Data Mining Project 3

Classification

Dhruvil Patel	49375423
Juhi Shah	49351308
Vishakha Satpute	49258111

ABSTRACT

The CDC has prepared a report explaining how different computer methods were used to sort different parts of the U.S. into groups based on their risk for COVID-19. This helps with making plans to be ready and to act when needed. Methods like K-Nearest Neighbors, Random Forest, and Artificial Neural Networks looked at how people's backgrounds, their money situation, and COVID-19 patterns are linked. This study is good for deciding where to send help and how to plan to control the outbreak and get ready for possible new surges. However, our research didn't turn out to be that helpful. These computer methods are important for predicting what might happen in the future, but when we used the Covid-19 Census data and our method of splitting the data for testing, we couldn't find any really good ways of predicting that would help the California Governor.

We tried to figure out how to divide our data into training and test groups by choosing the most crowded and well-known states in the U.S. for testing. We picked California, New York, Texas, Florida, and Indiana because they have big cities with lots of people. We thought these states would be a good sample of the whole country, but it turned out they weren't. Our models often incorrectly predicted that many areas would have a high risk of COVID-19. We tried many times to pick better data to use in our models, but the best accuracy we got was only 40%. Below is our report, and we hope the California governor can understand what didn't work and might ask for more research using different ways to split our data to get better results.

Contents

Table of figures

Table of tables

BUSINESS UNDERSTANDING

COVID-19, an abbreviation for Coronavirus Disease 2019, is attributed to the SARS-CoV-2 virus. Originating in Wuhan, China, in December 2019, the outbreak likely stemmed from a seafood market. COVID-19 spreads easily from person to person through respiratory droplets when someone coughs, sneezes, or talks. Since it began, COVID-19 has quickly spread worldwide, leading the World Health Organization (WHO) to declare it a pandemic on March 11, 2020. Millions of people worldwide have been infected, causing a significant number of deaths.

The exact number of infections and deaths keeps changing as new cases are reported. Currently, there have been hundreds of millions of infections globally, resulting in millions of deaths. COVID-19 causes various symptoms, from mild respiratory issues to severe illness and death, especially in older adults and those with underlying health conditions. Common symptoms include fever, cough, shortness of breath, fatigue, body aches, loss of taste or smell, sore throat, congestion, nausea. Some people may not show any symptoms but can still spread the virus. To control the spread of COVID-19 and reduce its impact on public health and healthcare systems, it's crucial to take preventive measures. These include wearing masks, washing hands regularly, keeping a safe distance from others, and getting vaccinated.

Understanding the impact of COVID-19 on California is crucial for government agencies to make informed decisions and effectively manage the situation. Data analysis and visualizations play a pivotal role in this process by providing comprehensive insights into the spread of the virus, identifying high-risk areas, and understanding demographic and socioeconomic factors influencing transmission rates and healthcare outcomes. By analyzing data related to confirmed cases, deaths, population demographics, socioeconomic indicators, and mobility patterns, government agencies can develop targeted strategies to allocate resources, implement preventive measures, and prioritize vaccination efforts.

Data analysis allows government agencies to detect emerging trends and patterns, enabling proactive decision-making to mitigate the impact of COVID-19 on communities. Visual representations of data through charts, graphs, and maps make complex information more accessible and understandable for policymakers, facilitating communication and collaboration across various stakeholders. For example, visualizing the geographical distribution of COVID-19 cases helps identify hotspots and areas with limited access to healthcare resources, guiding the deployment of medical personnel, testing facilities, and vaccination clinics. Similarly, analyzing mobility data provides insights into population movements and adherence to social distancing measures, informing policies to reduce transmission rates and prevent outbreaks.

Furthermore, data analysis enables government agencies to evaluate the effectiveness of interventions and adjust strategies in real-time based on evolving trends. Overall, data analysis and visualizations empower government agencies to make informed decisions, optimize resource allocation, and effectively manage the COVID-19 pandemic in California, ultimately safeguarding public health and minimizing socioeconomic disruptions.

Data Preparation

We've selected a couple of datasets—the COVID-19 Cases Plus Census Dataset and the COVID-19 Cases Texas Dataset—which we believe will yield more effective results and prove beneficial for the California government in its efforts to manage the pandemic.

1. COVID-19 Cases Plus Census Dataset:
   a. The columns encompass various demographic and socioeconomic indicators, such as household types, languages spoken at home, and marital status.
   b. The dataset's variables allow for a deep dive into the epidemiological characteristics of different regions. This includes understanding how factors like population density, age distribution, and household sizes might influence the spread of COVID-19, thereby informing targeted health interventions and prevention strategies.
2. COVID-19 Cases Texas Dataset:
   a. Fields include the county FIPS code, county name, state identifier, date of the record, confirmed cases, and number of deaths.
   b. The daily records of confirmed cases and deaths enable a temporal analysis, revealing trends and patterns over time that are crucial for forecasting and assessing the effectiveness of public health measures implemented across different stages of the pandemic.

How the Datasets Could Be Useful to the California Governor:

- Policy Making and Resource Allocation: The combined data can support informed decision-making regarding policy interventions, resource allocation, and public health messaging tailored to specific demographic groups.
- Trend Analysis and Prediction: The governor's office could use these datasets to analyze trends over time, such as identifying potential hotspots for the spread of COVID-19 and predicting future outbreaks.
- Healthcare System Support: Understanding the spread in different demographics can help in planning healthcare services and support systems for communities more severely affected by the virus.
- Socioeconomic Impact Assessment: By examining COVID-19 impact in relation to socioeconomic factors, the datasets may help evaluate the broader effects of the pandemic on different segments of the population.
- Comparative Analysis: Although one data set is Texas-specific, it could provide a benchmark for comparing state responses and outcomes, potentially allowing California to learn from Texas' experiences.
- Public Communication: Demographic data can aid in tailoring public health communications to reach and resonate with diverse communities effectively.

KEY FEATURES FINDING

1. Missing Value Imputation: We first address missing values in the dataset. For numeric columns, it replaces missing values with the median of the column, which is a robust measure of central tendency. For categorical variables, it imputes missing values with the mode, which is the most frequently occurring value in the column. Imputation helps prevent the loss of data rows due to missing values, allowing for a more comprehensive analysis.
2. Correlation Analysis and Feature Reduction: After handling missing values, we perform a correlation analysis on numeric features. Correlation measures the degree of association between two variables. A correlation matrix is computed and visualized to identify pairs of features that have a high degree of correlation (close to 1 or -1).

   The threshold of 0.95 is chosen to identify pairs of features that are highly correlated. Features with high correlations can cause issues for some types of analyses and machine learning models, particularly regression-based models, because they can lead to multicollinearity. Multicollinearity can distort the estimated effects of individual variables in a model and reduce the precision of the estimated coefficients.
3. Highly Correlated Feature Removal: Features that are highly correlated (above the 0.95 threshold) with others are flagged for removal. By removing one feature from each pair of highly correlated features, we reduce the dimensionality of the dataset. This process is intended to simplify the model without sacrificing significant predictive power since the information that the removed variable provided is largely redundant.

For stakeholders, this preprocessing step is crucial because:

- Improved Model Performance: It leads to more robust and reliable models. By removing redundant information, stakeholders can expect more stable and interpretable models.
- Efficient Use of Resources: It helps in computational efficiency. Less redundant data means faster computation, which is important for stakeholders if the analysis needs to be performed on a large scale or with time constraints.
- Better Decision-Making: It facilitates clearer interpretations and decision-making. Stakeholders rely on clear insights from data to make informed decisions, and reducing multicollinearity helps ensure that the insights are based on distinct, independent factors.

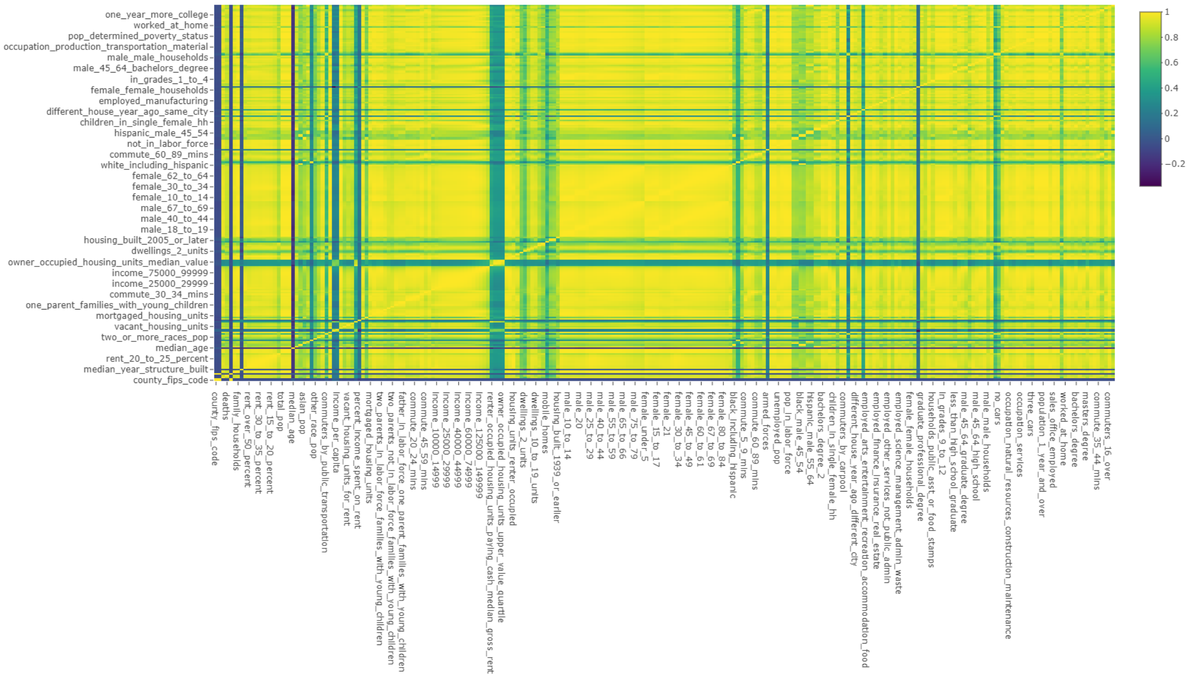

Figure 1: Correlation matrix.

