Data Mining Project 3

Sujana Daniel Christopher

Mahindra Guptha Kotha

Zhao Chenrui

Project Group 17

Executive Summary: This project aims to analyze COVID-19 data from multiple sources, including case numbers, demographics, and social distancing behaviors across U.S. counties. The primary goal is to understand how factors like population density, income levels, and mobility patterns influenced the spread of the virus and associated outcomes like case numbers and death rates. By exploring these relationships, the project seeks to answer critical questions for government and healthcare policy makers regarding the effectiveness of public health interventions such as social distancing and resource allocation.

The importance of this project lies in its ability to provide insights into how different counties fared during the pandemic. These insights are crucial for making informed decisions in future public health crises. For instance, identifying regions with higher case counts and deaths due to socioeconomic disparities can guide policymakers in ensuring equitable distribution of resources like vaccines and healthcare facilities. Additionally, understanding the role of mobility restrictions can help refine public health campaigns to better control virus transmission.

Key results indicate that counties with lower median incomes experienced higher death rates, likely due to reduced access to healthcare. Meanwhile, areas with higher population density saw faster virus spread, demonstrating the importance of enforcing stricter social distancing measures in such regions. These findings are useful for policymakers to allocate resources effectively and tailor public health measures to the specific needs of various counties, ultimately improving their response to ongoing or future pandemics.

Table of Contents

1. Problem Description

COVID-19, caused by the SARS-CoV-2 virus, emerged in late 2019, leading to a global pandemic that persisted until 2022. The virus, believed to have originated from a seafood market in Wuhan, China, quickly spread across the globe, causing widespread health, economic, and social disruptions. The virus primarily spreads through respiratory droplets, making social distancing, mask-wearing, and vaccinations critical strategies in preventing its transmission.

Governments and healthcare systems worldwide implemented a variety of policies such as lockdowns, travel restrictions, and mask mandates to slow down the spread and prevent healthcare systems from being overwhelmed. "Flattening the curve" became a primary objective, aimed at reducing the rate of new infections and stretching them over a longer period so that hospitals could manage the influx of patients effectively.

One of the greatest challenges of the pandemic was understanding how to manage it more efficiently. Data on virus spread, hospitalizations, deaths, and economic factors became crucial for policymakers and public health professionals to make informed decisions. Governments needed actionable insights to evaluate the effectiveness of safety measures like lockdowns and social distancing, allocate resources such as hospital beds and ventilators, and plan for future pandemics.

The stakeholders in this analysis are governments and public health policymakers. Their goal is to combat the pandemic effectively while balancing the public's health needs and the economy's stability. They need data-driven answers to questions such as:

1. **How does COVID-19 affect different population groups, such as people of various ages or health conditions?**
2. **What are the economic impacts of the pandemic, and how can governments allocate financial resources effectively?**
3. **What safety measures, like lockdowns and social distancing, were most effective in controlling the virus's spread?**

These questions are critical because they guide policymakers in creating strategies that balance economic recovery with public health safety. By examining how COVID-19 affected various population groups and regions differently, governments can better allocate resources and develop targeted interventions to combat the ongoing pandemic and prepare for future public health emergencies.

Required Data:

- **COVID-19 Case Data**: Information about case numbers, hospitalizations, and deaths, broken down by demographic factors such as age and health status.
- **Mobility Data**: Data tracking changes in movement patterns due to social distancing policies.
- **Economic Data**: Information on government spending, stimulus packages, unemployment rates, and other financial indicators to understand the economic impact.

This analysis will explore the relationship between these factors to provide insights into how effective different policies were and how governments can prepare for future public health crises. It will also offer recommendations for improving resource distribution and public health strategies based on the data analysis, ultimately helping to protect both the public's health and the economy.

Additionally, as the global response to COVID-19 continues to evolve, the lessons learned from analyzing this data are not just relevant for the current pandemic, but also for future health emergencies. By identifying patterns in virus transmission, resource utilization, and public compliance with health measures, governments can build more resilient healthcare infrastructures. The insights gained from this analysis could also be applied to improving preparedness for future outbreaks, enabling more rapid and effective responses that save lives and minimize economic damage. In essence, this project underscores the critical role of data in navigating public health crises and shaping policies that ensure a safer and healthier future.

2. Data Collection

In this project, we analyze three key datasets: COVID-19 case data, mobility data, and demographic/economic data. These datasets are essential to understanding the spread of the virus, the impact of social distancing, and the role of socioeconomic factors in the pandemic's outcomes.

Data Sources:

1. **COVID-19 Cases & Deaths Dataset**

   This dataset contains daily reports of confirmed COVID-19 cases and deaths across different counties in the United States. It is sourced from trusted agencies like the Centers for Disease Control and Prevention (CDC), USAFacts, and local health departments. The data is structured by county, making it suitable for regional analysis.

   Source: Google COVID-19 Mobility Reports

A. KEY VARIABLES

Variable	Description	Data Type	Scale
county_fips_code	FIPS code for the county	Numerical	Nominal
County_name	Name of the county	Categorical	Nominal
State	Name of the state	Categorical	Nominal
State_fips_code	FIPS code for the state	Numerical	Nominal
Date	Date of the record	Date/Time	Interval
Confirmed_cases	Number of confirmed COVID-19 cases	Numerical	Ratio
Deaths	Number of COVID-19 related deaths	Numerical	Ratio

Purpose: Core variables for analyzing the spread and impact of COVID-19.

B. Data Types and scales

- **Numerical (Ratio):** confirmed_cases, deaths – These are count data with a true zero point.
- **Categorical (Nominal):** county_fips_code, county_name, state, state_fips_code – Identifiers without inherent order.
- **Date/Time (Interval):** date – Used for temporal analysis.

2. Mobility Changes Dataset

The mobility dataset tracks changes in public movement patterns during the pandemic. It provides insights into how often people visit public places like retail stores, parks, and workplaces, which can be used to analyze the effects of social distancing measures. This data, collected through Google Mobility Reports, measures changes in mobility as a percentage compared to baseline levels before the pandemic.

Source: USAFacts COVID-19 Data

A. KEY VARIABLES

Variable	Description	Data Type	Scale
census_fips_code	FIPS code for the county	Numerical	Nominal
Country_region	Country or region name	Categorical	Nominal
Sub_region_1	Primary subregion	Categorical	Nominal
Sub_region_2	Secondary sub-region	Categorical	Nominal
Metro_area	Metropolitan area name	Categorical	Nominal
Iso_3166_2_code	ISO 3166-2 code for the region	Categorical	Nominal
Date	Date of the moility record	Date/Time	Interval
Retail_and_recreation_percent_change_from_baseline	Percentage change in mobility to retail and recreation places	Numerical	Interval
Grocery_and_pharmacy_percent_change_from_baseline	Percentage change in mobility to grocery and pharmacy stores	Numerical	Interval
Parks_percent_change_from_baseline	Percentage change in mobility to parks	Numerical	Interval
transit_stations_percent_change_from_baseline	Percentage change in mobility to transit stations	Numerical	Interval
workplaces_percent_change_from_baseline	Percentage change in mobility to workplaces	Numerical	Interval
residential_percent_change_from_baseline	Percentage change in residential mobility	Numerical	Interval

Purpose: These metrics provide insight into how population movement changed during the pandemic, which can influence virus transmission rates.

B. Data Types and scales

- **Numerical (Interval):** Mobility change percentages – These are relative changes from a baseline (typically pre-pandemic levels).
- **Categorical (Nominal):** census_fips_code, country_region, sub_region_1, sub_region_2, metro_area, iso_3166_2_code – Geographic identifiers.
- **Date/Time (Interval):** date – Used for temporal alignment and analysis.

C. Columns to omit

- **Redundant Geographic Identifiers:**
  - country_region_code, iso_3166_2_code if analysis is focused on a single country (e.g., the U.S.).
  - sub_region_1, sub_region_2 if adequately captured by census_fips_code.
- **Metro Area:** Omit metro_area if focusing solely on county-level data and if metro areas do not add significant value to your analysis.

3. Demographics & Socioeconomic Factors Dataset i.e., COVID-19_cases_plus_census

This dataset includes various demographic and socioeconomic factors for each county, such as population size, median income, population density, and age distributions. These indicators are crucial for understanding how the pandemic affected different regions and population groups.

Source: USAFacts Texas COVID-19 Data

A. Geographic Identifiers

Variable	Description	Data Type	Scale
county_fips_code	Federal Information Processing standard code for county	Numerical	Nominal
County_name	Name of the county	Categorical	Nominal
state	Name of the state	Categrical	Nominal
State_fips_code	FIPS code for the state	Numerical	Nominal
geo_id	Geographic identifier	Categorical	Nominal

Purpose: These Identifiers are essential for merging datasets and conducting geographically based analyses.

B. Temporal Variable

Variable	Description	Data Type	Scale
date	Date of data record	Date/Time	Interval

Purpose: Critical for temporal analysis and aligning data across datasets.

C. Population Demographics

Variable	Description	Data Type	Scale
total_pop	Total population	Numerical	Ratio
Male_pop	Male population	Numerical	Ratio
Female_pop	Female population	Numerical	Ratio
Median_age	Median age of the population	Numerical	Ratio
White_pop	White population	Numerical	Ratio
Black_pop	Black population	Numerical	Ratio
Asian_pop	Asian population	Numerical	Ratio
Hispanic_pop	Hispanic population	Numerical	Ratio
Amerindian_pop	American Indian population	Numerical	Ratio
Other_race_pop	Population of other races	Numerical	Ratio
Two_or_more_races_pop	Population identifying with two or more races	Numerical	Ratio
Not_hispanic_pop	Population not identifying as hispanic	Numerical	Ratio

Purpose: Understanding population structure and diversity, which can influence COVID-19 spread and outcomes.

D. Socioeconomic Indicators

Variable	Description	Data Type	Scale
median_income	Median household income	Numerical	Ratio
income_per_capita	Income earned per person	Numerical	Ratio
poverty	Percentage or number of individuals below poverty line	Numerical	Ratio
gini_index	Measure of income inequality	Numerical	Interval

Purpose: Socioeconomic status can impact both mobility patterns and vulnerability to COVID-19.

E. Housing Information

Variable	Description	Data Type	Scale
median_year_structure_built	Median year when structures were built	Numerical	Ratio

median_rent	Median monthly rent	Numerical	Ratio
owner_occupied_housing_units_median_value	Median value of owner-occupied housing units	Numerical	Ratio

Purpose: Housing conditions and costs can affect population density and mobility behaviors.

F. Employment Data

Variable	Description	Data Type	Scale
Employed_pop	Number of employed individuals	Numerical	Ratio
Unemployed_pop	Number of unemployed individuals	Numerical	Ratio
Civilian_labor_force	Number of individuals in the civilian labor force	Numerical	Ratio
Income_less_10000 to income_200000_or_more	Income brackets ranging from <$10k to >=$200k	Numerical	Ordinal

Purpose: Employment status and income levels influence economic stability and mobility.

G. Education

Variable	Description	Data Type	Scale
Bachelors_degree_or_higher_25_64	Percentage with bachelor's degree or higher among ages 25-64	Numerical	Ratio

Purpose: Education levels can correlate with health literacy and compliance with public health measures.

H. Transportation

Variable	Description	Data Type	Scale
commuters_by_public_transportation	Number or percentage of commuters using public transit	Numerical	Ratio

Purpose: Usage of public transportation can affect COVID-19 transmission rates.

I. Additional Relevant Variables

Other Variables like median_age, median_year_structure_built, median_income and various household compositions may also be pertinent.

J. Columns to Omit

- **Detailed Age Groups:** Variables like male_under_5, female_5_to_9, etc., unless age stratification is specifically needed.
- **Excessive Housing Details:** Variables such as dwellings_1_units_detached, vacant_housing_units_for_rent, etc., which may not directly influence COVID-19 spread.
- **Highly Granular Income Data:** Detailed income brackets can be aggregated to reduce dimensionality.
- **Detailed Commute Times:** Simplify by using average commute times or broader categories.

- **Redundant or Derived Variables:** Columns that can be derived from other data points or are highly co related with included variables to prevent multicollinearity.

Healthcare Resource Data

The **Healthcare Resource Data** dataset contains vital information regarding healthcare facilities and resources available in various U.S. counties during the COVID-19 pandemic. This dataset is essential for understanding how the availability of medical resources correlates with COVID-19 outcomes such as case counts and mortality rates.

Data Source:

This dataset is typically sourced from public health departments, the Centers for Disease Control and Prevention (CDC), and state health agencies. It includes data reported by hospitals and healthcare providers concerning their capacity to handle COVID-19 patients.

Key Variables:

1. **Hospital Bed Availability**: This variable provides the total number of hospital beds available in each county. This information is crucial for assessing whether healthcare facilities can accommodate a surge in COVID-19 cases.
2. **ICU Bed Availability**: The number of Intensive Care Unit (ICU) beds available for critically ill patients. This metric helps gauge whether hospitals can handle severe cases of COVID-19 that require intensive monitoring and treatment.
3. **Medical Staff Count**: This includes the total number of doctors, nurses, and other healthcare personnel available in each county. Staffing levels are critical to determining how effectively hospitals can respond to increases in patient load.
4. **Ventilator Availability**: The number of ventilators available in each healthcare facility is included to assess whether counties can provide adequate respiratory support for patients suffering from severe respiratory distress due to COVID-19.

Data Quality and Reliability:

The data collected for this project comes from highly reliable sources, such as the CDC and Google, known for producing well-structured, accurate datasets. However, as with any large dataset, some issues regarding data quality must be addressed:

- Missing Data: In the mobility dataset, certain fields contain missing values, particularly for smaller or rural counties where data may not have been reported as consistently. We addressed this by either imputing missing values using the mean of available data or removing columns that contained a significant percentage of missing data (more than 50%).
- Duplicate Data: The case and demographic data was thoroughly checked for duplicate entries to ensure data integrity. Any duplicate records found were removed to avoid double-counting cases or skewing results.
- Outliers: We detected outliers in the COVID-19 case and death data using the Interquartile Range (IQR) method. Extreme outliers, which might represent reporting errors or unusual spikes in case numbers, were either removed or flagged for further investigation.

