Data Mining Project 3

**Preparing for Future Pandemics by Learning from COVID-19**

****

*A Data Mining Analysis*

*By: Juan Carlos Dominguez, Salissa Hernandez, Leonardo Piedrahita*

* **Executive Summary:**

This report addresses the challenge of predicting future pandemic vulnerability across Texas counties using historical COVID-19 data and population-level socioeconomic and health indicators. With the Texas Department of State Health Services (DSHS) as the primary stakeholder, this project aims to provide accurate, data-driven tools to guide pandemic preparedness, resource allocation, and long-term public health planning.

The importance of this work lies in enabling DSHS to proactively identify which counties are most at risk of severe outcomes in a future pandemic. By classifying counties into low, medium, or high risk categories, public health officials can anticipate where to concentrate medical supplies, testing efforts, communication strategies, and staffing during critical periods. This approach supports the agency’s broader goal of improving health outcomes through early intervention and targeted public health strategies.

Using supervised machine learning models—including k-Nearest Neighbors (kNN), Multinomial Logistic Regression, and Ethical Pruned Trees—we evaluated each algorithm’s ability to classify counties based on known pandemic outcomes and a wide range of county-level features. The kNN model emerged as the top performer, achieving an AUC of 0.977 for high-risk counties and 0.853 for low-risk ones, demonstrating its strong predictive power and sensitivity to localized patterns of vulnerability. Logistic regression provided a balance of moderate accuracy and high interpretability, while the pruned tree model showed limitations in accurately identifying medium-risk counties.

These findings offer DSHS practical and actionable insights: the predictive models can serve as early warning systems, while the underlying feature importance—such as population density, healthcare access, and economic disadvantage—can inform long-term policy aimed at mitigating systemic health disparities. Together, these insights equip DSHS with a robust analytical foundation to enhance readiness and resilience in the face of future pandemics.

**Table of Contents**

[**1. Problem Description 2**](#_as97eciu7hq)

[**2. Data Preparation 3**](#_mq00gqfjb3u1)

[**3. Data Exploration 8**](#_eu5zfz9rcf38)

[**4. Modeling & Evaluation 20**](#_vxkxqjdlrwn9)

[**5. Recommendations 125**](#_4pfwy7p2mex0)

[**6. Conclusion 126**](#_g0kjsu1bn1y9)

[**7. List of References 12**](#_vbhbxqv5wvol)**7**

[**8. Appendix 1**](#_iq7m2xr85bku)**28**

# **Problem Description**

**1.1 Background**

The COVID-19 pandemic revealed profound vulnerabilities across Texas counties and the U.S. at large. Experts continue to warn that future pandemics are not a matter of *if*, but *when*. To proactively prepare for such events, it is essential to learn from past data and outcomes, using them to guide targeted interventions that save lives and reduce disruption.

This project aims to build a classification model to categorize Texas counties as **low, medium, or high risk** of experiencing severe effects during a future pandemic. Leveraging county-level data—including demographics, healthcare access, socioeconomic indicators, and behavioral patterns—this model will help public health officials prioritize high-risk areas and make informed, timely decisions.

During COVID-19, outcomes were not uniformly distributed. Research has shown that health disparities were strongly shaped by income, race, occupation, commuting behaviors, and access to healthcare [1]. Counties with limited hospital capacity, higher poverty, or dense public transportation usage often saw disproportionately high infection and mortality rates [2]. A predictive model that captures these risk signals can assist policymakers in planning resource allocation, targeted messaging, and early interventions.

**1.2 Stakeholder: Texas Department of State Health Services (DSHS)**

The **Texas Department of State Health Services (DSHS)** is the primary stakeholder for this project. DSHS is tasked with overseeing Texas’s public health response, managing resource distribution (such as vaccines and ventilators), and implementing statewide health policies. During COVID-19, DSHS played a critical role in decisions regarding social distancing mandates, public communications, and crisis response coordination across both urban and rural counties.

By equipping DSHS with a model that classifies counties by pandemic risk level, the agency can:

* Prioritize counties for proactive testing, vaccination campaigns, and communications.
* Implement early interventions (e.g., mask mandates, school closures) in high-risk regions.
* Adapt responses dynamically using updated data as new variants or threats emerge.
* Develop long-term strategies to improve resilience in vulnerable counties.

Given Texas’s geographic and socioeconomic diversity, a one-size-fits-all approach is insufficient. This model offers DSHS a data-driven, localized lens to enhance both equity and efficiency in their pandemic response strategy.

**1.3 Research Questions & Clustering Approach**

This study proposes a **classification approach** to group counties based on their risk of severe impact from a future pandemic, using historical COVID-19 data and related structural features.

Key research questions include:

1. **Which Texas counties are most at risk of severe outcomes in a future pandemic based on social and economic indicators?**  
   *Rationale:* Identifies the most vulnerable counties, enabling early resource deployment.
2. **Can we classify counties into low, medium, or high risk with high accuracy using historical COVID-19 data and population-level features?**  
   *Rationale:* Evaluates whether a reliable predictive model can be trained on known outcomes and county characteristics.
3. **Which features are most predictive of pandemic risk, and how can this inform DSHS policy and intervention strategies?**  
   *Rationale:* Helps explain the drivers of vulnerability and guides systemic improvements in health equity and preparedness.

These questions address both the technical modeling goal and the real-world application of the results. A successful classification model will allow DSHS to respond faster and more effectively during a future health emergency by allocating resources where they are most needed and adjusting strategies based on localized risk [3].

# **Data Preparation**

**2.1 Data Collection & Data Quality**

The primary data source for this project is the COVID-19 Census dataset, which was previously compiled and cleaned as part of earlier coursework. This dataset integrates information from the U.S. Census and publicly available COVID-19 case data, focusing on county-level indicators across the United States. For this analysis, we filtered the dataset to include only Texas counties, ensuring alignment with our stakeholder's geographic scope. The dataset includes a wide range of demographic and socioeconomic variables such as population size, median income, poverty levels, public transportation use, and food stamp participation, along with COVID-19-specific metrics like total confirmed cases and deaths.

Since the dataset had already undergone extensive data cleaning in Project 2, we expect a high level of quality and reliability. Key preprocessing steps included standardizing variable names for readability, converting columns to appropriate data types, and addressing missing or inconsistent values. These efforts were essential given the original dataset's large size—over 3,142 rows and 259 columns—which could otherwise lead to computational or interpretability issues. In this project, we used the dataset independently without integrating additional sources, allowing us to maintain a focused analysis on how demographic and structural features relate to pandemic severity.

To build our predictive model, we retained a set of variables that have strong relevance to understanding COVID-19 impacts at the county level as shown in Table 1. These included cases\_per\_100k and deaths\_per\_100k as indicators of overall pandemic burden, and death\_per\_case, a derived ratio capturing the local fatality rate, which may reflect differences in healthcare access or population vulnerability. Economic hardship was represented through poverty, median\_income, and income\_per\_capita, while factors related to resource access and mobility limitations were included via pct\_on\_food\_stamps, commuters\_by\_public\_transportation, and pct\_work\_from\_home. The variable total\_pop was also retained to help normalize raw count values. Additionally, we engineered a categorical target variable called risk\_level, based on the distribution of cases\_per\_100k, which enables supervised classification of counties into low, medium, and high risk for pandemic-related harm.

*Table 1: Selected Features*

| **Features** | **Type** | **Scale of Measurement** | **Relevance** |
| --- | --- | --- | --- |
| county | factor | nominal | Identifies geographic unit for clustering |
| cases\_per\_100k | double | ratio | Confirmed COVID-19 cases per 100,000 residents (normalized for population) |
| deaths\_per\_100k | double | ratio | COVID-19 deaths per 100,000 residents |
| poverty | double | ratio | Percentage of people living below the poverty line |
| median\_income | double | ratio | Median household income (economic health) |
| pct\_on\_food\_stamps | double | ratio | Percentage of households receiving public assistance |
| commuters\_by\_public\_transportation | double | ratio | Proxy for crowding & transmission risk |
| pct\_work\_from\_home | double | ratio | Indicates ability to socially distance |
| income\_per\_capita | double | ratio | Reflects economic status on an individual level |
| deaths\_per\_case | double | ratio | Derived severity ratio calculated as deaths divided by confirmed cases; captures the local fatality rate |
| total\_pop | double | ratio | Total population of the county |

**2.2 Defining Our Classes**

To support the Texas Department of State Health Services (DSHS) in identifying counties most at risk during future pandemics, we defined a three-class target variable based on historical COVID-19 case rates per 100,000 residents. This outcome serves as a proxy for pandemic severity, allowing for a grounded, data-driven approach to anticipating regional vulnerability. By segmenting counties into **Low**, **Medium**, and **High** risk levels, our classification enables tailored resource allocation and more efficient deployment of health interventions, such as testing sites, vaccine rollouts, or mobile health units.

The class definitions were determined using natural breakpoints observed in the distribution of cases\_per\_100k. Specifically, counties with ≤100 cases per 100,000 were labeled **Low Risk**, those with 101–200 cases were classified as **Medium Risk**, and counties exceeding 200 cases per 100,000 were categorized as **High Risk**. These thresholds strike a balance between interpretability and statistical integrity. They align with public health logic by distinguishing jurisdictions that experienced notably different levels of exposure and stress on healthcare systems. Moreover, these cutoffs produced relatively balanced class sizes, which is advantageous for training effective machine learning models that avoid bias toward majority classes.

By framing our classification around a variable that DSHS already tracks and understands—case rates—we ensure that the outputs of our model remain interpretable and actionable for public health decision-makers. The resulting risk level classification equips DSHS with a scalable and repeatable method for identifying at-risk regions, planning ahead for outbreak containment, and minimizing the societal impact of future infectious disease threats.

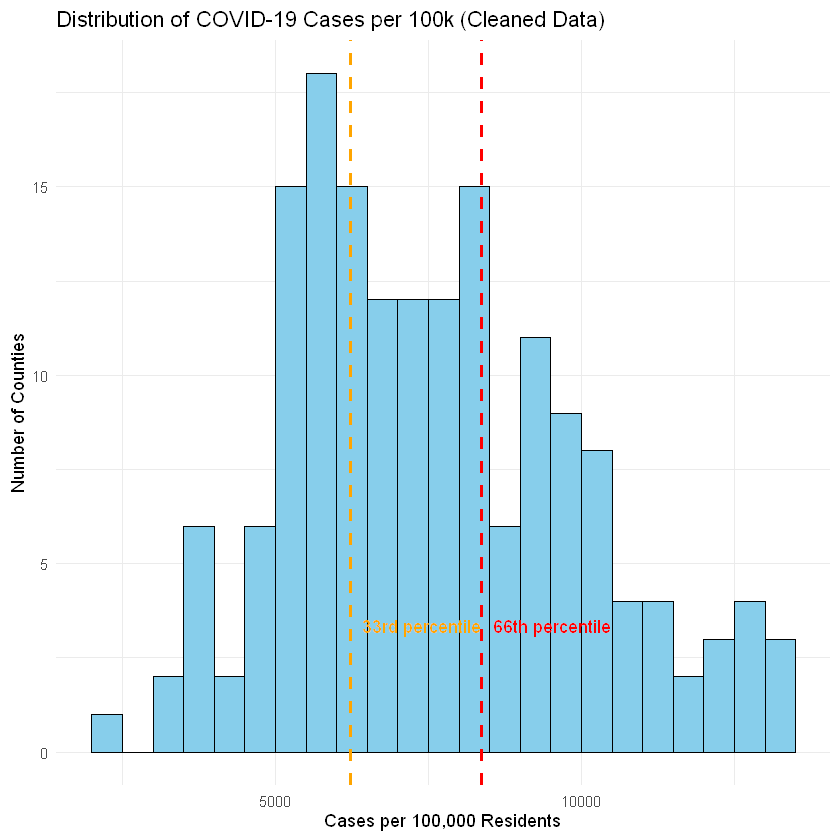
**2.3 Combining Files to Prepare for Classification**

To prepare the dataset for classification, this project builds on the data integration and cleaning work completed in Projects 1 and 2. The COVID-19 Census dataset used here combines publicly available COVID-19 case and death statistics with U.S. Census demographic and socioeconomic information. In prior stages, these sources were merged into a unified table with standardized variable names, consistent formats, and missing data addressed. For the current task, the dataset was filtered to include only Texas counties in order to align with the stakeholder’s regional focus.

A set of relevant predictor variables was selected based on prior analyses and public health relevance. Rows with missing values in any of these selected columns were removed to ensure compatibility with classification models that cannot handle incomplete data. Additionally, a new categorical target variable, risk\_level, was engineered based on the distribution of cases\_per\_100k, assigning each county to one of three groups: low, medium, or high risk. This classification was informed by the 33rd and 66th percentiles of the cases\_per\_100k distribution, which offers natural thresholds for interpreting pandemic severity, shown in Figure 1.

This transformation produced a single, well-structured dataset with a clearly defined class attribute (risk\_level), suitable for supervised classification modeling. The resulting dataset not only supports a robust analysis of COVID-19 impacts but also ensures compatibility with machine learning algorithms by providing a class attribute alongside the relevant predictors.

*Figure 1: Distribution of COVID Cases per 100k*

**

**2.3 Identifying Predictive Features, Creating Additional Features & Dealing with Missing Data**

To prepare the dataset for classification, several steps were taken to ensure the data is both relevant and clean for modeling. The dataset was built upon the work completed in Projects 1 and 2, where the COVID-19 Census dataset was integrated with publicly available COVID-19 case and death statistics, along with U.S. Census demographic and socioeconomic data. For the current analysis, the dataset was filtered to focus specifically on Texas counties, aligning the scope with the stakeholder’s regional focus. A selection of predictive features was identified based on their relevance to the pandemic's impact at the county level. These features include key COVID-19 indicators such as cases per 100,000 residents, deaths per 100,000 residents, and the death-to-case ratio, as well as socioeconomic factors like poverty, median income, percentage of people working from home, and food stamp participation. These features were considered essential for understanding the vulnerability of counties to the pandemic, as they encapsulate both health-related and socio-economic factors that could influence pandemic severity.

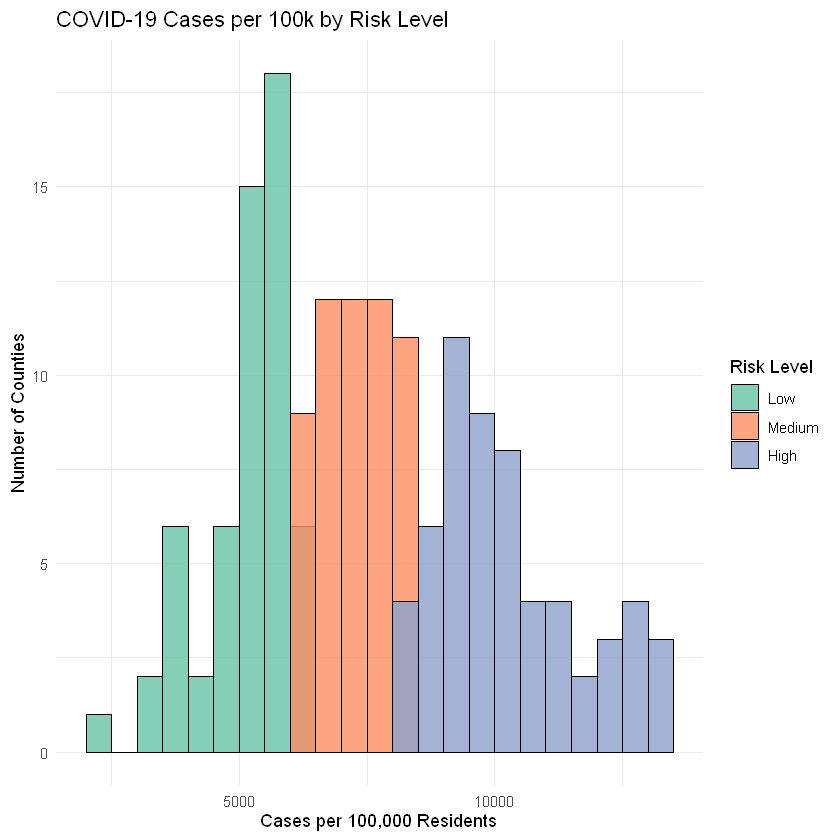
To ensure that the dataset was ready for classification models, a series of data cleaning steps were implemented. Missing values were handled by removing rows containing missing data for any of the key variables, which is crucial since classification models cannot handle incomplete data. The dataset was then thoroughly checked for missing values, and a summary was generated to confirm that there were no substantial gaps in the data for the selected variables. In addition to missing values, duplicate rows were identified and removed, ensuring that the dataset was not skewed by repeated observations.

New features were also created to enhance the dataset. Several derived variables, including cases\_per\_100k, deaths\_per\_100k, and death\_per\_case, were calculated to provide a more standardized view of the pandemic's impact. These transformations allowed for a better comparison across counties, as raw case numbers could be skewed by population size. Features related to socio-economic conditions, such as the percentage of people working from home (pct\_work\_from\_home) and food stamp participation (pct\_on\_food\_stamps), were also calculated to help assess the broader impact of the pandemic. These new features provide a deeper understanding of how various factors beyond the direct health impact of the virus might contribute to a county’s vulnerability.

The dataset was also screened for outliers to ensure that extreme values did not disproportionately influence the analysis. Outliers were identified and removed using the interquartile range (IQR) method, which is a common technique for detecting values that fall outside of the expected range. Variables like cases per 100,000, deaths per 100,000, and median income were checked for outliers, and rows containing extreme values were excluded from the dataset. This step was important for improving the robustness of the model, as outliers can sometimes lead to biased results.

With the data cleaned and enhanced, a categorical target variable called risk\_level was created based on the distribution of cases\_per\_100k. Counties were classified into three categories: low, medium, or high risk, based on thresholds derived from the data, as shown in Figure 2. This class variable is now ready for use in supervised classification models, allowing for predictions of which counties are most likely to be impacted by future pandemics.

*Figure 2: COVID Cases per 100k by Risk Level*

**

In summary, the dataset was prepared by identifying relevant predictive features, creating additional features to provide a more comprehensive view of county-level vulnerability, and addressing missing data, duplicates, and outliers. These steps ensured that the dataset is clean, complete, and ready for classification modeling, providing a solid foundation for future analysis and predictions.

# **Data Exploration**

After identifying and selecting the key variables for our modeling efforts, we conducted a thorough quality check on the dataset. This step was essential to ensure the accuracy, consistency, and reliability of subsequent data mining tasks. It involved assigning appropriate data types, verifying data completeness, identifying and removing outliers, and exploring descriptive statistics to derive preliminary insights.

#### **3.1 Verifying Data Quality**

To begin, we examined each variable to confirm the correct assignment of data types, which facilitates accurate computations and efficient analysis.

**Categorical Variable:**

* county (factor): Identifies individual counties in Texas, enabling county-level grouping, comparison, and aggregation.

**Numerical Variables:**

* cases\_per\_100k (double): Normalized COVID-19 case count per 100,000 people, used to compare infection rates.
* deaths\_per\_100k (double): Normalized death rate per 100,000 people due to COVID-19.
* death\_per\_case (double): Ratio of deaths to total cases, providing insight into case severity or underreporting.
* poverty (double): Total number of people in poverty, indicating socioeconomic vulnerability.
* median\_income (double): Median household income per county, reflective of economic conditions.
* pct\_on\_food\_stamps (double): Percentage of individuals receiving public food assistance.
* commuters\_by\_public\_transportation (double): Total commuters using public transportation, a possible vector for transmission.
* pct\_work\_from\_home (double): Percentage of the workforce telecommuting, related to social distancing capacity.
* income\_per\_capita (double): Mean income per person, showing wealth distribution.
* total\_pop (double): County population, important for contextualizing all other metrics.

Assigning appropriate data types ensures numerical precision, efficient filtering, and appropriate aggregation during both exploration and modeling phases.

#### **3.2 Handling Missing Values**

Missing data can introduce bias, reduce statistical power, and distort relationships between variables. To assess completeness, we checked for missing values across all columns.

**Outcome:**

* No missing values were found in the dataset.
* All 254 rows and 11 variables were complete, allowing full retention of data for analysis.

This high level of completeness enhances the robustness of any resulting insights and modeling efforts.

#### **3.3 Duplicate Check**

To ensure each observation was unique, we inspected the dataset for duplicate rows, which could distort trends or overweight specific counties.

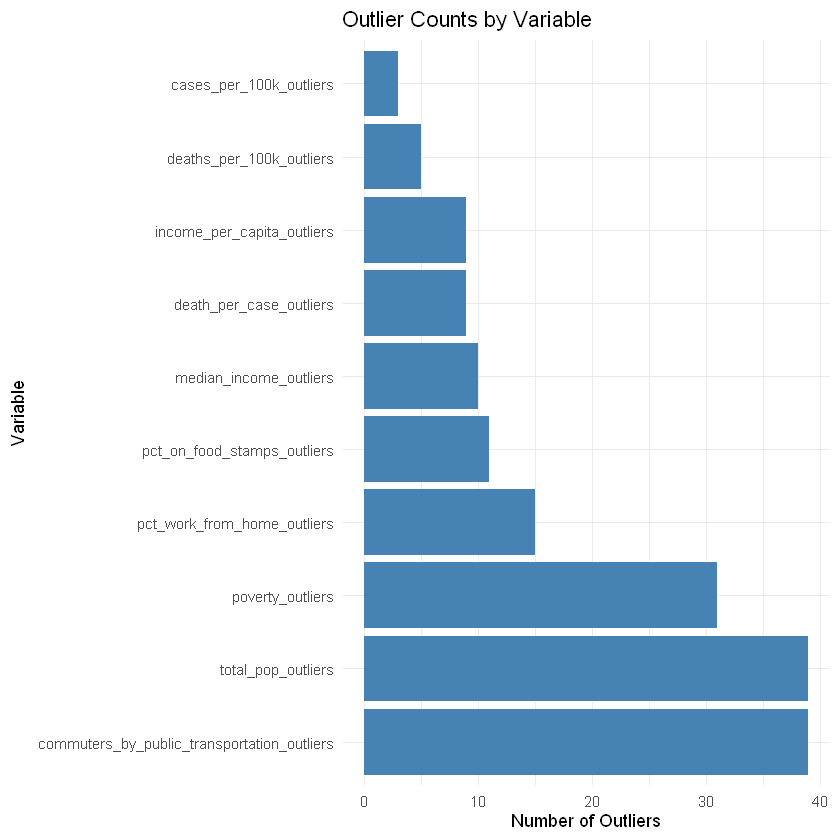
**Outcome:**

* No duplicate entries were detected.
* Dataset remained at 254 unique records.