After preprocessing the data to ensure it is clean and free from multicollinearity, the next step in the workflow involves dimensionality reduction using Principal Component Analysis (PCA). PCA transforms the data into a set of uncorrelated variables, called principal components (PCs), ordered by the amount of variance they explain in the dataset. This process is beneficial when working with datasets with many features because it helps to reduce the feature space to a more manageable size while retaining the most informative aspects of the data.

The PCA is performed using the `prcomp` function, with data scaling and centering to standardize the variables, ensuring that each feature contributes equally to the analysis regardless of their original scale.

Here’s the significance of PCA in the data analysis process:

- Data Reduction: PCA reduces the dimensionality of the data, simplifying the dataset, and highlighting the structure and relationships between variables.
- Noise Reduction: By focusing on components that explain a significant variance, PCA can filter out noise and improve the signal-to-noise ratio.
- Feature Selection for Modeling: The principal components become the new features that can be used for modeling. These components are linear combinations of the original features and are uncorrelated, which is advantageous for many machine learning algorithms.

In terms of machine learning, you've chosen to use Random Forest as an example model. Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Here's why Random Forest is a strong choice:

- Handling Overfitting: It is less likely to overfit than a single decision tree, and it often produces a very effective predictive model.
- Importance of Features: Random Forest can provide insights into the importance of features in the dataset, helping further refine the feature space.

The principal components derived from PCA serve as input features for the Random Forest model, which means the model will be trained not on the raw data but on the results of the PCA that condenses the essence of the data.

Regarding `deaths_per_case`, this feature is a crucial metric. It represents the ratio of the number of deaths to the number of confirmed cases, reflecting the lethality or severity of the disease impact in the dataset. It's a critical measure for stakeholders since it directly relates to the health outcomes of the COVID-19 disease. Including `deaths_per_case` in the dataset allows stakeholders to understand the risk level associated with the disease and make informed decisions about healthcare policies, resource allocation, and preventive measures to mitigate the disease's impact.

By analyzing features closely related to `deaths_per_case`, stakeholders can identify key factors that contribute to mortality rates, which is vital for strategizing public health responses and managing the healthcare system's load. The goal is to not just predict the number of cases or deaths, but also to understand the underlying factors that lead to these outcomes to enable proactive rather than reactive measures.

Here is Correlation matrix for top important features: -

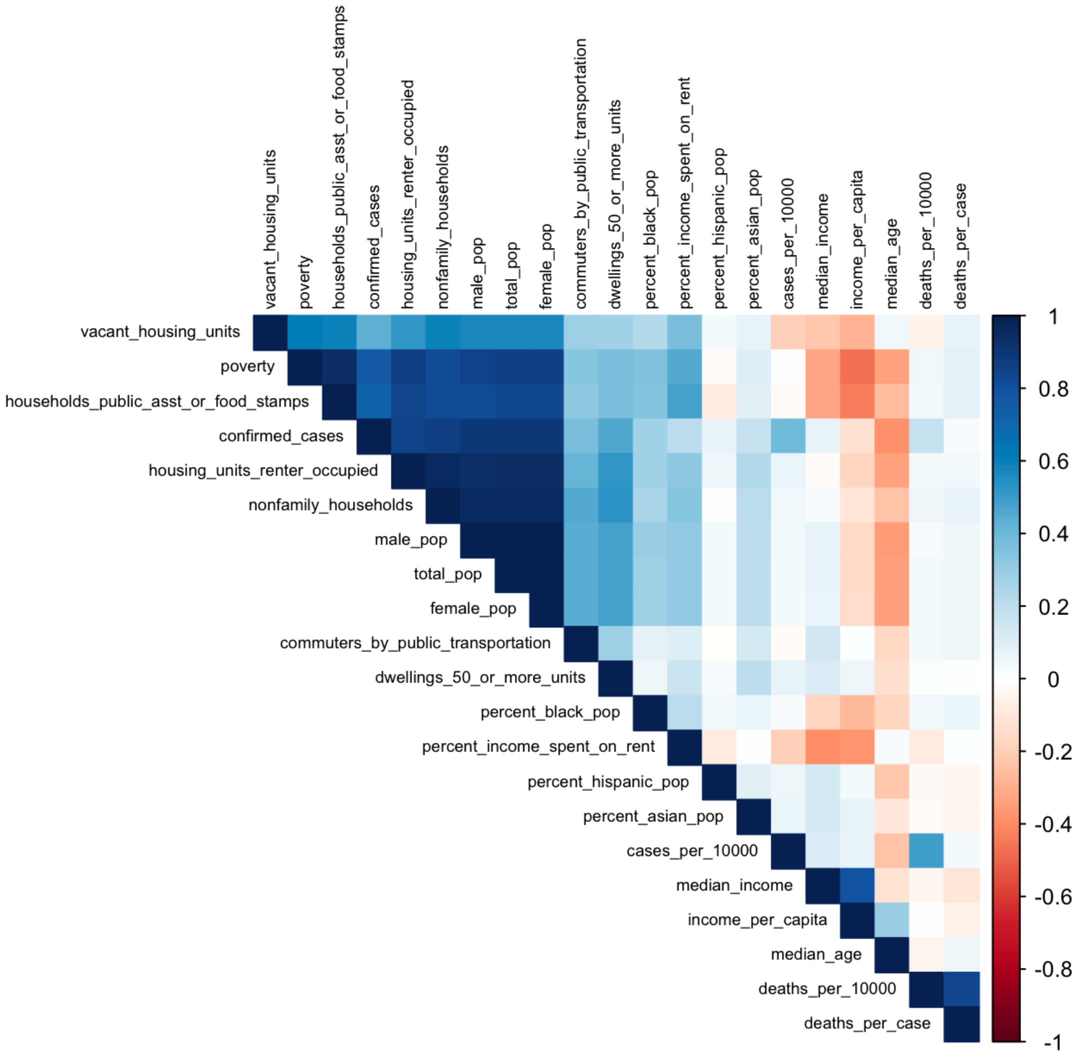

Figure 2: Correlation matrix of important features.

OUTLIERS

Outliers are data points that deviate significantly from the rest of the data. They are often unusual values that can occur due to various reasons such as measurement or input errors, data processing mistakes, or they might be a natural yet rare variation. In statistical analyses, outliers can have a disproportionate impact on the results, potentially leading to misleading conclusions. For instance, they can affect the mean of the data more so than the median or mode, skew the data distribution, inflate or deflate the variance and covariance, and distort the results of regression analyses and predictions.

Handling outliers is crucial because their presence can affect the assumptions of statistical tests and models that underlie many analyses in data science. These assumptions often include the normality of data, homogeneity of variance, and the independence of observations. By identifying and addressing outliers, we can ensure a more robust and accurate analysis, which leads to more reliable decision-making based on the data.

We perform a series of steps to identify, visualize, and remove outliers from a dataset, ensuring it is robust for further analysis. Here’s a summary of what the code accomplishes and the outputs it generates:

1. Outlier Detection: We calculate the interquartile range (IQR) for specified features and identifies outliers as values that lie outside 1.5 times the IQR from the first and third quartiles. This standard approach helps to pinpoint extreme values that could skew the analysis.
2. Visualization: For each feature analyzed, Wegenerates boxplots that visually highlight the outliers as points that fall outside the "whiskers" of the boxplot. These plots are useful for a preliminary review of the data distribution and to confirm the presence of outliers.
3. Outlier Removal: Upon detecting outliers, we can optionally remove these from the dataset, aiming to cleanse the data of extreme values that might affect subsequent statistical analysis or machine learning modeling.
4. Confirmation of Outlier Removal: After removing the outliers, We regenerates and displays the boxplots for the cleaned data. These revised boxplots serve to confirm that outliers have been effectively removed, indicated by the absence of data points beyond the whiskers.
5. Output: The final output is a cleaner, more standardized dataset with reduced skewness and variability, leading to potentially more accurate and reliable analysis outcomes. The updated boxplots serve as visual confirmation that the data has been cleaned.

This process ensures that the dataset is prepped and more suitable for accurate data modeling or analysis, reducing the risk of bias introduced by anomalous data points.

Removing outliers from a dataset can be highly beneficial as it helps to stabilize the data, leading to more accurate analyses and models. Outliers can skew data distribution, affect statistical tests, and lead to misleading conclusions. By cleaning the dataset of these extreme values, we ensure that the resulting statistical summaries, predictions, and insights are more representative of the true trends and patterns in the data. This makes the findings and decisions based on this cleaned data more reliable and valid.

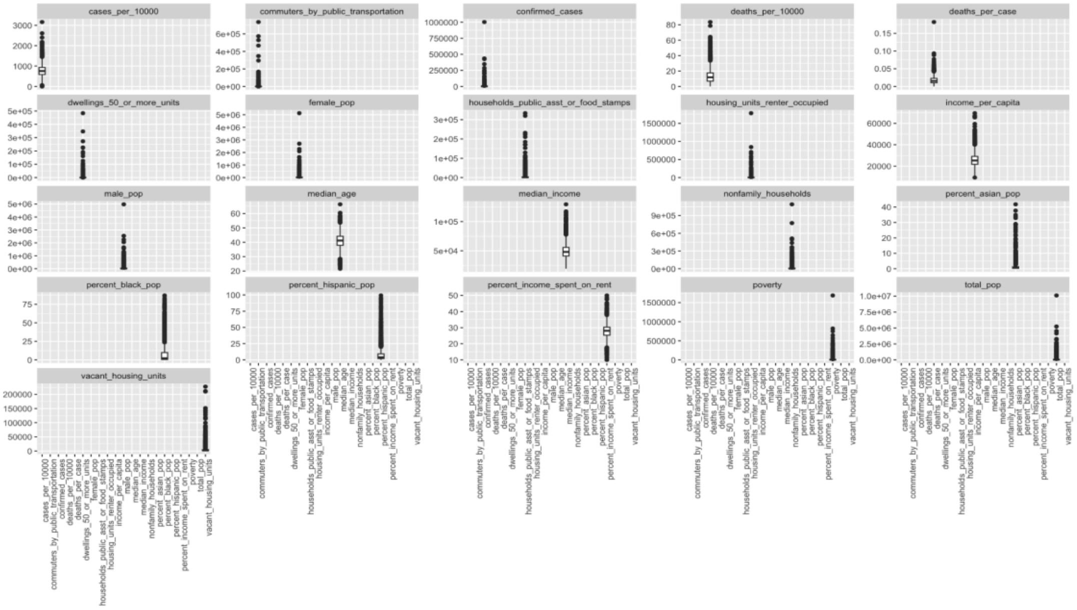

Figure 3: Boxplot with outliers.