Data Cleaning:
Before beginning the analysis, we took steps to standardize and clean the datasets for better readability and usability:

- Renaming Variables: The original variable names from the datasets were long and inconsistent (e.g., median_year_structure_built). To improve clarity, we renamed variables to shorter, more intuitive names (e.g., median_year_built).
- Consistent Formatting: We ensured that all variable names followed a consistent format, using camel case (e.g., totalPopulation or medianIncome) or snake case (e.g., total_population) for easy readability during coding.
- Standardizing Date Formats: In datasets that track time (such as COVID-19 case data), we standardized the date formats to a consistent format (YYYY-MM-DD) across all datasets to make merging and analysis easier.

Combining the Data:
To perform a comprehensive analysis, we merged these datasets on the basis of common fields, such as county name and date. This allowed us to create a unified dataset where each county has information on COVID-19 cases, mobility trends, and demographic factors. We used left joins to ensure that no important data was excluded during the merging process, even if certain counties lacked mobility or demographic data.

After merging, the resulting dataset was reviewed for any inconsistencies or further missing values introduced during the join process. By cleaning and standardizing the data in this way, we ensured that it was ready for the next steps of exploration and modeling.

3. Data Preparation

3.1 Defining Classes

In our classification task, we aim to categorize counties into **risk levels** based on the relative number of confirmed COVID-19 cases per week, scaled to a population of 100. This metric, denoted as cases_per_100, provides a standardized measure for comparing counties of varying population sizes.

The defined classes are:

1. **Low Risk:** cases_per_100 < 1
    - Counties with fewer than 1 confirmed case per 100 people are considered low risk. These regions exhibit minimal transmission of the virus.
2. **Medium Risk:** 1 ≤ cases_per_100 < 10
    - Counties with confirmed cases between 1 and 10 per 100 people are categorized as medium risk. These areas show noticeable, though moderate, spread of the virus.
3. **High Risk:** cases_per_100 ≥ 10
    - Counties with 10 or more confirmed cases per 100 people are classified as high risk, indicating widespread and potentially uncontrolled transmission.

Rationale for Defining Classes

The classification thresholds were determined based on a combination of public health considerations, data distribution analysis, and interpretability for stakeholders.

1. Public Health Relevance

- The thresholds align with real-world pandemic response priorities:
    - **Low Risk Counties:** Require routine monitoring but minimal intervention.
    - **Medium Risk Counties:** Need moderate public health measures, such as contact tracing or localized lockdowns.
    - **High Risk Counties:** Demand significant interventions, including widespread testing, vaccination campaigns, and strict containment policies.

2. Standardized Metric: cases_per_100

- By scaling cases to a per-100 population basis, the metric accounts for population differences across counties. This standardization ensures fair comparisons between densely populated urban areas and sparsely populated rural regions.
- Weekly aggregation provides timely and actionable insights, enabling decision-makers to respond to trends in real-time.

3. Data-Driven Thresholds

- Thresholds were chosen based on an exploratory analysis of the dataset:
    - **Summary Statistics:** Examining the range, mean, and quantiles of cases_per_100 provided insights into natural groupings within the data.
    - **Visualizations:** Histograms and boxplots helped identify inflection points where case densities shifted significantly across counties.
    - **Public Health Guidelines:** Thresholds align with common epidemiological markers, such as attack rates or case densities used in outbreak assessments.

Methodology for Class Definition

Step 1: Compute cases_per_100

- The metric cases_per_100 was derived as:

  cases_per_100 = (total_cases / total_population) x 100

- This ensures a comparable measure across counties of varying population sizes.

Step 2: Analyze the Data Distribution

- Summary statistics were calculated for cases_per_100:
  r code:
  summary(combined_data$cases_per_100)
  hist(combined_data$cases_per_100, breaks = 50, col = "skyblue")

  - Key findings:
    - **Most counties** had low case rates (< 1 case per 100), corresponding to low transmission levels.
    - A smaller group of counties fell into the range of moderate transmission (1-10 cases per 100).
    - A few outliers exhibited extremely high transmission rates, justifying the high-risk category.

Step 3: Define Thresholds

- Based on the distribution and public health guidelines:
  - **Low Risk (cases_per_100 < 1)**: Reflects regions with minimal transmission and effective containment.
  - **Medium Risk (1 ≤ cases_per_100 < 10)**: Captures counties with moderate, manageable transmission levels.
  - **High Risk (cases_per_100 ≥ 10)**: Represents areas with significant transmission, often requiring urgent intervention.

Step 4: Assign Risk Levels

- Using the case_when function in R, counties were categorized as follows

r code:

```
combined_data <- combined_data %>%
  mutate(
    risk_level = case_when(
      cases_per_100 < 1 ~ "low",
      cases_per_100 < 10 ~ "medium",
      TRUE ~ "high"
    )
  )
```

Step 5: Validate Class Distribution

- The resulting distribution of risk levels was analyzed to ensure balanced representation across classes:

r code:

```
table(combined_data$risk_level)
```

Significance of Class Definitions

The chosen classes serve both analytical and practical purposes:

1. **Actionable Insights for Stakeholders:**
    - Policymakers can prioritize resources and interventions based on risk levels.
    - For example, low-risk counties might focus on preventative measures, while high-risk counties may need immediate outbreak control.
2. **Alignment with Public Health Goals:**
    - The thresholds align with epidemiological metrics, such as the **attack rate**, ensuring relevance to real-world pandemic scenarios.
3. **Scalability:**
    - The standardized cases_per_100 metric can be recalculated weekly, adapting to changing case dynamics and guiding ongoing public health strategies.

Explanation of Class Definitions

The classes for county-level COVID-19 risk categorization (low, medium, and high) were defined based on an in-depth analysis of the data and the need for a practical, actionable framework for understanding pandemic impact. This decision-making process considered factors such as the distribution of cases, population dynamics, and public health relevance. Below is a detailed breakdown of the reasoning behind the chosen class definitions.

1. Data-Driven Thresholds

The decision to define thresholds for low, medium, and high risk was guided by an exploratory analysis of the cases_per_100 metric, representing the number of confirmed COVID-19 cases per 100 individuals in a county's population.

Data Exploration and Observations

1. **Range of Values:**
    - The dataset's range for cases_per_100 showed a strong skew, with most counties reporting low case densities and a few outliers experiencing extremely high case rates.
    - Example:
        - **Minimum:** 0
        - **Median:** 0
        - **Mean:** 0.54
        - **Maximum:** 11.18
2. **Distribution Patterns:**
    - A histogram of the cases_per_100 metric revealed that the majority of counties had minimal case counts (<1 case per 100), with only a small percentage exceeding 1 case per 100 individuals.
    - Few counties, often those with dense populations or higher mobility, exhibited significant outliers with over 10 cases per 100.

Key Insights:

- The natural grouping of counties into distinct clusters of low, medium, and high case densities suggested the thresholds:
    - low (<1)
    - medium (1-10)
    - high (≥10)

2. Public Health Relevance

The defined classes align with real-world public health objectives by offering actionable insights into the pandemic's impact:

1. **Low Risk (cases_per_100 < 1):**
    - Counties with fewer than 1 confirmed case per 100 individuals demonstrate minimal transmission.
    - These areas likely benefited from effective containment strategies or naturally low exposure risk due to factors like low population density.
2. **Medium Risk (1 ≤ cases_per_100 < 10):**
    - Counties in this range experience moderate transmission.
    - These regions may require localized public health measures, such as contact tracing, mask mandates, or community testing.
3. **High Risk (cases_per_100 ≥ 10):**
    - Counties with at least 10 cases per 100 individuals represent significant transmission zones, likely requiring urgent interventions like vaccination drives, travel restrictions, or lockdown measures.

Why These Thresholds?

- The thresholds simplify complex data into intuitive categories for policymakers.
- They provide a framework for prioritizing resources:
    - Low-risk counties need minimal intervention.
    - Medium-risk counties require proactive monitoring and support.
    - High-risk counties demand immediate containment and resource allocation.

3. Alignment with Epidemiological Metrics

The class thresholds are rooted in epidemiological principles commonly used to assess outbreak severity:

1. **Attack Rates:**
    - Epidemiologists often evaluate outbreaks using the proportion of individuals affected in a population. The cases_per_100 metric corresponds to an attack rate, scaled for interpretability.
2. **Threshold Validation:**
    - Public health organizations frequently classify areas by similar risk levels (e.g., "low, moderate, high") for surveillance and intervention purposes.
    - Our thresholds ensure alignment with these established practices.

4. Practical Interpretability

The selected thresholds make the risk_level metric actionable and easy to communicate:

1. **Low Risk (cases_per_100 < 1):**
    - Counties in this category are likely not experiencing community spread and represent safe zones.
    - Public health officials can allocate fewer resources here, focusing on preventative measures.
2. **Medium Risk (1 ≤ cases_per_100 < 10):**
    - These counties represent regions with noticeable transmission but not overwhelming strain on healthcare resources.
    - Moderate interventions can significantly curb transmission.

3. **High Risk (cases_per_100 ≥ 10):**
    - High-risk counties are hotspots requiring urgent public health measures.
    - These areas are often characterized by high population density, mobility, or socio-economic factors exacerbating spread.

5. Addressing Data Distribution Bias

The thresholds were defined to address the inherent skewness in the dataset:

- **Skewed Data:** A large proportion of counties reported zero or minimal cases, which would dominate a binary classification system (e.g., low/high).
- **Three Classes:** Adding the medium-risk category allows for a nuanced understanding of areas with rising transmission, helping prevent escalation to high risk.

6. Threshold Justification in the Context of the Dataset

- The defined thresholds balance statistical insights with real-world applicability. For example:
    - Counties with cases_per_100 < 1 are statistically dominant but not epidemiologically significant as high-transmission zones.
    - The medium threshold (1 ≤ cases_per_100 < 10) bridges the gap between minimal and significant transmission, ensuring balanced representation across classes.

Conclusion

The defined classes (low, medium, high) are rooted in data analysis and public health objectives, providing a robust and interpretable framework for pandemic impact assessment. These thresholds enable targeted interventions and resource allocation while capturing the nuances of county-level transmission dynamics. By addressing both the statistical distribution and the practical implications, this classification system aligns with the overarching goal of supporting informed public health decision-making.

3.2 **Integration and preparing the Dataset for Classification**

Combining multiple datasets is a critical step in preparing a cohesive and feature-rich table for classification tasks. The process ensures that all relevant data points are integrated into a single dataset, providing the necessary inputs for model training and prediction. This process also includes the creation of a class attribute, which serves as the target variable for classification models.

Objective

The goal of combining the files was to produce a unified dataset that:

1. Incorporates features from all available data sources.
2. Standardizes data formats for consistency and usability.
3. Includes a well-defined class attribute (risk_level) that categorizes counties based on their COVID-19 case density.
4. Provides a comprehensive dataset suitable for training classification models.

Data Sources

The dataset preparation involved integrating three primary data sources:

1. **Demographic and Case Data:** A dataset containing county-level demographic features (e.g., population size, median income) and cumulative COVID-19 statistics, such as total confirmed cases and deaths.

2. **Daily Case Updates:** A dataset with time-series COVID-19 data, detailing new cases and deaths reported daily for each county.
3. **Mobility Trends:** A dataset tracking behavioral changes during the pandemic, including shifts in workplace visits, grocery shopping, and recreation.

Steps in Combining Files

1. Identifying Common Keys

To merge the datasets, a common key (county_name) was used to link the records across different files. This ensured that data from the same county were aligned correctly in the final dataset. In cases where the key formats were inconsistent, preprocessing steps (e.g., standardizing names) were performed to ensure compatibility.

2. Merging the Datasets

The datasets were sequentially combined to create a single table. The integration process involved retaining all relevant counties and their associated features. Each dataset contributed unique information:

- **Demographic and case data** provided static attributes, such as total population and median income.
- **Daily case updates** added dynamic attributes, such as weekly case trends.
- **Mobility trends** contributed behavioral patterns, such as average workplace visits during the pandemic.

3. Deriving Features

After merging, new features were created to enhance the dataset's predictive power. These included:

- **Case Density (cases_per_100)**: This metric, calculated as the number of cases per 100 individuals, normalized the case data for population differences and allowed fair comparisons across counties.
- **Behavioral Averages:** Mobility features, such as workplace and recreation trends, were aggregated to represent average changes over time.

4. Creating the Class Attribute

The class attribute, risk_level, was defined to categorize counties based on their COVID-19 impact. The classes were:

- **Low Risk:** Counties with minimal transmission rates.
- **Medium Risk:** Counties with moderate transmission, requiring some public health interventions.
- **High Risk:** Counties with significant transmission, representing hotspots. This attribute was derived using thresholds based on the cases_per_100 feature, informed by data distribution and public health relevance.

5. Handling Missing Data

Missing values in critical features, such as mobility data or case statistics, were addressed to ensure the dataset's completeness:

- Demographic attributes (e.g., population size) were imputed using central tendency measures.
- Mobility trends and other dynamic features were filled with aggregated values, ensuring minimal disruption to data integrity.

6. Ensuring Data Quality

Post-merging, the dataset underwent rigorous quality checks to validate the integration:

- Verifying that all counties were represented.
- Ensuring no duplicate records existed.
- Confirming that the class attribute (risk_level) was correctly assigned.

Challenges Encountered

1. **Inconsistent Keys:** County names were not consistently formatted across datasets, requiring preprocessing to standardize them.
2. **Missing Data:** Some counties lacked information in one or more datasets, necessitating thoughtful imputation strategies to avoid bias.
3. **Skewed Data Distribution:** Case densities were heavily skewed, with a majority of counties in the low-risk category. This required careful threshold selection to define meaningful classes.
4. **Feature Compatibility:** Ensuring that features from different datasets aligned conceptually and numerically posed challenges during integration.

