SEIS 763-02: Group 2 Project Report

Tony Challeen
School of Engineering
University of St. Thomas
Minneapolis, MN
chal7675@stthomas.edu

Emma DePape
School of Engineering
University of St. Thomas
Minnetonka, MN
depa7174@stthomas.edu

Elizabeth Izaguirre
School of Engineering
University of St. Thomas
Minneapolis, MN
izag9154@stthomas.edu

Abanoub Lamie
School of Engineering
University of St. Thomas
Saint Paul, MN
lami6174@stthomas.edu

Abstract—**This project uses data from the High School Longitudinal Study of 2009 [1], a nationally representative dataset tracking more than 20,000 students across multiple survey waves, to model student outcomes by predicting postsecondary GPA. We apply classification techniques to predict postsecondary GPA, including logistic regression, support vector machines, and various ensemble models. The analysis incorporates feature selection, dimensionality reduction, and hyperparameter tuning to evaluate model performance using accuracy and F1 score as metrics. Features include high school academic performance, STEM identity, socioeconomic background, and educational aspirations. The final model aims to uncover early predictors of long-term academic success.**

I. INTRODUCTION

The High School Longitudinal Study of 2009 (HSLS:09) is a nationally representative dataset designed to track the educational, social, and career trajectories of US students beginning in ninth grade. The study followed more than 23 thousand ninth graders from 944 schools throughout their secondary and postsecondary years of schooling. Surveys of the students, their parents, math and science teachers, school administrators, and school counselors were conducted to evaluate the students, resulting in a large dataset containing over 3,000 variables. The variables relate to academic performance, attitudes towards school, expectations, school context, and family background.

The overarching purpose of HSLS:09 is to provide insight into how students' early academic experiences and personal characteristics shape their long-term educational outcomes. This project leverages this comprehensive dataset to investigate the conditions that impact how a student performs academically in postsecondary education. By applying machine learning methods to this dataset, the project aims to identify early predictors of postsecondary academic success. The outcome will contribute to a deeper understanding of the factors that influence students' educational pathways.

II. RELATED WORK

There have been numerous analyses done with the data derived from HSLS:09. Abt Global, a research and consulting firm, was contracted to analyze the data and provide insights to families, educational professionals, as well as policymakers to help promote interest in participating in future longitudinal studies by IES. Two reports were published that highlighted the connection between high school students' perceptions of college affordability and their future college enrollment or employment status three years after high school, and the connection between high school counseling, the completion of the Free Application for Federal Student Aid, and college aid [2][3]. Both reports were prepared for the National Center for Education Statistics (NCES).

The National Archive of Data on Arts and Culture also provides references to over 200 publications made using the HSLS:09 data. Analyses range from the connection between underrepresented students' motivational attitudes toward mathematics to the impacts of school spending on civic engagement. The breadth of analyses proves how versatile the HSLS:09 dataset is [4]. Our analysis stands apart from others by casting a wider net as we attempt to predict postsecondary GPA. This was done by selecting features believed to have a strong correlation with postsecondary performance.

III. DATASET DESCRIPTION

A. Dataset Origin

This data comes from the National Center for Education Statistics (NCES), an independent research, statistics, and evaluation arm of the U.S. Department of Education. This dataset in particular is a high school longitudinal study conducted in 2009, its base year. The study team then followed up with the students in 2012 and 2016. Also included are postsecondary transcripts and student financial aid records between 2016 and 2017.

The study is significant because it fulfills a congressional requirement for collecting, analyzing, and reporting on the condition of education in the United States. The subsequent analysis and statistics allow the federal government to assist state and local educators and agencies in improving their own data collection processes and, most importantly, help address critical education data needs. Data is reported at all levels: the U.S. Department of Education, states, policymakers, educators, Congress, and the public.

A few of the key research topics this dataset strives to address include high school to college transition plans, how those plans might change, and students' interest in STEM-related studies and career paths. Equally relevant is the topic of addressing educational and social factors that may influence their plans in postsecondary education, work, and life [5].