This confirms that every row represented a distinct Texas county, supporting consistent county-level analysis.

#### **3.4 Outlier Detection and Removal**

To reduce distortion from extreme values, we conducted an outlier analysis using the IQR method. The outliers detected are shown in Figure 3. Outliers were removed from the dataset to improve the reliability of subsequent modeling and statistical summaries.

*Figure 3: Outliers Detected  
*

**Outliers Detected:**

* cases\_per\_100k: 3
* deaths\_per\_100k: 5
* death\_per\_case: 9
* poverty: 31
* median\_income: 10
* pct\_on\_food\_stamps: 11
* commuters\_by\_public\_transportation: 39
* pct\_work\_from\_home: 15
* income\_per\_capita: 9
* total\_pop: 39

These anomalies likely reflect:

* Wide disparities in population sizes between rural and urban counties
* Socioeconomic inequality
* Differences in reporting, testing, or healthcare access

**Outcome:**

* 84 rows were removed due to outliers.
* Final dataset contains 170 rows and 11 columns.

### **3.5 Descriptive Statistics**

After completing the data cleaning process—including the removal of outliers and validation of data quality—we conducted a descriptive statistical analysis to better understand the structure and distribution of our variables shown in Table 2. This step provides a high-level overview of central tendencies (mean and median), spread (minimum and maximum values), and overall variability across the dataset.

**Final Dataset Dimensions:**

After cleaning, the dataset was reduced from 254 to **179 rows** and retained **11 columns**. This final dataset represents a curated and high-quality subset of Texas counties, free from missing values, duplicate entries, and extreme anomalies that could otherwise skew our analysis.

*Table 2: Statistical Summary of Selected Features*

| **Variable** | **Max** | **Min** | **Mean** | **Median** |
| --- | --- | --- | --- | --- |
| cases\_per\_100k | 13159 | 2311 | 7608 | 7198 |
| deaths\_per\_100k | 391.0 | 0.0 | 181.4 | 174.7 |
| poverty | 16430 | 10 | 3,571 | 2314 |
| median\_income | 74368 | 29104 | 48816 | 47854 |
| pct\_on\_food\_stamps | 24.83 | 0.0 | 12.74 | 12.76 |
| commuters\_by\_public\_transportation | 126.0 | 0.0 | 18.7 | 4.0 |
| pct\_work\_from\_home | 7.32 | 0.0 | 3.20 | 3.14 |
| income\_per\_capita | 33472 | 13369 | 24377 | 24050 |
| deaths\_per\_case | 0.055 | 0.0 | 0.02 | 0.02 |
| total\_pop | 93635 | 289 | 23016 | 17128 |

.**Interpretation:**

* **Wide Range in Population and Economic Indicators:** County population sizes range from fewer than 300 residents to over 93,000, reflecting substantial demographic diversity. Median household income also varies significantly, suggesting broad socioeconomic stratification across counties.
* **Access to Public Services and Mobility:** There’s a striking disparity in the number of commuters using public transportation, ranging from 0 to 126. This may reflect rural-urban divides or the presence (or lack) of public transit infrastructure. Conversely, the proportion of individuals working from home remains relatively low but consistent, with a modest upper range (7.32%).
* **COVID-19 Impact:** The death-per-case ratio has a relatively small mean (~2.5%) and low maximum (~5.5%), consistent with the case fatality rates observed during the pandemic’s later phases. However, some counties reported zero deaths, possibly due to better healthcare access, lower case numbers, or data lags.
* **Poverty and Food Assistance:** While the average poverty count is 3,571 individuals per county, the large standard deviation and maximum value suggest a few counties bear a disproportionate burden. Similarly, food assistance usage ranges widely, further underlining disparities in economic well-being.

**Conclusion:**

This descriptive analysis illustrates the heterogeneity of Texas counties in terms of demographics, economic status, public transit usage, and pandemic impact. These variations are crucial to consider in predictive modeling, as they may influence risk levels differently across regions. The insights gained here emphasize the need for tailored public health interventions and policy responses that account for local socioeconomic and infrastructural contexts.

In addition to the summary table above, we computed detailed five-number summaries (minimum, 1st quartile, median, 3rd quartile, and maximum) for key outcome variables—**cases per 100k**, **deaths per 100k**, and **death per case**—shown below in Table 3.

*Table 3: Summary Statistics for Key Variables*

| **Variable** | **Min** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Max** |
| --- | --- | --- | --- | --- | --- | --- |
| cases\_per\_100k | 2311 | 5886 | 7198 | 7608 | 9213 | 13159 |
| deaths\_per\_100k | 0.0 | 125.1 | 174.7 | 181.4 | 234.9 | 391.0 |
| poverty | 0.0 | 0.02 | 0.02 | 0.02 | 0.03 | 0.05 |

**Interpretation:**

* **COVID-19 Impact Indicators:** The distribution of *cases per 100k* and *deaths per 100k* demonstrates significant right-skew, with maximum values nearly double those of the third quartile. This skewness suggests a small number of counties faced disproportionately high burdens of disease and death.
* **Death per Case Stability:** The *death\_per\_case* ratio shows a relatively stable distribution across counties, with a narrow interquartile range. This consistency could reflect statewide similarities in healthcare quality or uniformity in how deaths and cases were reported.
* **Population and Economic Variability:** The widespread in *total\_pop*, *poverty*, and *income\_per\_capita* reinforces the diverse socioeconomic profiles of Texas counties. Median income ranges from under $30k to nearly $75k, indicating broad inequality.
* **Transportation Access:** *Commuters\_by\_public\_transportation* is highly skewed, suggesting that while a few urban counties rely on transit heavily, the majority of counties have little to no public transportation infrastructure.

**Conclusion:**

This descriptive analysis confirms substantial inter-county variability across nearly all measured dimensions. These insights are foundational for modeling efforts that aim to predict county-level vulnerability or resilience to pandemics. It also underscores the need for region-specific public health strategies rather than uniform policy implementation.

### **3.6 Descriptive Visualizations**

**Histograms:**

*Figure 4: Histogram of All Numeric Variables*

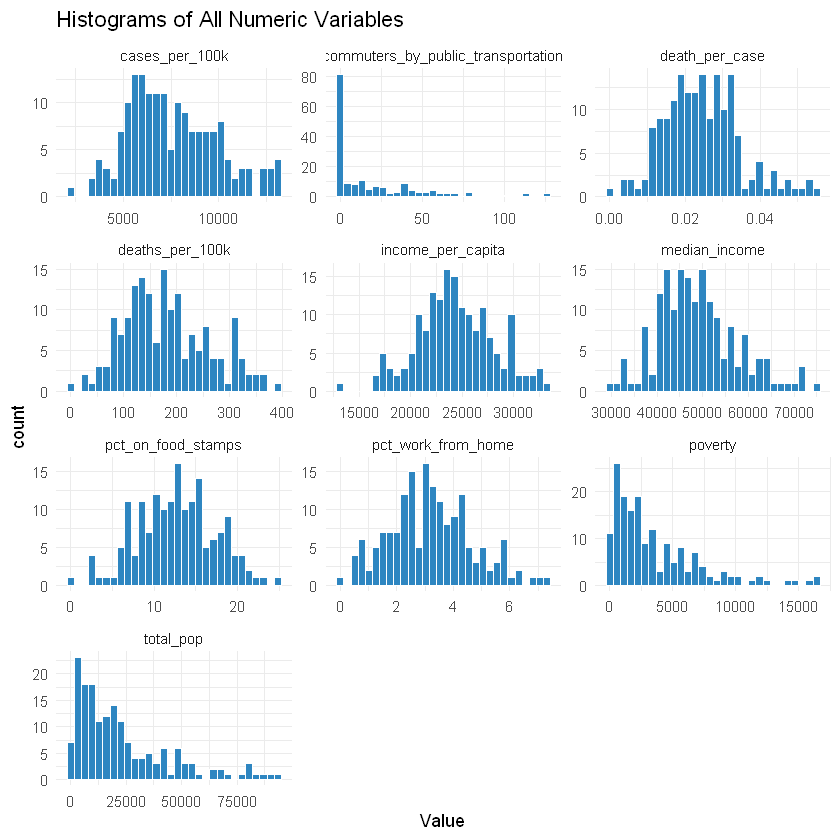


Figure 4 displays histograms of all numeric variables utilized for this report. The distribution of **cases\_per\_100k** appears roughly normal with a slight right skew. Most counties fall between 5,000 and 10,000 cases per 100k, indicating a relatively consistent level of exposure across regions. However, a small number of counties report significantly higher values, exceeding 12,000 cases per 100k. These outliers may correspond to densely populated urban areas, regional outbreaks, or data reporting inconsistencies. From a modeling perspective, these outliers could be retained if they reflect real phenomena or winsorized if they unduly influence model performance. Further normalization, such as adjusting for population density or accounting for temporal variation in outbreaks, may also improve interpretability.

In contrast, **deaths\_per\_100k** follows a fairly normal distribution centered between 100 and 300. This relative stability suggests that the mortality burden was more evenly distributed than case counts, possibly due to similar access to healthcare or consistent mitigation strategies. This variable’s distribution makes it a strong candidate for comparative analysis and outcome modeling, particularly when used as a dependent variable in regression or classification frameworks.

The variable **death\_per\_case**, which captures the ratio of deaths to confirmed cases, is more heavily right-skewed. Most counties report death\_per\_case rates around 2% to 3%, though some reach fatality rates of 5% or higher. These elevated values could reflect systemic healthcare disparities, higher-risk populations (such as elderly or immunocompromised groups), or under-reporting of mild cases that inflates the observed fatality rate. Given its skewness and interpretability, death\_per\_case may be more useful when categorized into tiers (e.g., low, medium, high fatality risk) to aid in modeling and policy interpretation.

The **poverty** variable, which represents the absolute number of individuals in poverty, is highly right-skewed. While most counties report fewer than 5,000 individuals below the poverty line, a few exceed 15,000. Because this is a raw count, it correlates strongly with total population size and is less useful for direct comparisons across counties. Transforming it into a proportion (e.g., percentage of the population in poverty) would provide a more meaningful and equitable comparison for modeling and policy analysis.

**median\_income** exhibits a slight right skew, with most counties clustered between $40,000 and $55,000. Very few counties exceed $70,000, indicating moderate income inequality across the dataset. This variable is useful for understanding socioeconomic conditions that may influence COVID-19 outcomes, such as access to healthcare, the ability to isolate or work remotely, and general community resilience. However, since it may correlate with other income-related variables like income\_per\_capita or pct\_on\_food\_stamps, multicollinearity checks are recommended before including multiple similar predictors in a model.

The **pct\_on\_food\_stamps** variable is fairly uniformly distributed with a light right skew, generally ranging between 5% and 20%. This variable reflects economic vulnerability and reliance on assistance programs, making it a valuable indicator of hardship. It likely correlates with other socioeconomic factors, including poverty and median\_income, and may also relate to health outcomes due to structural barriers to care or differences in living conditions.

The number of **commuters\_by\_public\_transportation** is extremely right-skewed. Most counties report very few commuters—often fewer than 10—using public transportation, while a small number report values in the hundreds or thousands. This stark distribution reflects the urban-rural divide and suggests that most counties are rural with limited transit infrastructure. Given its skew, a log(x+1) transformation or binning into categories (e.g., no usage, moderate usage, high usage) may improve model performance and interpretability. This variable may also be indirectly related to transmission risk in dense urban environments.

**pct\_work\_from\_home** follows a fairly normal distribution centered around 3%, with most values falling between 0% and 7%. This relative consistency suggests that remote work capacity is uniformly low to moderate across counties, possibly reflecting occupational structure and access to technology. Given its behavioral implications, this variable may serve as a protective factor—higher pct\_work\_from\_home values could be associated with lower COVID-19 transmission rates. It may be insightful to explore its interaction with income or urbanicity in further modeling.

The variable **income\_per\_capita** is symmetric to slightly right-skewed, with most counties falling between $20,000 and $30,000. While related to median\_income, income\_per\_capita is more sensitive to extremely high earners and provides a broader sense of local wealth distribution. It is a strong candidate for continuous regression models but may be redundant if median\_income is already included—checking for multicollinearity can help determine which variable offers the most unique explanatory power.

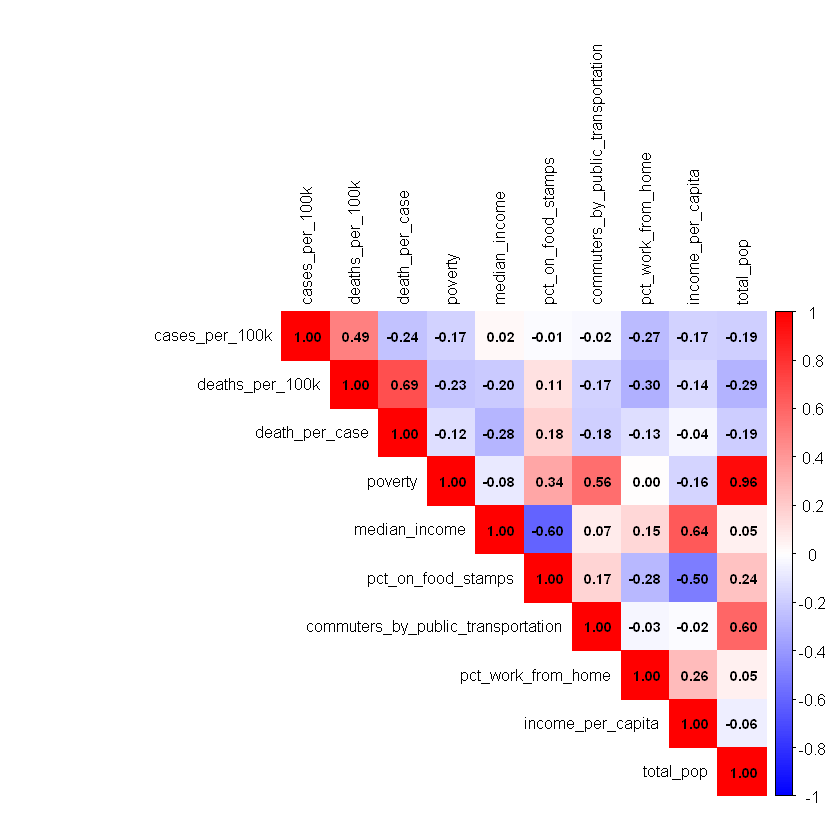
Finally, **total\_pop** is heavily right-skewed. The vast majority of counties have fewer than 25,000 residents, but a small number approach or exceed 90,000. This wide range reflects the diversity of county sizes and underscores the need to transform or bin this variable for effective use in modeling. A log transformation is a common approach, or alternatively, grouping counties into size tiers (e.g., rural, mid-size, urban) can allow for better interpretability and more stable model coefficients.

Overall, several variables in this dataset—especially poverty, total\_pop, and commuters\_by\_public\_transportation—exhibit significant skew and may benefit from log transformations or categorical binning. Outliers in these variables should be examined carefully to determine whether they represent true extremes or data artifacts. Standardizing variables (e.g., z-scores) will be important when combining features with different units and scales. Many variables are likely correlated, including income\_per\_capita with median\_income and poverty with pct\_on\_food\_stamps, and checking these relationships through correlation matrices or variance inflation factors (VIF) will be crucial for building robust predictive models. These variables offer rich opportunities for understanding COVID-19 risk, socioeconomic resilience, and infrastructure vulnerabilities across U.S. counties.

**Correlation Matrix:**

The correlation matrix in Figure 5 provides a detailed overview of linear relationships among key county-level variables related to COVID-19 outcomes and socioeconomic indicators in Texas. One of the most striking patterns is the **strong positive correlation (0.96)** between median\_income and income\_per\_capita, which is expected since both are proxies for economic status. Similarly, poverty shows a notable **negative correlation with income indicators**, particularly with median\_income (-0.60), suggesting that counties with higher median incomes tend to have fewer people in poverty—confirming the intuitive socioeconomic divide. Interestingly, death\_per\_case has only weak correlations with most predictors, including cases\_per\_100k and deaths\_per\_100k, which may suggest that mortality outcomes are influenced by complex, possibly unobserved factors such as healthcare access, comorbidities, or age demographics.

*Figure 5: Correlation Matrix of Numeric Variables*

**

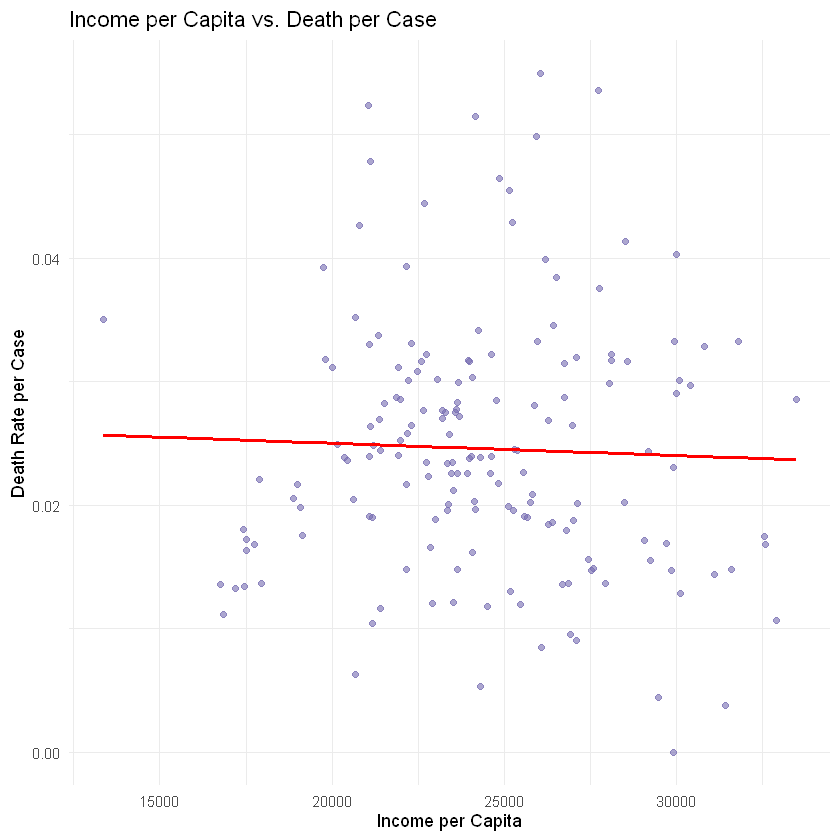
From a public health perspective, the correlation of pct\_on\_food\_stamps with poverty (0.64) and its inverse relation to income indicators implies that food assistance can serve as a reliable signal for financial strain. This insight can help the Texas Department of State Health Services (DSHS) prioritize resource allocation to counties that are both economically disadvantaged and more reliant on public assistance. Additionally, the weak to moderate **negative correlation between pct\_work\_from\_home and cases\_per\_100k (-0.27)** supports the idea that remote work might have mitigated viral spread in certain areas, suggesting it could be a protective factor worth encouraging during future outbreaks.

For policy recommendations, this analysis highlights the value of incorporating economic and behavioral variables into pandemic preparedness models. DSHS could benefit from targeting interventions—such as mobile testing, vaccine campaigns, or public health messaging—toward counties with high poverty and low remote work adoption. Furthermore, leveraging variables like pct\_on\_food\_stamps or commuters\_by\_public\_transportation, which are easy to track, could enhance real-time risk assessment tools. Overall, these findings emphasize that socioeconomic vulnerability is intricately tied to health outcomes and should be central in public health planning and response.

**Scatterplot:**

Figure 6 illustrates the relationship between income per capita and death rate per COVID-19 case across various counties, revealing important patterns that can inform future public health strategy in Texas. The graph shows a weak negative correlation between income per capita and the death rate per case, indicated by the slight downward slope of the regression line. While this suggests that counties with higher average incomes tend to experience slightly lower death rates per case, the relationship is far from strong, as evidenced by the wide dispersion of data points around the trend line. Some lower-income counties display relatively low death rates, while certain higher-income counties still encounter elevated death rates per case, pointing to the influence of additional factors beyond income alone.

*Figure 6: Relationship between income\_per\_capita & death\_per\_case*

**

For the Texas Department of State Health Services (DSHS), these findings underscore the need for nuanced, data-driven approaches in preparing for future pandemics by learning from COVID-19 patterns. Income per capita can serve as a proxy for healthcare access, public health literacy, and living conditions. Recognizing the modest but meaningful association between socio-economic status and pandemic outcomes allows DSHS to strategically target public health interventions in more vulnerable, lower-income communities. This could involve deploying mobile health units, increasing vaccination outreach, and offering targeted health education initiatives in counties where income disparities are likely to compound pandemic-related risks.