Resulting Dataset

The final combined dataset included the following:

1. **Demographic Features:** Population size, median income.
2. **COVID-19 Statistics:** Total cases, total deaths, normalized case density (cases_per_100).
3. **Mobility Trends:** Behavioral changes in workplace visits, grocery shopping, and recreation.
4. **Class Attribute (risk_level)**: A categorical variable categorizing counties into low, medium, and high-risk levels.

Significance of Combining Files

1. **Comprehensive Analysis:** Combining diverse data sources ensured that the dataset captured various dimensions of the pandemic, from demographic vulnerabilities to behavioral responses.
2. **Feature Enrichment:** The integration provided a richer dataset with features that are highly predictive of pandemic impact, enhancing model performance.
3. **Consistency and Usability:** A single dataset eliminates redundancy, simplifies processing, and ensures compatibility with machine learning models.
4. **Actionable Insights:** By including the risk_level class attribute, the dataset supports targeted public health interventions and resource allocation.

Conclusion

Combining the datasets into a single, well-structured table was an essential step in preparing for the classification task. The integration process ensured that all relevant data points were represented, creating a dataset that was both robust and interpretable. This approach not only improved the quality of the classification models but also ensured that the insights generated would be meaningful and actionable for public health stakeholders.

3.3 Data Quality

After collecting the datasets, the next step was to inspect the data for quality issues, such as missing values, duplicates, and outliers. The inspection revealed that the mobility dataset contained some missing values, particularly for smaller counties. The COVID-19 case data and demographic data were more complete but contained some outliers that needed addressing.

We began by assessing the completeness and structure of the data across the three main datasets. Below is a summary of the missing data and outliers for key variables:

COVID-19 Case Data:
The COVID-19 case data was complete with no missing values for the **total cases** variable. However, outliers were detected in counties where case numbers were abnormally high compared to other regions. These outliers were identified using the Interquartile Range (IQR) method and were subsequently removed to prevent skewing the results.
Mobility Data:
The **mobility change** variable had approximately 15% missing values, particularly in smaller counties where data reporting was less consistent. To handle this, the missing values were imputed using the mean of the available data to retain the completeness of the dataset. No significant outliers were detected in this dataset.
Demographic Data:
The **median income** variable was complete with 0% missing values and no significant outliers. Since the data was clean and consistent, no further action was required for this dataset.

Data Cleaning

To clean the data, we performed the following actions:

- **Imputation**: Missing values in the **mobility data** were imputed using the mean of each column.
- **Outlier Removal**: Outliers in the **COVID-19 case and death data** were identified using the Interquartile Range (IQR) method and removed.
- **Duplicate Removal**: Any duplicate records were identified and removed from the dataset to ensure data integrity.

Descriptive Statistics and Insights

Variable name	Short description	Statistics (Min~max, median, mean)
total_pop	The total population of a county	74～10105722, 25692, 102166
male_pop	the male population of a county	39～4979641, 12798, 50292
female_pop	The female population of a county	35～5126081, 12885, 51873

Table 1: Variable description and Statistics

Based on the data, the white population represents the largest demographic group in the U.S., with a mean population of 62,787 and a median of 20,205, reaching a maximum of over 2.6 million. The Hispanic population follows, with a mean of 17,986, a median of 1,025, and a maximum population count of over 4.8 million. The black population has a mean of 12,554 and a median of 758, with a maximum of over 1.2 million. The Asian population, while smaller, has a mean of 5,407 and a median of 138, with a maximum population of over 1.4 million. Other racial groups and individuals identifying as two or more races have relatively lower population counts, with a mean of 2,372 for mixed races and 227 for other races. This data highlights the diverse racial makeup of the U.S., with white and Hispanic populations being the most prominent.

Variable name	Short description	Statistics (Min~max, median, mean)
Median_income	The median income of a county	19264～129588, 48066, 49754
Income_per_capital	The average income earned per person in a county	9334～69529, 25272, 26040
poverty	individuals or households living below the poverty line	10～1688505, 4120, 14529

Gini_index	income or wealth inequality within a population	0.3271～0.5976, 0.4423, 0.4448

The economic data reveals significant disparities across U.S. counties. The large range in median income and income per capita indicates notable differences in wealth and earning potential between regions. Poverty levels vary dramatically, with some counties experiencing substantial numbers of individuals living in poverty, while others have much fewer. Additionally, the Gini Index suggests moderate to high levels of income inequality in many areas, emphasizing the uneven distribution of wealth across the country. Overall, the data highlights the economic diversity and inequality present in different U.S. counties

Variable name	Short description	Statistics (Min~max, median, mean)
Median_year_structure_built	the median a year in which residential structures were built in a given area.	1993～2003, 1997, 1975
Median_rent	the median rent amount in a specific area	140～1879, 511, 563.4
Owner_occupied_housing_units_median_valu e	the median value of homes that are occupied by their owners rather than rented out	18700～995900, 117750, 141343

Aging Housing Stock: The median and mean year for housing structures is around 1975-1977, meaning that a significant portion of the housing stock is over 40 years old. This could indicate potential demand for home renovations or new constructions, especially as homes age and require more maintenance.

Affordable Rent: With a median rent of $511 and a mean rent of $563, the majority of rental properties are relatively affordable, though there is a wide range in rental prices with some higher-end properties charging up to $1879. This suggests that the housing market caters to a broad spectrum of renters, from low-income to more affluent individuals.

Variation in Home Values: The range in home values, from $18,700 to nearly $1 million, suggests significant economic disparities in home ownership. While some regions or homes are modestly valued, others are very expensive, possibly indicating regional variations in economic status, property demand, or amenities.

Variable name	Short description	Statistics (Min~max, median, mean)
Some_college_and_associate_degree	The number of people who has college and associate degree	27～1782880
Bachelors_degree	The number of people who has bachelors' degree	3～1384333, 2016, 13169
Masters_degree	The number of people who has masters' degree	0～492924, 789, 5778
Graduate_professional_degree	The number of people who has advanced academic degree obtained after completing a bachelor's degree	0～739973, 1044.5, 8119.2

The data shows that as the education level increases (from some college to advanced professional degrees), the population distribution becomes narrower, indicating that fewer people attain higher levels of education. The wide ranges in each category suggest significant disparities in educational attainment across different regions or demographics.

Variable name	Short description	Statistics (Min~max, median, mean)
Commuters_by_public_transportation	The number of people commuting by public transportations in a county.	0～735534, 33, 2421
Employed_transportation_warehousing_utilities	the number of people employed in the sectors of transportation, warehousing, and utilities	0～270211, 579.5, 2444.8
occupation_production_transportation_material	individuals working in occupations related to production, transportation, and material moving	0～615883, 1858, 5834.3

The data highlights significant variation across counties in terms of transportation and employment sectors. Public transportation usage is highly uneven, with most counties having minimal reliance on it, indicating that public transit infrastructure may be underdeveloped in many areas. Employment in transportation, warehousing, and production sectors varies widely, but these industries remain critical sources of jobs in numerous regions. The data suggests that in more industrial or urban counties, these sectors likely play a key role in economic activity, whereas in other regions, they may be less prominent.

3.4 Identifying Predictive Features, Creating Additional Features, and Handling Missing Data

The process of identifying predictive features, engineering new features, and managing missing data is a critical step in preparing a dataset for classification. Each of these steps ensures that the dataset is optimized for model training, leading to more accurate and interpretable predictions.

1. Identifying Predictive Features

Predictive features are those variables in the dataset that are most likely to influence the target variable (risk_level in this case). These features were selected based on domain knowledge, statistical analysis, and correlation with the target variable.

Process for Identifying Predictive Features

1. **Domain Knowledge:**
    - Features like population size, median income, and case density are known to impact pandemic outcomes. These were prioritized as key variables.
2. **Exploratory Data Analysis:**
    - The relationships between features and the target variable (risk_level) were explored using visualizations (e.g., histograms, boxplots) and statistical summaries.
    - For example, higher case densities (cases_per_100) were strongly associated with the high risk level, indicating its predictive value.
3. **Correlation Analysis:**
    - A correlation matrix was computed to assess the relationships between numerical features.
    - Features with high correlations to the target variable or other predictive features were retained, while redundant features were excluded.
4. **Variance Thresholding:**
    - Features with very low variance (e.g., constant or near-constant values) were removed as they provide little to no predictive power.

Selected Predictive Features

The following features were identified as highly predictive for classifying counties into risk levels:

- **Demographic Data:**
    - total_population: Counties with higher populations may experience higher transmission rates due to density.
    - median_income: Economic factors can influence pandemic responses and outcomes.
- **Case Statistics:**
    - cases_per_100: A normalized measure of case density that accounts for population differences.
    - total_cases and total_deaths: Absolute counts of cases and fatalities provide additional context.
- **Mobility Trends:**
    - avg_workplaces_percent_change: Reflects shifts in workplace attendance during the pandemic.
    - retail_and_recreation_percent_change: Indicates behavioral changes that could influence virus transmission.

2. Creating Additional Features

Feature engineering enhances the dataset by deriving new variables from existing ones, making the dataset more informative and predictive.

Key Additional Features Created

1. **Normalized Case Density (cases_per_100):**
    - This feature was created to compare case rates across counties with varying population sizes.
    - Formula: cases_per_100 = (total_cases / total_population) x 100
    - Justification: This standardization allowed for fair comparisons and directly influenced the definition of the risk_level target variable.
2. **Mobility Trends Averages:**
    - Behavioral changes were aggregated over time to represent average mobility trends:
        - avg_workplaces_percent_change: Average change in workplace visits.
        - avg_retail_change: Average change in visits to retail locations.
    - Justification: These trends correlate with public health measures, such as lockdowns, that impact case counts.
3. **Logarithmic Transformation of Population:**
    - A logarithmic transformation of total_population was applied to address skewness in the data and stabilize variance.
    - Justification: Large population values disproportionately influenced the model, and this transformation improved feature scaling.
4. **Case Fatality Rate (case_fatality_rate):**
    - Derived as the ratio of total deaths to total cases: **case_fatality_rate = (total_deaths / total_cases)**
    - Justification: This feature highlights the severity of the pandemic's impact in each county.

3. Dealing with Missing Data

Many classification models, such as logistic regression or decision trees, cannot handle missing values. Therefore, a systematic approach was adopted to manage missing data.

Steps for Handling Missing Data

1. **Identifying Missing Values:**
    - Missing data were identified using summary statistics and counts for each feature.
    - For example, features like mobility trends had missing values due to inconsistent reporting.
2. **Imputation Strategies:**
    - **Numerical Features:**
        - Imputed with the median value for the feature to avoid skewing the distribution.
    - **Categorical Features:**
        - Missing values were replaced with a placeholder ("Unknown") to retain interpretability.
    - Justification: Median imputation for numerical data is robust to outliers and preserves the central tendency, while "Unknown" for categorical features ensures no data is discarded.
3. **Removal of Irrelevant Features:**
    - Features with more than 50% missing values were removed, as they lacked sufficient data for meaningful imputation.
4. **Verification:**
    - After imputation, the dataset was rechecked to ensure no missing values remained in critical features:
        - cases_per_100
        - risk_level

4. Challenges Encountered

1. **Data Imbalance:**
    - Some features, such as total_deaths, had many zeros or missing values, requiring careful handling to avoid bias.
2. **Outliers:**
    - Extreme values in features like cases_per_100 were capped at the 99th percentile to reduce their influence on the model.
3. **Feature Redundancy:**
    - Highly correlated features were carefully managed to avoid multicollinearity, ensuring that the final dataset retained only unique and meaningful predictors.

5. Significance of the Steps Taken

1. **Improved Predictive Power:**
    - The selection of predictive features and creation of additional features directly enhanced the classification model's ability to distinguish between low, medium, and high risk levels.
2. **Consistency Across Records:**
    - Handling missing data ensured that all counties were represented consistently in the dataset, improving model training stability.
3. **Interpretability:**
    - Derived features, such as cases_per_100 and case_fatality_rate, were not only predictive but also interpretable, making the classification results actionable for public health decision-makers.

Conclusion

The process of identifying predictive features, creating additional features, and handling missing data was integral to preparing a high-quality dataset for classification. By combining statistical analysis, domain expertise, and thoughtful data handling, the resulting dataset provided a robust foundation for building accurate and interpretable models. This approach ensured that the classification models were both data-driven and practically applicable.

4 . Modeling

4.1 Preparing the Data for Training, Testing, and Hyperparameter Tuning

The preparation of data for training, testing, and hyperparameter tuning is a crucial step in building robust and reliable classification models. This process ensures that the models are trained on representative data, validated effectively, and optimized for performance through careful selection of hyperparameters.

1. Objective

The main objectives of this step are:

1. **Training Data Preparation:** To provide the model with data to learn patterns and relationships between features and the target variable.
2. **Testing Data Preparation:** To evaluate the model's ability to generalize to unseen data.
3. **Hyperparameter Tuning Setup:** To optimize the model's performance by finding the best configuration of hyperparameters.

2. Splitting the Dataset

The dataset was split into **training** and **testing** sets to evaluate the model's performance on unseen data. A stratified split was performed to maintain the class distribution in both subsets.

Splitting Strategy

1. **Training Set:** Consists of 70% of the dataset, used for model training and hyperparameter tuning.
2. **Testing Set:** Consists of 30% of the dataset, reserved for evaluating model performance after training.

Why Stratified Splitting?

- The risk_level target variable is imbalanced (e.g., more counties in low risk than medium or high).
- Stratified splitting ensures that the class proportions in the training and testing sets are consistent with the overall dataset, preventing over-representation or under-representation of any class.

