Identifying COVID-19 Hotspots in Florida

By: Sushmita Persaud and Justin Prince

Background on COVID-19

- COVID-19 is caused by SARS-CoV-2
    - First reported cases in Florida on March 1, 2020
    - A sharp increase in cases occurred in June 2020
- Airborne disease
    - This means it is easily communicable especially in densely populated areas
- Prevention Measures include:
    - Curfews
    - Limited human interaction
    - Rapid testing
    - Hygiene Practices

Research Questions

- Main question:
  - What are the hotspot counties in Florida that there is an increased risk of contracting the virus?
  - Is it possible to identify the zones for early response to future outbreaks?
- Dataset
  - Classification problem
  - Explores various external factors
    - Gender
    - Age Distribution
    - Population Density
    - Income

State of the Art

- Previous Research
  - In an article published by NIH, they used R-programming to visualize the hotspots in the United States (Rimal Y, 2021)
  - The research utilizes multiple libraries such as “*(leaflet), (tidy verse), (ggmap), (htmltools), (leaflet. extras), (maps), (ggplot2), (mapproj), (mapdata), (spData)*” (Rimal Y, 2021)
  - They were able to filter through the dataset and identify clusters throughout the states and identify the hotspots through red dots on the map to separate the clusters

Materials and Methods

- Datasets
  - There were two datasets used in this project
  - The Florida Covid Cases dataset was sourced from ArcGIS Hub and archived by the University of South Florida Libraries
  - It contained information on Covid instances and testing metrics for every Florida county
  - The second dataset was sourced from the Covid-19 Data Repository at Johns Hopkins University
  - This dataset contained detailed demographic and population data for every county in the United States
  - Both of our chosen datasets contained numerical data however some records had missing data which meant they were disregarded for the purpose of this study

Materials and Methods

- Methods
  - Preprocessing the data involved removing records and features which were not relevant to the study (we made use of 30 of the 348 features in the JHU dataset)
  - Using the Florida Covid dataset, we determined the infection rate per 1000 residents in each Florida county as of May 2020
  - A rate of 3.52 cases per 1000 residents was one SD above the mean
  - Counties with an infection rate greater than 3.52 cases per 1000 people (as of May 2020) were labeled as hotspots (24 of the 67 counties)
  - The label was added to the JHU dataset allowing us to evaluate the counties by their demographic data to determine the correlation with the hotspot status
  - The addition of the label made this a supervised learning exercise

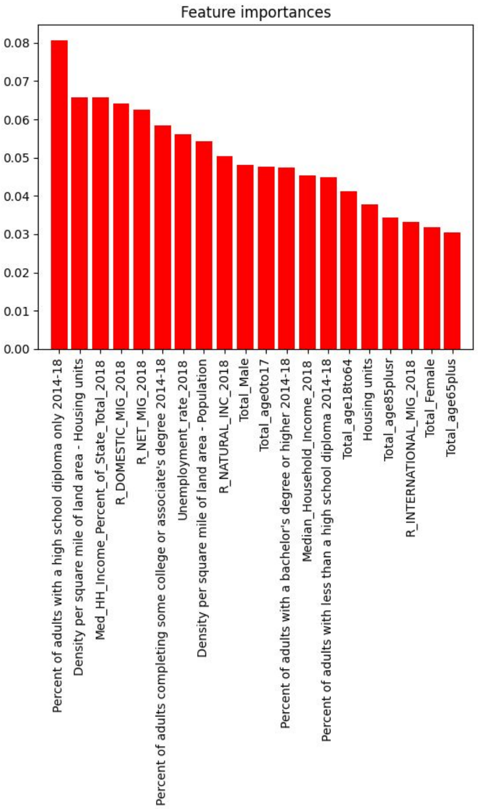
Materials and Methods

- Evaluation
    - The first model applied to the dataset was a Logistic Regression with the goal of identifying hotspot areas. The model was then tuned using cross validation
    - A Support Vector Machine allowed us to rank which counties were the biggest Covid hotspots
    - The SVM was adjusted with grid search and used to predict which areas had outstanding infection rates (above 3.52/1000)
    - A Random Forest Classifier was used to identify the features with the greatest importance when assessing the target variable
    - The Random Forest Model was retrained to identify hotspots based on the population data and optimized with a grid search for the best parameters

Results

- Most Viral Areas
  - The following counties were identified as the top four Covid hotspots in Florida:
    - Dade
    - Broward
    - Palm Beach
    - Hillsborough
- Key Takeaways
  - The three chosen models had high yet varying degrees of accuracy when determining which counties had high infection rates
  - Apart from population density, the most significant factors appear to be socioeconomic conditions
  - Education level, median income and unemployment rate had the highest correlation with high Covid case rates

Results



Hotspot Prediction Performance

Model	Tuning	Score
Random Forest	Grid Search	92.9% (93% after GS)
SVM	Grid Search	78.6 % (86% after GS)
Logistic Regression	Cross Validation	85.7% (75% after CV)

Using grid search to optimize parameters allowed the SVM to adapt to the data without underfitting or overfitting however, the LRM performs worse with cross-validation which suggests that it may not generalize well beyond the original train-test-split

The best prediction model was a RF with parameters: {'max_depth': 3, 'min_samples_leaf': 10, 'min_samples_split': 2, 'n_estimators': 50}

Conclusion and Future Work

- Conclusion
  - It is possible to determine which areas are Covid hotspots based on US census demographic data with a high degree of accuracy
- Future Research Questions
  - Is this research applicable to other airborne viral infections?
  - To what extent does the vector of a disease determine the regions where it spreads most rapidly?