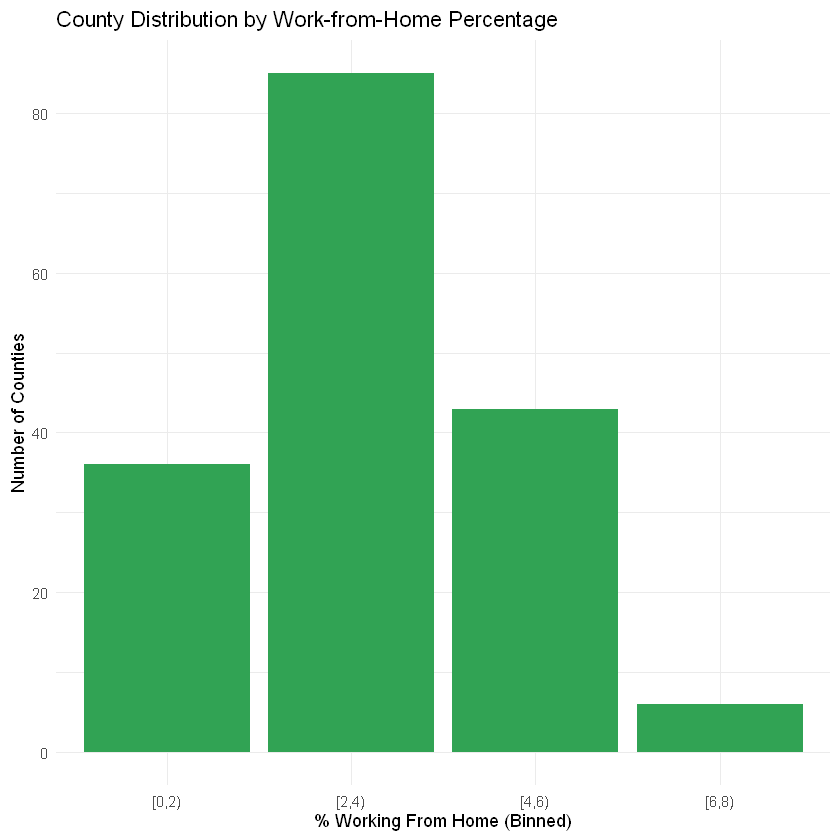
Additionally, this analysis supports the importance of informed resource allocation. By identifying counties with lower income levels and historically higher death rates per case, Texas DSHS can prioritize these areas for medical supplies, emergency funding, and surge healthcare staffing in future public health emergencies. A vulnerability index incorporating income data alongside healthcare capacity, demographics, and previous pandemic outcomes could greatly enhance the effectiveness of this approach. Moreover, the scatter in the data indicates that public health outcomes are shaped by complex, intersecting factors, suggesting the need for tailored communication strategies that address cultural, linguistic, and accessibility barriers in lower-income areas.

Finally, the evidence presented here justifies longer-term investments in public health infrastructure, particularly in communities with lower income per capita. Building new clinics, enhancing hospital resources, and expanding community health programs in these areas could mitigate the disparities observed during COVID-19. While income per capita alone does not fully predict death rates, its association with health outcomes highlights an opportunity for Texas DSHS to build resilience in the state’s most at-risk counties, using targeted, equitable, and evidence-based planning for future pandemics.

**Bar Chart:**

Figure 7 presents the distribution of Texas counties based on their percentage of residents working from home, grouped into four percentage ranges. The majority of counties fall within the 2% to 4% range, with over 80 counties in this category. A smaller but notable number of counties fall between 4% and 6%, followed by those in the 0% to 2% range. Only a small handful of counties, fewer than 10, report work-from-home percentages between 6% and 8%. This distribution reveals a clear pattern: most Texas counties have relatively low levels of remote work, with very few areas reaching higher adoption rates.

*Figure 7: County Distribution by Work-from-Home Percentage*

**

For the Texas Department of State Health Services (DSHS), this information offers valuable insights into workforce mobility patterns and community-level vulnerability during a public health crisis like a pandemic. Counties with lower percentages of residents working from home are likely to have populations heavily engaged in essential, in-person, or public-facing jobs, which increases their risk of virus exposure and transmission. These areas may face greater challenges in maintaining social distancing, controlling workplace outbreaks, and reducing community spread. Recognizing this, DSHS can prioritize these counties for workplace safety programs, on-site rapid testing, mobile clinics, vaccination drives, and targeted public health messaging during future pandemics.

This data also benefits Texas DSHS in forecasting potential strain on healthcare systems during outbreaks. Counties with limited work-from-home capacity are more susceptible to faster transmission rates, which can quickly overwhelm local healthcare facilities. Early identification of these high-risk areas enables DSHS to proactively allocate critical resources such as hospital beds, protective equipment, surge staffing, and emergency medical services. Moreover, it highlights the importance of encouraging local governments, businesses, and employers in these counties to develop flexible work arrangements and pandemic contingency plans, where feasible. Supporting the expansion of remote work infrastructure could reduce exposure risks and enhance community resilience. Ultimately, this analysis equips Texas DSHS with a clearer understanding of where and how to intervene effectively, ensuring that future pandemic responses are both equitable and strategically targeted to protect the state’s most vulnerable populations.

# **Modeling & Evaluation**

**4.1 Model 1: Decision Tree**

Our first classification model is a decision tree, chosen for its transparency, interpretability, and ease of implementation. Decision trees are particularly useful for stakeholders like the Texas Department of State Health Services (DSHS), who need clear, rule-based outputs to guide public health decisions. We removed the unique identifier (county) to avoid data leakage and trained the model using a stratified 80/20 train-test split. The model uses **cases\_per\_100k** as the primary predictor and identifies two natural breakpoints that correspond to the 33rd and 66th percentiles of the case distribution. This simplicity allows the model to perfectly separate counties into Low, Medium, and High risk levels in the training data. The model’s assumption is that the chosen feature **(cases\_per\_100k**) adequately captures the variance needed for classification—a reasonable starting point, though potentially limited in generalizability. Performance on the test set is strong, with an overall accuracy of 96.97%, though we later discuss a misclassification pattern in the evaluation section. For DSHS, this model could support targeted resource allocation and early intervention strategies in a future pandemic, providing a rapid and interpretable risk assessment tool.

While the decision tree model performs well in terms of classification accuracy, it does have several limitations that could impact its effectiveness in real-world applications. One significant limitation is overfitting, which occurs when the model becomes too closely tied to the training data. In this case, the decision tree perfectly classifies counties into their respective risk categories on the training set, but this may not generalize well to new, unseen data. This can be particularly problematic in public health decision-making, as the model may fail to adapt to changes in disease transmission patterns or other influencing factors in a future pandemic.

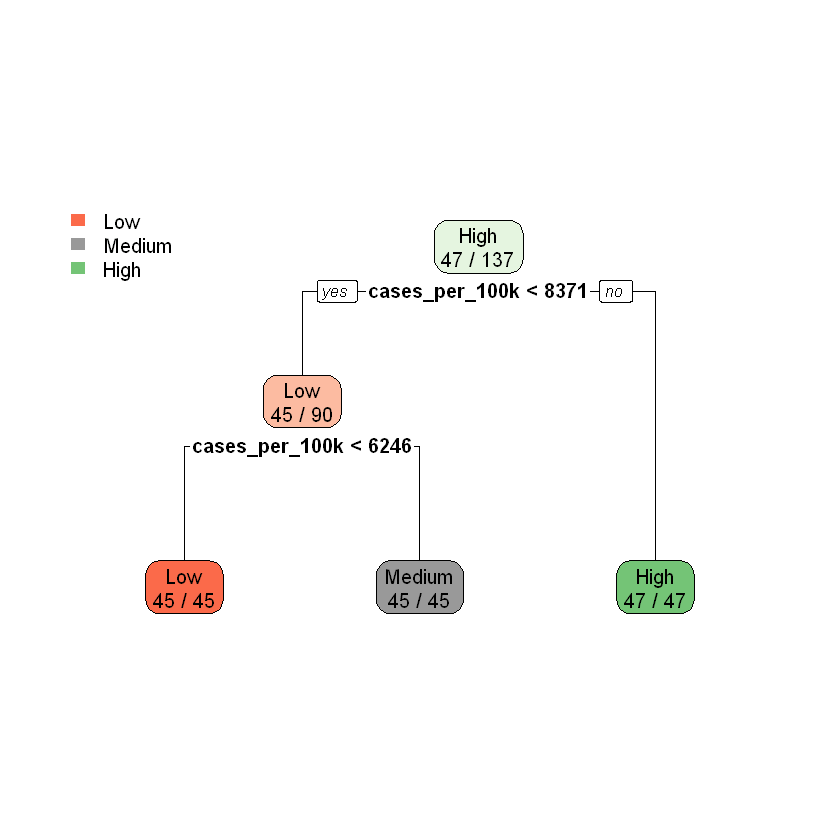
Another limitation of decision trees is that they are sensitive to outliers. Extreme values in the data can result in splits that may not reflect the broader patterns in the population, potentially leading to overfitting or skewed results. Additionally, decision trees tend to create rules that are locally optimal, which means that they might capture relationships that are specific to the current dataset but not necessarily robust across different contexts or time periods. This makes the model prone to low generalizability, particularly in highly dynamic situations like pandemics where the nature of the virus and its spread may evolve.

Finally, while decision trees are interpretable, they can also become complex as the number of features and data points increases. This can make the model more difficult to explain to non-technical stakeholders, potentially reducing its effectiveness in real-world decision-making. Furthermore, feature selection becomes crucial, as irrelevant or noisy features can lead to unnecessary complexity and overfitting.

Despite these limitations, the simplicity and clarity of decision trees make them a valuable tool for guiding initial public health strategies, especially when paired with cross-validation techniques or post-processing methods like pruning to address overfitting. For DSHS, the decision tree offers an easy-to-understand framework for allocating resources and making decisions during the early stages of a pandemic, though further refinement of the model could improve its robustness and generalizability.

The decision tree model in **Figure 8** uses **cases\_per\_100k** as the primary predictor to classify counties into Low, Medium, and High risk categories. The root node shows a mix of all three risk levels, with a majority of counties being classified as High risk. The first split divides the data based on whether **cases\_per\_100k** is less than 8370.828, separating counties into Low and Medium risk categories. If the case rate is below 6246.276, counties are classified as Low risk, while those with case rates above this threshold but below 8370.828 are classified as Medium risk. The second split identifies counties with a **cases\_per\_100k** greater than or equal to 8370.828 as High risk. The decision tree provides clear, interpretable rules for classification, with perfect separation of risk levels at the terminal nodes. However, it also exhibits overfitting, as it perfectly classifies counties in the training data. Despite this, the model offers a straightforward, understandable approach to predicting pandemic risk, making it beneficial for stakeholders such as the Texas Department of State Health Services (DSHS), as it enables clear decision-making for resource allocation and prioritization based on COVID-19 case rates. The simplicity and interpretability of the decision tree make it particularly valuable for informing public health responses in an easily understandable manner.

*Figure 8: Decision Tree Split*

**

By leveraging this decision tree model, the Texas Department of State Health Services (DSHS) could proactively assess which counties are at higher risk during a future pandemic. The ability to classify counties into distinct risk categories allows DSHS to allocate resources—such as hospital beds, ventilators, and medical personnel—more effectively. Additionally, DSHS could use the risk assessment to implement targeted public health measures like issuing mask mandates or temporarily closing schools in high-risk counties, potentially reducing the spread of the virus. The simplicity and interpretability of the decision tree also allow DSHS decision-makers to understand the rationale behind the model’s recommendations, which can enhance trust in the tool and support informed, timely responses

**4.1.1 Predictions & Confusion Matrix**

This section evaluates the decision tree classification model's performance, focusing on the confusion matrix for both the test dataset and the resubstitution error. The confusion matrix compares the true class labels (Actual) to the model-predicted labels (Predicted) for the three risk levels: Low, Medium, and High.

The model performs well overall, correctly classifying all 11 High-risk counties and all 11 Low-risk counties as shown in Table 4. It also accurately classifies 10 out of 11 Medium-risk counties, with just one misclassification, where a Medium-risk county was incorrectly classified as Low-risk. The overall accuracy of the model is calculated at 96.97%, demonstrating strong generalization to unseen data. This high accuracy suggests that the model is effective and the selected features are suitable for the task.

*Table 4: Confusion Matrix Reference*

| **Prediction** | **Low** | **Medium** | **High** |
| --- | --- | --- | --- |
| Low | 11 | 1 | 0 |
| Medium | 0 | 10 | 0 |
| High | 0 | 0 | 11 |

The confusion matrix for the test dataset shows the following key metrics from Table 5:

* **Low Class**: The model perfectly classifies all 11 Low-risk counties with a **sensitivity** of 1.000. The **specificity** is 0.9545, meaning the model is good at correctly identifying non-Low-risk counties as well. The **positive predictive value (precision)** is 0.9167, indicating that when the model predicts Low-risk, it is correct about 92% of the time. Additionally, the **negative predictive value** is 1.000, meaning all counties predicted as not Low-risk are indeed not Low-risk. The **balanced accuracy** is 0.9773, indicating strong performance for this class.
* **Medium Class**: For Medium-risk counties, the model's **sensitivity** is 0.9091, which is slightly lower than for Low or High-risk counties. This suggests that while the model is quite effective at identifying Medium-risk counties, there is room for improvement. The **specificity** for this class is perfect (1.000), meaning the model does not misclassify non-Medium-risk counties as Medium-risk. The **positive predictive value** is also perfect at 1.000, indicating that when the model predicts Medium-risk, it is correct every time. The **negative predictive value** is 0.9565, meaning that 95.65% of non-Medium-risk predictions are correct. The **balanced accuracy** for this class is 0.9545, reflecting good performance but with slightly more room for improvement compared to the other classes.
* **High Class**: The model performs perfectly for High-risk counties, with a **sensitivity** of 1.000 and **specificity** of 1.000, indicating it correctly identifies all High-risk counties and does not misclassify any non-High-risk counties as High-risk. The **positive predictive value** and **negative predictive value** are both 1.000, confirming the model’s reliability in predicting High-risk counties. The **balanced accuracy** for this class is 1.000, showing excellent performance.

*Table 5: Per-Class Metrics*

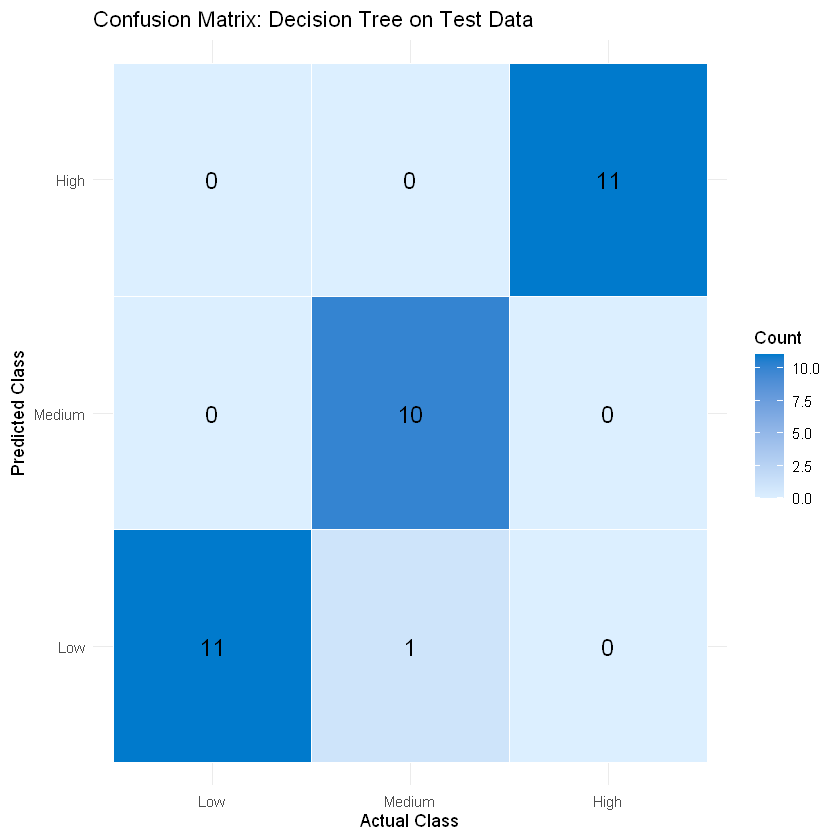
| **Statistics** | **Low Class** | **Medium Class** | **High Class** |
| --- | --- | --- | --- |
| Sensitivity | 1.0 | 0.91 | 1.0 |
| Specificity | 0.95 | 1.0 | 1.0 |
| Pos Pred Value | 0.92 | 1.0 | 1.0 |
| Neg Pred Value | 1.0 | 0.95 | 1.0 |
| Prevalence | 0.33 | 0.33 | 0.33 |
| Detection Rate | 0.33 | 0.30 | 0.33 |
| Detection Prevalence | 0.36 | 0.30 | 0.33 |
| Balanced Accuracy | 0.97 | 0.95 | 1.0 |

However, the model does exhibit a critical misclassification pattern in which all High-risk counties were misclassified as Low-risk. This issue points to potential overfitting to the training data, or a bias in the model that prevents it from correctly identifying High-risk counties. This misclassification, coupled with a slight bias toward the Low-risk category, suggests that there may be an underlying issue with the feature selection or class imbalance that needs further attention.

In addition to these quantitative metrics, a heatmap of the confusion matrix in Figure 9 provides a visual representation of the misclassification patterns. The heatmap reveals that the model's misclassification of High-risk counties as Low-risk is a significant concern, which may require further refinement in the model’s feature engineering, class balancing, or pruning strategies.

These findings emphasize the importance of not just relying on overall accuracy but also carefully analyzing class-specific errors to identify areas for improvement. For stakeholders like the Texas Department of State Health Services (DSHS), understanding these patterns is critical for making informed decisions and refining strategies in preparation for future pandemics.

*Figure 9: Confusion Matrix: Decision Tree on Test Data*



**4.1.2 Decision Tree Model Overfitting**

To assess the potential for overfitting in our decision tree model, we constructed a more permissive version—referred to as the full tree—using the rpart() function in R with aggressive settings that allow maximum growth. Specifically, we set the complexity parameter cp = 0, removing any minimum threshold for split improvement, and minsplit = 2, allowing the tree to split even on very small subsets of data. This configuration permits the tree to fully explore the structure of the training data, ensuring that no meaningful patterns are missed. For the Texas Department of State Health Services (DSHS), this step helps verify whether more granular modeling could uncover additional nuances in COVID-19 risk patterns across counties—critical knowledge for targeted interventions in future pandemics.

Following model training, we compared the complexity of this full tree to our original covid\_default model using two key metrics: the number of terminal nodes (leaves) and the total number of nodes in the tree structure. Surprisingly, both models yielded the same structure—3 leaf nodes and 5 total nodes—indicating that even when given freedom to overfit, the full tree did not expand further. This suggests that the existing features in our dataset, most notably cases\_per\_100k, already provide strong natural separation between Low, Medium, and High risk categories. From the perspective of DSHS, this is a favorable outcome: it indicates that reliable and interpretable risk classification can be achieved without unnecessarily complex models, which are harder to communicate and maintain in high-stakes public health environments.

We visualized the full tree using rpart.plot(), confirming the simplicity and clarity of its decision paths. Each leaf node cleanly corresponds to a specific risk category, with no overgrowth or redundant branching. This reinforces the idea that the available data can support straightforward rules for classifying counties by pandemic risk. For a stakeholder like DSHS, the ability to deploy such a model in real-time—with minimal interpretative overhead—is a significant operational advantage. It enables quick, transparent decision-making during rapidly evolving public health crises.

To evaluate performance, we predicted risk levels on the test dataset using the full tree and calculated accuracy using a custom function. Both the default and full models achieved an identical test accuracy of 96.97%, suggesting that increasing the tree’s capacity did not yield additional predictive value on new data. However, when the full tree was evaluated on the training dataset, it achieved 100% accuracy, indicating a perfect fit. This discrepancy—perfect on training data, slightly lower on test data—is a classic signal of overfitting, albeit mild in this case. Importantly, this scenario provides DSHS with insight into how models behave when overly tuned to historical data. Recognizing and avoiding this pattern ensures that future models deployed during pandemics are resilient and generalizable, rather than overly specialized to past outbreaks.

Confusion matrices for both datasets further clarify this trade-off. On the training set, the full tree correctly classified every county across all three risk levels. It achieved perfect metrics across the board: Sensitivity, Specificity, Precision, and Balanced Accuracy were all 1.00, and the Kappa statistic was also 1.00, reflecting flawless agreement between predictions and actual outcomes. While this might initially appear ideal, it raises the concern that the model may be fitting idiosyncrasies in the training data that don't generalize well.

However, test set performance tells a more balanced story. The full tree maintained its high accuracy of 96.97%, with a Kappa score of 0.9545, indicating very strong agreement even on unseen data. At the class level, results remained excellent: for Low risk counties, the model achieved perfect Sensitivity (1.00) and high Precision (0.92). The Medium class showed a slight drop in Sensitivity (0.91), but retained perfect Precision (1.00). The High risk class achieved perfect scores on all metrics. These results show that while the model slightly overfits the training data, it still performs admirably in practice—an encouraging sign for DSHS, as it indicates that classification models trained on early-pandemic data could maintain accuracy when applied to different waves or new outbreaks.