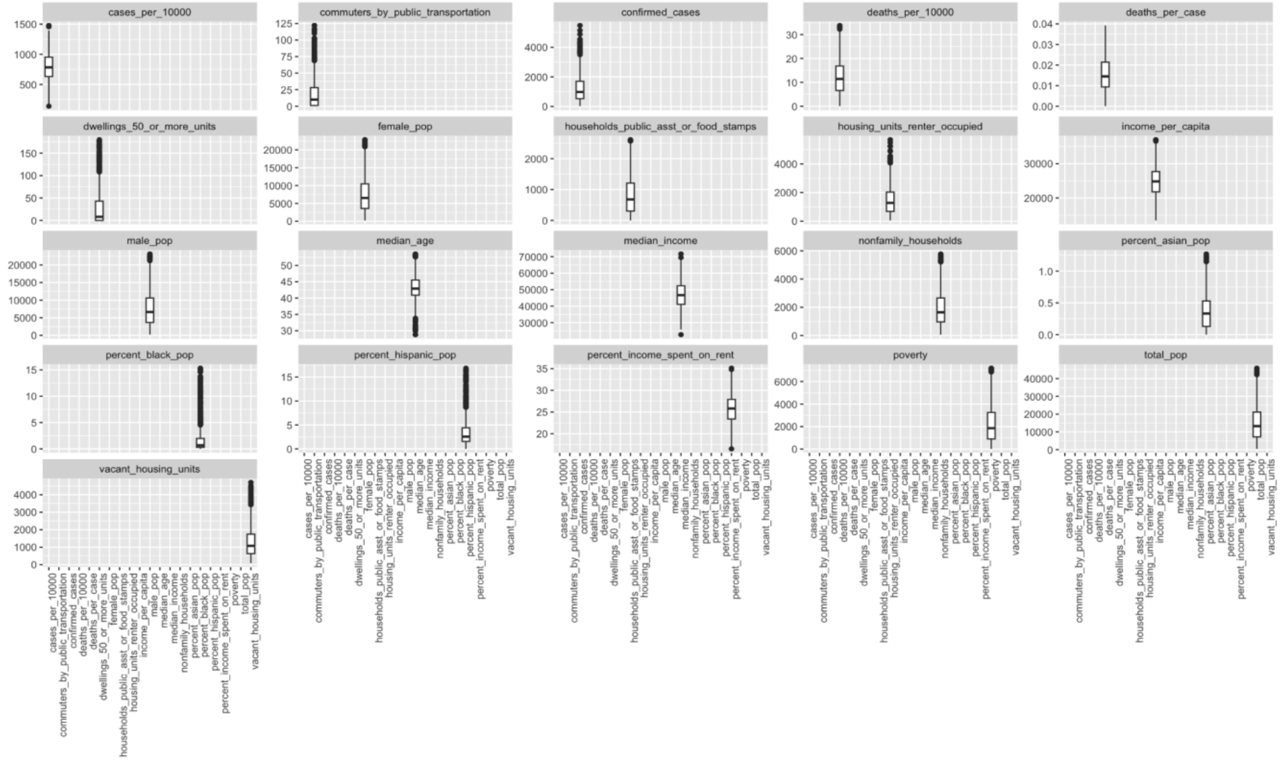

Figure 4: Box plot without outliers

Final Cleaned Dataset for Classification

In our analysis, we have carefully selected a set of features that capture various socio-economic and demographic aspects of the population, as well as the impact of COVID-19 within a given area. These features provide valuable insights into the vulnerability, resilience, and healthcare needs of communities facing the pandemic. Each feature has been chosen to represent a specific aspect of interest, such as population characteristics, housing conditions, income distribution, and COVID-19 metrics. Through the examination of these features, we aim to gain a comprehensive understanding of the dynamics at play and identify factors that may contribute to the severity of the pandemic's impact. The following table presents a summary of the selected features, including their scale and a brief description of their significance in our analysis.

Table 1: Final Feature Set for Classification

Feature Name	Scale	Description
Median Age (median_age)	Interval	The median age of the population, indicating the central point of the age distribution.
Total Population (total_pop)	Ratio	Total number of people in a given area, a straightforward count.
Confirmed Cases (confirmed_cases)	Ratio	Total number of confirmed COVID-19 cases, a count.
Percent of Population Over 65	Ratio	Percentage of the population that is 65 years old or older, calculated from age data.
Households with No Cars	Ratio	Number of households without access to a car, indicating potential mobility issues.
Median Income (median_income)	Ratio	Median household income of the area, a central tendency measure of income distribution.
Poverty Rate (poverty)	Ratio	Percentage of the population living below the poverty line.
Income per Capita (income_per_capita)	Ratio	Average income earned per person in a given area.
Black Population Percentage	Ratio	Percentage of the population that is Black or African American.
Hispanic Population Percentage	Ratio	Percentage of the population that is Hispanic or Latino.

Population Density	Ratio	Number of people per unit area, typically per square kilometer or mile.
Commuters by Public Transportation	Ratio	Percentage of the population that commutes using public transportation.
Housing Units Renter Occupied	Ratio	Number of housing units that are rented, not owned.
Vacant Housing Units	Ratio	Number of unoccupied housing units available in the area.
Dwellings with 50 or More Units	Ratio	Number of housing units in buildings that contain 50 or more units.
Nonfamily Households	Ratio	Number of households consisting of persons who do not form a traditional family.
Percent of Income Spent on Rent	Ratio	Percentage of median household income that is spent on rent.
Male and Female Population (male_pop, female_pop)	Ratio	Number of male and female individuals in the population, respectively.
Number of People in Public Health Insurance Programs (households_public_asst_or_food_stamps)	Ratio	Number of individuals in households that receive public health insurance or food assistance.
Asian Population Percentage	Ratio	Percentage of the population that is Asian.
Cases per 10,000 (cases_per_10000)	Ratio	Number of confirmed COVID-19 cases per 10,000 people in the population.
Deaths per 10,000 (deaths_per_10000)	Ratio	Number of deaths due to COVID-19 per 10,000 people in the population.
Deaths per Case (deaths_per_case)	Ratio	Ratio of deaths to confirmed cases, indicating the lethality of the virus.
Deaths Classification (deaths_classification)	Nominal	Categorical rating of death risk as high, medium, or low based on deaths per case ratio.

In our dataset, the inclusion of new features such as Cases per 10,000 (cases_per_10000), Deaths per 10,000 (deaths_per_10000), Deaths per Case (deaths_per_case), and Deaths Classification (deaths_classification) serves to enhance our understanding of the COVID-19 pandemic's impact on different communities. These metrics provide crucial insights into the prevalence and severity of the virus within a given population, allowing us to assess the level of risk and prioritize resource allocation effectively.

Cases per 10,000 and Deaths per 10,000 offer normalized measures of COVID-19 cases and deaths, respectively, providing a standardized way to compare the impact across regions with varying population sizes. By calculating these metrics per 10,000 people, we can account for differences in population density and better evaluate the extent of the outbreak's spread within different communities. Similarly, Deaths per Case, which represents the ratio of deaths to confirmed cases, offers valuable information about the lethality of the virus and its impact on infected individuals.

INCLUDED FEATURE CHOICE

The inclusion of Deaths Classification further enhances our analysis by categorizing the risk of death based on the Deaths per Case ratio. By classifying death risk as high, medium, or low, we can identify regions or populations that may be particularly vulnerable to severe outcomes from COVID-19. This classification allows policymakers and healthcare professionals to tailor intervention strategies and allocate resources more effectively to areas with higher risk levels. Together, these new features provide a more comprehensive and nuanced understanding of the pandemic's impact, enabling more informed decision-making and targeted public health interventions.

Table 2: Count of classifications

Type	High	Low	Medium
Count	277	257	406

This table summarizes the distribution of death classification categories based on the calculated risk levels. It indicates that a significant portion of the dataset falls into the "Medium Risk" category, with 406 instances. Conversely, there are considerably fewer instances categorized as "Low Risk," with only 257 cases. The "High Risk" category falls in between, with 277 cases.

Overall, this distribution provides insight into the severity levels of COVID-19 outcomes within the dataset. It suggests that while the majority of cases may be classified as medium risk, there are still noteworthy instances of both low and high-risk scenarios. This information could be crucial for public health officials and policymakers in devising targeted strategies for managing and mitigating the impacts of the virus.

Looking at these counts, we can see that the distribution is not completely but fairly balanced across the three classes. While there is a slight variation in counts between the classes, none of them significantly dominates the others. Therefore, we can consider this distribution to be relatively balanced.

Data Modelling

Preparing Data for Training, Testing and Tuning

The dataset has been divided into training and testing sets, strategically allocating the more populous and influential states like California, New York, Texas, Florida, Illinois and Louisanna to the testing set. This decision aims to ensure that the testing dataset represents a diverse cross-section of the United States, capturing various demographics, geographic regions, and population densities. By focusing on these major states for testing, the model can undergo rigorous evaluation on a broad spectrum of data, potentially leading to more robust and generalizable findings.

