Forecasting Influenza Outbreaks Using Machine Learning Models

Wan, Aaron (School: The Mississippi School for Mathematics and Science)

Accurate flu forecasts are needed to limit the spread of outbreaks. Current flu forecasts have not incorporated several environmental, demographic, and transport factors. This study looks at temperature, precipitation, wind speed, air quality, population density, underage population, elderly population, median household income, vaccination rate, and enplanements to determine which factors influence the spread of influenza. It was hypothesized that using these factors alongside machine learning models can pinpoint the severity of the monthly “Influenza-Like-Illness” (ILI) rates in all fifty states. The data was fit to a LASSO regression model to determine which factors most impacted the spread of the flu. The correlation between each of the factors and ILI rate was calculated. A p-test was performed to determine whether the correlation was statistically significant. The data was incorporated into a K-Nearest Neighbors Classifier, a Random Forest Classifier, a Gaussian Naïve Bayes Classifier, and a Decision Tree Classifier. The K-NN Classifier, the most accurate model, was 99.25% accurate on training data and 89.39% accurate on test data. This model was able to predict the severity of the ILI rates in each of the fifty states in February 2019 with 90% accuracy. This study sheds light on the effect of the studied factors to on flu outbreaks. A statistically significant, negative correlation was found between air quality and ILI rate, meaning that poor air quality should be addressed when tackling flu outbreaks. The ten data sources in this study should be incorporated into other models to improve the accuracy of flu forecasts.