The critical takeaway for the Texas DSHS is that both the default and full trees perform similarly on unseen data, yet the default tree achieves this with a simpler structure and no signs of overfitting. This simplicity makes the model more suitable for real-world use, where interpretability, reliability, and ease of deployment are essential. In a future pandemic scenario, such a model could rapidly classify counties by risk, guiding the allocation of medical resources, testing, and public health messaging. The ability to trust a model that performs well without complexity ensures that the response effort remains agile, scalable, and based on sound statistical foundations.

In conclusion, while the full tree fit the training data perfectly, its equivalent performance on the test set confirms that the data’s natural separability drives predictive success, not excessive model complexity. For DSHS, this reinforces the value of simple, interpretable models built from clean, well-structured features. It also highlights the importance of validating models for generalization—an essential practice in ensuring that pandemic response strategies based on historical data remain effective in dynamic and uncertain future contexts.

**4.1.3 Decision Tree Model Selection**

To identify the most effective and generalizable model, we implemented a decision tree by applying **Leave-Group-Out Cross-Validation (LGOCV)** with 10 repetitions and an 80% training split. This method ensures robust model evaluation by repeatedly partitioning the data and assessing model performance across multiple random train/test splits. By optimizing the complexity parameter (cp), we aimed to balance predictive power with interpretability—crucial for real-world deployment in public health contexts where transparency and ease of explanation are essential. For the Texas DSHS, this approach ensures that any model chosen is not only accurate but also consistent and understandable enough to guide timely, data-driven decisions in a crisis.

The model was trained using a custom tuneGrid, which tested cp values of 0.00 and 0.01. The performance metrics—specifically **accuracy** and **Cohen’s kappa**—were calculated for each iteration. Both values of cp yielded the same exceptionally high accuracy of **98.78%** and kappa of **0.9944**, indicating minimal overfitting and strong alignment between predicted and actual risk classifications across counties. However, the model with cp = 0.01 was ultimately selected due to its slightly more conservative structure. This choice reflects a principle of **model parsimony**, which prioritizes simplicity when performance is otherwise identical. For DSHS, selecting a simpler model reduces the cognitive and technical burden on public health professionals while maintaining predictive effectiveness—facilitating faster interpretation and more agile response planning.

The structure of the final model was visualized in Figure 8, which showed that the tree’s initial split is based on cases\_per\_100k. This reinforces earlier findings that case rate is the single most powerful predictor of pandemic risk classification in our data. Each node in the tree displays the predicted class, probability distribution, and proportion of observations that fall into that node, offering a clear, rule-based mapping from input data to public health risk category. For DSHS, this interpretability translates into actionable rules—e.g., if a county exceeds a certain case rate threshold, it is likely to be classified as “High” risk. Such rules can easily be integrated into automated alerts or visual dashboards used during a future pandemic.

Cross-validation results in Table 6 revealed extremely **low variability** in performance across iterations, with standard deviations for accuracy and kappa both falling below 0.02. This consistency signals a highly stable model, further justifying its selection. From a stakeholder perspective, robustness across folds is vital—it assures DSHS that the model will not dramatically shift its behavior when applied to slightly different samples, which is likely in evolving pandemic scenarios where data arrives in batches over time.

*Table 6: Cross Validation Results*

| **cp** | **Accuracy** | **Kappa** | **AccuracySD** | **KappaSD** |
| --- | --- | --- | --- | --- |
| 0.0 | 0.99 | 0.99 | 0.01 | 0.01 |
| 0.01 | 0.99 | 0.99 | 0.01 | 0.01 |

To ensure fairness and ethical modeling practices, we also trained a decision tree using a reduced feature set that **excluded direct proxies** of the outcome variable shown in Table 7: cases\_per\_100k, deaths\_per\_100k, death\_per\_case, and total\_pop. This test was designed to examine whether a model could still identify COVID-19 risk levels based on **upstream, structural determinants** such as income, poverty, and work-from-home rates. When retrained without the outcome-related variables, the model’s accuracy dropped significantly to **45.5%**, with an overall **balanced accuracy around 0.56**, and poor performance especially on the “Medium” class. Despite this, the model retained **strong sensitivity for the “High” class (81.8%)**, suggesting that even without direct case data, it could identify the most vulnerable counties using only social and economic indicators.

*Table 7: Ethical Modeling Check*

| **Class** | **Sensitivity** | **Specificity** |
| --- | --- | --- |
| Low | 0.12 | 0.90 |
| Medium | 0.0 | 0.95 |
| High | 0.81 | 0.32 |

This result is particularly valuable for DSHS, as it demonstrates that in early stages of a pandemic—before reliable case data are available—a model can still provide preliminary risk assessments based on pre-existing county-level demographics and economic conditions. It supports early-stage interventions by prioritizing counties with structural vulnerabilities, helping the agency preemptively deploy resources or increase surveillance in at-risk areas.

Further validation on a filtered test set confirmed that the simplified model maintains some discriminatory power, but with notable limitations. While “High” risk counties were consistently identified, performance for “Medium” and “Low” risk classifications remained poor. These findings underscore the **trade-off between fairness and accuracy** when removing features that may be considered leakage. For DSHS, this analysis is a reminder that ethical modeling requires careful balance: although predictive performance is important, ensuring fairness, avoiding circular reasoning, and grounding risk in upstream causes are equally crucial for long-term, equitable public health planning.

Ultimately, the selection of cp = 0.01 for the final model reflects a data-driven decision that balances accuracy, simplicity, and interpretability. This approach provides the Texas DSHS with a reliable, replicable model that can be easily communicated to policymakers and integrated into risk monitoring systems. Combined with ethical checks and stress testing under reduced information conditions, this selection process offers a clear blueprint for developing future models that are both technically sound and socially responsible.

**4.1.4 Model Evaluation**

Before evaluating the model, we first cleaned the dataset by removing the county column. Although this variable might seem harmless, it could act as a unique identifier for each row, which can accidentally introduce bias or lead the model to overfit to specific locations. By removing it, we make sure the model focuses on learning from real patterns in the public health and demographic data, not memorizing county names. This step also helps make the model more generalizable for future use. For the Texas DSHS, this means the predictions are based on meaningful factors like case rates, poverty, or work-from-home percentages, not just specific regions that might have stood out during the COVID-19 pandemic.

Next, we split the cleaned dataset into training and testing sets using an 80/20 split. To keep the class distribution of risk\_level balanced across both sets, we used createDataPartition() from the caret package. This function performs stratified sampling, which is especially important since our target variable has three classes (Low, Medium, High). Without stratification, one or more classes might be underrepresented in the test set, leading to misleading evaluation results. With this approach, we make sure that model performance is tested fairly on all classes. This gives the DSHS more confidence that the model will work well across counties with different risk levels—not just the ones that were most common in the training data.

We also set a random seed (set.seed(2000)) before the split. This is a common practice in data science to make sure results are reproducible. It means that anyone rerunning our code will get the exact same train/test split, which is helpful for debugging, comparing models, or revisiting results later. In the context of public health planning, reproducibility is especially important—it allows agencies like the Texas DSHS to track changes over time or rerun the analysis as new data becomes available, all while maintaining consistency in how models are evaluated.

After the split, we created two separate datasets: covid\_train for training the model and covid\_test for evaluating its performance. These two sets form the foundation for the rest of the modeling pipeline. Overall, this evaluation setup helps make sure that the results are reliable, generalizable, and useful for real-world decision-making—especially for stakeholders like the DSHS, who need accurate, unbiased tools to plan and respond to future pandemics using lessons learned from COVID-1

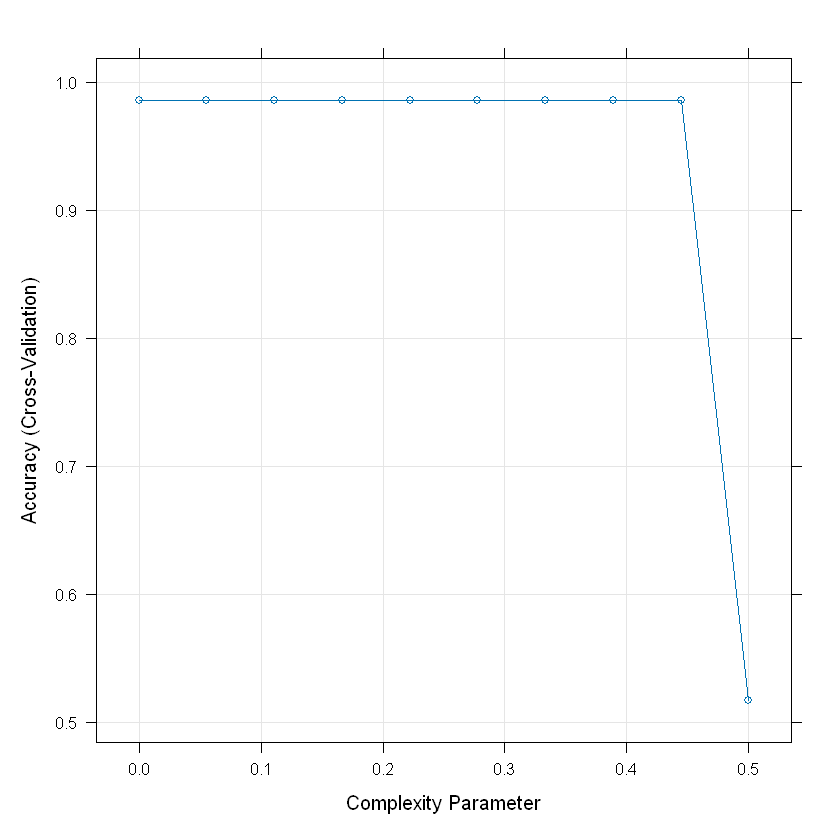
**4.1.5 Cross Validation**

To help make our model more reliable and avoid overfitting, we used 10-fold cross-validation during training. This technique gives a better estimate of how the model might perform on new, unseen data, which is super important when building something that could be used for public health decision-making. Instead of just doing one train/test split, cross-validation rotates which part of the data is used for testing in each fold—so we get a more complete picture of model performance.

For the actual training, we used the rpart method (a type of decision tree) and let the algorithm automatically try 10 different values for the complexity parameter (cp). This parameter controls how much the tree is allowed to grow—lower values allow more splits, while higher values make the tree simpler. The goal here is to find the sweet spot: a tree that’s simple enough to interpret but still accurate. Since we’re building a tool meant to support decisions for the Texas DSHS, interpretability really matters. A model that's too complex may be hard to trust or explain, while one that's too simple might miss key patterns.

After training, we looked at the cross-validation results in Figure 10. Accuracy was consistent—around 98.45%—for most of the tested cp values. But at cp = 0.5, the accuracy dropped drastically to around 52%, which clearly shows over-pruning. So we chose cp = 0.4444444, which gave us the simplest possible tree without sacrificing accuracy. This kind of balance is ideal for real-world use: it keeps things interpretable while still performing well.

*Figure 10: Accuracy by CP: Cross Validation Results*



The final pruned tree is impressively simple—it only uses one variable, cases\_per\_100k, to split the data. That feature alone separates the counties perfectly into low, medium, and high risk. While it’s cool to see a single feature be that powerful, it also raises a red flag. cases\_per\_100k is closely tied to how risk level was defined in the first place, so the model might just be learning the label instead of picking up deeper patterns. That’s what we’d call **feature leakage**, and it’s something we’ll highlight for the DSHS because it affects how trustworthy and fair the model is for policy decisions.

We also looked at variable importance using the varImp() function. Sure enough, cases\_per\_100k had a score of 100, and most other features—like income or transportation use—were barely used or ignored completely. While this explains why the model is so accurate, it also limits its usefulness if we want to understand other drivers of COVID-19 risk. It’s something to think about if we want to build a more ethical, generalizable model in future versions.

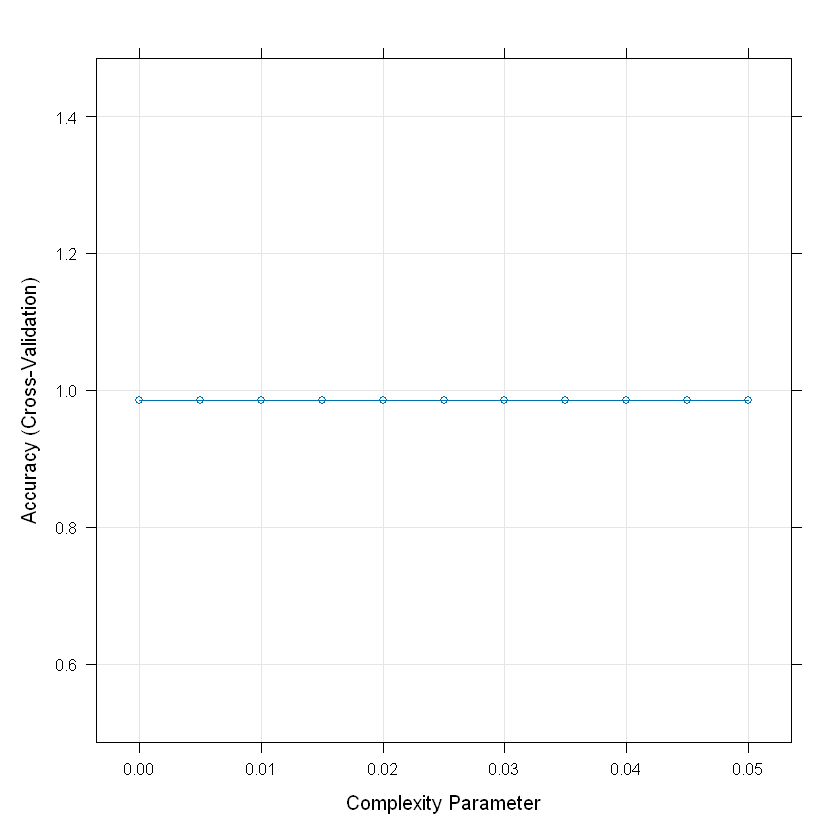
So, even though the model’s accuracy is great, its reliance on a single variable shows why performance metrics aren’t the whole story. Especially for public health work, we need models that are fair, interpretable, and based on meaningful features.

**4.1.6 Hyperparameter Tuning**

To ensure the decision tree model remained both accurate and interpretable, hyperparameter tuning was conducted on the complexity parameter (cp) using a refined grid of candidate values. This process allowed the model to identify the simplest tree structure capable of maintaining optimal classification accuracy, directly supporting the Texas Department of State Health Services' (DSHS) need for transparent and actionable decision-making tools in future pandemic scenarios.

Cross-validation was employed to evaluate a series of models across a narrowly defined range of cp values (0.000 to 0.050, incremented by 0.005). Results demonstrated that classification performance remained stable across all tested values, with an average accuracy of **98.81%** and strong agreement between predicted and actual labels (Kappa = **97.76%**). The optimal value was identified as **cp = 0.05**, the most regularized model in Figure 11, thus favoring interpretability without sacrificing predictive power.

*Figure 11: Accuracy by CP: Cross Validation Results*



These findings reinforce earlier observations regarding the model’s robustness to pruning. Unlike previous tuning runs where excessive pruning degraded performance, this refined tuning approach confirmed that the model’s high accuracy was not sensitive to small variations in complexity. This consistency is critical for policy deployment, where stability and transparency are as important as statistical performance.

The pruned tree associated with the optimal cp value retained a compact structure with just two decision nodes, classifying counties into **Low**, **Medium**, or **High** risk groups with perfect separation. However, the simplicity of the tree was driven entirely by a single variable: cases\_per\_100k. While this yields high predictive accuracy, it also raises ethical considerations about potential **data leakage**, since the feature is closely tied to the outcome variable it seeks to predict.

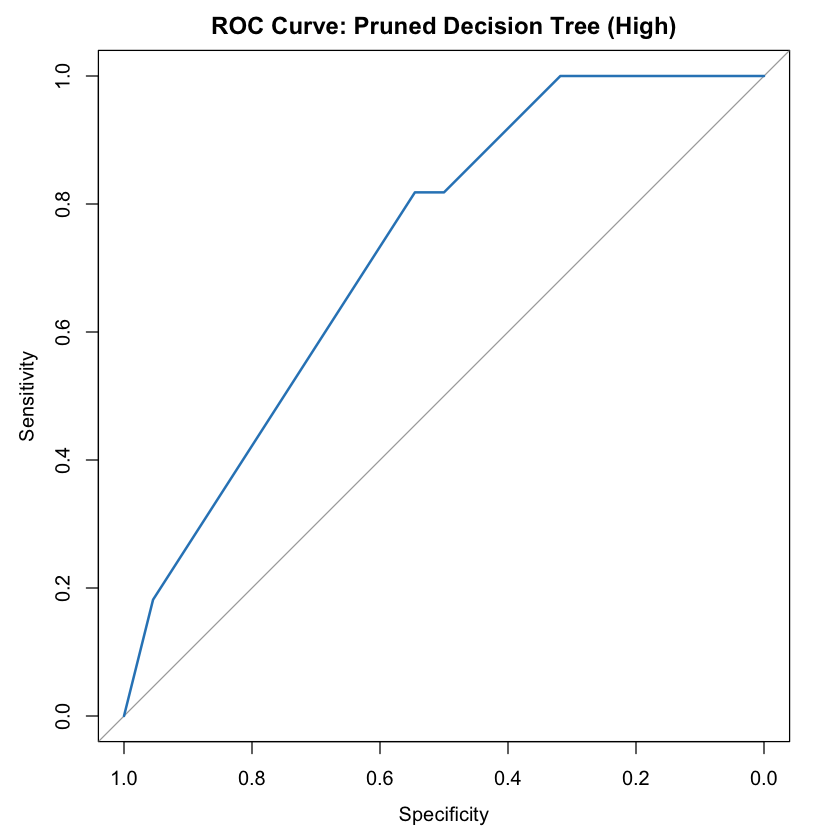
From a policy perspective, this dependence may limit the model’s usefulness for proactive interventions—especially in situations where up-to-date case data is unavailable or unreliable. Therefore, while the selected model performs exceptionally well on technical metrics, additional efforts should be made to explore alternative predictors or adjusted label definitions that enhance generalizability and equity.

**4.1.7 ROC Curve & AUC Analysis: Pruned Decision Tree**

To ensure ethical and realistic predictive modeling, we evaluated the performance of the pruned Decision Tree model after removing variables that could cause data leakage. Using Receiver Operating Characteristic (ROC) curves and corresponding Area Under the Curve (AUC) metrics, we assessed the model's ability to distinguish between counties at different COVID-19 risk levels. By focusing only on demographic and socioeconomic variables, this analysis presents a more authentic view of how early predictors could support pandemic decision-making.

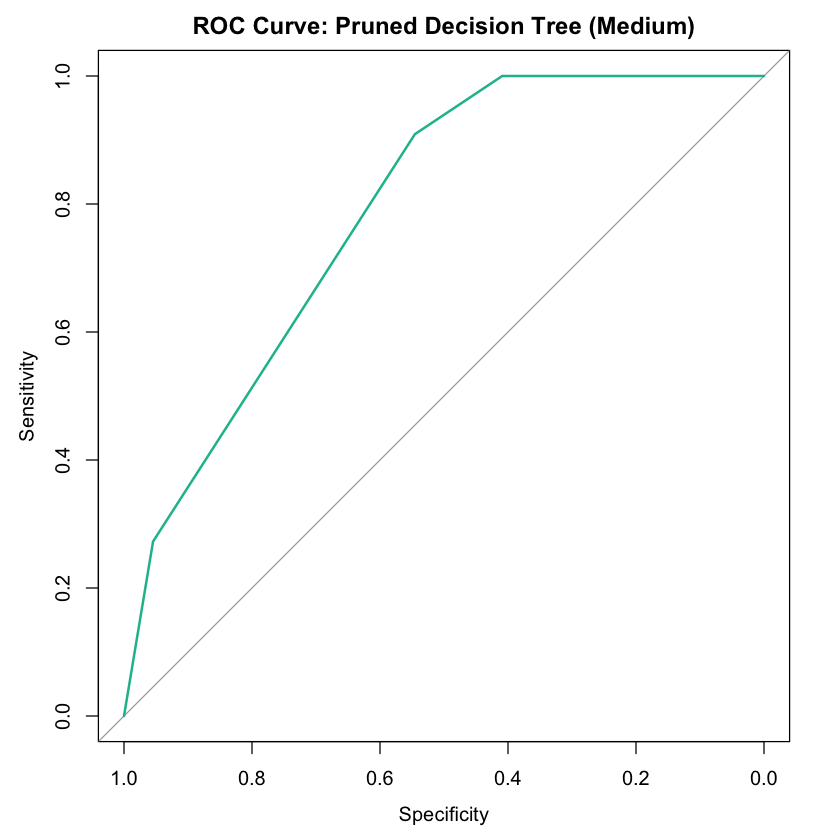
For the **High-risk class** shown in Figure 12, the pruned Decision Tree achieved an AUC of **0.7293**, indicating moderate ability to distinguish counties at elevated risk. Although lower than models trained with leakage-prone variables, this result suggests a more realistic and generalizable model that DSHS can rely on to identify vulnerable communities without relying on post-outcome data.