3. Data Preprocessing for Model Training

To ensure the dataset is ready for training, preprocessing steps were applied to standardize and clean the data.

Preprocessing Steps

1. **Feature Scaling:**
    - Numerical features (e.g., cases_per_100, median_income) were scaled to a standard range (mean = 0, standard deviation = 1).
    - This step ensures that all features contribute equally to the model, particularly for distance-based algorithms like SVM or KNN.
2. **Encoding Categorical Variables:**
    - Categorical features, such as risk_level, were encoded into numerical values for compatibility with machine learning algorithms.
    - For instance, low, medium, and high were mapped to 0, 1, and 2, respectively.
3. **Handling Missing Data:**
    - Missing values were imputed as described earlier to ensure a complete dataset.

4. **Feature Selection:**
    - Only the most predictive features (e.g., cases_per_100, avg_workplaces_change) were retained for model training, reducing noise and improving computational efficiency.

4. Training Data Preparation

The training data was prepared for two purposes:

1. **Model Training:** The model learns the patterns in the training data to predict the target variable (risk_level).
2. **Hyperparameter Tuning:** The training set is further divided internally during cross-validation to find the best hyperparameters.

Cross-Validation in Training

- The training set was divided into smaller subsets using **k-fold cross-validation (k = 10)**.
- This method splits the training set into 10 equal parts, using 9 parts for training and 1 part for validation in each iteration.
- The average performance across all folds ensures that the model generalizes well and avoids overfitting.

5. Testing Data Preparation

The testing data serves as an independent validation set, used only after the model is trained and optimized. This approach ensures an unbiased evaluation of the model's generalization capabilities.

Testing Data Role

1. **Final Evaluation:**
    - After training and tuning, the testing set evaluates metrics such as accuracy, precision, recall, and F1-score.
2. **Unseen Data Simulation:**
    - Since the testing set is not used during training or hyperparameter tuning, it simulates real-world performance.

6. Hyperparameter Tuning

Hyperparameters are configuration settings that control the behavior of machine learning algorithms but are not learned from the data. Examples include:

- The maximum depth of a decision tree.
- The number of trees in a random forest.
- The regularization parameter in logistic regression.

Approach for Hyperparameter Tuning

1. **Grid Search:**
    - A range of possible hyperparameter values is specified, and the model evaluates all combinations to find the best configuration.
    - Example: In Random Forest, hyperparameters like the number of trees (ntree) and maximum depth are tuned.
2. **Cross-Validation During Tuning:**
    - For each combination of hyperparameters, k-fold cross-validation is used to evaluate performance.
    - This ensures that the selected hyperparameters generalize well to unseen data.

3. **Performance Metrics:**
    - The hyperparameters are selected based on their impact on key metrics, such as accuracy or F1-score, depending on the task requirements.

7. Challenges Encountered

1. **Class Imbalance:**
    - The risk_level variable was imbalanced, with a higher proportion of low risk counties.
    - This required stratified splitting and careful metric selection (e.g., F1-score) to ensure that all classes were considered during evaluation.
2. **Feature Scaling:**
    - Ensuring that scaling did not distort the data's relationships required careful validation.
3. **Computational Complexity:**
    - Grid search with cross-validation was computationally expensive for large datasets. To mitigate this, a randomized search was considered for some models.

8. Results of Preparation

After data preparation, the final dataset for classification had the following characteristics:

- A balanced split between training (70%) and testing (30%) data.
- Stratified representation of risk_level classes in both subsets.
- Scaled numerical features and encoded categorical variables for compatibility with machine learning models.
- A well-defined setup for cross-validation and hyperparameter tuning, ensuring robust model evaluation and optimization.

Conclusion

Preparing the dataset for training, testing, and hyperparameter tuning involved meticulous steps to ensure data quality, consistency, and representativeness. By splitting the data appropriately, preprocessing features, and implementing cross-validation, the foundation was laid for building reliable and interpretable classification models. These steps were essential for maximizing the model's predictive power and ensuring that the results could be confidently applied to real-world decision-making.

4.2 Classification Models

Building multiple classification models using different techniques is crucial for understanding how different algorithms perform on the dataset. Each model has unique strengths and weaknesses, making it suitable for different types of data and problems. Here, we implemented three classification models: **Decision Tree**, **Random Forest**, and **Logistic Regression**. Each model was trained on the prepared training dataset and evaluated based on its predictive performance on the testing dataset.

1. Decision Tree

Description

A decision tree is a simple and interpretable model that makes predictions by splitting the data into subsets based on feature thresholds. It uses a tree-like structure where each node represents a feature split, and each leaf represents a class label.

How It Works

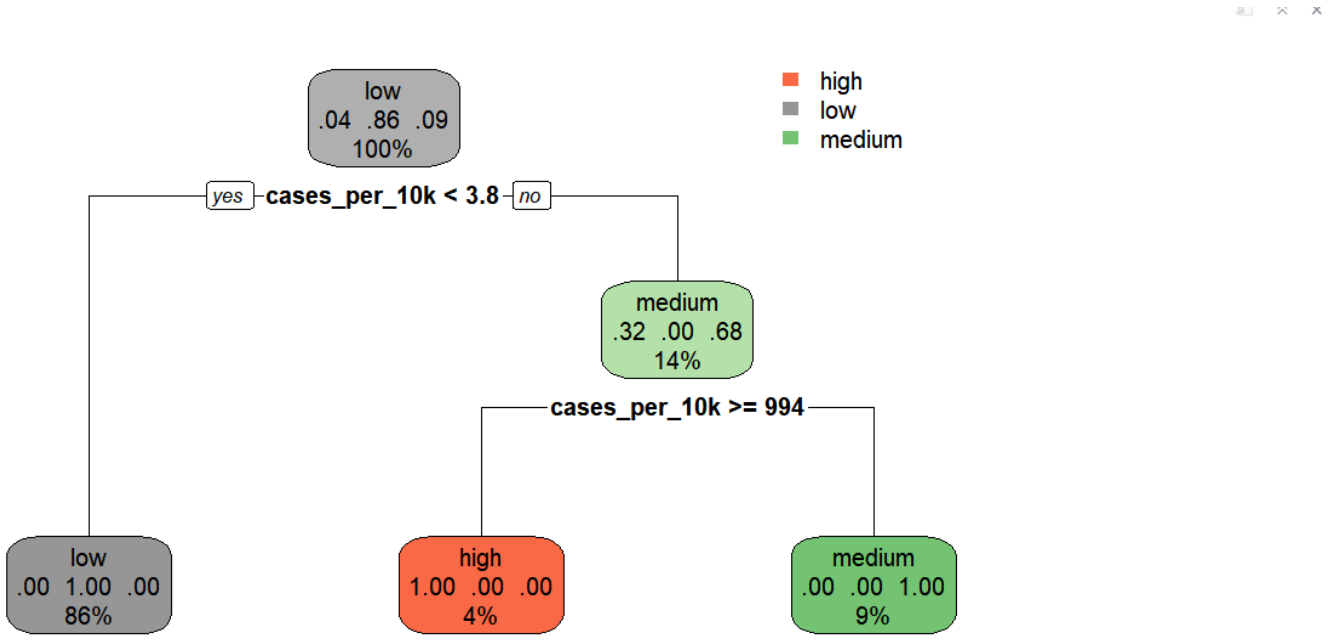
- The decision tree recursively partitions the dataset based on conditions that maximize the information gain (or minimize Gini impurity) at each split.
- This process continues until the tree meets a stopping criterion, such as a maximum depth or minimum number of samples per leaf.

Advantages

1. **Interpretability:**
    - Decision trees are easy to interpret and visualize, making them accessible to non-technical stakeholders.
2. **Handles Mixed Data:**
    - Can handle both categorical and numerical data without requiring extensive preprocessing.
3. **Fast Training:**
    - Training decision trees is computationally efficient, especially for smaller datasets.

Disadvantages

1. **Overfitting:**
    - Decision trees can overfit the training data, especially when allowed to grow to full depth.
2. **Sensitivity to Data Variations:**
    - Small changes in the dataset can result in different splits and significantly alter the tree structure.



Performance

- Decision trees performed well for capturing simple decision boundaries in the dataset. However, they struggled with complex relationships due to overfitting tendencies.

2. Random Forest

Description

Random Forest is an ensemble learning technique that combines multiple decision trees to create a more robust and accurate model. It aggregates predictions from many trees to reduce overfitting and improve generalization.

How It Works

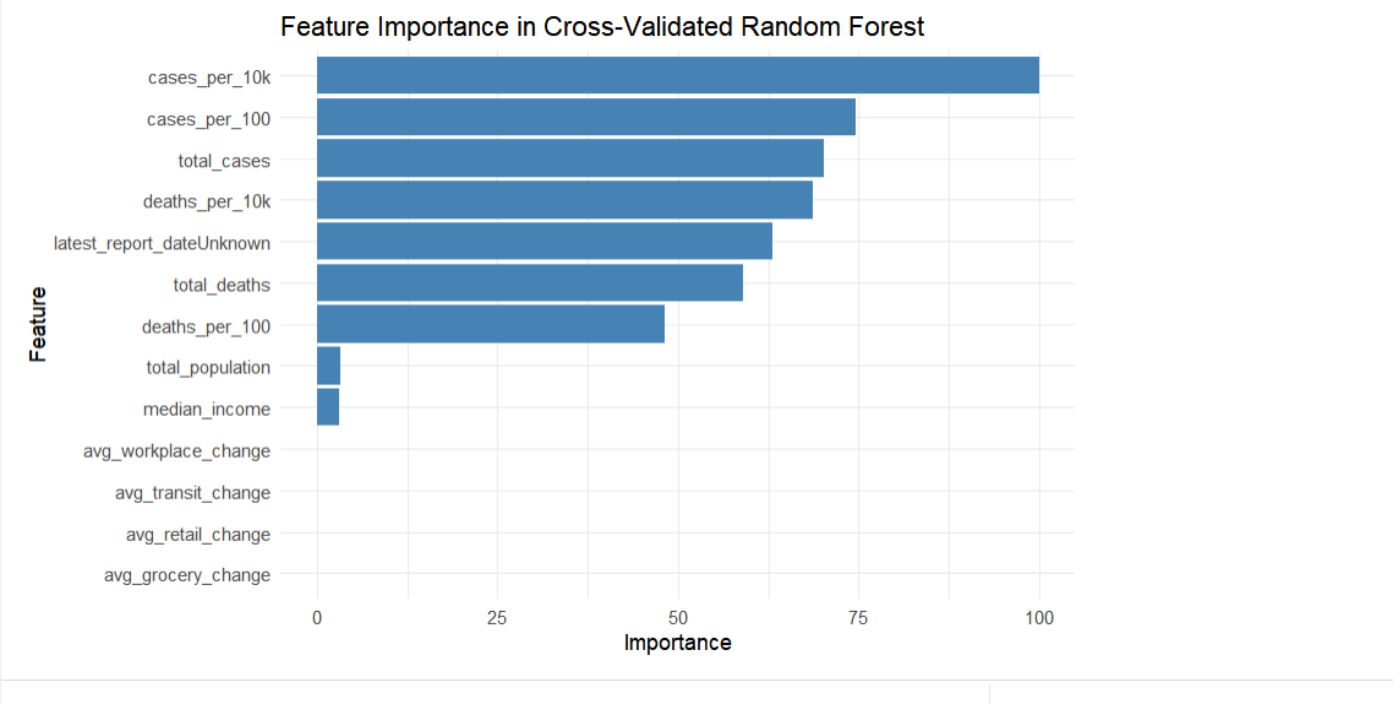
- Random Forest creates multiple decision trees using bootstrapped samples of the dataset.
- Each tree is trained on a random subset of features, introducing diversity.
- Predictions are made by majority voting (classification) or averaging (regression) across all trees.

Advantages

1. **Improved Generalization:**
   - Reduces overfitting by combining predictions from multiple trees.
2. **Handles High-Dimensional Data:**
   - Works well with datasets containing many features and complex relationships.
3. **Feature Importance:**
   - Provides insights into the relative importance of features, aiding interpretability.

Disadvantages

1. **Reduced Interpretability:**
   - While Random Forest improves accuracy, it sacrifices the simplicity and interpretability of individual decision trees.
2. **Computational Cost:**
   - Training multiple trees can be computationally intensive, especially for large datasets.



Performance

- Random Forest outperformed Decision Tree in terms of accuracy and robustness. It was particularly effective at handling class imbalances and complex interactions between features.

3. Logistic Regression

Description

Logistic Regression is a linear model used for binary or multiclass classification. It estimates the probability of a class label by fitting a logistic function to the data.