Although HSLS:09 is designed to be nationally representative, several forms of bias may influence both the dataset and the predictive modeling results. The data is from a longitudinal study which requires sustained engagement over many years, which may attract certain types of students and families more than others. The potential for selective participation could skew the sample toward certain classes of students. Another potential form of bias is due to this study being survey-based. Survey data collection assumes the respondents understand the questions, feel comfortable answering them, and feel motivated to answer all of them. As a result, certain perspectives or experiences may be underrepresented in the dataset. Finally, since the study relies on self-reported attitudes, behaviors, and perceptions, the responses are vulnerable to students overstating positive behaviors. Collectively, these sources of bias highlight the challenges of using survey-based data for predictive modeling.

B. Data Collection

Data was collected from multiple respondent groups, including students, parents, math and science teachers, school administrators, and school counselors. From these sources, the dataset presents more than 3,000 variables capturing academic performance, attitudes and expectations, family background, school context, and postsecondary outcomes. For this project, data was used from the following HSLS:09 stages: Base Year, First Follow Up, 2013 Update/High School Transcripts, Second Follow Up, and our target variable was obtained from the Post Secondary Transcripts stage. All variable names are a combination of a letter, number, and shorthand description. Table 1 summarizes the variable naming convention used throughout HSLS:09.

TABLE 1: HSLS:09 VARIABLE NAMING CONVENTIONS

Character 1	Component Identifiers	X = Composite Variables S = Student P = Parent M = Math Teacher N = Science Teacher A = Administrator C = Counselor T = Transcripts
Character 2	Stage Identifier	Numbers 1-5 1 stands for the Base Year, 2 stands for First Follow Up, etc.
Characters 3-12	Indicates a descriptive name for the variable	

There are exactly 23,503 instances in the dataset. Many of the features contain missing values, and the researchers gave these instances a negative value that corresponds to a particular scenario. The scenarios include data suppression for privacy, questions not applicable to the student, nonrespondents, or true missing values.

TABLE 2: HSLS:09 MISSING VALUE CODES

Code	**Meaning**
-5	"**Data Suppressed**"—indicates values that are available on the restricted-use data file but suppressed on the public-use data file.
-7	"**Item legitimate skip/NA**"—indicates items that are programmatically skipped based on rules in the questionnaire and are not applicable to those respondents.
-8	"**Nonrespondent/component NA**"—indicates that data are not available because of unit nonresponse or the interview component did not apply (e.g., the student has no math class; thus, the math teacher interview does not apply).
-9	"**Missing**"—indicates item-level missing where the question may apply to the respondent, but it is not answered, or the question is not administered because the gate/introductory question is not answered.

The dataset is comprised of both categorical and continuous values, the majority of which are categorical. All categorical variables are pre-encoded, for which the data source provides a codebook to understand the meaning of each encoding.

C. Variable Selection

Selecting the features we eventually used during our analysis was done with our hypothesis in mind. We knew we were interested in predicting postsecondary GPA, and therefore the variables we chose to explore initially were those we believed might have a direct impact. As stated previously, the dataset contained more than 3,000 variables, therefore the variable selection process was one of the more time-consuming stages.

We decided to divide the variables amongst ourselves based on their prefixes, for instance X1-X3, X4-X5, S1-S4, etc. During this exploratory phase of variable selection, we utilized a codebook provided by the researchers that described each variable in much detail to learn more about the variable, its distribution, and other factors like missing values. It was during this process that we realized we could not use some variables provided by the accompanying high school dataset due to privacy and data suppression measures taken by the researchers. With this understanding, we chose to focus on the student dataset and use only features found within it.

Finally, with postsecondary GPA in mind as our target label, we narrowed our list of potential features down to 40 variables. The final selected variables represented academic performance, STEM interest, socioeconomic background, school engagement, and behavioral indicators.

IV. Preprocessing

The following preprocessing steps were applied to prepare the dataset for machine learning. To start, we split the dataset into training and testing sets using a 75/25 stratified split to preserve the class distribution.

A. Missing Values

Multiple methods were used to handle missing values. First, rows with a missing value in the target variable field were removed. The size of this dataset allowed 12,588 samples to remain. Next, imputation methods were used for the missing values in the remaining rows. The mean was imputed for numerical variables, and the mode was imputed for categorical and Boolean variables.