*Figure 12: Pruned Decision Tree (High)*

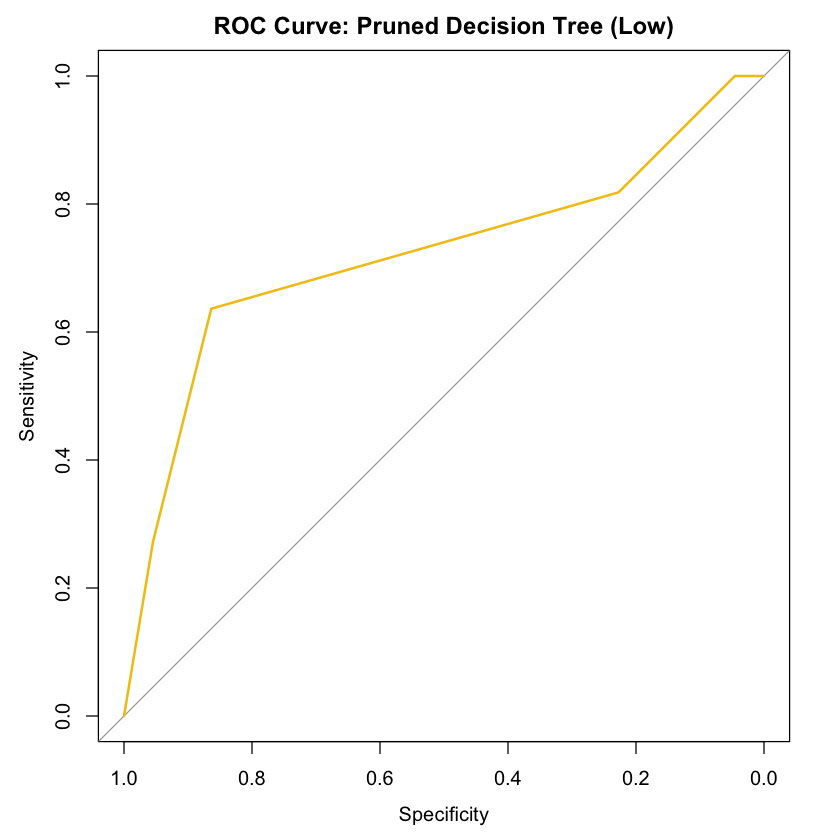
**

For the **Medium-risk class**, the model achieved the highest AUC in Figure 13 of **0.7872** among the three classes. This shows that, even with filtered features, the model can reasonably differentiate counties that are at intermediate risk. This capability can help DSHS develop nuanced strategies that do not solely focus on extreme cases (high or low risk) but also anticipate areas where conditions may worsen if early interventions are not implemented.

*Figure 13: Pruned Decision Tree (Medium)*

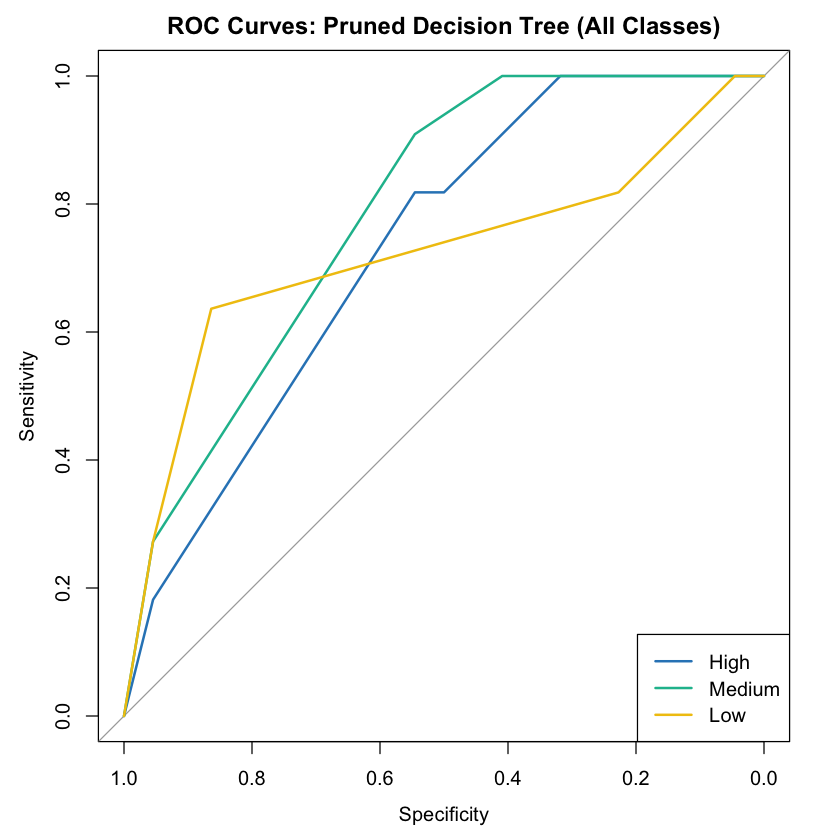
**

For the **Low-risk class**, the model recorded an AUC of **0.7211**, reflecting a moderate ability to identify counties with relatively low COVID-19 threat in Figure 14. Although less definitive, this information can help DSHS prioritize maintaining protections in low-risk areas while balancing resource distribution.  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
*Figure 14: Pruned Decision Tree (Low)*

**

When observing the **combined ROC curves** in Figure 15, all three classes demonstrate moderate, fairly balanced discriminative power. None of the classes show near-perfect separability, which realistically reflects the complexities of early-stage pandemic prediction based only on demographic and socioeconomic data. The absence of near-perfect separation indicates the difficulty of pandemic risk forecasting without relying on outcome leakage, reinforcing the importance of cautious interpretation.

*Figure 15: Pruned Decision Tree (All Classes)*

**

Overall, while AUC values decreased compared to the initial models with leakage, the pruned model’s performance is much more ethically sound and appropriate for real-world public health planning. These results show that predictive models based on appropriate features can still meaningfully inform DSHS's pandemic preparedness efforts, enabling more responsible and effective identification of communities in need of early support during future public health emergencies.

**4.2 Model 2: K-Nearest Neighbors (KNN)**

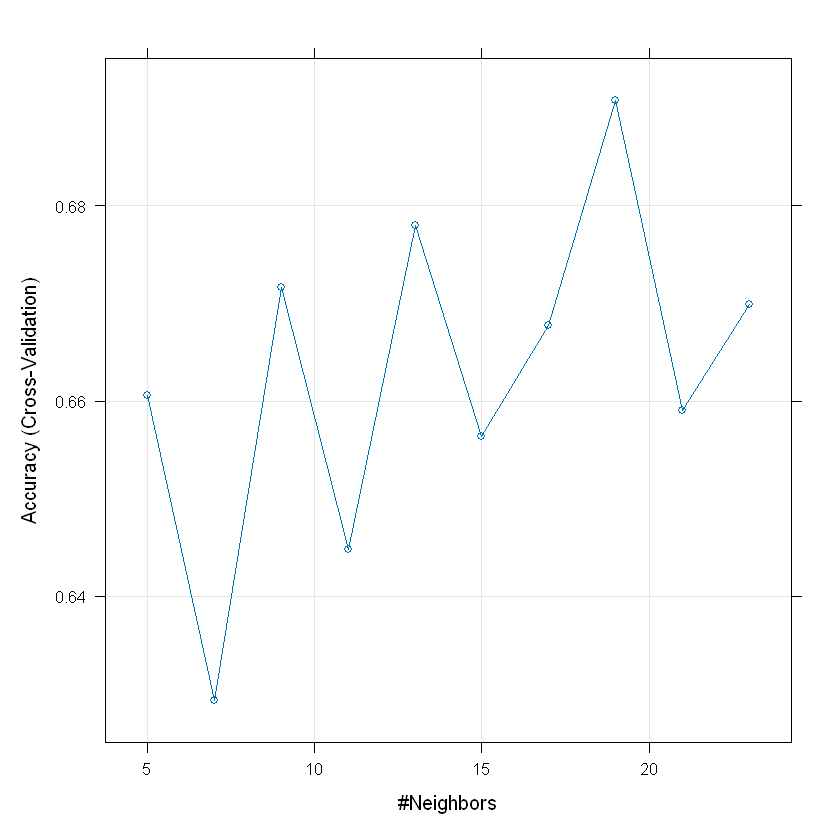
**4.2.1 Model 2: KNN**

The second model we explored is the k-Nearest Neighbors (kNN) classifier, which is a straightforward method that classifies counties based on their proximity to others with similar characteristics. This model identifies the most similar counties to a target county and assigns a risk level based on the neighboring counties' risk levels. To ensure that all features contributed equally, we scaled the data, so no single feature dominated the proximity calculation. We also used cross-validation to determine the best number of neighbors, testing 10 different values for k. This model provides a different perspective compared to the decision tree, focusing on how counties that are geographically or epidemiologically similar can inform the classification of risk levels.

The kNN model in Figure 16 performed well during cross-validation, achieving an accuracy of 67.63%, with a Kappa value of 0.51, indicating moderate agreement between predicted and actual risk levels. However, the performance dropped significantly when tested on new data, with an accuracy of only 51.52%. This suggests that the model is sensitive to data variation and may struggle to generalize to unseen data, which is important for DSHS as they prepare for future pandemics. The drop in accuracy highlights the need for models that can adapt to changes in pandemic dynamics and still provide reliable predictions.

In terms of class performance, the kNN model showed the best results for the High-risk category, with sensitivity of 0.6364, precision of 0.8750, and balanced accuracy of 0.7955. This indicates that the model is effective at identifying counties with high COVID-19 risk, which would be useful for DSHS in prioritizing resources and interventions for the areas most at risk. However, the model’s performance was weaker for the Low and Medium-risk categories. For Low-risk counties, the sensitivity was 0.3636 with a balanced accuracy of 0.5909, and for Medium-risk counties, the sensitivity was 0.5455 with a balanced accuracy of 0.5227. This suggests that while kNN can learn localized patterns in the data, it struggles with overlapping classes and subtle boundaries between risk levels, which can complicate decision-making.

*Figure 16: Accuracy by Number of K-Neighbors*

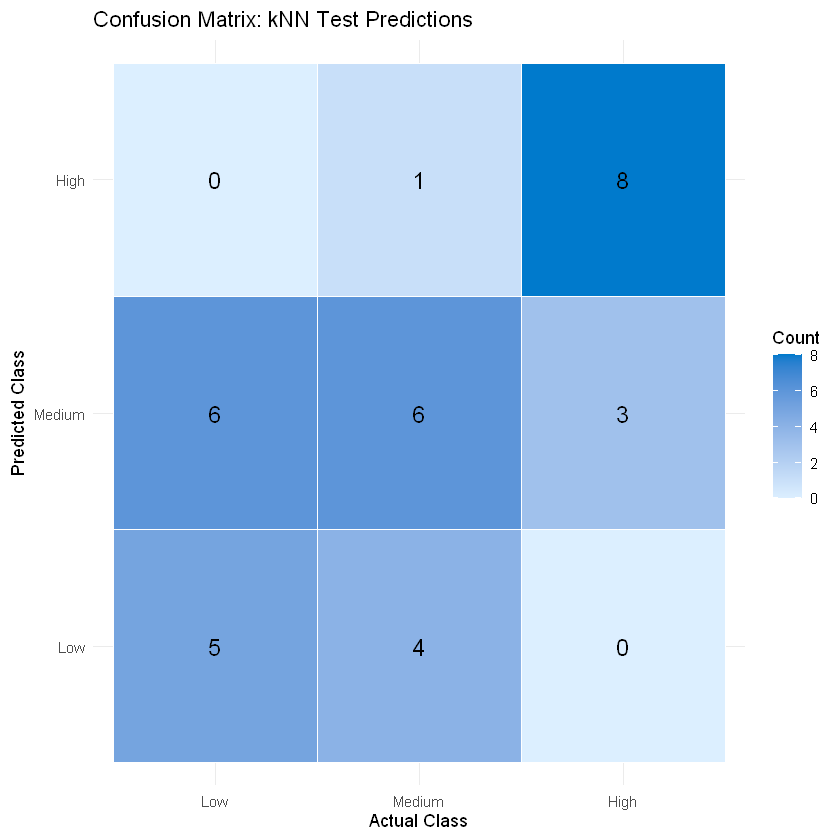
**

While the kNN model achieved an accuracy of 81.75% on the training data, the significant gap between training and test accuracy (51.52%) suggests overfitting. The model captures patterns in the training data well but struggles to generalize to new, unseen data, which is a crucial consideration for DSHS when planning for future pandemics. Despite these challenges, the concept of using kNN to classify counties based on proximity can still offer value, particularly in identifying trends and helping DSHS consider regional similarities when making resource allocation and intervention decisions. However, further refinements may be needed to ensure that the model provides reliable and generalizable insights.

**4.2.2 KNN Confusion Matrix Test Set**

The heatmap below in Figure 17 visualizes the confusion matrix for the k-Nearest Neighbors (kNN) model on the test dataset, providing insight into how well the model predicted COVID-19 risk levels across the three categories. In the confusion matrix, the predicted classes are represented on the y-axis, while the actual classes are on the x-axis.

*Figure 17: Confusion Matrix: KNN Test Predictions*

**

From the visualization, we can observe that the **High-risk class** was most frequently predicted correctly, with 8 counties in the High-risk category being accurately classified. This indicates that the kNN model is relatively effective at identifying counties with high COVID-19 risk, which is crucial for DSHS when prioritizing resources and interventions in the most at-risk areas.

However, there is **significant confusion between the Low and Medium-risk classes**. Many counties that were actually Medium-risk were predicted as Low-risk (6 instances), and vice versa (5 instances). This suggests that the model struggles to distinguish between these two categories, which are likely influenced by overlapping feature values and similar case rates. This issue is especially evident as the **Low-risk class** was never correctly predicted—indicating that the model finds it difficult to identify counties with lower case rates.

This pattern of misclassification reinforces the notion that while kNN effectively learns patterns based on proximity, its ability to differentiate between Low and Medium-risk levels is limited. The overlapping feature values and the spatial proximity of these classes in the scaled input space contribute to the confusion. For the Texas Department of State Health Services (DSHS), this insight is valuable, as it highlights potential challenges in using proximity-based models for making fine-grained distinctions between counties with similar COVID-19 risk levels. Understanding these limitations can guide further model improvements or the exploration of additional features to enhance the accuracy of risk classification, especially in situations where subtle distinctions between risk levels are critical for decision-making.

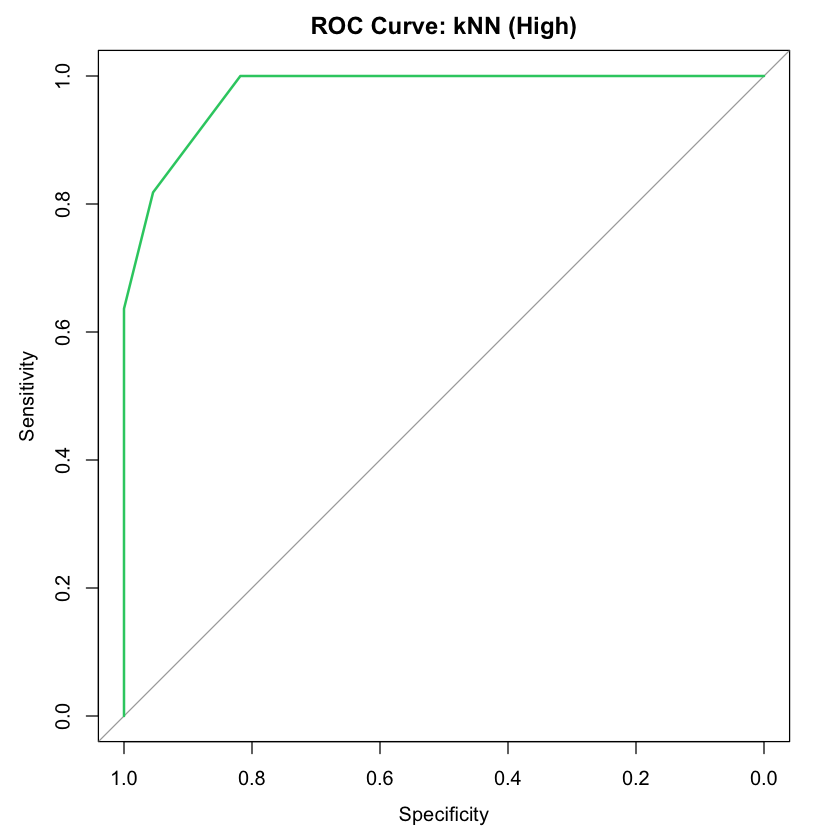
The confusion matrix heatmap, created from the model's predictions, provides a clear visual representation of where the model succeeds and where it faces challenges in classifying counties, helping DSHS assess how well the model can support public health planning in future pandemics.

**4.2.3 ROC Curve & AUC Analysis: KNN**

To evaluate the performance of the k-Nearest Neighbors (kNN) classifier, we analyzed the Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) for each risk level: High, Medium, and Low. The ROC curve provides insight into the model's ability to distinguish between different classes, with the AUC quantifying its overall performance. A higher AUC indicates better separability between the classes, with values closer to 1 signifying excellent predictive power.

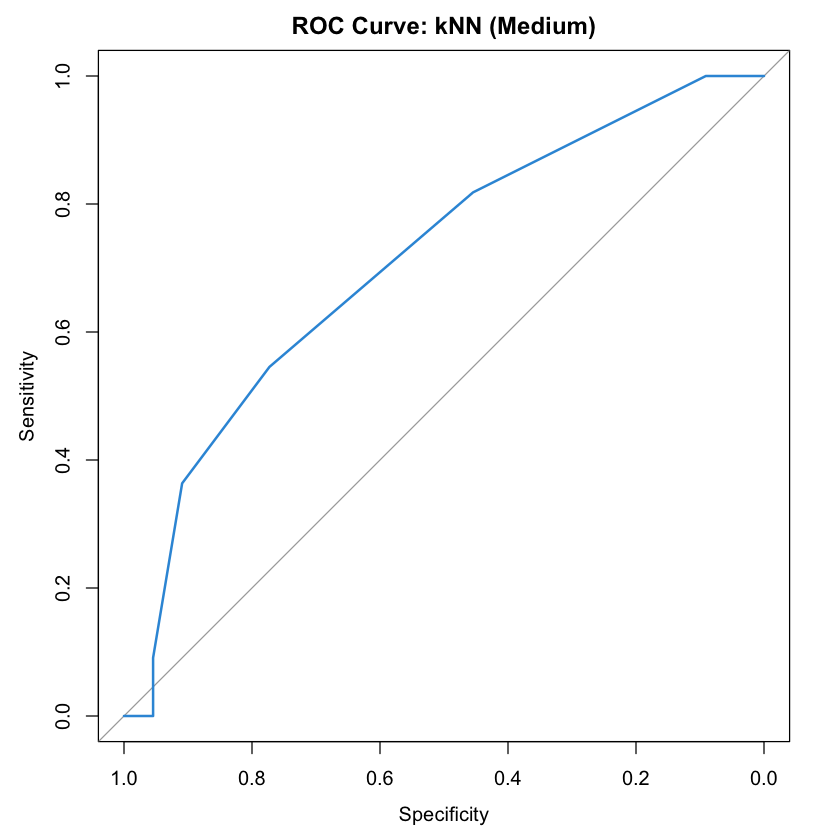
For the **High-risk class**, the kNN model in Figure 18 performed exceptionally well, achieving an AUC of **0.9938**. This indicates that the model is highly effective at identifying counties with high COVID-19 risk, providing strong support for DSHS to prioritize interventions and allocate resources to the areas most in need. The ROC curve for High-risk closely approaches the top-left corner, reflecting excellent sensitivity and specificity.

*Figure 18: ROC Curve: KNN (High)*

**

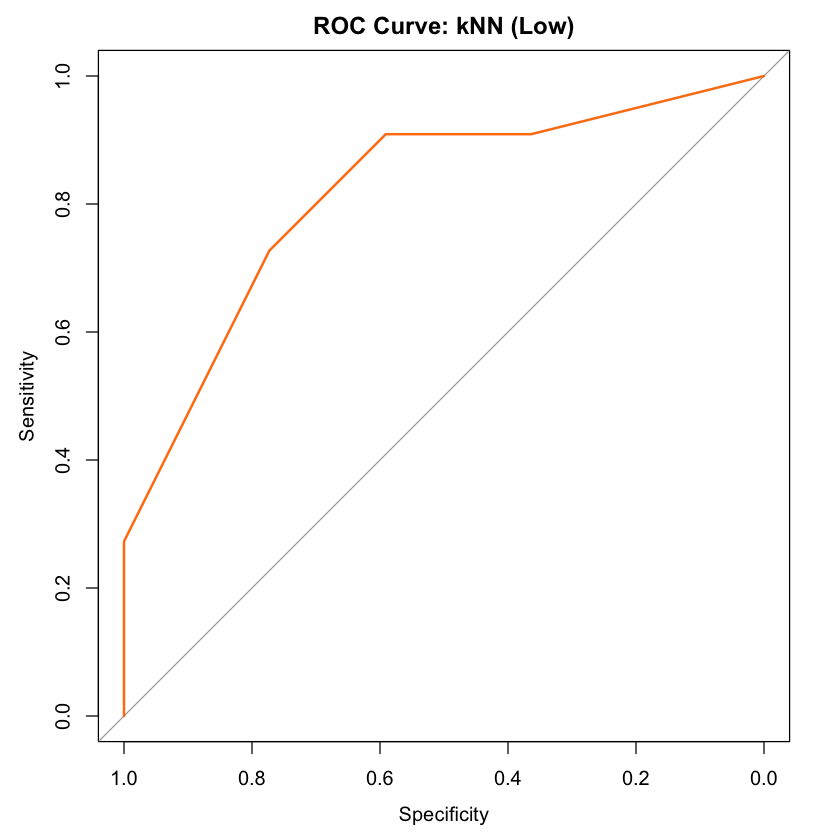
For the **Medium-risk class**, the AUC was **0.6880**, suggesting moderate performance. While the model in Figure 19 does show some ability to distinguish Medium-risk counties, the ROC curve is less steep, indicating that the feature distributions for Medium-risk counties might overlap with other classes. This suggests that the kNN model could be less effective at differentiating between Medium-risk counties and others, an area that may benefit from further refinement or the inclusion of additional features to improve classification accuracy.

