**Authors**: Bigyan Bhandari, Tyler Stephens, Michael Bednar, Ruth Kanyinda

**Abstract**

**Problem Statement**

The problem we decided to work on was how to determine whether a school was poverty stricken or not. We wanted to see if we could create a model that was able to predict if a school was poverty stricken as accurately or possibly more accurately than the current Title-1 criteria predicts.

**Data Collection**

We gathered our data from the National Center for Education Statistics (NCES) website as it had data covering many years. However, we limited it to the years 1987 and onwards as years before that did not contain much collected data. The categories we started with were all data on students enrolled by race, teacher employments status and student free lunch eligibility and school locale.

**Data Processing**

We utilized pandas to process the data by first removing nan values as attempting to fill them in would result in skewing of the data and inaccuracy in the results. Then, we created new categories by calculating the percentage of students of the categories. We then created a correlation matrix and mapped it to a heatmap to narrow down our features. The features we chose after our exploratory analysis were: percentage of American Indian students, percentage Asian students, percentage of Hispanic students, percentage of African American students, percentage of Caucasian students, school state, school locale, teacher employment status, and total students. For modeling purposes we further classified high poverty and strict poverty of school districts. 1-High, 2-Strict poverty. We attained these standards by looking at various factors ranging from income, free lunch eligibility and a schools title one status. A interesting thing to note is that with just title one being an indicator location doesn’t have much of a correlation to title one status but with our new classifiers we found that a large portion of large city school districts were in strict poverty.

We further processed our features using: Random Forest Classifier (50 leaf, 1000 trees), Mutual Gain, and another Correlation Heatmap. After further processing we chose to drop the teacher employment status and percentage of American Indian students.

**Machine Learning Approaches**

Before committing to a model, we ran exploratory training on 4 different models: Logistic Regression, KNN (50 neighbors), Random Forest, and Logistic Regression with SGD training. The Most accurate models were KNN and Random Forest and we chose KNN because we thought there would be clustering with two of our features being school state and school locale.

After Deciding on a KNN model, we ran more tests with different attributes. There ended up being no significant difference in model accuracy however the model with the highest accuracy and the one we ended up using was a KNN model with 25 neighbors and weight applied by distance.

**Conclusion**

After finishing deciding on a model and what attributes to give it, we trained our model with data from 2014-2019 and then tested it on data from 2019-2020. Our model ended up being very good at correctly predicting that a school was poverty stricken however it was not as good at predicting that a school was not poverty stricken.