B. Scaling

Normalization was applied to the continuous variables to ensure that all numerical features contributed proportionally to the model training process. Numerical values were scaled using a Standard Scaler, a standardization method that transforms each feature by subtracting the mean and dividing it by the standard deviation. This procedure centers the data around zero with a standard deviation of one.

Standardization was necessary due to the nature of several of the machine learning models used in this project and their sensitivity to differences in feature magnitude. Without scaling, features with larger numeric ranges might dominate the models and negatively impact performance and interpretability. The Standard Scaler was only applied to the numerical columns to preserve the numerically encoded categorical variables and the Boolean variables.

C. Encoding

The target values were originally presented in increments of 0.1, ranging from 0.0-4.0. These values were grouped and encoded into four categories using LabelEncoder from scikit-learn: 0.0-1.9 = 0, 2.0-2.9= 1, 3.0-3.4= 2, and 3.5-4.0= 3. The four categories were then mapped to a label of "Failing," "Average," "Good," or "Excellent," respectively, to enhance explainability.

TABLE 3: TARGET VARIABLE ENCODING

Class	**Code**
Failing	0
Average	1
Good	2
Excellent	3

D. Class Imbalance

The class distribution was analyzed and determined to be imbalanced. Imbalanced datasets can cause majority class bias during model training and poor generalization during model testing. Some of the classifiers selected here are especially sensitive to or rely on balanced data. Upsampling was chosen as the method to address this concern, to avoid eliminating more data. SMOTE was performed on the minority classes and resulted in a balanced distribution. SMOTE was applied to the training set only to avoid data leakage.

a) Before balancing classes

TABLE 4: IMBALANCED TARGET CLASSES

GPA Category	**Count**
Failing	2,418
Average	4,107
Good	3,164
Excellent	2,899
Total	**12,588**

b) After balancing classes

TABLE 5: BALANCED TARGET CLASSES

GPA Category	Count
Failing	3,080
Average	3,080
Good	3,080
Excellent	3,080
Total	**12,320**

E. Feature Selection

To reduce model complexity and identify the most informative predictors, a wrapper-based backward elimination procedure was applied. This method recursively evaluates a Logistic Regression model to determine the statistical contribution of each feature to the target variable. At each step, the feature with the highest p-value is removed, and the model is refit until all remaining predictors meet the specified significance threshold.

Backward elimination was selected because it evaluates the full model at each step, allowing the procedure to account for interactions among all correlated features. Forward selection may prematurely exclude variables that become important only in the presence of other variables.

Two elimination criteria were tested. Using a p-value threshold of 0.10, the procedure retained 12 features out of the original 54. When applying a stricter threshold of 0.05, the final model retained 8 features. This reduction intends to help remove noise and ensures that only statistically significant predictors are included in subsequent modeling stages.

F. Feature Engineering

Feature engineering was intentionally not performed in the project. The HSLS:09 dataset originally contained more than 3,000 variables derived from multiple respondent groups and survey stages. Many of those variables are composite indicators already engineered by the original research team. The variables we selected for use in our predictive model were already a majority of the composite or engineered variables. The primary challenge of this analysis was not the lack of available predictive attributes but rather the need to reduce dimensionality and identify more informative predictors from the already extensive feature space. For this reason, we abstained from performing feature engineering.

G. Dimensionality Reduction

Three dimensionality reduction techniques were evaluated to improve model efficiency and explore lower dimensional representations of the data: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Kernel PCA.

For PCA, the cumulative explained variance curve was examined to determine the number of components that captured a sufficient proportion of the variance in the original feature space. As shown in Figure 1, the PCA curve showed a point of diminishing returns at 18 components, which was selected as the optimal balance between dimensionality reduction and information retention.

