun. 2017. [Online]. Available:

https://trepo.tuni.fi/bitstream/handle/10024/101646/GRADU-1498472565.pdf?sequence=1&isAllowed=y

[14] M. Quadril and N. Kalyankar, “Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques,” in Global Journal of Computer Science and Technology, vol. 10, Apr. 2010, pp 2-5.

[15] National Center for Educ. Statistics, “Dropout Rates.” nces.ed.gov. https://nces.ed.gov/fastfacts/display.asp?id=16 [Accessed Sep. 3, 2020].

[16] Scikit-learn, “3.1 Cross-validatoin: evaluating estimator performance.” scikit-learn.org. https://scikitlearn.org/modules/cross_validation.html/ [Accessed Sep. 12 2020]