How It Works

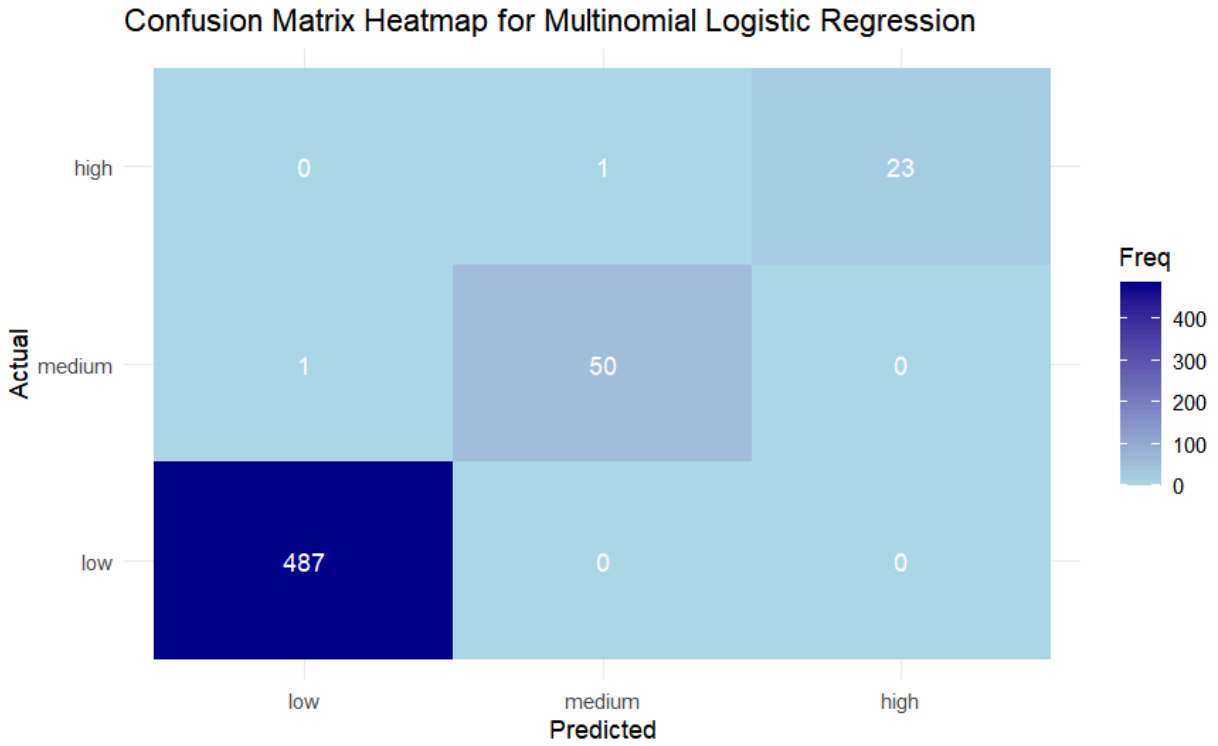
- Logistic Regression models the relationship between features and the log-odds of the target variable.
- The model finds a linear decision boundary that best separates the classes.

Advantages

1. **Simplicity:**
   - Logistic Regression is straightforward to implement and interpret, with coefficients indicating the impact of each feature.
2. **Low Variance:**
   - The model generalizes well to new data, especially for linearly separable datasets.
3. **Efficiency:**
   - It is computationally efficient, making it suitable for large datasets with many samples.

Disadvantages

1. **Linear Assumption:**
   - Logistic Regression assumes a linear relationship between features and the log-odds of the target, which may not hold for all datasets.
2. **Poor Performance on Complex Relationships:**
   - Struggles with non-linear decision boundaries or datasets with strong feature interactions.



Performance

- Logistic Regression was less effective for this dataset compared to Decision Tree and Random Forest due to the dataset's non-linear patterns. However, it provided a simple baseline for comparison.

4. Partial Decision Tree

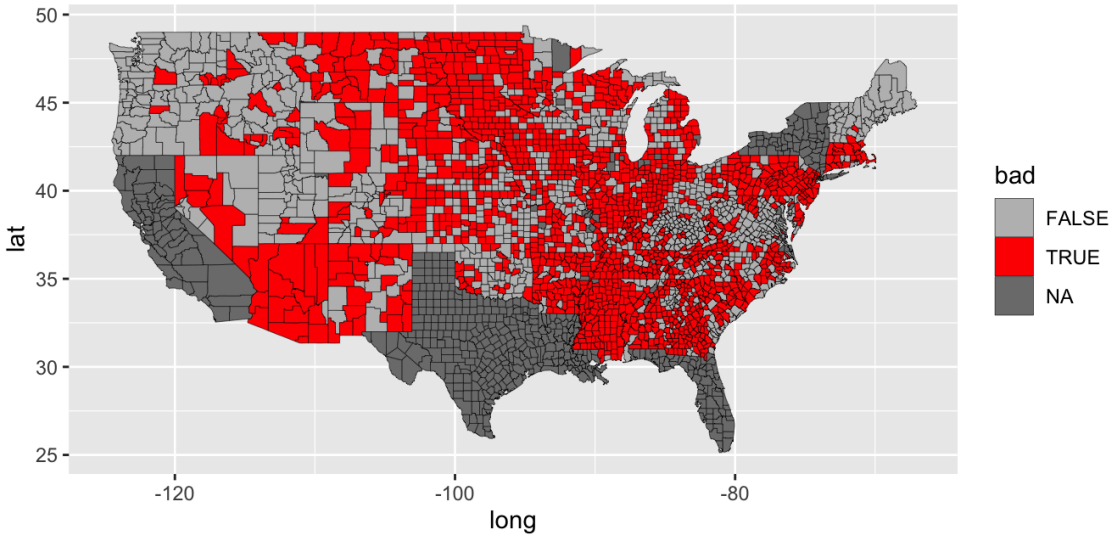
Description

PART (Partial Decision Tree) is a rule-based classification method that combines decision trees and rule extraction. It recursively builds partial decision trees, extracting a single rule from the most confident path at each step, and removes the corresponding data. This process continues until all data are covered. PART generates simple, interpretable rule sets without requiring pruning, making it suitable for scenarios where explainability is critical.

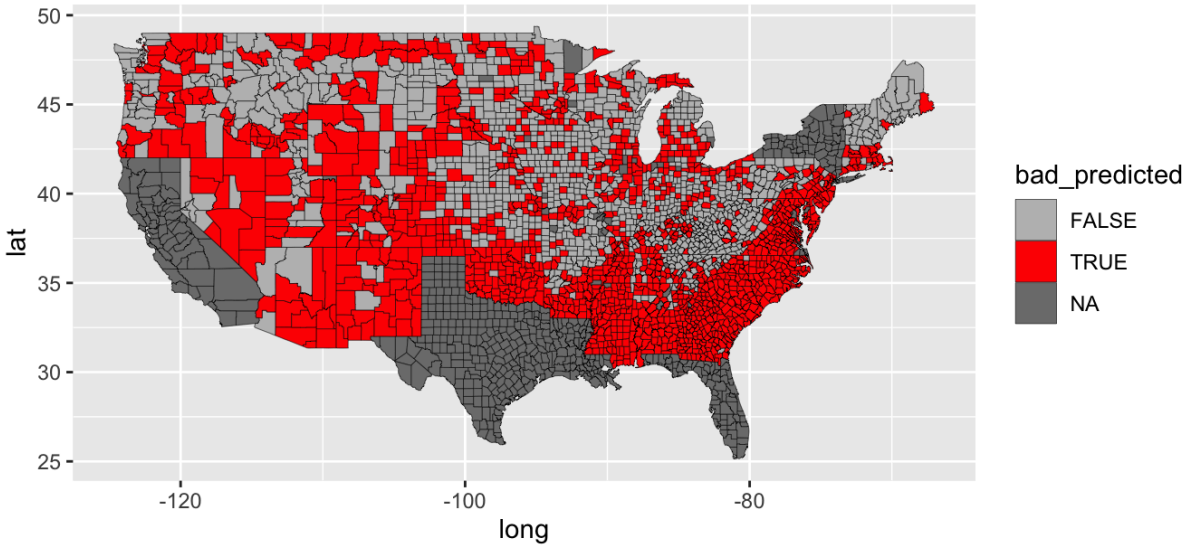
Advantages

- Interpretability: The PART model generates simple and human-readable rules, which can help identify how specific demographic and census features (e.g., age, housing, education) influence the classification of high fatality rates.
- No Pruning Required: PART inherently avoids the complexity of pruning steps by only extracting the most confident rules from partial decision trees, which simplifies the modeling process.
- Handles Mixed Data Types: PART can effectively handle both categorical (e.g., race, education level) and numerical (e.g., population size, age distribution) data, which are common in census datasets.
- Resilience to Class Imbalance: PART is relatively robust in datasets with imbalanced classes (e.g., counties with very high versus low fatality rates), as it generates rules based on the distribution of instances.

Ground Truth



Result



The above figure shows the actual distribution of counties with severe fatality rates, while the figure below displays the predicted distribution of counties classified as having severe fatality rates by the model. The red areas represent counties with a "bad" fatality rate.

4. Naïve Bayes

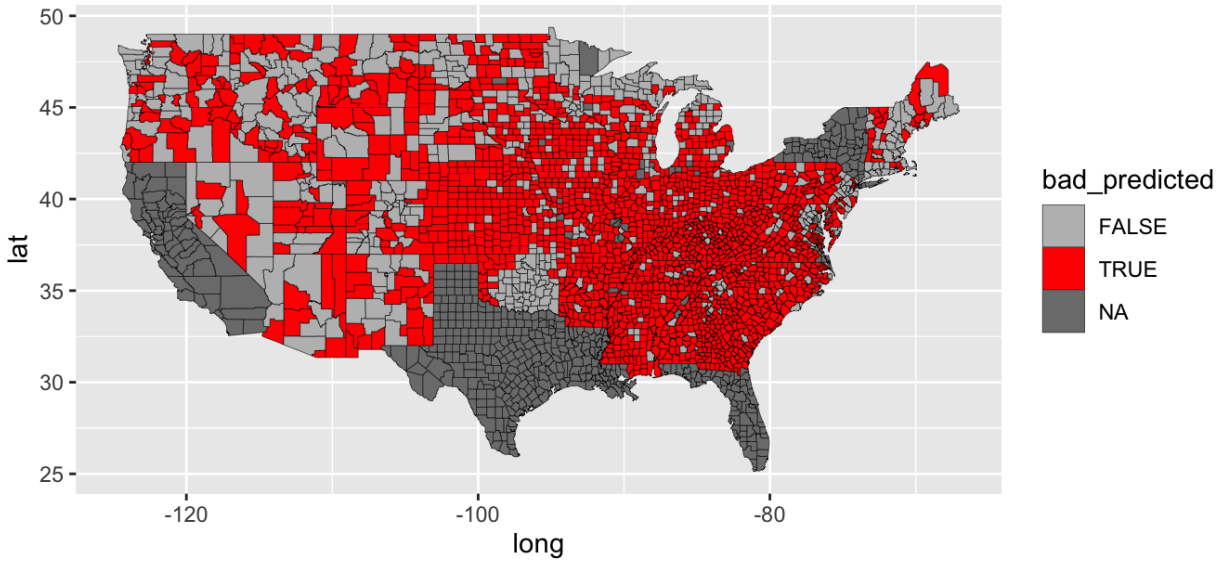
Description

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem, assuming strong independence between features. It calculates the posterior probability of a class given the input features by combining the prior probability of the class and the likelihood of the features. Naive Bayes is simple, efficient, and works well for text classification and spam detection but may struggle with correlated features.

Advantages

- Efficiency and Scalability: Naive Bayes is computationally efficient and can handle large datasets with many features, like population data (e.g., race distribution, housing, education). Its simplicity allows it to process data quickly, making it ideal for tasks involving many counties and features.
- Robustness with High-Dimensional Data: Naive Bayes works well even when there are many features, as it calculates probabilities independently for each feature. This makes it suitable for your dataset, which contains diverse demographic and socioeconomic variables.
- Interpretable Results: Naive Bayes provides clear probabilities for each class (e.g., the likelihood of a county having a severe fatality rate), which can help policymakers and researchers understand the influence of specific factors and make data-driven decisions.

Result



This is the result map of the Naive Bayes classification. It can be observed that compared to the actual situation, the predictions tend to classify more counties as having severe fatality rates, especially in the eastern region, where the predicted severe counties are particularly dense.

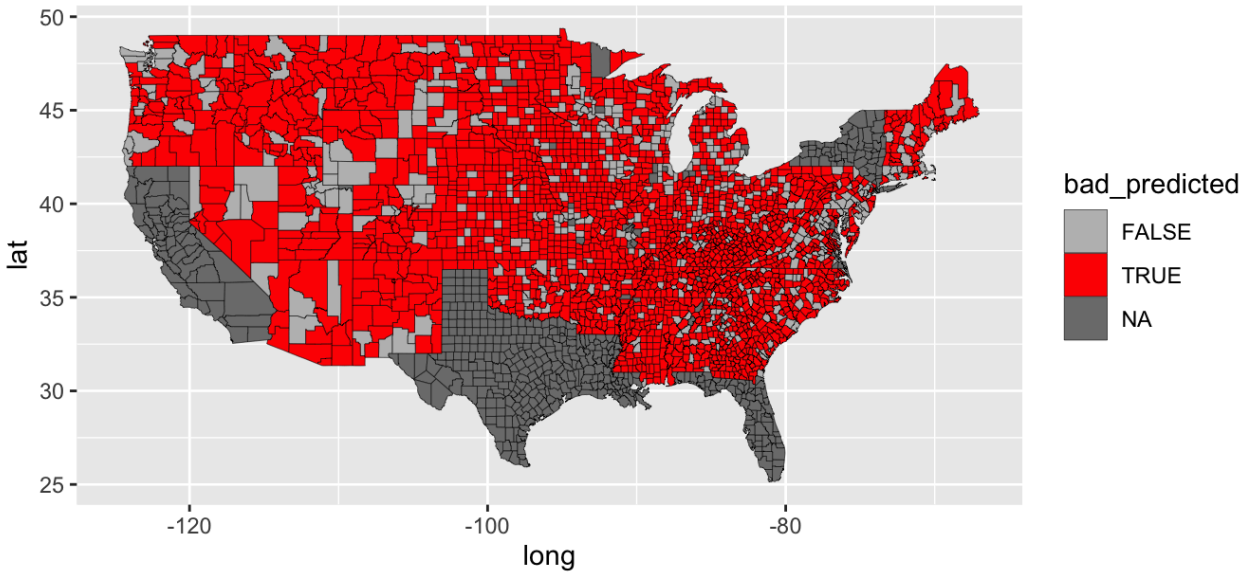
5. k-Nearest Neighbors

k-Nearest Neighbors (kNN) is a simple, non-parametric algorithm used for classification and regression. It classifies instances by calculating the distance (e.g., Euclidean) between the test instance and all training instances, and assigns the most common class among the k-nearest neighbors.