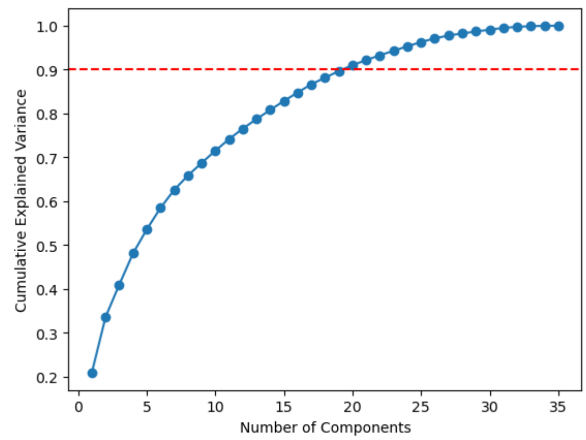

FIGURE 1: PCA CUMULATIVE EXPLAINED VARIANCE PLOT

For LDA, the number of components is constrained by the number of classes in the target variable. Since this classification task contains four categories, LDA can produce at most three discriminant components. As shown in Figure 2, at 3 components, the cumulative explained variance is near 1. Therefore, all three components were retained for subsequent model training.

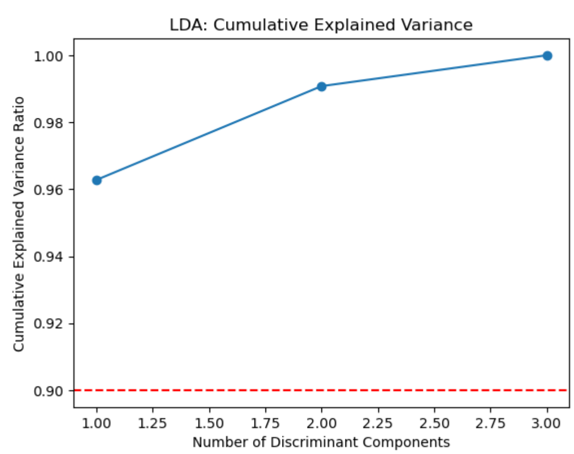

FIGURE 2: LDA CUMULATIVE EXPLAINED VARIANCE PLOT

Kernel PCA using the Radial Basis Function (rbf) kernel was performed, to capture potential non-linear relationships, and a grid search was performed over a range of component values to identify the configuration that produced the strongest model performance. The search indicated that 23 components provided the best results.

These dimensionality reduction methods allowed for a comparison between linear and nonlinear transformations of the feature space and an assessment of their impact on classification performance.

V. Hypotheses

Students who demonstrate stronger academic performance in high school, higher STEM identity and interest, and come from higher socioeconomic backgrounds will achieve higher postsecondary academic GPAs.

This hypothesis was tested using classification models, with the target variable as the student's cumulative postsecondary GPA grouped into four levels of GPA labeled as "Failing," "Average," "Good," and "Excellent". Those groups were GPAs of 0.0-1.9, 2.0-2.9, 3.0-3.4, and 3.5-4.0, respectively. The selected predictive features represent students' academic performance, STEM interest, socioeconomic background, school-based STEM enrichment opportunities, daily study and activity patterns, and behavioral or attendance indicators. These factors span multiple time points in the study and reflect both academic readiness and social context.

VI. ML Techniques

A range of supervised classification models were implemented to evaluate the predictive power of high school attributes on postsecondary GPA, with hyperparameter tuning applied where appropriate.

Logistic regression served as a baseline model, and additional linear, nonlinear, and ensemble methods were explored to assess whether more flexible algorithms could capture complex relationships in the data. All models were trained using the feature subsets and reduced dimensionality training sets generated during preprocessing.

All hyperparameter tuning was evaluated using a 5-fold cross-validation Grid Search to reduce variance across folds.

A. *Logistic Regression*

Logistic regression was used as the baseline classifier due to its interpretability and simplicity as a classification model. Logistic regression models have the ability to quantify linear relationships between predictors and target values. Our multiple logistic regression models established a reference point for evaluating whether more flexible algorithms provide meaningful improvements. The model was trained on the full feature set, on the subsets produced through backward elimination, and on the reduced feature spaces generated by PCA, LDA, and Kernel PCA.

To understand the contribution of individual predictors, the model coefficients were examined. Figure 3 illustrates the absolute coefficient magnitudes for each variable. The model reached about 40% accuracy, and this performance was determined to be dominated by one feature, X3TGPAACAD, the GPA for high school academic courses. When the model was retrained using only this feature, the results were identical to those obtained using the full feature set. This indicates that the remaining features' contributions were very low to the predictive value.

