The goal of our project was to investigate the MMR rate across schools in 32 states.

We used the Wall Street Journal’s (WSJ) US Measles Vaccination Data in our analysis: <https://github.com/WSJ/measles-data>. The dataset contains data on 32 states and their MMR rates, vaccination exemptions and enrollment count. Our group decided to pursue four algorithms to analyze the data. Kristin pursued decisions trees and random forests. Ben pursued KNN and I pursued linear regression. Each of us coded our own python files with respect to our chosen algorithms. We then combined our codes into one Pyqt5 file for presentation.

The dataset posed a challenge for the linear regression algorithm. The data contained many missing values which had to be imputed. Missing values for Oregon are shown below:

A close up of text on a black background

Description automatically generated

To address the missing values, the means of each column were imputed. I further subset the data to include only numeric variables. I then regressed mmr on enroll, overall, xrel, xmed and xper.

The linear model performed well for several states but performed poorly for others. I suspect the states where the linear model performed poorly are highly imputed. The lack of variation in the data may cause the linear model to fail.

Oregon is a state where the linear model performed well. The coefficient on enroll suggest a one student increase in total enrollment increases the MMR rate by 0.000106percent. The coefficient on the overall MMR rate increases the MMR rate by 0.0983 percent. The coefficient on xmed suggests a one student increase in medical exemption decreases the MMR rate by -0.146 percent. The signs of the coefficients on xrel and xper do not makes sense and suggest poor data and heavy bias for those variables.

A close up of a logo

Description automatically generated

Despite probable bias in xrel and xper, the model performs well and displays a score of 93 percent.

A close up of text on a black background

Description automatically generated

Montana is a state where the linear model performed poorly. The coefficient on enroll suggests that a one student increase in enrollment decreases the MMR rate by -0.0029 percent. The coefficient on overall, xrel and xper are zero because of high imputation. The coefficient on xmed suggests that a one student increase in medical exemption decreases the MMR rate by -10 percent.

A close up of a logo

Description automatically generated

The linear model for Montana does not perform well. It displays large errors and maintains a score of only 2.7 percent.

A close up of text on a black background

Description automatically generated

Overall the linear model performs well in states that are not heavily imputed and poorly in states that are. Given a state that is not heavily imputed, the linear model is able to predict accurately the MMR rate. The variables that affect the MMR rate the most are the exemption variables: xrel, xmed and xper. These exemption variables significantly decrease the MMR rate.

The linear model would benefit greatly from improvements in data quality. Obtaining data on all 50 states and decreasing the amount of missing data would be a great improvement. I used 5 percent of code from the internet and generate 95 percent of my own.