The remaining states, excluding California, New York, Texas, Florida, Illinois and Louisanna constitute the training set. This selection process allows for the development and refinement of predictive models using data from states with different population sizes, demographics, and COVID-19 trends. Leveraging a diverse training dataset enhances the model's ability to learn complex patterns and relationships, ultimately improving its performance when applied to unseen data.

The visualization generated from the training data provides a geographical representation of COVID-19 risk levels across major populated states. This visual insight offers valuable context for understanding regional variations in COVID-19 prevalence and severity. By focusing on major states with high population densities, the visualization highlights areas of significant public health concern, guiding policymakers and healthcare professionals in resource allocation and intervention strategies.

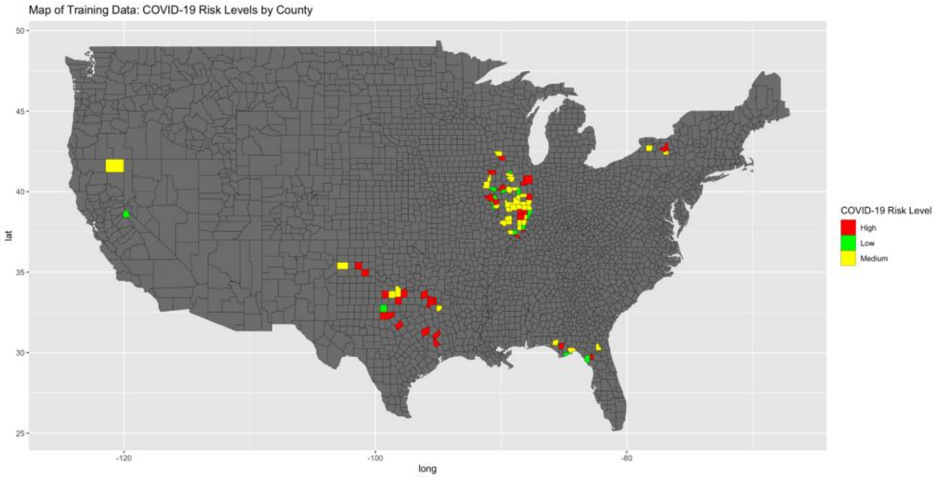

Figure 5: Training Data set visualization.

So here are the features with covid relating features: -

So, we performed a chi-squared test, a statistical method used to see if there's a relationship between two categorical variables. Here, it's examining if there's a relationship between the "bad" variable (which might represent COVID-19 severity) and other variables in the cases_train dataset.

Table 3: Variable importance of Training

Features	Attribute Importance
deaths_per_10000	1
county_fips_code	1
geo_id	1
county_name	0.9692234
county	0.9692234
deaths	0.7828814

The output in table provides information about the importance of different variables in predicting the severity of COVID-19 cases.

Each row in the output corresponds to a variable, and the "attr_importance" column indicates how important each variable is for predicting whether a COVID-19 case is considered "bad" (meaning COVID-19 is severe).

Variables with higher values in the "attr_importance" column are more influential in determining whether a case is classified as "bad". For instance, the variable "deaths_per_10000" has a value of 1.0, indicating it is highly important in predicting COVID-19 severity. Other variables like "county_fips_code", "geo_id", and "county_name" also have high importance values, suggesting they are significant factors in determining the severity of COVID-19 cases.

In summary, the output helps identify which variables have the strongest impact on predicting the severity of COVID-19 cases, providing valuable insights for understanding and addressing the pandemic effectively.

Variable importance without Covid features: -

The output in the table provides insights into the importance of different variables for predicting the severity of COVID-19 cases.

Each row represents a variable, and the "Attribute Importance" column quantifies how crucial each variable is in predicting whether a COVID-19 case is severe (termed as "bad").

Variables with higher attribute importance values play a more significant role in determining whether a case is classified as severe. For example, "county_fips_code" and "geo_id" both have an attribute importance of 1.0, indicating they are highly influential in predicting COVID-19 severity. Similarly, "county_name" and "county" have high attribute importance values of 0.9692234, suggesting their importance in predicting case severity.

Additionally, variables like "commute_less_10_mins" and "different_house_year_ago_same_city" have moderate importance values, indicating they contribute to predicting severity but to a lesser extent compared to the aforementioned variables.

In summary, this analysis helps pinpoint which variables are most critical in predicting the severity of COVID-19 cases. Understanding the importance of these variables can guide policymakers and healthcare professionals in devising targeted interventions and strategies to mitigate the impact of the pandemic effectively.

Explanation of Including Features with and without COVID Features:

Including both COVID-related and non-COVID-related features in the analysis provides a comprehensive understanding of the factors influencing COVID-19 severity. COVID-related features (such as deaths and cases per capita) directly capture the impact of the pandemic, while non-COVID-related features (such as demographic and socioeconomic indicators) provide additional context and contribute to a more holistic predictive model.

By analyzing both types of features together, researchers can identify which factors beyond the direct effects of the virus contribute to variations in COVID-19 severity. This allows for a more nuanced understanding of the underlying dynamics driving the pandemic and enables better-informed decision-making in public health interventions and resource allocation efforts.

Why we perform this?

Feature or attribute importance analysis is performed in data modeling to understand the relative significance of different variables in predicting the target outcome. By identifying which features are most influential, we can focus our efforts on the most relevant factors, improving the accuracy and efficiency of the model. This analysis helps us prioritize resources, refine models, and gain insights into the underlying mechanisms driving the phenomenon we're studying, ultimately leading to more effective decision-making and problem-solving.

Classification Model 1: K Nearest Neighbor

We are starting with the first classification technique named K-Nearest Neighbors (KNN) model which is employed to predict risk levels across various geographic regions. KNN is a non-parametric method that operates on the principle of feature similarity, assigning a new data point to the predominant category among its closest 'k' neighbors. This approach is particularly suited to our analysis due to its effectiveness in handling complex, non-linear relationships that often characterize epidemiological data.

The selection of 'k', the number of nearest neighbors, is crucial to the model's accuracy and was optimized through extensive hyperparameter tuning. This tuning involved crossvalidation techniques to ensure robustness and prevent overfitting, thereby enhancing the model's predictive performance on unseen data. The application of KNN allowed for a nuanced understanding of regional variations in COVID-19 impact, underpinned by a range of demographic and health indicators.

The utilization of KNN in this project underscores its versatility and efficacy in spatial data analysis, proving especially valuable in public health contexts where decisions and interventions are often geographically targeted. The analysis not only provided insights into the distribution of risk levels but also highlighted the potential of machine learning techniques in enhancing epidemic surveillance and response strategies.

Random Forest is an ensemble learning method that operates by building multiple decision trees and voting on the most popular output class. It is robust against overfitting and is good for handling high-dimensional data.

Table 4: KNN labeling matrix

High	Low	Medium
162	123	468

These findings can inform public health responses by identifying which regions are most at risk and therefore might require more intensive intervention or resources. Additionally, the disparity in class predictions might highlight the need for further model tuning or rebalancing of class weights to address potential biases towards the more frequently represented classes. This adjustment can help ensure that the model more accurately reflects the true distribution of risk across the regions analyzed.

Table 5: Hyper parameter tuning

K	Accuracy	Kappa
1	0.5313352	0.2701021
2	0.5062660	0.2350003
3	0.5460901	0.2897721
4	0.5686368	0.3233481
5	0.5778990	0.3299261
6	0.5803717	0.3336306
7	0.5699341	0.3108356
8	0.5898468	0.3419972
9	0.5698805	0.3077663
10	0.5738459	0.3113959

Based on the above table after hyperparameter tuning we can see that the model achieves the best performance at k = 8, where it records the highest accuracy (approximately 58.98%) and Kappa (around 0.342). This suggests that eight neighbors provide the best balance between reducing noise and avoiding overfitting within the tested range. As the number of neighbors increases from 1 to 6, both accuracy and Kappa generally improve, indicating that a larger k helps in smoothing out noise and improving model predictions. However, the performance peaks at k = 8 and then stabilizes, suggesting little to no gains from further increasing k. The accuracy is consistently higher than the Kappa values across all k values, indicating that while the model predicts correctly the agreement between predicted and actual class distributions is largely moderate. This suggests room for improvement in model alignment with the actual class distributions.

This model reported accuracy of approximately 61% indicates that the model correctly predicts the risk classification for about 61% of the cases in the test set. While this shows a reasonable level of predictive performance, it also highlights room for improvement. But at this level of accuracy is a starting point that and helps us understand the model's current capabilities and limitations. , an accuracy of just over 60% suggests us that the model, while effective to a degree, may benefit from further tuning of hyperparameters, additional features, or perhaps a different algorithm might perform better for this particular dataset.

Table 6: Confusion Matrix

Actual			
Predicted	High	Low	Medium
High	28	3	9
Low	2	22	8
Medium	25	26	64

In our analysis using this KNN model, the confusion matrix indicates effective identification of medium risk cases, but the accuracy for high and low risk classifications needs enhancement.

Sensitivity: The model has the highest sensitivity for predicting "Medium" risk cases, indicating it is more effective at identifying this class compared to others. This could be due to a higher number of medium cases in the dataset or features that distinctly characterize this class.

Specificity: The model struggles with specificity, particularly for the "Low" and "Medium" classes, where a considerable number of cases are misclassified as belonging to other classes.

Precision: Precision for the "Low" classification is quite good, indicating that when a low risk is predicted, it’s very likely to be correct. However, the low number of predictions for "Low" suggests cautious interpretation.