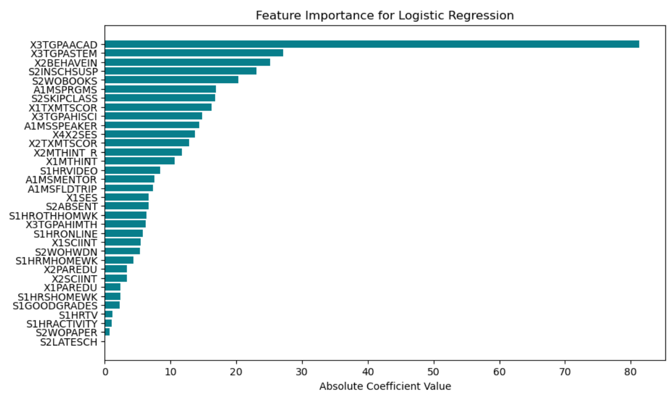

FIGURE 3: FEATURE IMPORTANCE FOR LOGISTIC REGRESSION

Upon testing the removal of the dominant feature, the model's performance degraded substantially, which confirmed the feature's central role in predicting postsecondary GPA.

Among the dimensionality reduction approaches, the LDA reduced feature set yielded slightly higher accuracy than PCA, Kernel PCA, and backwards elimination. For this reason, the LDA components were used as the input for subsequent models.

B. *Stochastic Gradient Descent*

Stochastic Gradient Descent (SGD) was applied to the retained LDA components because it is an efficient variant of gradient descent that incrementally updates model parameters and aims to minimize the loss function. Also, its use of randomly selected data points introduces stochasticity, which can help the model escape shallow local minima and improve generalization. This model was included to test whether an incremental optimization approach could better handle the high dimensional feature space and potentially escape local minima that limit logistic regression. The log_loss loss function was selected, as it penalizes confident but incorrect predictions more strongly, making it well-suited for classification tasks where calibrated probabilistic outputs are important.

C. Random Forest

A Random Forest classifier was trained on the retained LDA components, as this model is a well-suited model for capturing non-linear relationships and interactions among features. This model also recognizes hierarchical relationships among features, which could further enhance performance. A grid search was performed to optimize performance and tune the hyperparameters, specifically the criterion and max_depth parameters. The optimal configuration was found to be the Gini impurity measure for the criterion and a maximum tree depth of 13.

D. Support Vector Classifier

Support Vector Classification (SVC) was evaluated due to its strong performance in high-dimensional spaces. Additionally, it was used to evaluate whether nonlinear decision boundaries could capture more complex relationships among the predictors. To optimize performance, GridSearchCV was used to tune hyperparameters, specifically the kernel, number of support vectors (C), and gamma parameters. The search identified RBF as the best kernel, 10 as the best number of support vectors, and a gamma of scale as the most effective combination. This indicated that a nonlinear decision boundary provided the best fit.

E. XGBoost

XGBoost, a gradient boosted tree ensemble known for strong performance on structured data, was trained on the LDA components. This model was included to see if sequential boosting could extract additional signals from weaker predictors in the feature space. Although XGBoost is typically competitive in classification tasks, its performance in this project did not exceed that of the next model, AdaBoost.

F. AdaBoost

To improve the baseline logistic regression model, AdaBoost was applied using logistic regression as the base estimator. AdaBoost iteratively increases the weight of misclassified samples, allowing the base model to focus on the difficult cases. AdaBoost models can provide insight into whether the dataset benefits more from boosting weak learners than from complex nonlinear models. Hyperparameter tuning was completed using GridSearchCV which identified 200 estimators and a learning rate of 0.01 as the most effective hyperparameters. AdaBoost ultimately achieved the strongest overall performance among the models tested.

G. Naive Bayes

Naive Bayes was evaluated to test the effect of independence assumptions on classification performance. Because Naive Bayes relies on the naive assumption that all features are conditionally independent, we explored whether this model could benefit from a simplified feature space. The model was trained using only the least influential features identified in earlier models to test whether removing dominant predictors would improve performance. This experiment provided insight into how feature dependence affected classification accuracy.