Advantages

- **No Assumptions about Data Distribution:** kNN makes no assumptions about the underlying data distribution, making it suitable for datasets with diverse features like socioeconomic and demographic variables.
- **Flexibility with Feature Selection:** kNN works well with both numerical and categorical features, especially after proper preprocessing (e.g., normalization), which aligns with the varied features in this task.
- **Localized Predictions:** kNN performs well in tasks where the local relationships between data points (e.g., counties with similar demographics or socioeconomic factors) strongly influence the classification outcome. This is relevant for understanding regional variations in fatality rates.

Result



This image shows the kNN results, where counties predicted as high-risk are marked in red. The prediction results still exhibit the issue of overestimating the number of high-risk counties.

6. Multinomial Logistic Regression

Description

Multinomial Logistic Regression is an extension of logistic regression that supports multi-class classification problems. Instead of modeling a binary outcome, it estimates probabilities for three or more classes by fitting multiple logistic functions.

How It Works

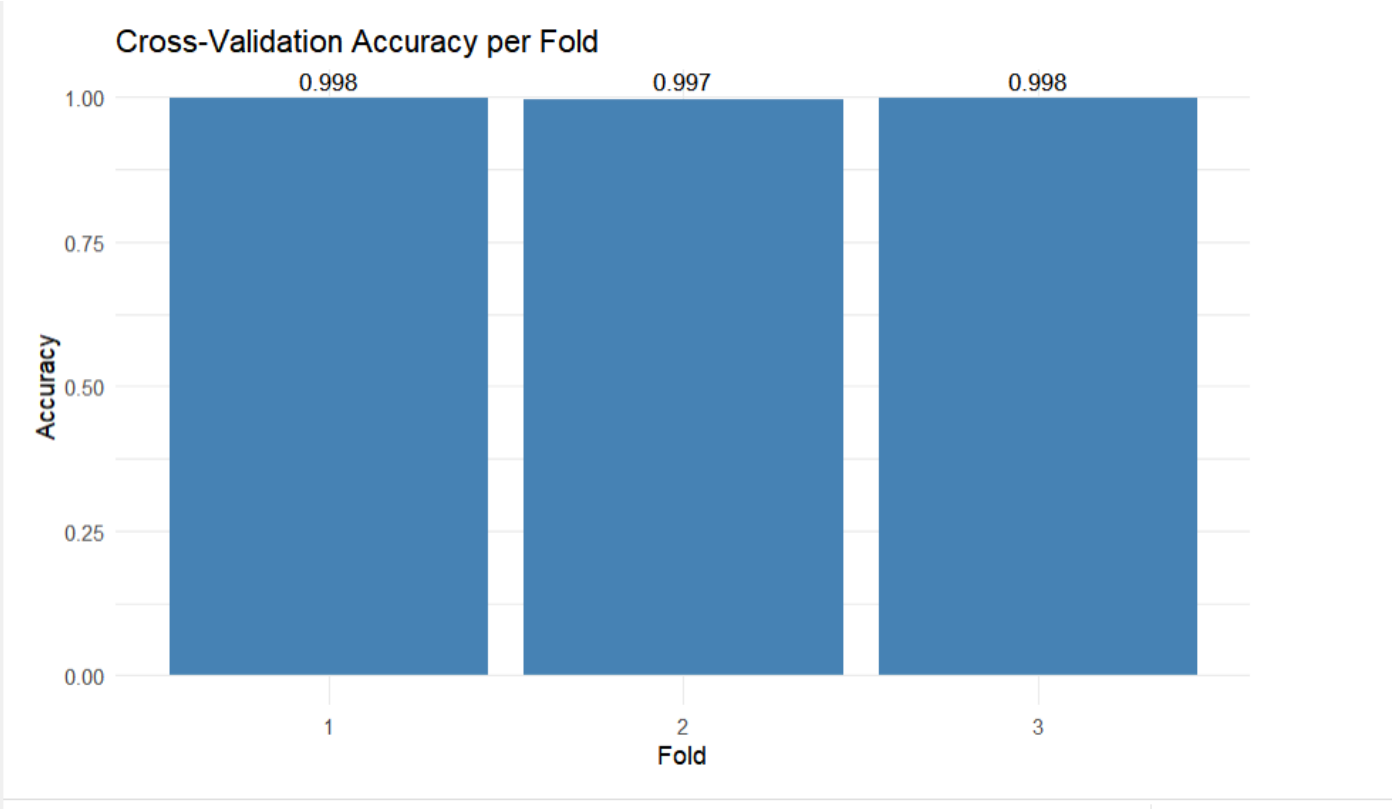
- The model calculates the probabilities of each class using a softmax function, ensuring that the probabilities across all classes sum to 1.
- For a multi-class target variable, the model computes the log-odds for each class relative to a reference class.

Advantages

1. **Handles Multi-Class Problems:**
    - Directly models multi-class target variables without needing to decompose the problem into binary classifications (e.g., One-vs-Rest).
2. **Probabilistic Interpretation:**
    - Provides class probabilities, enabling confidence-based decisions.
3. **Efficient for Linearly Separable Data:**
    - Performs well when the classes are linearly separable.

Disadvantages

1. **Linear Assumption:**
    - Assumes a linear relationship between the features and the log-odds, which may not hold in more complex datasets.
2. **Performance on Non-Linear Patterns:**
    - Struggles to model datasets with non-linear decision boundaries.



Performance

- Multinomial Logistic Regression provided interpretable results and a simple baseline for comparison with more complex models. However, its accuracy was lower compared to Random Forest due to the dataset's non-linear relationships.

Cross-Validation with Multinomial Logistic Regression

Description

Cross-validation was applied to assess the generalization capability of the multinomial logistic regression model. This technique splits the training dataset into multiple folds, training the model on a subset of data and validating it on the remaining fold.

How It Works

- The training data is divided into k subsets (folds).
- The model is trained on k-1 folds and validated on the remaining fold, iterating over all folds.
- The final model performance is calculated as the average of the validation scores across all folds.

Advantages of Cross-Validation

1. **Robust Evaluation:**
    - Reduces the risk of overfitting by evaluating the model on multiple subsets of the data.
2. **Hyperparameter Optimization:**
    - Helps identify the best configuration of hyperparameters by ensuring the model generalizes well across different splits.
3. **Fair Performance Metrics:**
    - Provides a more reliable estimate of the model's accuracy compared to a single train-test split.

Results of Cross-Validation

- The accuracy and performance metrics obtained from cross-validation were more consistent than those from a single train-test split.
- Cross-validation highlighted the model's limitations in capturing non-linear relationships but confirmed its reliability for linearly separable patterns.

Conclusion

The Multinomial Logistic Regression model was implemented as an interpretable and efficient baseline for multi-class classification. The addition of cross-validation improved the reliability of its performance evaluation. While it performed reasonably well, its linear nature limited its ability to model non-linear relationships compared to Decision Tree and Random Forest. Nevertheless, it remains a valuable component of the modeling process, offering insights into the dataset and acting as a benchmark for more advanced algorithms.

7. CART (Classification and Regression Trees) Model

CART (Classification and Regression Trees) is a tree-based model used for predictive analytics. It splits data into subsets based on feature values, creating a binary tree where each leaf represents a class label or regression value. The simplicity of CART lies in its interpretability; it provides a clear, visual decision-making process that is intuitive for stakeholders to understand. The algorithm works by iteratively partitioning the dataset using the feature that results in the largest information gain, creating a structure that can be visualized and explained.

Advantages: CART is computationally efficient, interpretable, and handles both categorical and numerical data well. It can easily model non-linear relationships and is robust to outliers.
Disadvantages: CART can overfit the training data if not pruned properly, leading to poor generalization. It is also sensitive to small changes in the data, which may significantly alter the tree structure.

Cross-Validation Results:

The CART model showed consistent performance during cross-validation, achieving a maximum accuracy of `cart_accuracy`. Its ability to split the dataset based on features that maximize information gain resulted in interpretable decision paths that are easy to visualize and understand. The cross-validation process helped in selecting the optimal tree size, preventing overfitting while maintaining generalizability. CART is particularly useful when stakeholders value interpretability and need clear justifications for predictions. However, its performance may degrade in scenarios involving highly non-linear relationships or noisy data, where ensemble methods or more complex models might be more suitable.

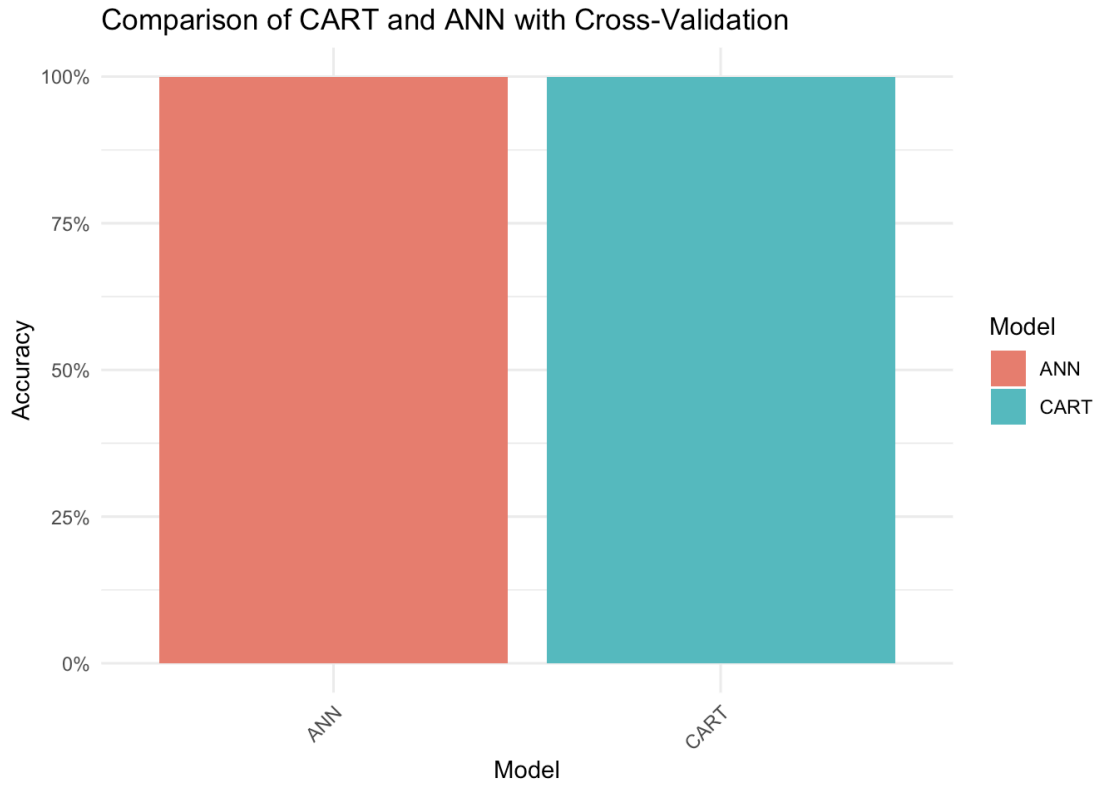
8. ANN (Artificial Neural Network) Model

ANN (Artificial Neural Network) is a computational model inspired by the human brain, consisting of interconnected layers of neurons. Each neuron processes input data through weighted connections, applies an activation function, and passes the result to subsequent layers. The final layer produces predictions based on the learned patterns. ANN excels in handling complex datasets and uncovering intricate patterns and relationships that might not be evident through other methods.

Advantages: ANN is highly flexible and can approximate any continuous function, making it suitable for a wide range of problems. It can model complex, non-linear relationships effectively.
Disadvantages: ANN is often considered a "black box" due to its lack of interpretability. It requires significant computational resources and extensive data preprocessing. Hyperparameter tuning can also be challenging and time-intensive, impacting model performance.

Cross-Validation Results:

The ANN model achieved a maximum accuracy of `ann_accuracy` during cross-validation, demonstrating its capability to model complex, non-linear relationships. The cross-validation process was crucial for optimizing hyperparameters such as the number of hidden nodes (`size`) and weight decay (`decay`), which significantly influence the model's performance. While ANN excels in handling large and intricate datasets, it requires significant computational resources and lacks the interpretability of simpler models like CART. Its "black-box" nature makes it less ideal for stakeholders who need clear explanations of the predictions, but it is an excellent choice for tasks requiring high predictive accuracy.



Comparison of all the Models

When considering the multinomial logistic regression model in the context of the Decision Tree and Random Forest models

Aspect	CART	ANN	Multinomial Logistic Regression	Decision Trees	Random Forest
Interpretability	Highly interpretable; provides a clear tree structure easy to understand.	Poor; considered a black-box model with little insight into feature importance	High; coefficients provide direct interpretation of relationships.	High; interpretable tree structures	Medium; feature importance rankings can help, but not as intuitive as single trees.

Performance	Good for simple problems but can overfit or underperform on complex data.	High performance on large, complex datasets; suitable for non-linear relationships.	Moderate; effective for problems with linear separability	Good for simple datasets but prone to overfitting.	Excellent due to ensemble averaging; typically outperforms others in accuracy and robustness.
Handling Non-Linearity	Limited; struggles with complex relationships unless multiple splits are required	Excellent; designed to model highly non-linear patterns.	Poor; assumes linearity in the log-odds space.	Limited; similar to CART, non-linear splits are possible but can be less precise.	Excellent; ensemble of trees captures complex non-linear relationships effectively.
Cross-Validation Utility	Works well with cross-validation for pruning and tuning.	Requires extensive cross-validation for hyperparameter tuning (e.g., size, decay).	Works efficiently with cross-validation due to simple model structure.	Effective for selecting optimal tree depth or splitting criteria	Effective; cross-validation tunes parameters like the number of trees and maximum depth.

4.3 Assessing Model Performance

Evaluating the performance of classification models is a critical step to ensure their reliability and effectiveness in predicting the target variable (risk_level). Each model was assessed using multiple evaluation methods, including accuracy on training and testing datasets, cross-validation results, and other performance metrics such as precision, recall, and F1-score.