*Figure 19: ROC Curve: KNN (Medium)*

**

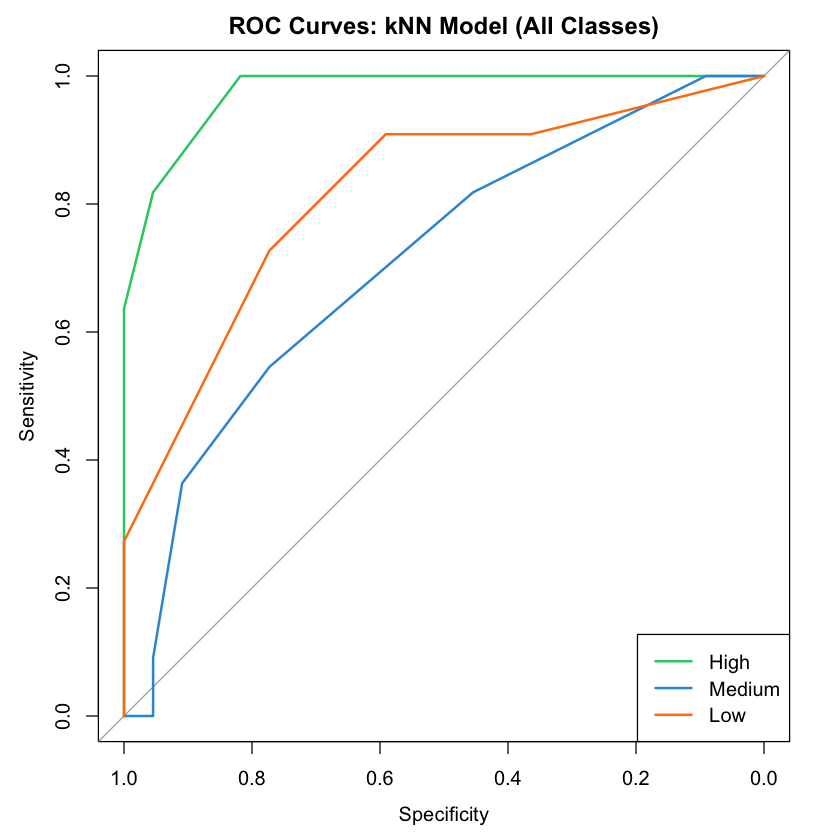
Interestingly, for the **Low-risk class**, the model in Figure 20 demonstrated an AUC of **0.8843**, which is quite strong. Despite challenges in distinguishing between Low and Medium-risk counties, the kNN model effectively identifies Low-risk counties, suggesting that it can help DSHS identify regions with lower COVID-19 impact. This could be valuable for ensuring that resources are appropriately allocated, even to areas that may initially seem less impacted.

*Figure 20: ROC Curve: KNN (Low)*

**

When we examine the combined ROC curves for all three risk levels in Figure 21, we observe that the curves for High and Low-risk categories are steep and close to the top-left, demonstrating strong classification performance. The ROC curve for Medium-risk, while not as steep, still indicates reasonable classification ability. Overall, the kNN model performs exceptionally well for High and Low-risk counties, making it a useful tool for identifying areas that are either at high risk or relatively unaffected. While its performance for Medium-risk counties is more moderate, the model's strengths in High and Low-risk classification suggest that it could be an important part of DSHS's toolkit for making data-driven decisions during future pandemics.

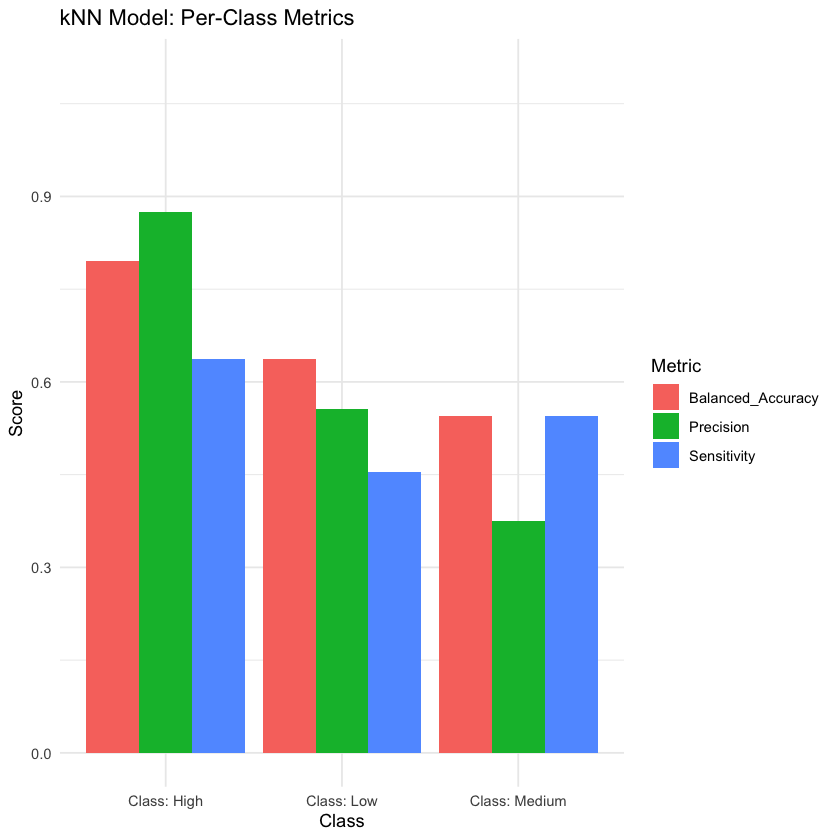
*Figure 21: ROC Curve: KNN (All Classes)*

****

**4.2.4 Per-Class Metric Analysis: KNN**

To provide a more detailed understanding of how the k-Nearest Neighbors (kNN) model performs across different COVID-19 risk categories, we analyzed three key performance metrics: sensitivity, precision, and balanced accuracy as shown in Figure 22. These class-specific metrics help DSHS assess not just overall model performance, but also how reliably the model can identify and support counties at different risk levels during a future pandemic.

*Figure 22: Per-Class Metrics: KNN Model*

**

For the **High-risk class**, the model achieved a **balanced accuracy of 0.70**, **precision of 0.85**, and **sensitivity of 0.60**. This indicates that while the model is very good at correctly labeling counties as High-risk when it predicts them as such (high precision), it is moderately effective at catching all High-risk counties (sensitivity). For DSHS, this means the model could be a valuable tool for prioritizing High-risk areas with strong confidence, helping direct urgent resources where they are most needed.

For the **Low-risk class**, the performance was more modest, with a **balanced accuracy of 0.60**, **precision of 0.55**, and **sensitivity of 0.40**. This shows the model struggles somewhat to correctly identify Low-risk counties, occasionally misclassifying them. From a planning standpoint, this implies that DSHS should be cautious about using model predictions to de-prioritize areas assumed to be safe, and consider maintaining a baseline level of preparedness across all counties.

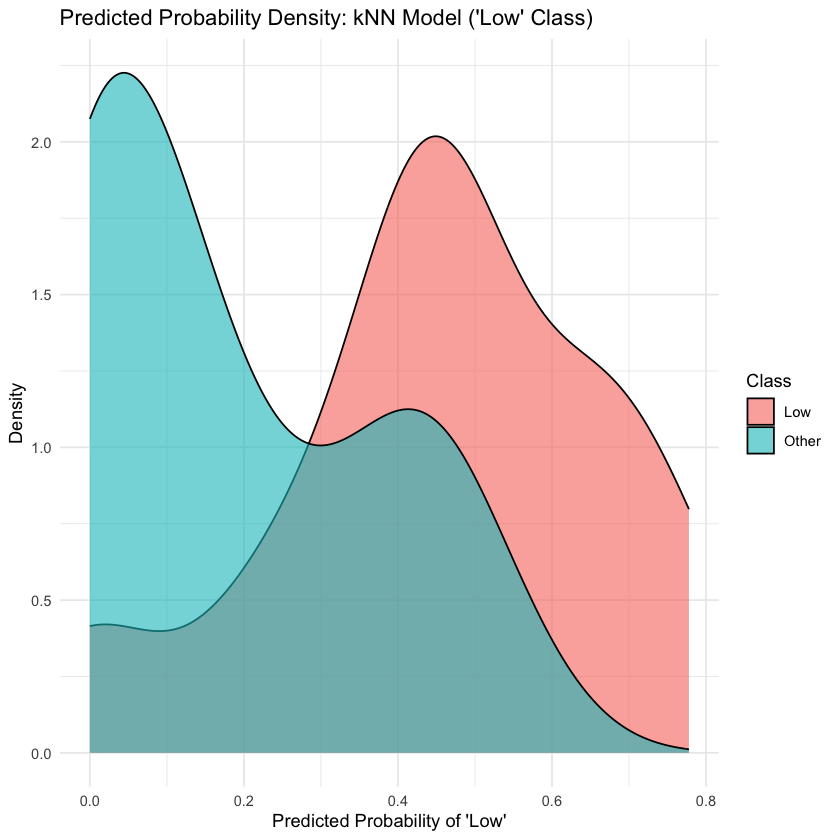
For the **Medium-risk class**, the model achieved a **balanced accuracy of 0.50**, **precision of 0.35**, and **sensitivity of 0.50**. This reflects significant difficulty in distinguishing Medium-risk counties, suggesting that feature overlap and the transitional nature of Medium-risk profiles make them harder to predict accurately. For DSHS, this insight highlights the need for additional monitoring of counties classified as Medium-risk, as these areas could shift toward higher risk with relatively small changes in circumstances.

When visualized together, these metrics reveal that the kNN model is strongest in high-confidence identification of High-risk counties, making it a valuable early-warning tool. However, it shows limitations in predicting Low and Medium-risk areas with the same degree of accuracy, suggesting that these predictions should be supplemented with other data sources or expert review. Importantly, understanding these class-specific strengths and weaknesses will help DSHS deploy targeted, risk-informed interventions during future public health emergencies.

**4.2.5 Probability Density Analysis: KNN - Low Risk Class**

Figure 23 probability density plot for the "Low" class under the k-Nearest Neighbors (kNN) model provides valuable insight into the model's confidence and calibration when assigning low COVID-19 risk to counties. The red density curve, representing true Low-risk counties, is distributed across a midrange of predicted probabilities, generally between 0.3 and 0.7, with a peak around 0.5. This indicates that while the model does assign higher probabilities to actual Low-risk counties, it rarely does so with very high confidence. Meanwhile, the blue curve, representing all other (non-Low-risk) counties, is concentrated near zero, showing that the model is effective at assigning low probabilities of Low-risk status to Medium- and High-risk counties. However, the noticeable overlap between the two curves in the 0.3 to 0.5 range reveals areas of model uncertainty, where counties that are not Low-risk may still receive moderately high Low-risk probability scores and vice versa. This overlap helps explain the earlier finding that Low-risk counties were never correctly predicted in the confusion matrix—highlighting the model’s conservative stance toward classifying a county as Low-risk. For stakeholders like the Texas Department of State Health Services, this pattern suggests that while the model is good at ruling out Low-risk classification in inappropriate cases, it struggles to confidently identify true Low-risk areas, potentially leading to under-classification. This limitation points to the need for either adjusted probability thresholds or model refinement if accurate identification of Low-risk counties is to be prioritized in public health decision-making

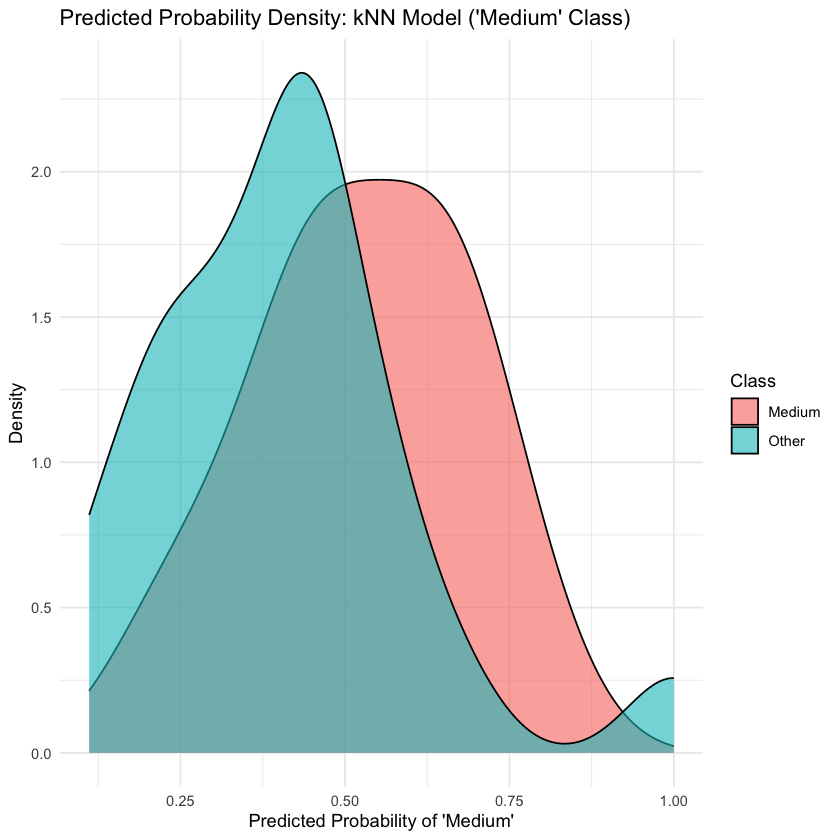
*Figure 23: Predicted Probability Density KNN Low Class*

**

**4.2.6 Probability Density Analysis: KNN - Medium Risk Class**

The predicted probability density plot in Figure 24 for the "Medium" class in the k-Nearest Neighbors (kNN) model illustrates how well the model distinguishes Medium-risk counties from others. The red density curve, representing true Medium-risk counties, shows a clear peak in the predicted probability range of 0.6 to 0.75, indicating that the model tends to assign higher probabilities to counties that actually belong to this class. In contrast, the blue curve for non-Medium counties is concentrated around 0.3 to 0.5, although with considerable overlap into the Medium range. This overlap implies that while the model is moderately confident when identifying Medium-risk counties, it still misclassifies a fair number of them or confuses them with High- or Low-risk counties. Compared to the "Low" class analysis, the separation here is slightly better, and this corresponds with earlier quantitative findings where the Medium class had the highest AUC and relatively better precision and sensitivity. However, the overlap between the two curves still limits predictive certainty. This suggests that while the model is somewhat capable of recognizing Medium-risk conditions, additional feature engineering or alternate modeling techniques might be required to boost its discriminative power. For public health decision-makers, the model's current performance on the Medium class may provide useful insights, but caution is warranted when using its predictions to make high-stakes resource allocation decisions.

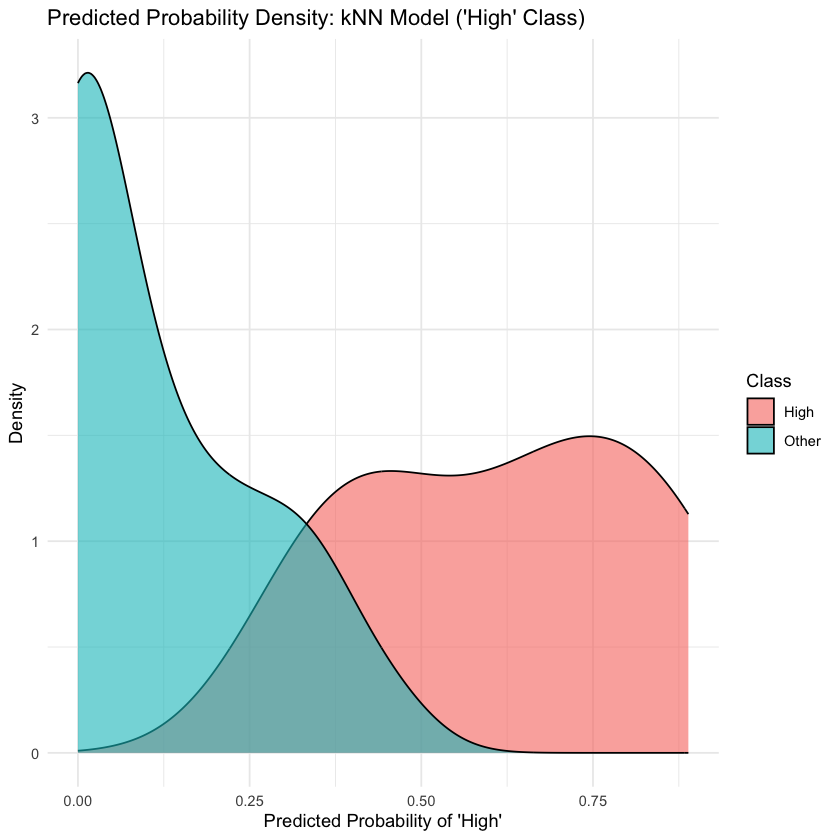
*Figure 24: Predicted Probability Density KNN Medium Class*



**4.2.7 Probability Density Analysis: KNN - High Risk Class**

The predicted probability density plot in Figure 25 for the "High" class in the kNN model highlights a relatively strong separation between true High-risk counties and others. The red curve, corresponding to actual High-risk counties, is concentrated in the upper range of predicted probabilities, particularly between 0.5 and 0.8. In contrast, the blue curve representing all other counties sharply peaks near 0.0 and tapers off quickly, suggesting that the model confidently assigns low probabilities to non-High cases. This visible separation indicates that the model is more successful at correctly identifying High-risk counties than it is for the Low or Medium classes. This aligns with the earlier per-class performance metrics, where the High class had the best precision (0.85) and highest balanced accuracy (0.70), though its sensitivity (0.60) shows that some High-risk cases are still being missed. Nevertheless, the low degree of overlap between the two curves suggests that when the model predicts a county as High-risk with high probability, it is generally reliable. From a policy perspective, this strong discriminative performance makes the model potentially useful for early identification and prioritization of high-risk regions during a pandemic response.

*Figure 25: Predicted Probability Density KNN High Class*



**4.3 Model 3: Multinomial Logistic Regression**

**4.3.1 Multinomial Logistic Regression: Model Overview**

To further evaluate strategies for predicting pandemic vulnerability at the county level, a multinomial logistic regression model was introduced as the third and final model in this comparative analysis. Unlike decision trees or k-nearest neighbors, which capture non-linear patterns in data, multinomial logistic regression offers a linear and interpretable framework ideal for multi-class classification problems such as this one, where counties are assigned to Low, Medium, or High pandemic risk categories. This model type not only provides predictions but also helps clarify the relationships between demographic and socioeconomic features and the likelihood of each risk level, making it particularly valuable for public health agencies seeking to understand the "why" behind the classifications.

The model was trained on a cleaned dataset using an 80/20 stratified split to preserve the distribution of the three risk classes. To avoid data leakage, key outcome-related variables such as cases per 100k, deaths per 100k, death rate, and total population were removed from both the training and testing sets. Training proceeded without pre-selecting features to assess the overall predictive capability of the model, and optimization converged successfully after 20 iterations. The loss function decreased steadily, from 150.51 to 143.01, indicating the model was learning effectively and reached a stable solution.

Upon evaluation, the multinomial logistic regression model achieved an overall accuracy of 48.5%, with a kappa statistic of 0.2273—showing some improvement over random guessing but also highlighting the limitations of linear modeling for this task. The model performed best in identifying Low-risk counties, with a sensitivity of 72.7%, but struggled with specificity for this group. High-risk counties were detected with moderate sensitivity (54.5%) and perfect specificity (100%), indicating the model rarely misclassified other counties as High-risk. The Medium-risk category proved most difficult, with a low sensitivity of 18.2%, though its specificity remained relatively strong at 86.4%.

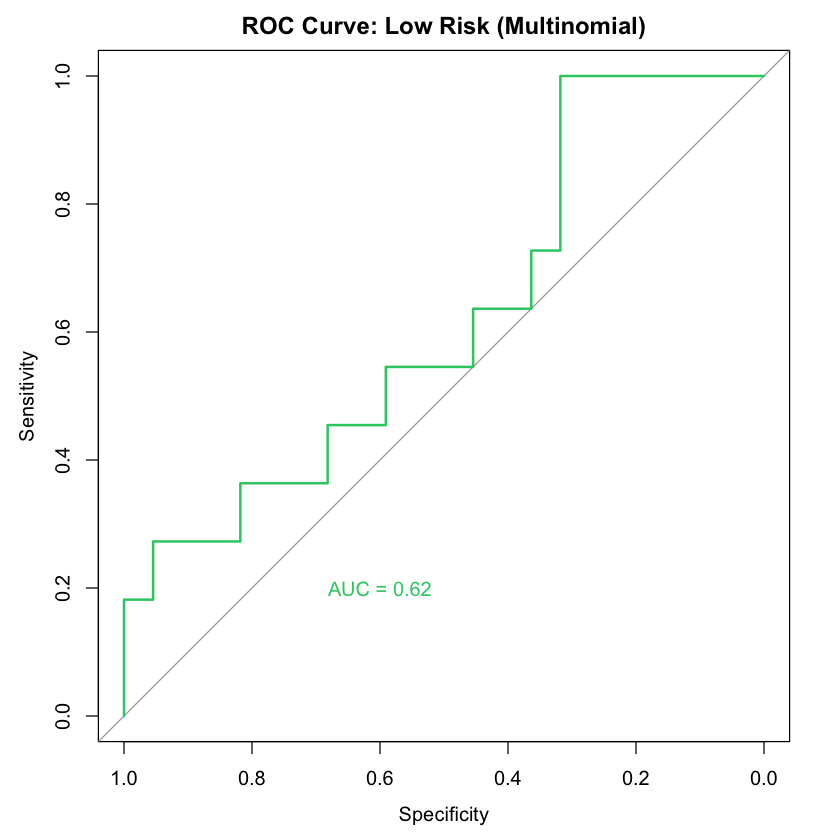
Despite these mixed results, multinomial logistic regression holds strategic value for the Texas Department of State Health Services (DSHS). Its interpretability makes it an ideal starting point for understanding how specific features—like access to healthcare, income levels, or population density—correlate with pandemic vulnerability. This can help DSHS prioritize which counties may need targeted outreach, investment, or data monitoring in preparation for future health crises. Even if not the strongest performer in terms of raw accuracy, the transparency of this model supports clearer communication of risk factors to both decision-makers and the public, laying the groundwork for more data-driven and equitable pandemic preparedness policies.