H. Voting Classifier

Finally, a VotingClassifier was evaluated to aggregate predictions from multiple models. The expectation was that combining diverse learners might improve robustness and reduce variance across the GPA classes. However, the ensemble did not outperform the strongest individual model.

I. Summary

Across all models, the approaches taken allowed for comparison between linear, nonlinear, and ensemble-based methods using a consistent set of reduced features. Logistic regression provided a baseline for interpretability, with SGD, SVC, and Naive Bayes offering alternative linear and probabilistic perspectives. Random Forest, AdaBoost, XGBoost, and the VotingClassifier enabled evaluation of whether ensemble strategies could capture additional structure in the data. By applying hyperparameter tuning and training each model on the LDA reduced feature space, a fair basis for comparison was established, and the algorithms most capable of leveraging the available high school predictors were identified. These models form the foundation for the performance evaluation in the Results section.

VII. Results

Model performance on the GPA classification task is evaluated using accuracy, macro-F1 scores, and confusion matrices. These metrics reveal how each model handled the four GPA categories and where misclassification patterns emerged.

To start, the performance of the baseline Logistic Regression model was tested on three dimensionality-reduced feature sets. This resulted in the selection of the LDA reduced components to be used for all future model training.

TABLE 6: DIMENSIONALITY REDUCED MODEL METRICS

Metric	PCA Model	LDA Model	KPCA Model
Precision	0.42	0.42	0.41
Recall	0.45	0.45	0.44
F1 Score	0.42	0.42	0.42
Accuracy	42.20%	42.55%	42.04%

To ensure that the model with the highest F1 score was selected, Table 7 presents a comparison of the evaluation metrics for all models.

TABLE 7: MODEL METRIC COMPARISON

Model	Precision	Recall	F1 Score	Accuracy
ADABoost	0.45	0.44	0.44	43%
RandomForest	0.42	0.44	0.42	42%
SVC-rbf	0.42	0.44	0.42	42%
VotingClassifier	0.41	0.44	0.42	42%
NaiveBayes	0.41	0.45	0.41	41%
XGBoost	0.40	0.42	0.40	40%
StochasticGradient	0.37	0.43	0.35	39%

Across all models, accuracy ranged from 39% to 43%, with F1 scores following a similar pattern. These results indicate that while the predictors contain meaningful signals, the overall predictive power remains modest.

As is evident from Table 7, AdaBoost achieved the highest F1 score and accuracy together, indicating that boosting slightly improved predictive performance compared to logistic regression alone.

To assess class specific performance, we examined the confusion matrix for the AdaBoost model.

TABLE 8: ADABOOST CONFUSION MATRIX

True \ Predicted	Failing	Average	Good	Excellent
Failing	410	76	224	317
Average	103	320	29	273
Good	207	7	310	80
Excellent	198	189	83	321

The confusion matrix provided a detailed view of how the AdaBoost model classified students across the four GPA categories. While the model captures meaningful structure in the data, its performance varies substantially by class, with clear strengths and weaknesses.

The model demonstrates its strongest performance for the Good and Excellent classes, correctly classifying 310 and 321 students, respectively. These relatively high diagonal values suggest that the model is more effective at identifying higher performing students. However, misclassifications for these classes were still relatively high.

Performance for the Failing class is mixed. While 410 students were correctly identified, substantial misclassification occurs in the Good and Excellent classes. This pattern reflects a systematic upward bias, in which low performing students are frequently predicted to belong to higher GPA classes.

The Average class also exhibits mixed classification performance. While the model correctly classified 320 students, there were still large proportions misclassified to the Failing and Excellent classes. This distribution indicates that mid-range academic performance is not well separated in the feature space, leading to considerable overlap with both lower and higher categories.

Overall, the confusion matrix indicates that the model performs more effectively at distinguishing high performing students, while lower and middle performing students are more frequently misclassified.

The following comparison provides context for the model's overall performance: the model's accuracy (43%) is significantly better than Random Chance (25%) and noticeably better than if the model only predicted the Majority Class (33%). This confirms that the features used have some predictive power. The results indicate that the model is identifying meaningful patterns rather than producing predictions purely by chance.