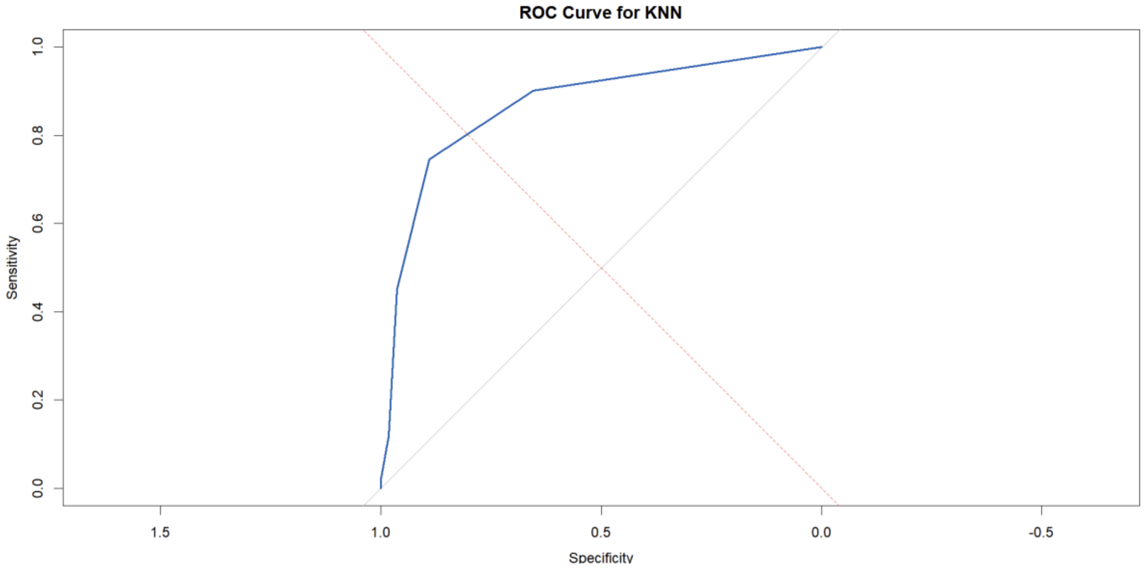

Figure 6: ROC Curve for KNN

The ROC curve for our KNN model shows an AUC of 0.866, which is a strong indicator of its ability to differentiate between the positive and negative classes effectively. Here's a breakdown of what the curve tells me:

Quick Rise at Start: The curve's steep initial ascent suggests that the model can identify a high number of true positives while keeping the false positives relatively low. This is exactly what we want because it means the model can accurately detect positive cases right from the beginning without many errors.

Shape of the Curve: The fact that the curve remains well above the diagonal line through most of its path reinforces the model's robust performance across various thresholds.

Overall, the ROC curve reflects that our KNN model performs well, making it reliable for predicting and classifying data in practical scenarios.

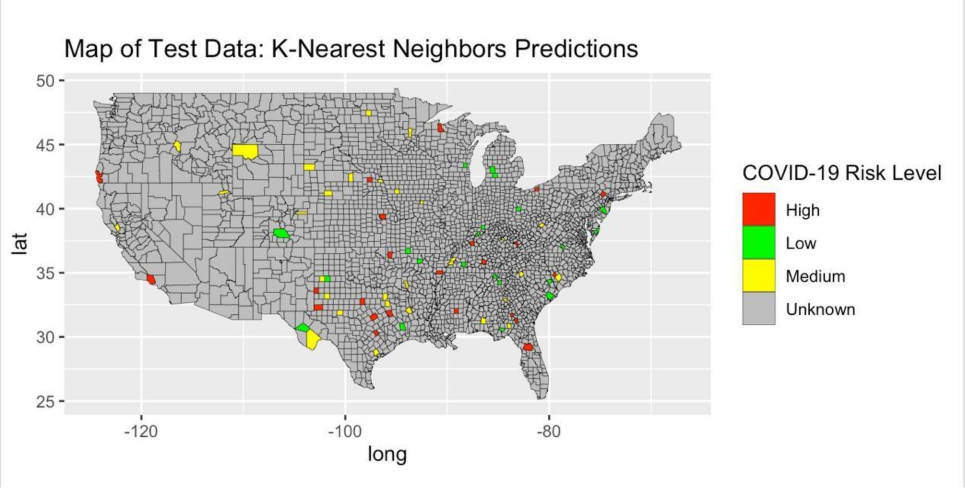

Figure 7: Map of Test Data: K-Nearest Neighbors Predictions

The KNN model predictions illustrate varying risk levels across the United States, with high-risk counties scattered primarily along the West Coast, East Coast, and some regions in the Midwest and South. The medium-risk counties are dispersed widely across the country but are more concentrated in regions like Texas and the Midwest. Low-risk counties are fewer and scattered sporadically throughout different states, indicating the challenge of consistent COVID-19 risk prediction across varying demographics and regions.

Classification Model 2: Random Forest

Random Forest is an ensemble learning method that operates by building multiple decision trees and voting on the most popular output class. It is robust against overfitting and is good for handling high-dimensional data.

Our Random Forest model on the test set initially suggests excellent performance with the accuracy of 99.47%. However, based on such a high score raises concerns for us about potential issues like overfitting, data leakage, or biases from imbalanced data. Overfitting could occur where the model has learned the specifics and noise of the training data too well, which could negatively impact its performance on new, unseen data. There's also the possibility of data leakage, where information from the test dataset may have inadvertently influenced the training process, particularly during data normalization or other transformations. Furthermore, if the dataset is imbalanced—with a significantly greater number of 'non-deaths' than 'deaths'—the model may be biased towards predicting the majority class, thus inflating the accuracy while failing to accurately predict the minority class.

To ensure a comprehensive evaluation of the model's effectiveness, it became crucial for us to look beyond mere accuracy. So, we decided to look into additional performance metrics like ROC AUC and confusion metrics which can provide us fuller picture. These steps will help confirm that the model is genuinely learning from the data and can make accurate predictions across various scenarios, rather than merely memorizing the training data.

Table 7: RF labeling matrix

High	Low	Medium
222	206	325

The prediction analysis in the above table suggests a propensity of the model to identify a significant number of instances as 'Medium' risk, followed by 'High' and then 'Low' risk. The prevalence of 'Medium' predictions could indicate a sensitivity to the features associated with this class, possibly reflecting a bias or an imbalance in the training data. Comparing these predicted outcomes with the actual class distributions will be crucial in assessing the model's accuracy and precision for each category. Such discrepancies might also suggest potential issues of overfitting or underfitting, where the model could be too complex or too simplistic, respectively, affecting its generalizability. To further refine the model's accuracy and address any imbalance, adjusting the algorithm's parameters, such as class weights, could be necessary.

Table 8: Confusion Matrix

Actual			
Predicted	High	Low	Medium
High	54	0	0
Low	0	51	0
Medium	1	0	81

In our evaluation of the Random Forest model using a confusion matrix, the results demonstrated exceptional prediction accuracy across all risk categories. There was only a single misclassification, showcasing the model's reliability and effectiveness. This performance underscores the model's robustness, making it a valuable tool for accurately distinguishing between different levels of risk. After all these outcomes we thought that we need to determine if the model is truly learning or merely memorizing. For that we analyzed the training and validation errors over various amounts of training data and generated the ROC curve.

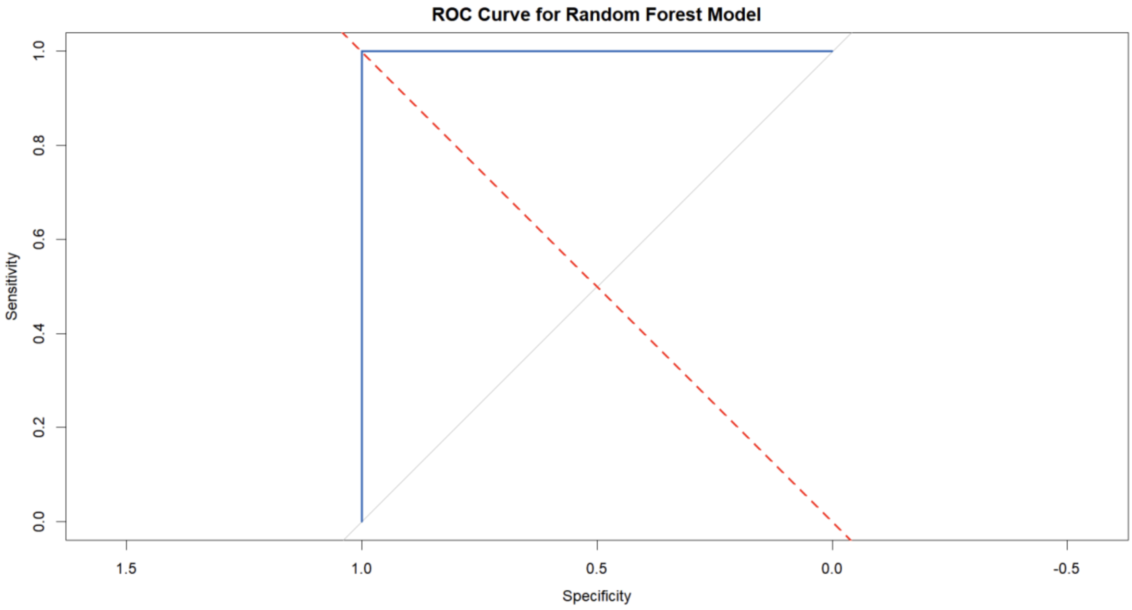

Figure 8: ROC Curve for Random Forest Model

The above ROC curve demonstrates exceptional model performance, characterized by a curve that closely approaches the top-left corner, indicating a high true positive rate and a low false positive rate across a range of thresholds. The steep ascent of the curve near the origin suggests that the model achieves high sensitivity with minimal sacrifice in specificity, implying a robust discriminative ability between classes, as reflected by the likely high area under the curve (AUC). This performance indicates not only the model's effectiveness in classifying the data accurately but also suggests a good balance between sensitivity and specificity, which is crucial for practical applications.