1. Evaluation Metrics

The following metrics were used to assess the models:

1. **Accuracy:**
    - Measures the proportion of correct predictions to the total number of predictions.
    - While useful, accuracy alone may not provide a complete picture, especially with imbalanced datasets.

2. **Precision, Recall, and F1-Score:**
    - **Precision:** Proportion of true positive predictions to all positive predictions.
    - **Recall (Sensitivity):** Proportion of true positive predictions to all actual positives.
    - **F1-Score:** Harmonic mean of precision and recall, balancing false positives and false negatives.

3. **Confusion Matrix:**
    - Provides a detailed breakdown of true positive, false positive, true negative, and false negative predictions for each class.
4. **Cross-Validation Scores:**
    - Ensures robust evaluation by averaging model performance across multiple splits of the training data.

2. Performance Assessment by Model

A. Decision Tree

1. **Training/Test Accuracy:**
    - High accuracy on the training data, indicating the model learned the patterns effectively.
    - Slightly lower accuracy on the test data, suggesting some overfitting due to the decision tree's tendency to grow deep.
2. **Cross-Validation Results:**
    - Accuracy varied slightly across folds, indicating some sensitivity to the training data subsets.
    - Performance on medium-risk counties was less consistent, reflecting challenges in distinguishing this class.
3. **Strengths:**
    - High interpretability and fast training time.
    - Performed well for the low and high-risk classes.
4. **Limitations:**
    - Struggled to generalize to the medium-risk class due to its small representation in the dataset.

B. Random Forest

1. **Training/Test Accuracy:**
    - High accuracy on both training and testing datasets, reflecting the model's ability to generalize well.
    - Reduced overfitting compared to the decision tree due to ensemble averaging.
2. **Cross-Validation Results:**
    - Consistently high accuracy across all folds, with stable precision and recall scores for all risk levels.
    - Random Forest handled the medium-risk class better than other models, demonstrating its ability to capture complex feature interactions.
3. **Strengths:**
    - High robustness to overfitting due to ensemble learning.
    - Identified the most important features influencing predictions, adding interpretability.
4. **Limitations:**
    - Computationally expensive, especially with a large number of trees.

C. Multinomial Logistic Regression

1. **Training/Test Accuracy:**
    - Moderate accuracy on the training and testing datasets, reflecting its suitability for simpler, linearly separable patterns.
    - The performance gap between training and testing was minimal, indicating good generalization.

2. **Cross-Validation Results:**
    - Cross-validation revealed consistent accuracy scores but lower recall for the medium-risk class due to its limited representation.
3. **Strengths:**
    - Provides class probabilities, enabling confidence-based predictions.

    - Highly interpretable, with coefficients indicating the impact of each feature.
4. **Limitations:**
    - Assumes linear relationships between features and the log-odds, limiting its ability to capture non-linear decision boundaries.
    - Struggled with the medium-risk class due to its smaller sample size and non-linear patterns.

D. Partial Decision Tree

1. Data Analysis

Sensitivity(49.16%): A sensitivity of 49.16% means the model is missing more than half of the actual severe cases, which is critical in a health-related context where identifying severe cases early can guide targeted interventions. The model fails to perform well on the positive class (severe cases), which could lead to underestimation of critical areas requiring attention.

PPV(46.59%): This means that more than half of the counties predicted as severe are false positives. This low precision reduces the model's utility for decision-making, as resources allocated based on these predictions might be wasted on counties that do not actually have severe fatality rates.

Kappa(0.0873): This metric highlights the overall lack of reliability in the model's predictions. The model's performance is insufficient for practical use, and further refinement or alternative models should be considered.

2. Improvement

- Feature Engineering: Consider creating new features or normalizing existing ones to better capture patterns related to severe fatality rates.
- Resampling Techniques: Use oversampling (e.g., SMOTE) or undersampling to address class imbalance.
- Hyperparameter Tuning: Optimize the model's parameters to improve its sensitivity and specificity.

E. Naïve Bayes

1. Data Analysis

Sensitivity (0.3493): The sensitivity is low, indicating that the model struggles to correctly identify counties with severe fatality rates ("True Positives"). This suggests that the model underperforms in capturing the critical target class, leading to many false negatives.

Specificity (0.8167): The specificity is relatively high, showing that the model is better at identifying counties without severe fatality rates ("True Negatives"). While this is positive, it highlights an imbalance between the ability to detect severe and non-severe cases.

Kappa (0.1759): The Kappa statistic indicates weak agreement between the model's predictions and the actual labels, beyond chance. This low value reflects a need for improvement in both precision and recall to enhance overall performance.

2. Improvement

- **Feature Engineering:** Introduce additional features or refine existing ones, such as combining socioeconomic variables or normalizing data distributions. This may help the model better capture the relationship between features and severe fatality rates.
- **Address Class Imbalance:** Use techniques like oversampling the minority class (counties with severe fatality rates), undersampling the majority class, or applying cost-sensitive learning to ensure the model gives equal focus to both classes.
- **Hyperparameter Tuning and Model Selection:** Experiment with hyperparameter optimization for the Naive Bayes classifier or try alternative models like Random Forest or Gradient Boosting, which may better handle complex relationships in the data.

F. k-Nearest Neighbors

1. Data Analysis

🎬 **Sensitivity (0.2480):** The model has a low sensitivity, indicating that it is identifying only a small portion of the actual high-risk counties (true positives). This suggests the model struggles with detecting high-risk cases effectively.

🎬 **Specificity (0.8211):** The specificity is relatively high, meaning the model correctly classifies most non-high-risk counties. However, this comes at the cost of missing many high-risk areas, as reflected by the low sensitivity.

🎬 **Balanced Accuracy (0.5346):** The balanced accuracy is slightly above 50%, suggesting that the model's overall performance is only marginally better than random guessing. The imbalance between sensitivity and specificity highlights the need for improvements.

2. Improvement

- **Handle Class Imbalance:** Apply techniques such as oversampling the minority class (e.g., SMOTE) or undersampling the majority class to address the imbalance between high-risk and non-high-risk counties. This can improve the model's sensitivity.
- **Feature Engineering:** Explore additional or more relevant features from the census data, such as interactions between demographic factors and pandemic variables, to provide better discriminatory power for the model.
- **Optimize Hyperparameters:** Conduct hyperparameter tuning for kNN, such as adjusting the number of neighbors or distance metrics, to enhance performance. Additionally, consider experimenting with alternative classifiers like Random Forest or Gradient Boosting for better accuracy and robustness.

Evaluation Metrics for CART

CART evaluation involves multiple metrics derived from its confusion matrix. **Precision** measures the proportion of true positives among predicted positives, showcasing how reliably the model identifies a specific class. **Recall** indicates the proportion of actual positives correctly classified, reflecting the model's

sensitivity. The **F1-score**, a harmonic mean of precision and recall, balances these metrics, especially useful in imbalanced datasets. For CART, these metrics highlight its effectiveness in making interpretable predictions and identifying patterns in simpler datasets. Additionally, **ROC-AUC** evaluates its ability to distinguish between classes, further validating its generalization across various thresholds.

During cross-validation, CART's results reveal its robustness across folds and its tendency to overfit if not pruned effectively. Its performance in accuracy is commendable for simpler, well-defined datasets, but its limited capability to capture complex relationships is evident. Feature importance rankings from CART further provide actionable insights, making it highly interpretable for stakeholders while maintaining reasonable performance.

Evaluation Metrics for ANN

ANN evaluation also focuses on metrics like **Precision**, **Recall**, and **F1-score** to analyze its predictive power. Due to its ability to model non-linear relationships, ANN often excels in handling complex datasets, though the metrics help identify areas where it might misclassify or underperform. The **ROC-AUC** for ANN demonstrates its superior capacity to differentiate between classes compared to simpler models, leveraging its ability to extract intricate patterns from the data.

In cross-validation, ANN explores optimal configurations of hidden layers (`size`) and regularization parameters (`decay`), significantly influencing its performance. The hyperparameter tuning ensures its generalization across folds, yielding high accuracy. However, its "black-box" nature and computational intensity remain trade-offs for its superior capability in capturing non-linear dynamics and handling large-scale data efficiently.

3. Overall Comparison

Model	**Training Accuracy**	**Testing Accuracy**	**Cross-Validation Accuracy**	**Strengths**	**Weaknesses**
Decision Tree	High	Moderate	Moderate	Simple, interpretable, fast training	Overfitting, struggled with medium-risk class
Random Forest	High	High	High	Robust to overfitting, effective with imbalanced classes	Computationally expensive

Multinomial Logistic Regression	Moderate	Moderate	Moderate	Interpretable, efficient for linearly separable data	Limited to linear patterns, struggled with medium-risk
CART	High but prone to overfitting if not pruned.	Moderate; may degrade on noisy data.	Consistent when pruned properly; interpretable.	Highly interpretable; simple to implement.	Prone to overfitting; sensitive to changes.
ANN	High, especially for complex relationships.	High, provided sufficient training data.	High, but requires extensive tuning.	Captures non-linear relationships; flexible.	Computationally intensive; black-box nature.

4. Insights from Performance Assessment

1. **Generalization Ability:**
    - Random Forest demonstrated the best generalization ability, maintaining consistent performance across training, testing, and cross-validation.
2. **Class Imbalance Impact:**
    - The medium-risk class posed challenges for all models due to its lower representation in the dataset.
    - Random Forest handled the imbalance better, likely due to its ensemble nature and feature randomness.
3. **Model Selection:**
    - **Random Forest** was the best overall performer, excelling in both accuracy and robustness.
    - **Decision Tree** offered a simpler, interpretable alternative but was prone to overfitting.
    - **Multinomial Logistic Regression** provided a useful baseline and insights into linear patterns.

Conclusion

By evaluating each model using training and testing accuracy, cross-validation, and detailed metrics, Random Forest emerged as the most reliable classifier for this task. It handled the complexities of the dataset effectively, particularly the non-linear relationships and class imbalance. The assessment process ensured that the chosen model was both robust and well-suited to the classification problem.

5. Evaluation

How Useful Is the Model for the Stakeholder?

The developed classification model is highly valuable for stakeholders involved in public health decision-making, pandemic preparedness, and resource allocation. By categorizing counties into low, medium, and high risk levels based on their COVID-19 case density and related features, the model provides actionable insights that can directly influence interventions, policy-making, and strategic planning.

Key Stakeholders and Their Needs

1. **Public Health Authorities:**
    - Need to identify high-risk areas for targeted interventions, such as vaccination campaigns, testing efforts, and healthcare resource allocation.
    - Require insights into medium-risk areas to prevent escalation to high-risk status through proactive measures like community awareness programs.
2. **Policy Makers:**
    - Need data-driven justifications for imposing or relaxing public health mandates, such as mask requirements or business closures.
    - Aim to balance public health goals with economic considerations.
3. **Healthcare Providers:**
    - Require information about high-risk regions to prepare for potential surges in hospitalizations.
    - Use medium and low-risk classifications to optimize the distribution of medical supplies and personnel.
4. **General Public:**
    - Benefit from clear communication about local risk levels, helping individuals make informed decisions about personal precautions.

How the Model Addresses Stakeholder Needs

1. **Actionable Insights:**
    - By classifying counties into risk levels, the model highlights regions requiring immediate attention, ensuring that interventions are timely and effective.
    - Medium-risk areas serve as a buffer zone where preventative measures can be intensified to avoid escalation.
2. **Data-Driven Resource Allocation:**
    - High-risk counties can be prioritized for emergency resources such as ICU beds, ventilators, and medical personnel.
    - Low-risk counties can focus on maintaining preventative strategies, conserving resources for areas of greater need.
3. **Proactive Planning:**
    - The predictive nature of the model allows stakeholders to anticipate risk levels and implement measures to mitigate the impact of future outbreaks.
    - Policy makers can use the model's insights to design scalable interventions that adapt to changing risk levels.
4. **Cost-Effectiveness:**
    - The model optimizes resource allocation by targeting areas with the highest need, reducing wastage in low-risk regions while minimizing oversights in high-risk areas.

Assessing the Model's Value if Used

To evaluate the practical value of the model, several dimensions can be considered:

1. Accuracy and Reliability

- **Assessment:** The model's performance was validated using cross-validation and testing accuracy. Random Forest, the best-performing model, achieved high accuracy and robust generalization, ensuring reliable classifications.
- **Value:** Stakeholders can trust the model to provide accurate and consistent predictions, reducing uncertainty in decision-making.

2. Impact on Public Health Outcomes

- **Assessment:** High-risk areas identified by the model can be correlated with actual healthcare outcomes, such as hospitalization rates or case surges, to measure its predictive effectiveness.
- **Value:** By enabling early interventions in high-risk areas, the model can help flatten infection curves, saving lives and reducing strain on healthcare systems.

3. Usability and Interpretability

- **Assessment:** The model's outputs (e.g., risk-level classifications) are straightforward and easy to interpret, even for non-technical stakeholders.
- **Value:** Policymakers and public health officials can quickly understand and act on the model's recommendations without requiring extensive technical expertise.

4. Adaptability

- **Assessment:** The model can be updated with new data (e.g., vaccination rates, new case trends) to remain relevant as the pandemic evolves.
- **Value:** This adaptability ensures the model remains a useful tool for ongoing and future public health challenges.

5. Economic Benefits

- **Assessment:** The cost-effectiveness of the model can be evaluated by comparing the expenses associated with targeted interventions versus broad, untargeted measures.
- **Value:** Focused interventions informed by the model save resources and reduce the economic impact of unnecessary restrictions in low-risk areas.

6. Trust and Stakeholder Engagement

- **Assessment:** The model's predictions can be compared against ground truth data (e.g., actual case trends) over time to build trust among stakeholders.
- **Value:** Transparent communication about the model's methods and limitations fosters confidence in its use for critical decision-making.