TABLE 9: FINAL MODEL COMPARED TO CHANCE

Model Strategy	Total Accuracy
Our Final Model	**43%**
Majority Class	33%
Random Chance	25%

A. *Interpretation of Results*

Several factors likely contributed to these outcomes. First, many of the predictors available in the high school waves of HSLS:09, such as GPA, standardized test scores, and attitudinal measures, reflect performance within a structured and highly supported academic environment. Postsecondary settings differ significantly in autonomy, instructional practices, grading standards, and academic rigor. As a result, high school attributes may not fully capture the determinants of academic performance during postsecondary education.

Second, postsecondary GPA can be influenced by numerous factors that are not measured in the early waves of the study. These include college-specific study habits, time management, mental health, motivation, resilience, financial pressures, employment hours, and access to institutional support services. Institutional characteristics, such as university selectivity and grading culture, also play an important role but are not represented in the high school data used for modeling.

Taken together, these limitations suggest that the available predictors do not sufficiently represent the multidimensional factors that shape the postsecondary academic outcomes. Consequently, the models plateaued at modest performance levels, reflecting the inherent difficulty of forecasting postsecondary GPA using only high school information.

VIII. Conclusion

In conclusion, the results of this study indicate that high school attributes, even when combined with multiple machine learning techniques, provide only limited predictive power for postsecondary GPA. Although the models performed consistently across algorithms and tuning strategies, their modest accuracy and F1 scores demonstrate that early academic, attitudinal, and demographic variables do not fully capture the complexity of college academic performance. This outcome remains an important empirical finding. It underscores the inherent limitations of relying solely on high school data to forecast postsecondary outcomes and highlights the need for more comprehensive, context-specific information. Thus, the results provide only partial support for our hypothesis, as the selected high school predictors capture general trends but do not fully account for the complexity of postsecondary academic performance.

The analysis suggests that meaningful improvements in predictive performance will likely require the incorporation of additional variables that reflect students' postsecondary environments, behaviors, and support systems. Factors such as study habits, time management, mental health, financial pressures, institutional characteristics, and course-level rigor are not represented in the early HSLS:09 waves but are known to influence academic success. Future research that integrates these richer data sources or that explores alternative outcome measures beyond cumulative GPA may yield more robust insights into the transition from high school to college.

Overall, this work contributes to our growing understanding of the challenges involved in modeling educational trajectories. By identifying the limits of high school predictors, the study provides a foundation for developing more nuanced hypotheses and more informative models that better reflect the multifaceted pathways leading to college achievement.

IX. Future Work

Looking ahead, there are several directions we would pursue to strengthen this research. First, we would seek permission to use the Restricted Use Files (RUFs) from the study. These files contain richer variables, including more detailed postsecondary data, which could help us capture the missing factors that influence GPA. We could also test different hypotheses based on alternative target variables. For example, instead of predicting GPA, we might model outcomes such as degree completion or employment after graduation. These targets may align more closely with the predictors available in the dataset and yield stronger insights.

X. References

[1] National Center for Education Statistics, *High School Longitudinal Study of 2009 (HSLS:09),* U.S. Department of Education, Institute of Education Sciences, 2020. [Online]. Available: https://nces.ed.gov/surveys/hsls09/

[2] Abt Global, *College Affordability Views and College Enrollment,* Abt Global, 2022. [Online]. Available: https://www.abtglobal.com/insights/publications/report/college-affordability-views-and-college-enrollment

[3] Abt Global, *High School Counseling and College Financial Aid,* Abt Global, 2023. [Online]. Available: https://www.abtglobal.com/insights/publications/report/high-school-counseling-and-college-financial-aid

[4] National Archive of Data on Arts and Culture, *Publications, NADAC*, 2016. [Online]. Available: https://www.icpsr.umich.edu/web/NADAC/studies/36423/publications

[5] Institute of Education Sciences, *A First Look at the 2021 Postsecondary Enrollment, Completion, and Financial Aid Outcomes of Fall 2009 Ninth-Graders,* U.S. Department of Education, Institute of Education Sciences, 2020. [Online]. Available: https://nces.ed.gov/pubs2024/2024022.pdf