Introduction

The ICFES examination, also known as Saber 11, stands as a crucial high school exit examination conducted annually in grade 11 across Colombian high schools. This standardized exam, akin to the SAT and ACT in the United States, serves as a comprehensive assessment of students' academic skills and knowledge.

In this research, we want to find explanations of the overall performance of students and potentially predict scores based on socioeconomic features.

Materials and Methods

Dataset 1. Student Performance

The Student Performance Dataset encompasses two sets of data for each year, spanning from 2000 to 2022, covering thousands of students. Our initial research section focuses on the period from 2014 to 2022.

The datasets include variables across various categories such as personal information, contact details, socioeconomic background, school-related information, expectations module, school information, exam appointment details, and exam results.

Dataset 2. Geographical Data in Columbia

We leverage shapefiles obtained from HDX to delineate Colombia's administrative boundaries at levels 0 (country), 1 (department), and 2 (municipality), providing geographical context to our study.

For visualization, we create a choropleth map and a scatter plot to enhance our understanding of the relationship between the average overall score PUNT_GLOBAL in different departments. Specifically, linear regression indicates a negative correlation between the distance to the capital, Bogotá, and the average overall score PUNT_GLOBAL. This suggests that departments closer to the capital may have better educational resources or higher socioeconomic levels within families in these areas.

Dataset 3. Social Economic Indicator in Columbia

We acquire the Socioeconomic Indicator dataset in Colombia from Geoportal. During the data preprocessing phase, we observe that in the Socio-demographic Indicators - Ethnicity datasets, the upper portion displays the peracentage for each state, while the lower part shows the percentage for the entire country. For our research, we exclusively utilize the percentage data for each state. Regarding the dataset on Unmet Basic Needs (NBI) per category %, we specifically choose the data from the year 2018, aligning with our student performance period spanning from 2014 to 2022.

LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) regression is the first machine learning model we performed to predict a student's performance based on social economic indicator in each department in Columbia. Similar to linear regression, LASSO regression models the relationship between a continuous dependent variable and a collection of independent variables, but it additionally performs an L1-regularization to prevent overfitting and perform variable selection.

As a major problem in machine learning, overfitting happens when the model performs well on the training set but not so well on the test data. With a regularization technique, the overfitting problem could be addressed by shrinking the coefficient estimates towards zero. Since the L1-regularization will

move coefficients with the same step size regardless of the magnitude of the coefficient, it allows the coefficients to shrink exactly to zero, leading to sparsity in the model. In LASSO, alpha is the penalty term indicating the amount of shrinkage, with a larger value denoting the greater amount of penalization. Specifically, the objective function in LASSO includes both the residual sum of squares (RSS) in ordinary least square (OLS) regression and the absolute penalty term:

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Our dependent variable, the student's total SAT scores, is a continuous variable, and our independent variables are continuous variables. Therefore, a regression model would be appropriate. Since we are interested in which variables are predictive among all the information we obtained, a procedure of variable selection is necessary to reduce the search space. The L1-norm penalty and resulting sparsity will perform variable selection and modeling simultaneously, so LASSO is suitable for our problem.

In classical linear models, several assumptions such as linearity, independence, homoscedasticity, and normality are required. However, in LASSO regression, those assumptions are less important because LASSO gives a biased estimator by penalizing all model coefficients. To make the results more interpretable, we will use the variables selected by LASSO to further model relationships. 5-fold cross-validation is used to select the best alpha. R square, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are applied to measure the model performance. To assist with the interpretation, we further applied other methods such as decision trees, random forest, using variables selected by LASSO.

Random Forest

Random Forest regression is another powerful machine learning model employed in our analysis to predict student performance based on socioeconomic indicators at each department in Columbia. Unlike traditional linear regression, Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions to improve accuracy and control overfitting.

In the context of our study, Random Forest can effectively handle the complexity of the relationships between the dependent variable (student's total SAT scores) and a multitude of independent variables. The model aggregates predictions from multiple decision trees, providing robustness against overfitting and enhancing the overall predictive performance.

One advantage of Random Forest is its ability to capture non-linear relationships and complex interactions among features, making it suitable for scenarios where the true underlying relationships are intricate. This is particularly relevant in educational settings where student performance is influenced by a diverse set of socioeconomic factors.

Similar to LASSO, Random Forest does not heavily rely on strict assumptions such as linearity, independence, homoscedasticity, and normality. The model's robustness makes it less sensitive to violations of these assumptions, allowing for more flexibility in real-world applications.

To assess the performance of the Random Forest model, we employ common metrics such as RMSE, and MAE. These metrics provide insights into the model's accuracy, precision, and overall predictive

capability. Additionally, we leverage the variables selected by Random Forest for further interpretability and explore potential interactions between features.

Result and Finding

We investigate the correlation coefficients among various subjects for each time period. While the majority of subject pairs display correlation coefficients exceeding 0.7, our analysis reveals a strong linear relationship, particularly emphasized by correlation coefficients surpassing 0.9 between PUNT_RAZONA_CUANTITATIVO & PUNT_MATEMATICAS and PUNT_SOCIALES_CIUDADANAS & PUNT_COMP_CIUDADANA. PUNT_INGLES exhibits a slightly lower correlation coefficient ranging from 0.5 to 0.6 with respect to other subjects and the overall score, PUNT_GLOBAL. This discrepancy may be attributed to the PUNT_GLOBAL formula, which could mitigate the influence of PUNT_INGLES, as well as disparities in grammar or logic in the construction of exam problems compared to other subjects.

Furthermore, we perform ANOVA analyses on datasets for each period, revealing that the factor ESTU_DEPTO_RESIDE significantly influences the average overall score PUNT_GLOBAL. This suggests that the residential area has a notable impact on student academic performance.

Applying 5-fold cross-validation to select the best alpha (0.99901) for LASSO, the model identifies 10 variables from 58, including PERCENTAGE_OF_HOUSEHOLDS_WITH_ACCESS_TO_INTERNET_SERVICES_(HOME_OR_MOBILE)_VALOR , quantitative_housing_deficit_cohabitacion, and others. The R square is 28.55 in the training set and 20.94 in the test set. RMSE is 32.49 for the training data and 36.43 for the test data. MAE is 26.02 for the training data and 29.54 for the test data.

Using the best parameters (max_depth=3 and n_estimators=150) determined by 5-fold cross-validation for the Random Forest model, the top 3 important features are PERCENTAGE_OF_HOUSEHOLDS_WITH_ACCESS_TO_INTERNET_SERVICES_(HOME_OR_MOBILE)_VALOR , QUANTITATIVE_HOUSING_DEFICIT_COHABITACION, and PERCENTAGE_OF_HOUSEHOLDS_WITH_ACCESS_TO_WATER_SERVICES_VALOR. RMSE is 32.60 for the training data and 36.86 for the test data. MAE is 1063.06 for the training data and 1358.74 for the test data.