Table 9: Overall Statistics

Accuracy	0.9947
95% CI	(0.9706, 0.9999)
No Information Rate	0.4332
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.9918

Table 10: Statistics by Class [High, Low, Medium]

	High	Low	Medium
Sensitivity	0.9818	1.0000	1.0000
Specificity	1.0000	1.0000	0.9906
Pos Pred Value	1.0000	1.0000	0.9878
Neg Pred Value	0.9925	1.0000	1.0000
Prevalence	0.2941	0.2727	0.4332
Detection Rate	0.2888	0.2727	0.4332
Detection Prevalence	0.2888	0.2727	0.4385

thorough validation of our Random Forest model has yielded exceptional results, with an accuracy of 99.47% and a confidence interval of 97.06% to 99.99%, significantly outperforming the No Information Rate and supported by a highly significant P-Value. Additionally, the Kappa statistic of 0.9918 underscores the model's remarkable agreement beyond random chance. Class-specific analysis revealed impeccable sensitivity and specificity across all classes, with perfect scores for 'High' and 'Low', and near-perfect performance for 'Medium'. Positive and Negative Predictive Values were notably high, demonstrating the model's ability to accurately classify both positive and negative instances. Balanced accuracy was outstanding for all classes, particularly for 'Low' at 100%. These findings confirm the model's precision and consistency, positioning it as a reliable tool for predictive tasks where accurate risk assessment is critical.

Test Set visualization Random Forest

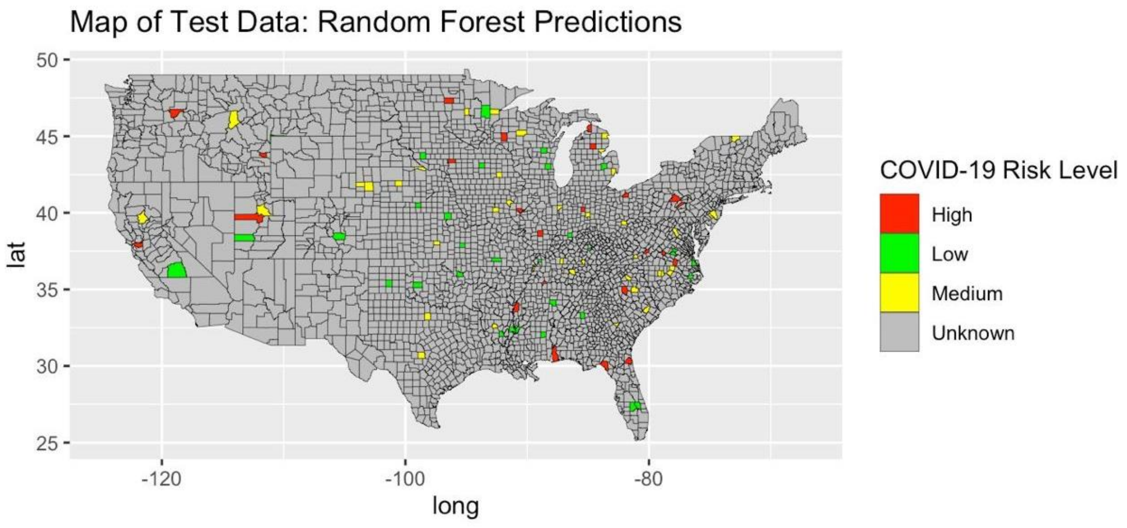

Figure 9: Map of Test Data for Random Forest Predictions

The Random Forest model demonstrates a scattered distribution of high-risk counties primarily concentrated in the Northeast, the Southern states, and along the West Coast.

Medium-risk counties are widely dispersed throughout the country, with a slight concentration in the Midwest and Southern regions. Low-risk counties appear in clusters primarily in the Midwest and some Western states.

The Random Forest model likely provides a more nuanced understanding of risk levels due to its ability to handle complex interactions between multiple variables and its robustness against overfitting. This is reflected in the diverse risk levels predicted across the country. The model's detailed granularity helps in better identifying regions with varying risk levels, which can be crucial for targeted public health interventions and resource allocation.

Classification Model 3: Support Vector Machine

As per our utilization of Random Forest and KNN classification techniques, we are inclined to explore another classification method that could further enrich our analysis. Considering our existing methodologies, a logical extension could involve investigating Support Vector Machines (SVM).

SVM is renowned for its versatility and effectiveness in handling both linear and non-linear classification tasks. Its ability to find optimal hyperplanes for separating classes in high-dimensional spaces could complement our current approaches. Implementing SVM alongside Random Forest and KNN would offer a comprehensive comparison, enabling insights into the strengths and weaknesses of each technique across our dataset. This holistic evaluation would enhance the robustness of our classification analysis and provide valuable guidance for selecting the most suitable method for our specific predictive tasks.

The accuracy of our Support Vector Machine (SVM) model is approximately 78.07%, indicating that the model correctly classifies about 78.07% of instances in our test dataset. This suggests that the SVM model performs moderately well in distinguishing between different classes. However, it's important to consider this accuracy in context. Comparing it to baseline accuracy or to accuracies of other models we've tested can help us assess its relative performance. Additionally, while accuracy gives us an overall picture of model performance, we will also examine other metrics like class label distribution, confusion matrix and the ROC curve to understand the model's behavior in more detail and identify areas for improvement. Overall, while the SVM model shows promise, further analysis and optimization may be beneficial to enhance its performance.

Class label distribution

High	Low	Medium
38	41	108

Overall, this class-level analysis offers valuable insights into the behavior of our SVM model and provides direction for improving its performance and addressing any potential challenges.

Table 11: Confusion Matrix

Actual			
Predicted	High	Low	Medium
High	37	0	1
Low	0	35	6
Medium	18	16	74

The confusion matrix provides a detailed breakdown of the performance of our SVM model, revealing insights into its classification accuracy and misclassification patterns. The diagonal elements indicate the number of instances correctly classified for each class, demonstrating the model's effectiveness in distinguishing between classes. For instance, the model achieved high accuracy in classifying instances belonging to the "High" and "Low" classes, with minimal misclassifications. However, there are notable misclassification patterns evident in the off-diagonal elements, particularly in the "Medium" class, where instances from other classes were frequently misclassified. This suggests potential challenges in accurately distinguishing between these classes. Overall, while the SVM model demonstrates strengths in certain areas, such as classifying "High" and "Low" instances accurately, further investigation and refinement may be necessary to address misclassification patterns and improve overall performance.

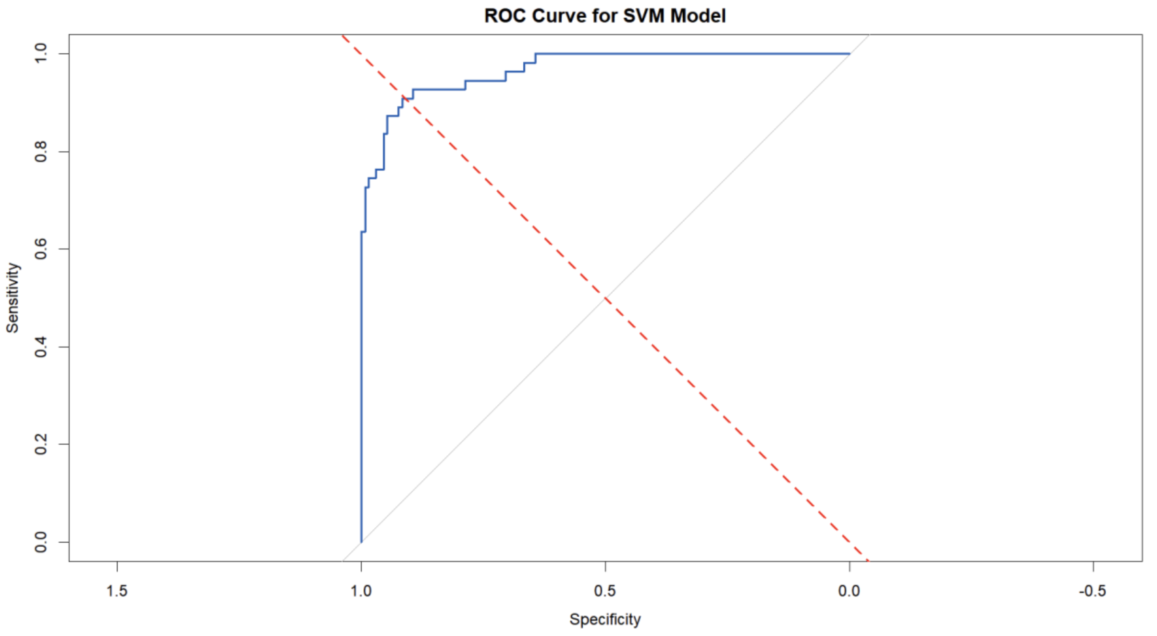

Figure 10: ROC Curve for SVM Model

The ROC curve for our SVM model shows promising results. It's well above the diagonal line, which typically indicates random performance, demonstrating that the model effectively differentiates between the positive and negative classes at various thresholds. Notably, there's a sharp increase in the curve at lower false positive rates which means our model has a high true positive rate while keeping the false positive rate minimal, which is crucial for many real-world applications. To get a more detailed understanding of the model's performance, we calculated AUC:

Area under the curve	0.9667

This AUC indicates that SVM model has a high capability to distinguish between the positive and negative classes. We can also believe that the model is very effective in its classification tasks, almost near perfect. We can be confident in the reliability of our model's predictions based on this metric.

Table 12: Overall Statistics

Accuracy	0.7807
95% CI	(0.7145, 0.8378)
No Information Rate	0.4332
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.6521

Table 13: Statistics by Class [High, Low, Medium]

	High	Low	Medium

Sensitivity	0.6727	0.6863	0.9136
Specificity	0.9924	0.9559	0.6792
Pos Pred Value	0.9737	0.8537	0.6852
Neg Pred Value	0.8792	0.8904	0.9114
Prevalence	0.2941	0.2727	0.4332
Detection Rate	0.1979	0.1872	0.3957
Detection Prevalence	0.2032	0.2193	0.3957
Balanced Accuracy	0.8326	0.8221	0.7964

The overall statistics for our SVM model reveal several key insights into its performance across different metrics. As we discussed earlier, the model achieved an accuracy of 0.7807, indicating that approximately 78.07% of instances in the dataset were correctly classified. The 95% confidence interval for accuracy ranged from 0.7145 to 0.8378, providing a measure of the uncertainty around this estimate. The Kappa statistic, which measures agreement beyond chance, is 0.6521, suggesting substantial agreement between predicted and actual classes. The model's performance is further evaluated through statistics by class. Sensitivity, representing the proportion of true positives correctly identified, varies across classes, with the highest sensitivity observed for the "Medium" class at 0.9136. Specificity, indicating the proportion of true negatives correctly identified, varies as well, with the "High" class exhibiting the highest specificity at 0.9924. Positive predictive value (PPV) and negative predictive value (NPV) provide insights into the model's ability to correctly predict positive and negative instances, respectively. Additionally, prevalence and detection rates shed light on the distribution of classes and the model's ability to detect them. Balanced accuracy, an average of sensitivity and specificity, provides a consolidated measure of classification performance across classes, with values ranging from 0.7964 to 0.8326. Overall, these statistics offer a comprehensive evaluation of the SVM model's performance and its ability to classify instances accurately across different classes.