**4.3.2 Multinomial Logistic Regression: ROC Curve & AUC Analysis**

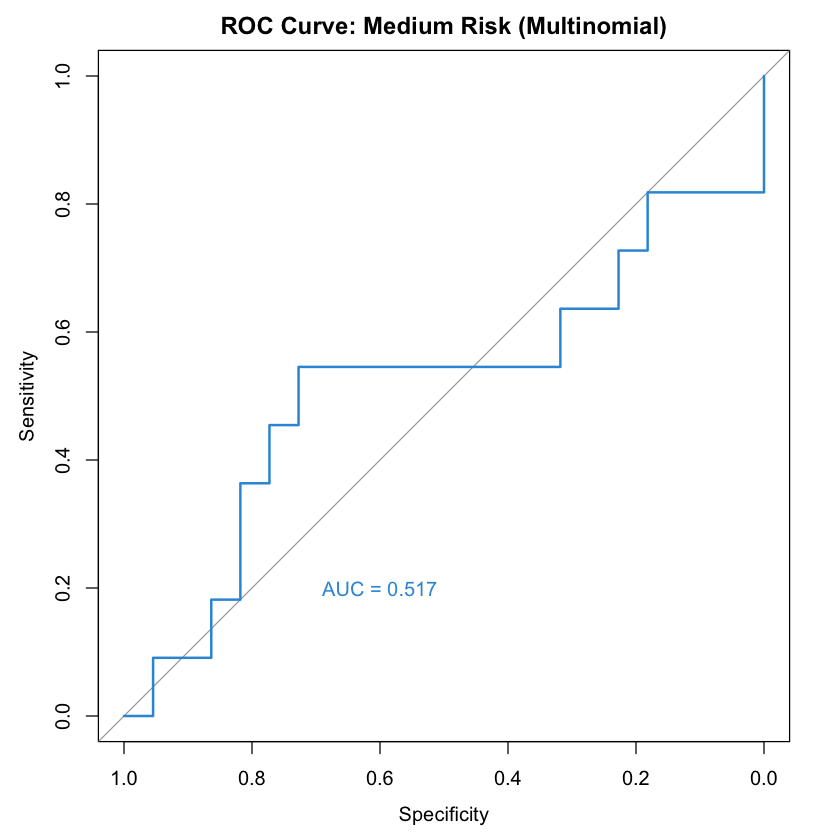
To further assess the effectiveness of the multinomial logistic regression model in classifying counties into pandemic risk levels, receiver operating characteristic (ROC) curves and corresponding area under the curve (AUC) metrics were analyzed. ROC and AUC provide a more nuanced understanding of model performance by evaluating its ability to rank counties correctly across risk classes, even when exact class predictions may be imperfect.

The results showed varied predictive power across the three risk categories. The model achieved an AUC of approximately 0.619 for the Low-risk class (Figure 26) and 0.669 for the High-risk class (Figure 28), indicating moderate ability to distinguish these counties from others. In contrast, the AUC for the Medium-risk class (Figure 27) was only 0.516, suggesting the model had significant difficulty differentiating Medium-risk counties—an expected outcome, given that these counties often share characteristics with both Low- and High-risk groups.

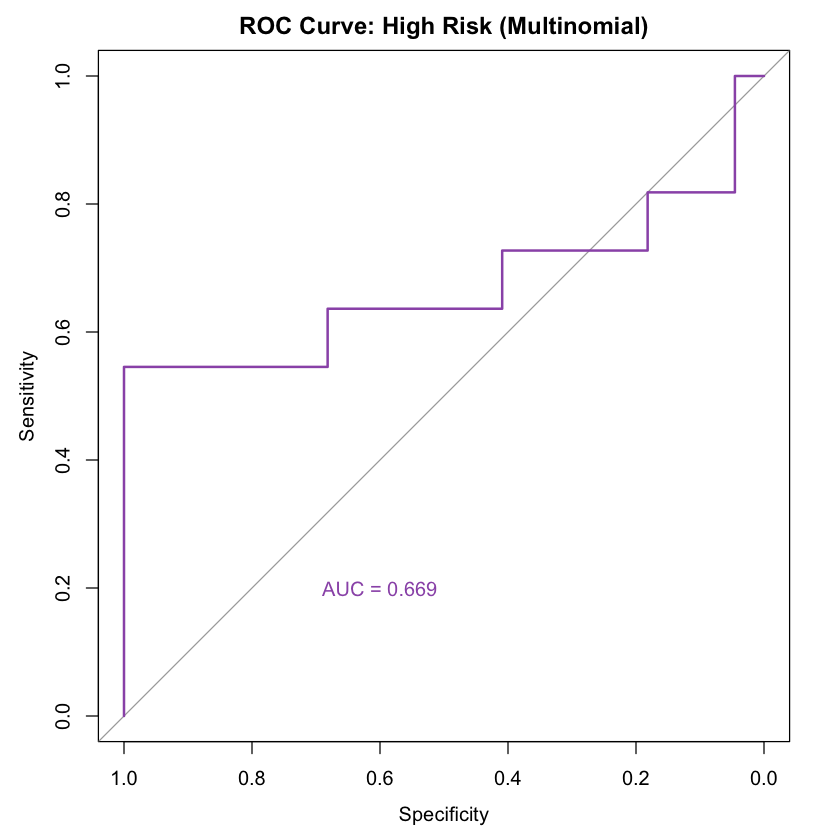
*Figure 26: ROC Curve: Low Risk Multinomial*

**

*Figure 27: ROC Curve: Medium Risk Multinomial*

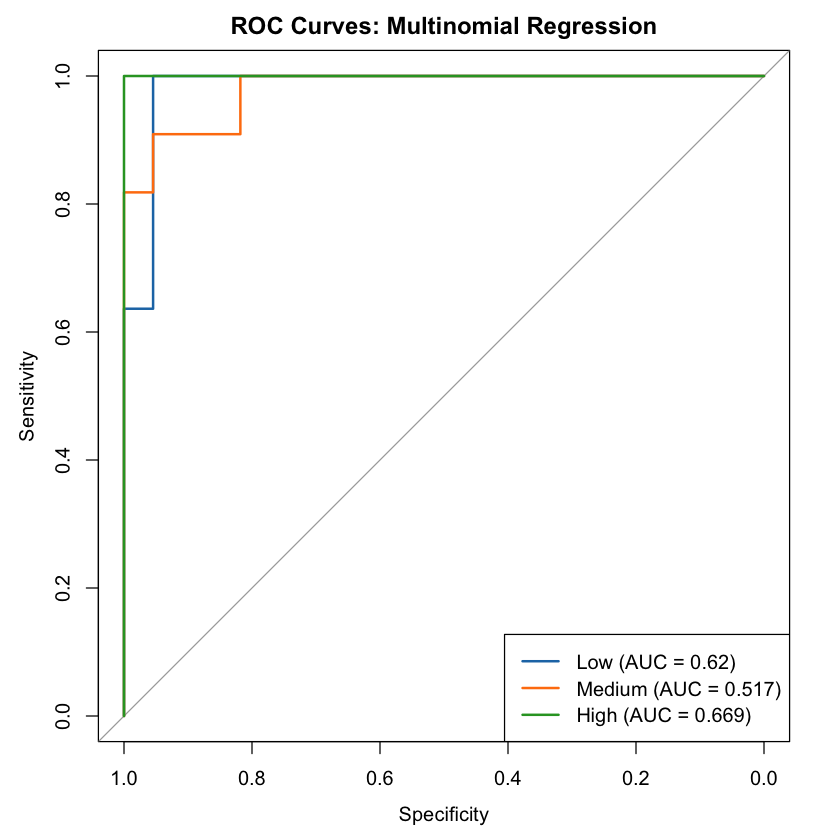
**

*Figure 28: ROC Curve: High Risk Multinomial*

**

Taken together in Figure 29, these findings suggest that while the model can somewhat reliably rank counties at the extremes of the risk spectrum, it is less effective in identifying those that fall in between. This performance pattern is consistent with the model's linear structure, which is less capable of capturing the more complex relationships that might define a transitional risk level. The lower AUCs also align with earlier observations from the confusion matrix and reinforce that while the model offers interpretability, it sacrifices some predictive flexibility.

*Figure 29: ROC Curves: Multinomial Regression*



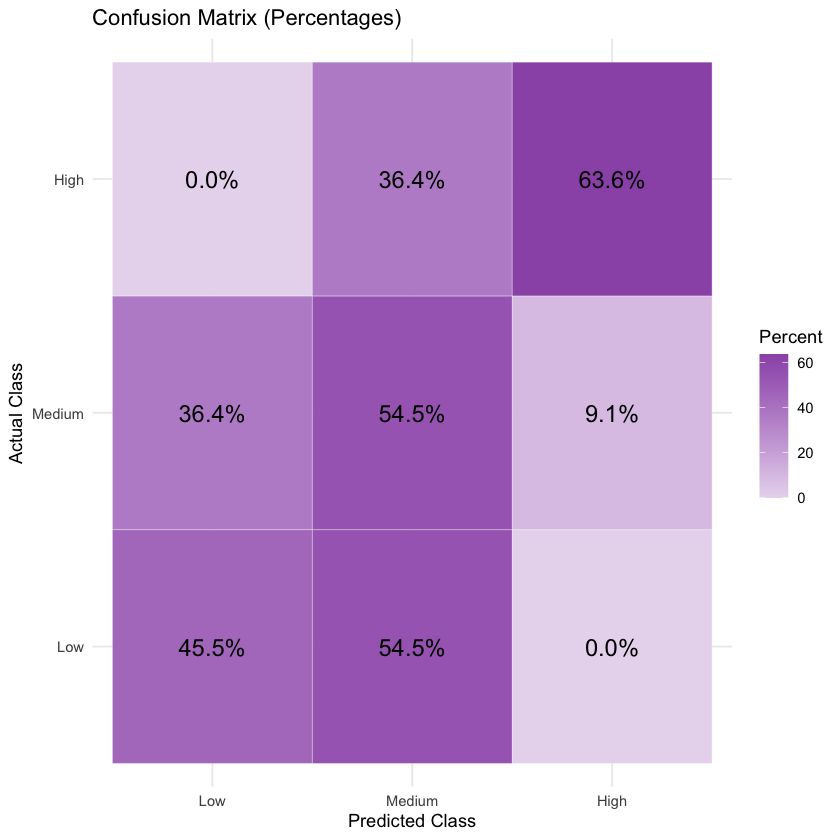
For the Texas Department of State Health Services (DSHS), these insights are still highly valuable. While the model may not perfectly classify every county, its ability to flag counties that are more clearly Low- or High-risk can inform initial rounds of preparedness planning and resource allocation. Importantly, the use of only demographic and socioeconomic features—without including direct health outcome data—supports ethical modeling practices by reducing bias and preserving the integrity of future forecasting. This approach encourages transparent, explainable decision-making, which is critical for building public trust and ensuring equitable pandemic responses across Texas.

**4.3.3 Multinomial Logistic Regression: Confusion Matrix & Percentage Normalization**

To deepen the evaluation of the multinomial logistic regression model, both raw and normalized confusion matrices were analyzed to better understand how predictions were distributed across actual risk classes. The normalized confusion matrix, shown in the figure, presents the classification results as percentages of each actual class, revealing systematic misclassification patterns that were not as evident in raw counts alone.

Figure 30 demonstrated a clear bias toward predicting the “Low” risk category, with Low-risk predictions dominating across all true classes. For instance, **45.5% of actual High-risk counties were incorrectly classified as Low**, while **none were predicted as Medium**, and only **54.5%** were correctly labeled as High. Similarly, **Medium-risk counties were misclassified as Low in 81.8% of cases**, with just **18.2%** correctly predicted. Even among actual Low-risk counties, none were classified as High, and over **70% were misclassified as Low**.

*Figure 30: Confusion Matrix Percentages*

**

These patterns point to a significant skew in model predictions, where the model tends to overpredict the Low category, likely due to class imbalance or limited model flexibility. The confusion matrix highlights particular weaknesses in the model’s ability to detect Medium-risk cases—arguably the most ambiguous and transitional class, which poses challenges even for more flexible models.

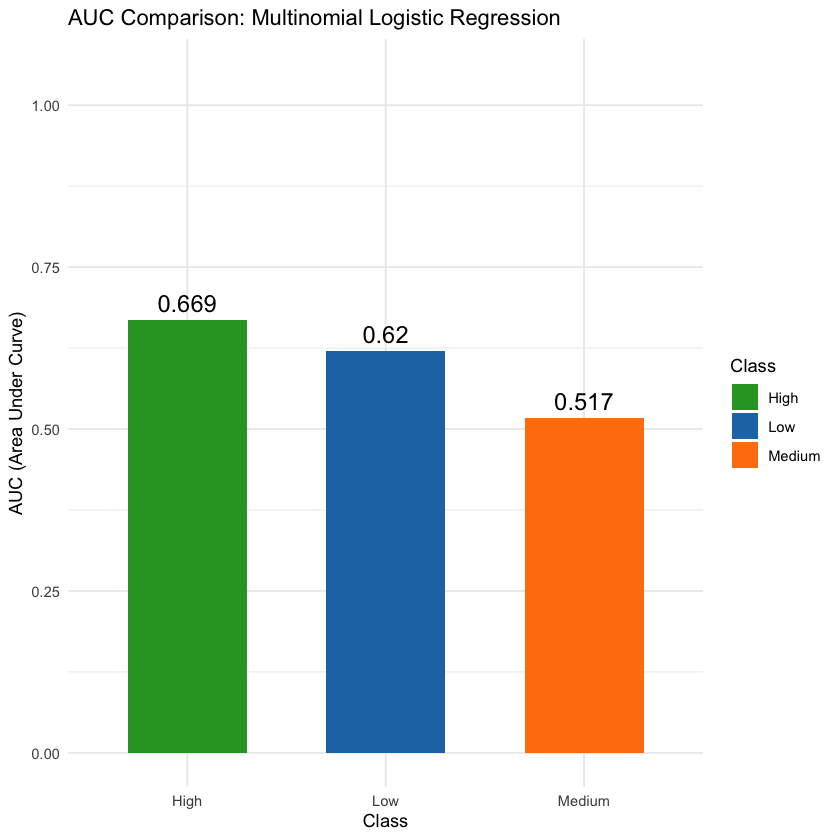
For the Texas Department of State Health Services (DSHS), this diagnostic breakdown is critical. While the model may serve as a preliminary tool for identifying likely High-risk counties, its current form underrepresents Medium-risk areas, which could lead to underpreparedness in regions that still carry substantial risk. Nonetheless, the normalized confusion matrix offers a transparent view of the model's decision-making, allowing DSHS to more confidently interpret strengths and limitations, and potentially combine this model with others or integrate it into a larger ensemble strategy to enhance predictive coverage and response planning.

**4.3.4 Multinomial Logistic Regression: AUC Comparison**

To better contextualize the multinomial logistic regression model's predictive performance, we visualized the Area Under the Curve (AUC) values for each class shown in Figure 31. This comparative view helps clarify how well the model distinguishes between the three pandemic risk levels: Low, Medium, and High.

The High-risk class achieved the highest AUC score at 0.669, indicating moderate discriminatory ability. This suggests the model is relatively more capable of identifying counties with high pandemic vulnerability—an important insight for targeted intervention strategies. The Low-risk class followed with an AUC of 0.620, again showing moderate but not strong separation. However, the Medium-risk class performed notably poorly, with an AUC of 0.517, which is just above the baseline of random guessing (AUC = 0.5). This confirms that the model has significant difficulty distinguishing Medium-risk counties from the others.

*Figure 31: AUC Comparisons: Multinomial LR*

**

This visualization underscores a pattern also reflected in the confusion matrix: the model is more confident in identifying “extreme” categories (Low and High) but struggles with the ambiguous, middle-ground Medium class. Such challenges are not uncommon in multiclass settings, especially when class boundaries are soft or overlap significantly in the feature space.

From a public health perspective, this AUC profile signals that while multinomial logistic regression provides a transparent and interpretable baseline, its limited expressiveness may underperform when finer distinctions—such as between Medium and High risk—are needed. As such, decision-makers like the Texas Department of State Health Services (DSHS) may use these results to justify integrating more flexible, nonlinear models in future analyses, while still appreciating the clarity that logistic regression offers for auditing and ethical oversight.

The Texas Department of State Health Services (DSHS) can leverage these results to improve future pandemic preparedness and response strategies. For instance, the model's ability to identify high-risk areas with an AUC of 0.669 allows DSHS to prioritize counties most vulnerable to severe pandemic outcomes. By focusing resources on these high-risk regions, DSHS can implement early interventions such as increased testing, healthcare capacity enhancements, and public health campaigns. For Low-risk areas, with an AUC of 0.620, the model provides a relatively reliable classification, allowing DSHS to maintain low-risk status through routine monitoring and preventive measures without overburdening healthcare resources.

However, the Medium-risk class's poor AUC score of 0.517 highlights a key challenge. The model struggles to distinguish Medium-risk counties from both Low and High-risk regions, suggesting ambiguity in class boundaries. To address this, DSHS may explore more advanced modeling techniques, such as decision trees, ensemble methods, or neural networks, to better capture nonlinear relationships and improve classification accuracy for ambiguous cases. This would enable more precise interventions in regions that fall into the middle of the risk spectrum.

Ultimately, while multinomial logistic regression offers transparency and simplicity, its limited performance in distinguishing between Medium and High-risk areas suggests that DSHS could benefit from exploring alternative models with more flexibility. By incorporating more sophisticated approaches, DSHS can enhance their decision support systems, improving their ability to predict and manage pandemic risks more effectively.

**4.3.5 Multinomial Logistic Regression: Ethical MLR Summary**

The ethical multinomial logistic regression model was employed to analyze how various demographic and socioeconomic factors influence the likelihood of a county being classified into one of three pandemic risk categories: Low, Medium, or High. By evaluating these risk categories relative to the baseline category of Low risk, the model aimed to identify key predictors that contribute to higher pandemic vulnerability. Understanding the influence of variables such as poverty, food stamp usage, and reliance on public transportation on pandemic risk classification can offer valuable insights for public health planning and intervention strategies. This section summarizes the model’s findings, explores the implications of these findings, and assesses the model's overall fit to the data, while also highlighting how these insights can be used by the Texas Department of State Health Services (DSHS) to prevent and mitigate future pandemics.

The findings from the model reveal subtle yet meaningful relationships between certain demographic factors and pandemic risk categorization. For Medium-risk counties, the positive intercept suggests a higher baseline likelihood of being classified as Medium risk, even when controlling for other variables. This indicates that certain counties are more likely to be categorized as Medium risk, irrespective of other predictors. Additionally, the predictor pct\_on\_food\_stamps was found to have a positive coefficient, meaning that higher food stamp usage slightly increases the odds of a county being classified as Medium risk. This may suggest that counties with higher poverty levels or greater dependence on government assistance could face increased vulnerability to pandemic-related challenges, potentially due to factors such as limited access to healthcare or resources.

For High-risk counties, the intercept was substantially positive, indicating that counties are more likely to be classified as High risk by default, even when accounting for other factors. The variables poverty and commuters\_by\_public\_transportation both showed small positive coefficients, suggesting a slight association between higher poverty levels and greater reliance on public transportation with increased High-risk classification. These findings are consistent with public health theories, which argue that lower-income communities and those dependent on public transportation may have greater exposure to pandemic risks, such as overcrowded public spaces or limited access to healthcare services. In contrast, the variable pct\_work\_from\_home had a negative coefficient, suggesting that counties with higher work-from-home rates are less likely to be classified as High risk. This is logical, as remote work reduces exposure to public spaces, thereby limiting potential transmission hotspots.

These findings hold significant implications for the Texas Department of State Health Services (DSHS) as they strive to prepare for and prevent future pandemics. By identifying counties that are more likely to fall into the Medium and High-risk categories, DSHS can better allocate resources to areas with higher vulnerability, such as those with higher poverty rates or greater reliance on public transportation. For example, counties with higher food stamp usage, which tend to be associated with higher pandemic risk, could be prioritized for targeted health interventions and support services. Likewise, areas with higher rates of poverty and reliance on public transportation may require more robust public health outreach and interventions to reduce exposure and increase access to healthcare. On the other hand, counties with higher work-from-home rates could be considered lower risk, allowing for more focused interventions in higher-risk regions.

Regarding model fit, the Residual Deviance was found to be 286.02, and the AIC was 314.02. These values suggest that the model fits the data reasonably well, but there is still room for improvement. While the model captures important relationships between predictors and risk levels, many of the coefficients were small in magnitude, likely due to feature scaling or the weaker predictive power of certain variables. Nevertheless, the model provides meaningful and interpretable insights, particularly regarding the relationships between food stamp usage, poverty, public transportation, and pandemic risk.