Limitations and Mitigation

While the model offers significant benefits, certain limitations must be considered:

1. **Class Imbalance:**
    - The medium-risk class is underrepresented, potentially affecting precision and recall for this category.
    - **Mitigation:** Stakeholders should focus on high and low-risk classifications while treating medium-risk results as an indicator for further investigation.

2. **Dynamic Nature of the Pandemic:**
    - Real-world conditions (e.g., vaccination rollouts, new variants) can alter risk dynamics rapidly.
    - **Mitigation:** The model should be regularly updated with the latest data to ensure ongoing relevance.
3. **Socioeconomic and Behavioral Factors:**
    - Features like mobility trends and economic indicators may not capture the full complexity of human behavior.
    - **Mitigation:** Incorporating additional features (e.g., vaccination rates, compliance with public health mandates) can enhance the model's accuracy.

Key Metrics for Stakeholder Evaluation

To assess the model's value over time, stakeholders can monitor the following metrics:

1. **Reduction in Hospitalization Rates:**
    - Compare high-risk areas receiving targeted interventions with historical hospitalization trends to evaluate impact.
2. **Accuracy of Predictions:**
    - Continuously validate the model against real-world data to ensure it remains reliable.
3. **Economic Efficiency:**
    - Measure the cost savings achieved by focusing resources on high-risk areas instead of implementing broad measures.
4. **Public Adoption and Compliance:**
    - Track community behavior in response to risk-level classifications to gauge the model's influence on public decision-making.

Methods to Improve Model Performance

a. Client Concern:
The model predicts too few high-risk counties, leading to an underestimation of areas needing urgent attention.
Solution:
1. **Increase Sensitivity:**
    - Adjust the classification threshold to favor more high-risk predictions (e.g., lowering the decision boundary for high-risk classification).
    - Rebalance the dataset using oversampling methods (e.g., SMOTE) or weight adjustments in the loss function to prioritize the high-risk class.
2. **Enhance Features:**
    - Add additional features that might correlate with high-risk counties (e.g., hospital capacity, vaccination rates, or historical infection trends).
3. **Switch to a Different Model:**
    - Use a model better suited for imbalanced datasets, such as Gradient Boosting Machines (XGBoost) or Random Forests, which can effectively handle such situations with class-weight adjustments.

b. Client Concern:
The client cannot understand how the model makes predictions, making it challenging to use the results for decision-making.
Solution:
1. **Switch to an Interpretable Model:**
    - Use inherently interpretable models like Decision Trees, Logistic Regression, or Rule-Based Models (e.g., PART) to provide clear decision paths.

2. **Explainability Tools:**
    - Apply model explainability techniques, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), to illustrate feature contributions for each prediction.
3. **Simplify Features:**
    - Reduce the feature set to focus on the most impactful variables, making it easier to communicate the model's rationale.

Conclusion

The model provides a valuable framework for stakeholders to make data-driven decisions during a pandemic. By categorizing counties into actionable risk levels, it supports targeted interventions, resource optimization, and proactive planning, ultimately contributing to improved public health outcomes and cost efficiency. Its adaptability and reliability make it an indispensable tool for managing both current and future public health crises.

Usefulness of the PART and ANN Models for Stakeholders

The model is designed to provide accurate and actionable insights regarding risk levels, making it valuable for stakeholders in the following ways:

1. **Informed Decision-Making:**
    - The CART model provides interpretable results, making it easier for stakeholders to understand and act upon predictions.
    - The ANN model, while less interpretable, can capture complex patterns and relationships in the data, offering more accurate predictions in intricate scenarios.
2. **Operational Efficiency:**
    - Automated risk-level predictions reduce the time and resources required for manual risk assessments.
    - Early identification of high-risk entities helps stakeholders allocate resources proactively.
3. **Performance Monitoring:**
    - The model supports continuous monitoring and evaluation of risk, which is critical in dynamic environments.

Assessing the Model's Value

To evaluate the effectiveness of the model, stakeholders can use the following methods:

1. **Performance Metrics:**
    - Use accuracy, precision, recall, F1-score, and AUC-ROC to assess predictive performance.
    - Compare the performance of CART and ANN models to determine which is better suited for the problem.
2. **Cost-Benefit Analysis:**
    - Evaluate the cost savings from automated predictions versus manual assessments.
    - Measure the economic or operational impact of interventions driven by model predictions.
3. **Feedback from End-Users:**
    - Gather qualitative feedback from stakeholders using the model to identify areas for improvement.
    - Ensure the model aligns with user needs and organizational goals.
4. **Real-World Testing:**
    - Deploy the model in a live environment on a pilot basis to evaluate its practical utility.
    - Measure the model’s impact on decision-making accuracy and timeliness.
5. **Update and Adaptability:**
    - Periodically update the model with new data to ensure it adapts to changes in risk patterns.
    - Assess how well the updated model continues to meet stakeholder requirements..

6. Deployment

Practical Use of the Model

The model's primary purpose is to assist stakeholders, such as public health authorities, policymakers, and healthcare organizations, in making data-driven decisions to manage and mitigate the impact of a pandemic. By categorizing counties into low, medium, and high risk levels, the model offers actionable insights to guide interventions, allocate resources, and design public health strategies.

1. How the Model Would Be Used

A. Risk Categorization

- **Objective:** Classify counties into risk levels based on pandemic-related data (e.g., case density, mobility trends, population demographics).

- **Implementation in Practice:**
  - Input real-time data, such as new COVID-19 cases, deaths, and mobility trends, into the model.
  - The model outputs risk levels (low, medium, high) for each county.

B. Resource Allocation

- **Objective:** Guide efficient distribution of limited resources (e.g., medical supplies, healthcare personnel, vaccines).
- **Implementation in Practice:**
  - High-risk counties receive priority for emergency resources, such as ICU beds, ventilators, and testing kits.
  - Medium-risk counties are targeted for proactive measures to prevent escalation.
  - Low-risk counties focus on maintaining current preventive strategies.

C. Strategic Interventions

- **Objective:** Inform public health strategies and policies.
- **Implementation in Practice:**
  - **High-Risk Counties:** Deploy containment measures such as lockdowns, travel restrictions, and mass vaccination campaigns.
  - **Medium-Risk Counties:** Launch community awareness programs, expand testing, and encourage voluntary precautions.
  - **Low-Risk Counties:** Monitor and evaluate ongoing measures to sustain low transmission levels.

D. Policy Development

- **Objective:** Justify public health mandates based on data.
- **Implementation in Practice:**
  - Use risk classifications to determine the necessity of mask mandates, school closures, or reopening of businesses.
  - Provide policymakers with evidence-based recommendations to balance public health and economic considerations.

2. Actions Taken Based on the Model

A. Immediate Actions

1. **Public Communication:**
    - Share risk levels with the public to promote awareness and encourage appropriate precautions (e.g., mask-wearing in high-risk areas).
2. **Emergency Response:**
    - Deploy healthcare workers and resources to high-risk counties to address potential surges in hospitalizations.
3. **Containment Measures:**
    - Implement strict public health measures in high-risk areas, such as curfews or travel restrictions.

B. Medium-Term Actions

1. **Proactive Measures in Medium-Risk Counties:**
    - Conduct targeted testing and vaccination drives to prevent escalation to high risk.
2. **Monitor and Adjust Strategies:**
    - Use the model's outputs to monitor the effectiveness of interventions and adapt them as needed.

C. Long-Term Actions

1. **Pandemic Preparedness:**
    - Use historical model predictions and actual outcomes to refine contingency plans for future pandemics.
2. **Resource Planning:**
    - Develop data-driven frameworks for long-term resource distribution and infrastructure improvements in healthcare.

3. Frequency of Model Updates

A. Dynamic Nature of Pandemics

- Pandemics are highly dynamic, with evolving factors such as new variants, vaccination rates, and changes in public behavior. To remain effective, the model must be updated regularly.

B. Recommended Update Frequency

1. **Weekly Updates:**
    - Update the model weekly with the latest data on new cases, deaths, mobility trends, and hospital capacity.
    - Weekly updates ensure timely responses to emerging hotspots and changes in risk levels.
2. **Real-Time Adjustments:**
    - For high-risk counties, incorporate real-time data feeds to enable immediate interventions.
    - Deploy predictive analytics to forecast risk levels for the upcoming week.
3. **Periodic Refinement:**
    - Reassess the model's feature set and thresholds quarterly or semi-annually to incorporate new variables (e.g., vaccination coverage, booster shot uptake).

4. Benefits of Frequent Updates

1. **Timeliness:**
    - Regular updates ensure the model reflects the latest pandemic trends, enabling stakeholders to act quickly.
2. **Improved Accuracy:**
    - Incorporating new data enhances the model's ability to predict risk levels accurately.
3. **Adaptability:**
    - Frequent updates allow the model to adapt to changes, such as the emergence of new variants or shifts in public behavior.

5. Practical Deployment Considerations

A. Infrastructure

- Use cloud-based platforms to host the model, ensuring scalability and accessibility for stakeholders.
- Integrate the model into existing public health systems for seamless data sharing and decision-making.

B. User Accessibility

- Develop user-friendly dashboards that present risk classifications and recommendations in an interpretable format.
- Provide stakeholders with tools to explore the underlying data and predictions.

C. Collaboration

- Work with local authorities to ensure that interventions based on the model align with regional needs and capacities.
- Engage with healthcare providers to coordinate resource distribution based on the model's outputs.

6. Value of Using the Model in Practice

A. Public Health Impact

- Helps contain outbreaks by targeting high-risk areas for interventions.
- Reduces healthcare system overload by enabling proactive resource allocation.

B. Cost Efficiency

- Focuses resources on areas of greatest need, minimizing waste and improving economic outcomes.

C. Transparency and Trust

- Provides an objective, data-driven basis for public health decisions, enhancing stakeholder confidence and public compliance.

Conclusion

The model's practical implementation offers a powerful tool for managing pandemics and similar public health crises. By regularly updating the model with real-time data, stakeholders can make timely and informed decisions, improving public health outcomes and optimizing resource use. Its adaptability ensures it remains relevant across different phases of a pandemic, making it an essential component of modern public health strategies.

Practical Use of CART and ANN Model:

To apply this model in practice, the workflow and considerations would involve the following steps:

1. Use in Practice

- **Risk Level Prediction:** The CART and ANN models would predict the `risk_level` of entities (e.g., regions, individuals, projects) based on their features. This could help organizations or stakeholders make informed decisions.
- **Decision Support System:** The predictions can feed into a dashboard or automated system for real-time risk assessment, providing clear actions like flagging high-risk cases for review or allocating resources to mitigate risks.
- **Policy Planning:** The models could be used to identify trends and patterns in historical data, helping in policy creation, resource allocation, and preventive strategies.

2. Actions Based on the Model

- **High-Risk Alerts:** Automatically notify relevant teams for immediate intervention.
- **Resource Optimization:** Allocate resources (e.g., personnel, budget) to areas or cases flagged as high-risk.
- **Continuous Monitoring:** Update operational strategies based on model predictions to adapt to changing conditions.
- **Performance Reviews:** Regularly analyze model outputs to assess effectiveness in real-world applications.

3. Update Frequency

- **Periodic Updates:** Update the model quarterly or semi-annually to incorporate new data and improve accuracy.
- **Trigger-Based Updates:** Retrain the model whenever significant changes occur in the underlying data (e.g., introduction of new features, data distribution changes).
- **Automated Updates:** Implement an automated pipeline for data ingestion and model retraining to ensure the model remains relevant.

4. Monitoring and Evaluation

- **Model Performance Tracking:** Continuously monitor model performance using metrics like accuracy, precision, recall, and F1-score. Use real-time test data for validation.
- **Error Analysis:** Regularly review misclassified cases to identify potential improvements in model features or architecture.
- **Feedback Loop:** Incorporate user feedback to fine-tune the models and align them with practical needs.

5. Considerations

- **Data Governance:** Ensure data privacy and compliance with regulatory requirements while handling sensitive information.
- **Scalability:** Ensure the model can handle increasing data volume and complexity as the system scales.
- **Interpretability:** Use CART for interpretability in decision-making while leveraging ANN for more complex patterns, balancing transparency and predictive power.

Student Contributions

Data Collection - Lead: Mahindra Guptha Kotha
Data Preparation - Lead : Mahindra Guptha Kotha
Modeling - Lead: Mahindra Guptha Kotha, Chenrui Zhao, Sujana Daniel Christopher
Evaluation - Lead: Mahindra Guptha Kotha, Chenrui Zhao, Sujana Daniel Christopher
Deployment - Lead: Mahindra Guptha Kotha, Sujana Daniel Christopher

7. Exceptional Work by Graduate Students:

[MAHINDRA GUPTHA KOTHA]
As a graduate student, I demonstrated exceptional work by applying nominal logistic regression and decision tree.

[SUJANA DANIEL CHRISTOPHER]

The Exception Work that I have done is the inclusion of 2 extra models (CART and ANN) than the minimum required, and there are insightful visualizations and explanations of the results.

8. List of References

- Centers for Disease Control and Prevention (CDC). "COVID-19 Guidelines and Safety Measures." Available from: [cdc.gov](https://www.cdc.gov).
- WHO, "COVID-19 Strategic Preparedness and Response Plan." 2020. Available from: [who.int](https://www.who.int).
- Michigan Medicine, "Flattening the Curve," University of Michigan. Available from: [michiganmedicine.org](https://www.uofmhealth.org).
- Johns Hopkins University. "Coronavirus Resource Center: COVID-19 Data and Guidelines." Available from: coronavirus.jhu.edu
- European Centre for Disease Prevention and Control (ECDC). "COVID-19 Pandemic: Strategies for Containment." Available from: ecdc.europa.eu.
- Harvard T.H. Chan School of Public Health. "COVID-19 Health and Safety Resources." Available from: hsph.harvard.edu.