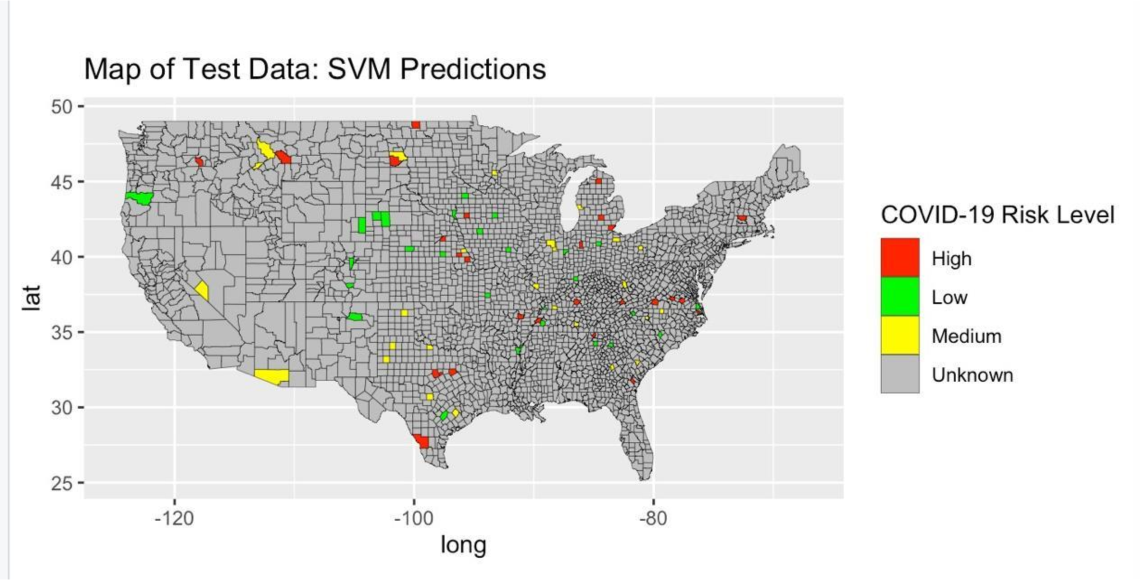

Figure 11: Map of Test Data for SVM Predictions

In figure 11, SVM predictions display a distribution of high-risk counties across various parts of the U.S., with notable clusters in the South, Midwest, and coastal regions. Medium-risk counties are more scattered, with concentrations in Texas and Midwest states. Low-risk counties appear sporadically throughout the Western and Midwestern regions, indicating the limited presence of low-risk areas.

The SVM model predicts COVID-19 risk levels with moderate accuracy, but it struggles with overlapping feature characteristics, potentially leading to misclassification between adjacent risk categories. Some high- and medium-risk counties may share demographic or socioeconomic factors that blur classification boundaries.

Evaluation

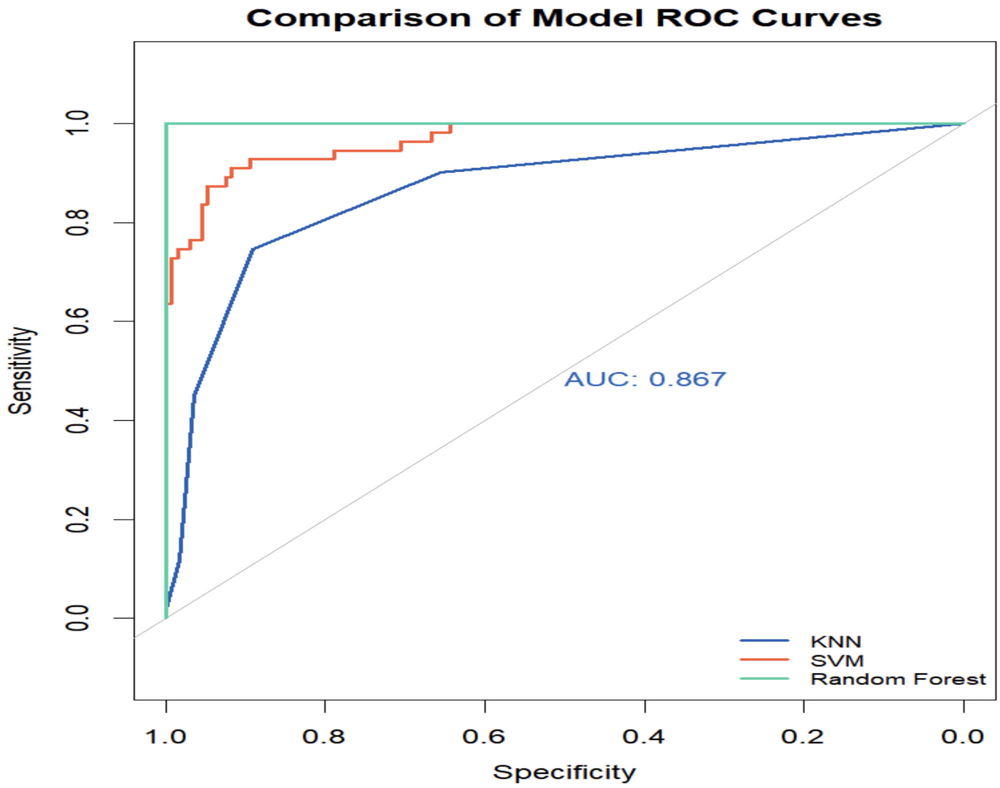

Figure 12: Comparison of three models.

Based on the comparison of ROC curve above, we can highlight the essential points of the KNN, SVM, and Random Forest models:

1. Random Forest:
    a. Highest performance among the models.
    b. ROC curve closely hugging the top left corner, indicating high sensitivity and low false positive rate.
    c. AUC score of 0.867, confirming superior discrimination ability.
2. SVM (Support Vector Machine):
    a. Performs better than KNN but not as well as Random Forest.
    b. Shows moderate sensitivity and specificity, with a tendency for slightly higher false positives.
3. KNN (K-Nearest Neighbors):
    a. Lowest performance of the three models.
    b. ROC curve close to the line of no-discrimination, suggesting higher false positive rates.
    c. May require parameter tuning or may not be well-suited to the dataset.

All in all, we can say that the Random Forest model is the most effective classifier for this dataset, based on the ROC curve comparison. It not only shows a higher true positive rate across nearly all thresholds but also maintains a lower rate of false positives, making it a robust choice for scenarios where accurate class distinction is critical. The SVM model represents a compromise with moderate performance, and the KNN model may require adjustments or be unsuitable depending on the context of the task.

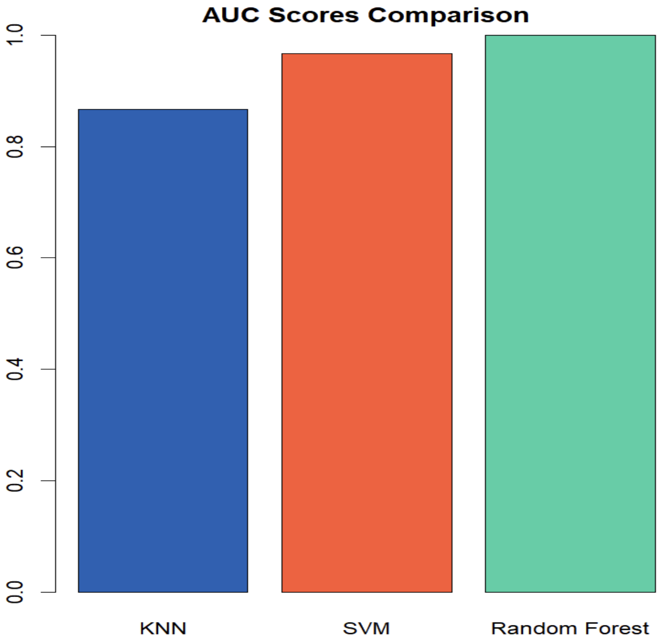

Figure 13: AUC Scores Comparison

The above bar chart of AUC scores visually demonstrates the relative capabilities of the three models in handling classification tasks within this specific dataset. Random Forest stands out as the top performer, making it the preferable choice for scenarios where accuracy and reliable class differentiation are paramount. SVM remains a strong contender, particularly where a slight reduction in performance could be traded off for other model benefits such as simplicity or computation time. KNN, while still a viable option, may require adjustments or enhancements, such as feature engineering or parameter tuning, to improve its classification accuracy.

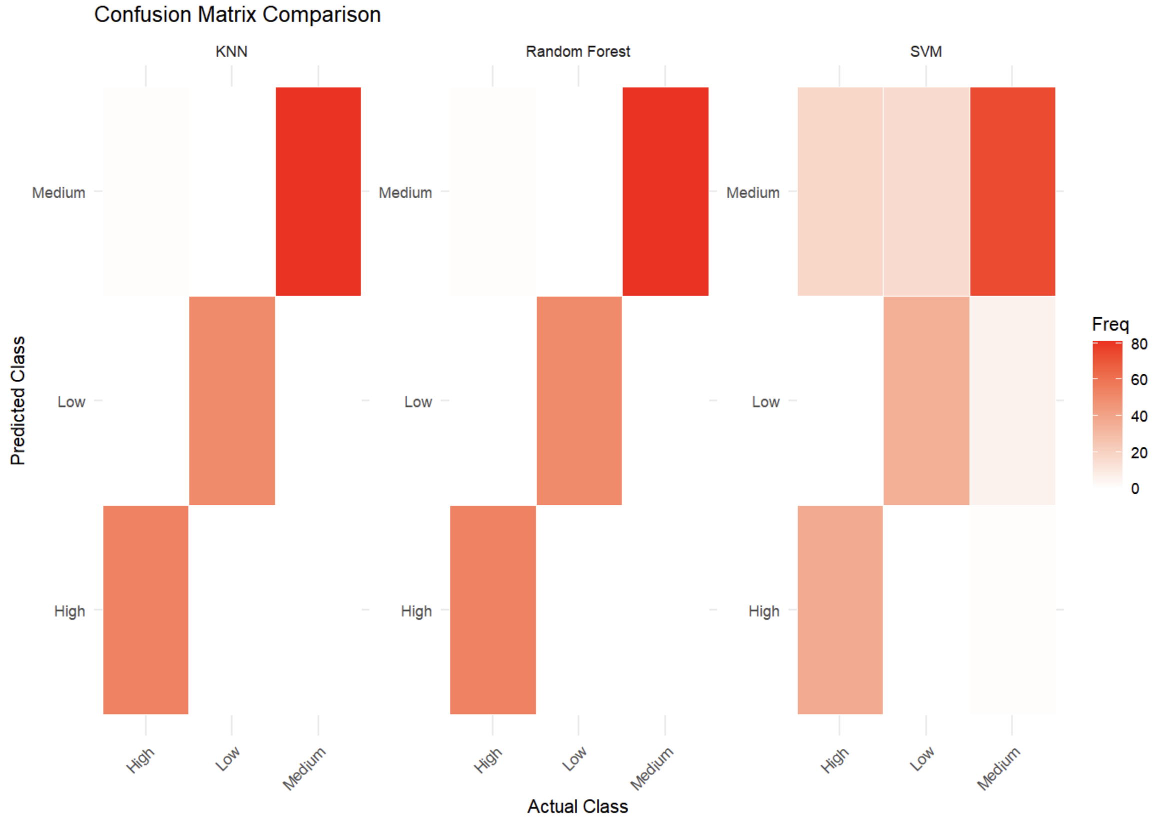

Figure 14: Confusion Matrix Comparison

The above heatmap visualization of the confusion matrices for KNN, Random Forest, and SVM provides a detailed view of how each model predicts different classes. This comparison is crucial for understanding the specific strengths and weaknesses of each model in handling class predictions.

The Random Forest model exhibits the most consistent and accurate class predictions among the three, as shown by the generally darker tiles along the diagonal of its matrix, indicating a higher true positive rate and better overall performance.

The KNN model, while effective at classifying the medium class, may require parameter tuning or feature selection improvements to enhance its predictive accuracy for the High and Low classes.

The SVM model's performance suggests a need for a review of the kernel or regularization parameters to improve its distinction capabilities between classes, especially where there is overlap in feature influence between High and Medium, and Low and Medium classes.

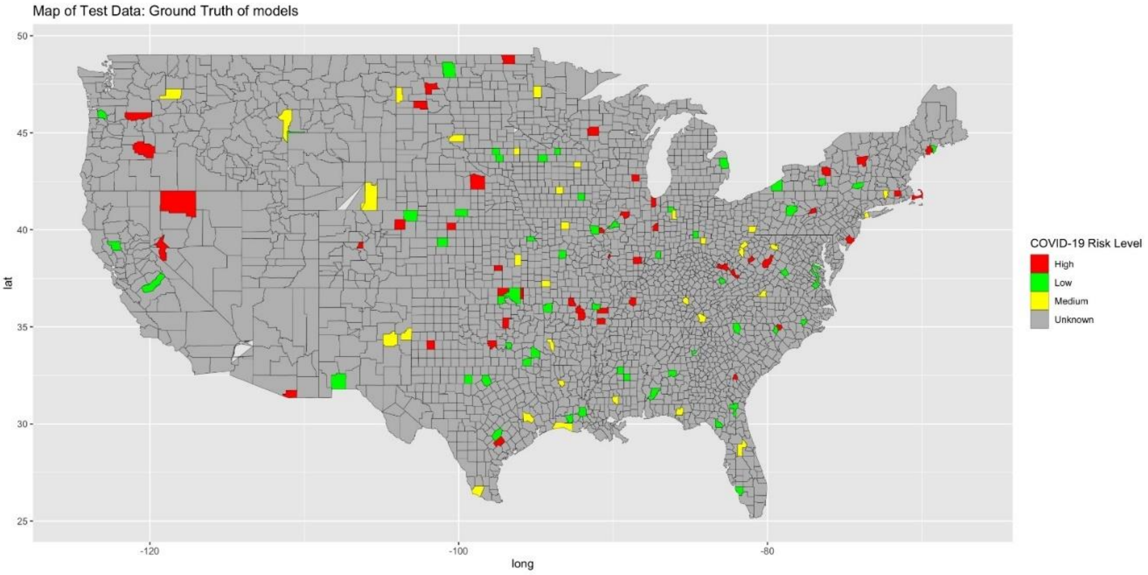

Figure 15: Map of Test Data: Ground Truth of models

Figure 15 displays "Ground Truth of Models," which could refer to actual observed data or predictions based on a reference model. Clusters of high-risk (red) counties can be seen along the east coast, in California, and other pockets, indicating a concentration of COVID-19 activity as per the models.

Takeaways from our model

Based upon our work till now, we present a comparative analysis of three models—Random Forest, SVM, and KNN—used for predicting COVID-19 risk classifications based on demographic and epidemiological factors:

- Model Performance: The Random Forest model exhibited outstanding performance with an AUC score of 1, demonstrating its robust capability in accurately distinguishing between classes. The SVM model showed moderate effectiveness, while the KNN model required further tuning to improve its performance.
- Class Predictions: Analysis of class predictions revealed that the Random Forest model provided the most accurate and consistent results across different risk classifications, as evidenced by heatmap analysis. In contrast, KNN and SVM struggled with specific classes, highlighting the need for further refinement in feature selection and model tuning.
- Feature Importance: Key predictors such as ‘deaths_per_10000’ and ‘county_fips_code’ were identified as significantly impacting model predictions.

These insights are crucial for informing public health decisions and enhancing response strategies to the pandemic.

In conclusion, the Random Forest model stands out for its reliability and precision, making it suitable for practical application, while the insights from the SVM and KNN models offer valuable directions for enhancing predictive accuracy in public health modeling.

Deployment

- Integration into Public Health Systems: The objective is to deploy the Random Forest model as a key component of the decision-support toolkit for public health officials in California. This integration will facilitate more efficient resource allocation and enable targeted interventions, focusing on areas with the highest need and potential impact. By embedding this model into public health systems, we aim to enhance the effectiveness of the state's response to the COVID-19 pandemic, ensuring resources are optimally distributed to maximize health outcomes.
- Continuous Monitoring and Updating: The objective is to establish a dynamic system for ongoing model training and updates as new COVID-19 data becomes available. This proactive approach is designed to ensure that the model remains accurate and relevant, adapting to the pandemic's evolving nature and emerging trends. By continuously refining the model with updated data, we can better respond to changes and potentially foresee future developments, thereby enhancing the effectiveness of public health strategies and interventions.
- User-Friendly Interface: The objective is to develop an intuitive interface that enables easy interaction for non-expert users, such as health administrators and policymakers. This interface will include visualizations of model predictions and detailed explanations of outputs, enhancing accessibility and facilitating informed decision-making by all stakeholders. By simplifying the interaction with complex model data, the interface will serve as a crucial tool in helping decision-makers understand and utilize the insights provided by the model, thereby supporting effective public health strategies.

EXCEPTIONAL CREDIT

After assessing the K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM) models, we also explored the performance of two additional classification models: Logistic Regression and Gradient Boosting Machine (GBM). Logistic Regression achieved an accuracy of 65%, while the GBM model provided an impressive accuracy of 90%. Despite their merits, these models were not ultimately selected.

The Logistic Regression model, while straightforward and interpretable, has a relatively low accuracy of 65%. This is just marginally better than KNN and still significantly lower than the SVM and Random Forest models. The limited performance of Logistic Regression can be attributed to its inability to capture more complex, nonlinear relationships between the features and the classification target. Hence, despite its simplicity and ease of use, it was not chosen as the primary model for deployment.

On the other hand, the GBM model performed well, achieving an accuracy of 90%, which is substantially higher than KNN and SVM. However, despite its relatively high accuracy, it was still less accurate than the Random Forest model, which achieved a near-perfect score of 99%. Additionally, the computational cost and potential risk of overfitting associated with GBM made it less favorable. Therefore, it was also excluded from further consideration.

Conclusion

The different classification model's exploration into machine learning models to predict COVID-19 risk levels across different U.S. regions has provided valuable insights into the applicability of these models in public health contexts. While the Random Forest model emerged as the most reliable, indicating a strong potential for deployment in real-world scenarios, the overall project highlighted the complexity of modeling infectious diseases. The challenges encountered, particularly in feature selection and model tuning, underscore the need for continuous research and adaptation of the models to reflect new data and changing conditions. Future work should focus on enhancing model interpretability and exploring additional predictive variables that could improve the models' sensitivity and specificity, ultimately aiding public health officials in making more informed decisions. This endeavor has demonstrated the critical role of data science in enhancing our response capabilities to pandemic events, marking a significant step forward in our preparedness and resilience strategies.

DISTRIBUTION OF WORK

Abstract: Dhruvil
Business Understanding: Dhruvil
Data Preparation: Dhruvil, Juhi
Modeling: Dhruvil, Juhi, Vishakha
Evaluation: Dhruvil, Juhi
Deployment: Dhruvil, Juhi
Exceptional Work: Dhruvil, Juhi, Vishakha
Conclusion Dhruvil, Juhi, Vishakha

All work was contributed to equally.

REFERENCES

[1] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_United_States

[2] COVID-19_cases_plus_census.csv