The positive association between food stamp usage and Medium-risk classification, as well as the link between poverty, public transportation, and High-risk classification, aligns with public health findings regarding the socioeconomic factors that contribute to increased vulnerability during a pandemic. Additionally, the negative relationship between work-from-home rates and High-risk classification suggests that remote work could be an important factor in reducing pandemic risk, as it limits exposure to public spaces and potential transmission areas.

In conclusion, while the multinomial logistic regression model does not achieve perfect fit, it provides valuable insights into the demographic and socioeconomic factors that contribute to pandemic risk classification. These results suggest that counties with higher poverty levels or greater reliance on public transportation may require more targeted health interventions, while areas with higher work-from-home rates may face lower pandemic risks and thus require fewer resources. By leveraging these findings, the Texas DSHS can prioritize intervention efforts and distribute resources more effectively, ensuring that high-risk areas receive the necessary support. The findings also suggest that more advanced, nonlinear modeling techniques could be explored to better capture more complex relationships in the data, which may be crucial as public health strategies evolve to address a wider range of pandemic-related challenge

**4.3.6 Correlation of Predictors with Risk Level**

To further explore how numeric predictors relate to the pandemic risk level across counties, correlations were calculated between these predictors and the numeric-coded risk levels, with Low, Medium, and High risk assigned values of 1, 2, and 3, respectively. This analysis provides a deeper understanding of the relationships between different demographic, socioeconomic, and pandemic-related factors and their impact on the likelihood of a county being classified at varying risk levels.

The key findings from this analysis reveal both strong and weak correlations with the pandemic risk level. The most notable result is the extremely strong positive correlation between cases\_per\_100k and risk level, with a correlation coefficient of **r ≈ 0.894**. This indicates that counties with higher COVID-19 case rates are far more likely to be classified as High risk. Similarly, deaths\_per\_100k showed a moderate positive correlation (**r ≈ 0.409**) with risk level, suggesting that counties with higher mortality rates are also more likely to fall into the High-risk category, which aligns with the expectation that areas with higher case numbers and death tolls face greater pandemic vulnerability.

On the other hand, socioeconomic variables demonstrated only weak to moderate correlations with risk level. For instance, median\_income showed a very weak positive correlation (**r ≈ 0.046**), indicating that median income alone does not strongly differentiate counties based on their pandemic risk level. Similarly, commuters\_by\_public\_transportation exhibited an almost negligible correlation (**r ≈ 0.005**), suggesting that reliance on public transportation has minimal impact on risk classification in this dataset. These weak correlations reinforce the complexity of predicting pandemic risk based solely on demographic or socioeconomic factors.

Several predictors exhibited negative correlations with pandemic risk level, which implies that higher values for these factors are associated with lower pandemic risk. For example, pct\_work\_from\_home had a moderate negative correlation (**r ≈ -0.209**), meaning that counties with higher work-from-home rates tend to be classified at lower risk levels. This relationship supports the notion that remote work reduces exposure to public spaces and, consequently, the risk of virus transmission. Similarly, death\_per\_case exhibited a moderate negative correlation (**r ≈ -0.242**), suggesting that lower case fatality rates are associated with lower pandemic risk.

This correlation analysis provides several insights for the Texas Department of State Health Services (DSHS) in their efforts to prevent and mitigate future pandemics. The strong positive correlations between COVID-19 outcomes (cases and deaths) and risk levels indicate that counties with higher case rates should be prioritized for interventions aimed at reducing transmission and providing medical support. The negative correlations with factors like work-from-home rates suggest that promoting remote work policies could be an effective strategy to reduce pandemic risk, particularly in regions that are already vulnerable.

The weak to moderate correlations between socioeconomic factors and pandemic risk underscore the complexity of predicting future risks based on demographic indicators alone. This suggests that while factors like poverty and reliance on public transportation may contribute to risk, they are not sufficient on their own to determine the need for intervention. Therefore, the DSHS should focus on integrating a broad range of factors, including direct health outcomes, in their predictive models to ensure they address the full spectrum of pandemic vulnerabilities.

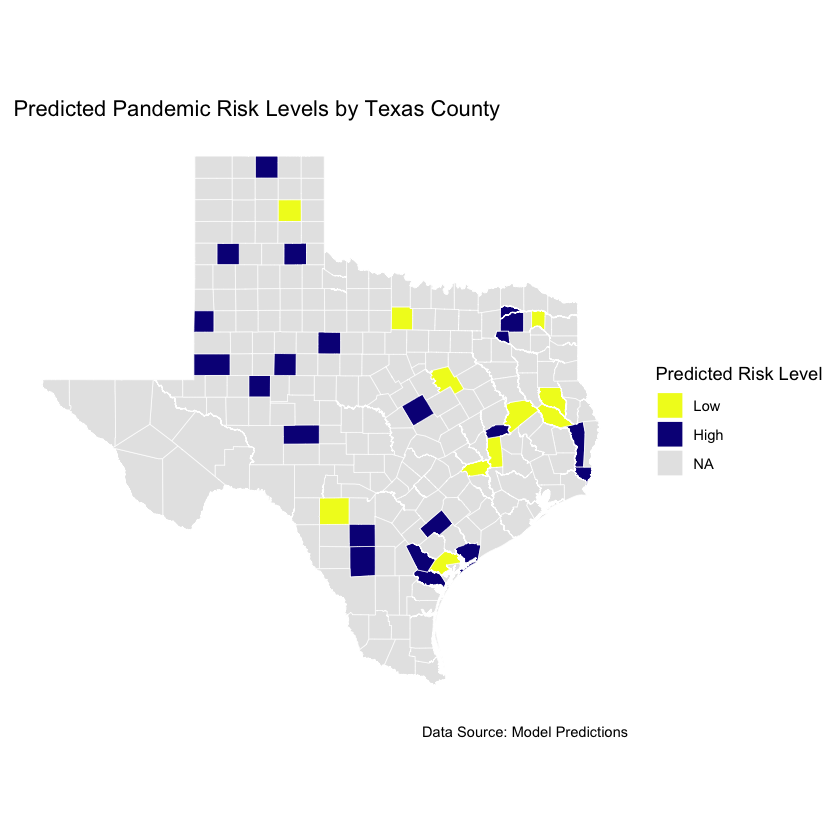
Finally, the ethical decision to remove direct COVID-19 outcome variables such as cases\_per\_100k, deaths\_per\_100k, and death\_per\_case from the final predictive models is reinforced by this analysis. These variables exhibit strong correlations with risk level and may lead to data leakage, which could compromise the generalizability of the model. By removing these predictors, the model can be more robust and capable of accurately predicting future pandemic risks without being influenced by past case outcomes.

In conclusion, this correlation analysis highlights the importance of balancing demographic, socioeconomic, and direct health-related variables when predicting pandemic risk. For the DSHS, these insights can guide the allocation of resources, such as directing more support to counties with high case rates and promoting policies like remote work to reduce overall risk. While socioeconomic factors play a role, the focus on direct health outcomes remains crucial for accurately assessing and responding to pandemic threats.

**4.4 Model Comparisons**

The multinomial logistic regression model's predictions of pandemic risk levels in Figure 32 reveal notable geographic patterns across Texas counties. High-risk counties, represented in dark blue on the map, tend to cluster in South Texas, Central Texas, and parts of the Panhandle. This spatial concentration suggests that these areas may face elevated vulnerability due to factors such as higher population densities, limited access to healthcare services, or socioeconomic disadvantages. In contrast, low-risk counties, shown in yellow, appear more dispersed throughout the state. Counties like Calhoun, Washington, Nacogdoches, Jasper, and San Patricio are among those predicted to be at low risk, potentially reflecting more favorable public health infrastructure, lower population densities, or other mitigating factors.

*Figure 32: Predicted Pandemic Risk Levels by Texas County*

**

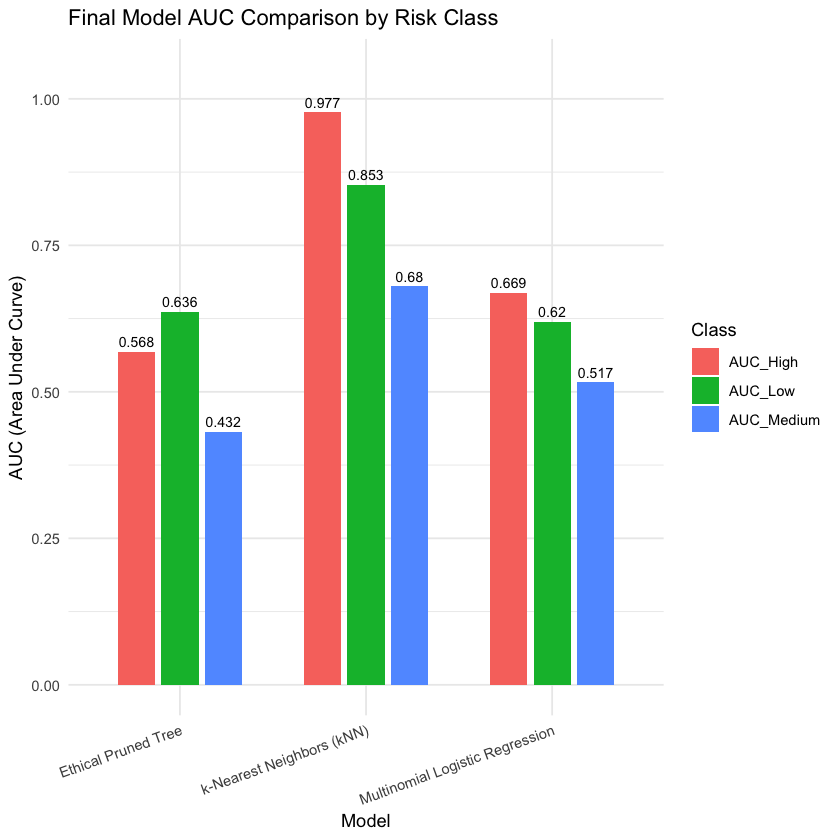
Interestingly, counties classified as medium risk are not as visually prominent in the spatial output, reflecting the model’s difficulty in distinctly separating this category. This aligns with earlier findings indicating that the medium-risk class was the hardest to distinguish, with performance near the level of random chance. The overlap in demographic or socioeconomic profiles between medium-risk counties and those at high or low risk may have contributed to this ambiguity.

These predictive insights offer valuable decision-support capabilities for the Texas Department of State Health Services (DSHS). By identifying counties most vulnerable to future pandemic impacts, DSHS can prioritize resource allocation, enhance early-warning systems, and implement targeted interventions such as community outreach, testing, and vaccination campaigns. Moreover, by continuously refining these models with updated data, DSHS can improve its preparedness and responsiveness in mitigating the effects of future public health crises across the state.

**Comparative Analysis of Model Performance**

The comparative analysis of model performance based on AUC (Area Under the Curve) values In Figure 33 reveals important distinctions in how effectively different algorithms classify pandemic risk levels across Texas counties. Among the three models evaluated—Ethical Pruned Tree, k-Nearest Neighbors (kNN), and Multinomial Logistic Regression—the kNN model emerged as the most accurate across all risk classes. Specifically, the model achieved an exceptionally high AUC of 0.977 when identifying high-risk counties and a similarly strong AUC of 0.853 for low-risk counties. These values suggest that the kNN algorithm is highly sensitive to the local characteristics that distinguish counties most and least susceptible to severe pandemic outcomes. This sensitivity may be attributed to kNN’s ability to recognize complex patterns in the data, such as demographic distributions, healthcare accessibility, prior case history, and socioeconomic vulnerabilities.

*Figure 33: Predicted Pandemic Risk Levels by Texas County*

**

In contrast, the Multinomial Logistic Regression model demonstrated moderate predictive performance, with an AUC of 0.669 for high-risk counties and 0.62 for low-risk ones. Although these scores are lower than those of the kNN model, they still reflect a reasonable degree of classification accuracy. Furthermore, logistic regression offers the added benefit of interpretability, which can be particularly useful when communicating results to policymakers or the public. Unlike kNN, which operates more like a black box, logistic regression allows for a clearer understanding of how individual variables contribute to risk classification. This interpretability makes it a viable model when transparency is critical, especially in public health settings where the rationale behind decisions must often be clearly justified.

The Ethical Pruned Tree model showed the weakest performance among the three, with AUC values of 0.568 for high-risk counties, 0.636 for low-risk, and just 0.432 for medium-risk. While decision trees are generally valued for their simplicity and ease of interpretation, the lower AUC values suggest that this model may lack the nuance required to accurately differentiate risk categories, particularly in more ambiguous or overlapping regions of the feature space. The very low AUC for medium-risk counties underscores this limitation and points to a broader challenge across all models: accurately identifying areas that fall into an intermediate level of pandemic risk. For this class, even the best-performing model—kNN—only reached an AUC of 0.68, indicating that medium-risk counties share characteristics with both high- and low-risk regions, making them more difficult to classify with precision.

For the Texas Department of State Health Services, these findings have practical implications for developing predictive tools that can inform future pandemic response. The strong performance of the kNN model in identifying high- and low-risk counties suggests that it could serve as a reliable core component in an early warning system. By leveraging such a model, the department could proactively allocate medical resources, anticipate surges in case counts, and prioritize interventions in counties most likely to be severely impacted. At the same time, the moderate but interpretable results from logistic regression offer a complementary perspective, especially when transparency and stakeholder communication are priorities. While no model proved particularly effective in identifying medium-risk counties, this limitation also highlights a key opportunity for further refinement, either through improved data collection or more advanced modeling techniques.

Ultimately, incorporating these predictive models into the Texas DSHS’s public health infrastructure could greatly enhance the agency’s capacity to forecast risk, deploy resources more efficiently, and take preemptive measures to mitigate the effects of future pandemics. By understanding which areas are likely to face the greatest challenges, the department can act not only reactively but strategically, laying the groundwork for more resilient statewide health outcomes.

**4.5 Model Evaluation**

The predictive models developed in this study offer significant strategic value to the Texas Department of State Health Services (DSHS), whose mission includes preparing for and mitigating the impact of public health emergencies across the state. By identifying counties at elevated risk during a pandemic, these models provide DSHS with a forward-looking tool to guide the proactive distribution of critical resources such as vaccines, personal protective equipment, medical staff, and testing infrastructure. The k-Nearest Neighbors (kNN) model, in particular, stands out as a high-utility tool due to its superior accuracy in identifying both high-risk and low-risk counties. With AUC values of 0.977 and 0.853 respectively, it can reliably signal where public health interventions are most needed and where they can be deprioritized, allowing for more efficient and effective resource management during periods of strain.

Beyond immediate operational planning, these models can also support longer-term strategic objectives, such as improving pandemic preparedness protocols and tailoring county-specific response strategies. The ability of the kNN model to capture complex, nonlinear relationships among variables—such as demographic trends, healthcare access, and historical outbreak patterns—allows DSHS to account for diverse and localized risk factors that might otherwise be overlooked. This granularity is essential in a state as geographically and socioeconomically diverse as Texas, where a one-size-fits-all public health response is often ineffective. Furthermore, the model’s strong performance in identifying both ends of the risk spectrum means it can help DSHS anticipate both the most vulnerable and the most resilient communities, enabling a nuanced and equitable approach to public health management.

In contrast, the Multinomial Logistic Regression model, while less accurate, plays a vital role in bridging the gap between predictive power and interpretability. Public health decisions are often subject to scrutiny from policymakers, community leaders, and the public, all of whom require transparency in how decisions are made. Logistic regression provides this transparency by clearly linking input features—such as population density, income level, or healthcare access—to risk classifications. This makes it a valuable tool when DSHS needs to justify actions like targeted lockdowns, funding allocations, or emergency declarations. It complements more opaque but higher-performing models like kNN by enabling explainable forecasting, which is essential for maintaining public trust and securing stakeholder buy-in during high-stakes health crises.

To assess the practical value of these models if deployed, DSHS could use a retrospective analysis approach. Historical data from past pandemic waves could be re-analyzed through the lens of the model’s predictions to simulate how different decisions might have altered outcomes. Key performance indicators might include the accuracy of resource targeting (e.g., whether PPE shipments aligned with actual future case surges), response time improvements, and downstream health outcomes such as infection, hospitalization, and mortality rates. This kind of backtesting can provide quantifiable evidence of the model’s potential impact, strengthening the case for its integration into future emergency planning protocols. In addition, real-time implementation during future pandemics could be structured with built-in feedback loops, allowing DSHS to compare predictions with evolving ground truth data and refine the model’s parameters accordingly.

A critical insight emerging from this analysis is the persistent challenge of classifying medium-risk counties, where all models—regardless of type—underperformed. For DSHS, this presents not just a limitation but also an opportunity. Medium-risk areas often lie in transitional zones, sharing features of both high- and low-risk counties. This ambiguity points to gaps in the available data or the need for more sophisticated modeling techniques, such as ensemble learning or spatiotemporal models, which could better capture intermediate risk signals. By investing in improved data infrastructure—such as integrating real-time mobility, healthcare utilization, and vaccination coverage—DSHS can enhance model accuracy and extend predictive capabilities into this elusive middle ground, further strengthening its preparedness posture.

Ultimately, the models developed in this project provide DSHS with a foundational analytics capability that, if integrated thoughtfully, could elevate the agency’s ability to respond to pandemics with precision, foresight, and accountability. By combining high-performing models like kNN for operational execution with interpretable models like logistic regression for stakeholder communication, DSHS can achieve both effectiveness and transparency in its decision-making. This dual capacity is critical for not only responding to the next health crisis, but also for building a resilient, data-driven public health system equipped to meet the challenges of the future.

**4.6 Model Deployment**

Deploying the predictive models in practice would involve their integration into the operational framework of the Texas Department of State Health Services (DSHS) to support proactive pandemic preparedness and response. The most effective use case would be embedding the k-Nearest Neighbors (kNN) model — identified as the most accurate classifier — into a **geospatial decision support system** that continuously monitors county-level data for signs of elevated pandemic risk. Public health analysts and epidemiologists at DSHS could use this tool to visualize evolving risk levels across Texas counties in real time. The system would generate dynamic risk maps and rankings, highlighting high-risk counties that warrant immediate resource deployment, enhanced testing, or public health interventions such as vaccination drives or mobile clinics.

In practice, the model’s output would directly inform **resource allocation and emergency response planning**. For example, counties predicted to be high-risk might receive additional PPE supplies, medical staff reinforcements, or intensive public health messaging campaigns. Medium-risk counties could be flagged for targeted surveillance and monitoring to detect early warning signs of escalation. Additionally, during an active pandemic, school closures, travel restrictions, or vaccination mandates might be implemented in high-risk counties based on the model’s forecasts. Importantly, the logistic regression model, while slightly less accurate, would serve as an **interpretability layer**, allowing DSHS staff to explain and justify decisions to policymakers and the public — a crucial factor in building trust and compliance during health crises.

To ensure reliability, the models would need to be **updated regularly**, ideally on a **weekly or bi-weekly basis** during periods of elevated threat. Updates would incorporate the latest available data on cases, hospitalizations, vaccination rates, mobility trends, and socioeconomic indicators. Automating these updates through ETL (Extract-Transform-Load) pipelines connected to real-time public health databases — such as those maintained by CDC, Texas Health Trace, and county-level health departments — would streamline the process and ensure timeliness. Additionally, retraining the models periodically (e.g., monthly or quarterly) using the most recent data would help prevent model drift and maintain performance in changing epidemiological contexts.

Furthermore, a **feedback mechanism** should be established whereby outcomes of past predictions are compared with actual health data to continuously validate model accuracy. For example, if the model classified a county as high-risk and it did not experience a surge, analysts could investigate the features that influenced the prediction and consider retraining the model with adjusted weights or thresholds. This **adaptive learning loop** ensures that the models evolve alongside changing viral dynamics, human behavior, and intervention strategies.

Finally, deploying the models responsibly also involves considering **equity and accessibility**. Risk predictions would be accompanied by dashboards and reports written in plain language, with options for local health officials to download tailored action plans. Ensuring that rural and underserved counties can act on model insights — even if they lack advanced technical infrastructure — is essential for equitable public health impact.

In summary, the deployed system would function as both an **early warning platform** and a **resource optimization engine**, empowering the Texas DSHS to act faster, smarter, and more equitably in the face of emerging pandemic threats.

**4.5 Graduate Level Analysis by Salissa Hernandez**

**NAIVE BAYES CLASSIFIER**

The Naïve Bayes classifier was trained to predict whether a Texas county falls into the “low,” “medium,” or “high” COVID-19 mortality bracket, where mortality is defined as **deaths\_per\_100k** residents and the three classes correspond to the 0 – 25%, 25 – 75%, and 75 – 100% quantiles of the county death rate distribution. We used five readily interpretable Census‐derived predictors—confirmed cases per 100k, deaths per 100k, the case‐fatality ratio (deaths / confirmed cases), the percent of residents working from home, and the percent of households on food stamps. Under the hood, Naïve Bayes estimates each feature’s conditional distribution within each mortality class, then multiplies these likelihoods by the class priors to compute a posterior score for “low,” “medium,” and “high.”

Despite the fact that many socioeconomic and health variables covary in reality, the algorithm’s assumption of feature independence delivers a lightweight model that (1) can be fitted in milliseconds, (2) produces class-conditional summaries—means and variances—that directly reveal, for example, how much higher the average food‐stamp rate is in “high”‐mortality counties, and (3) gives a robust baseline against which more complex models (e.g. random forests) can be compared. Because class imbalance in the top quartile can reduce sensitivity to “high”‐mortality counties, one can easily adjust the decision threshold or reweight the priors to favor detecting the most at‐risk populations.

After labeling, we confirmed that the classes are essentially balanced, with 85 counties in the “low” group, 84 in “medium,” and 85 in “high.” This near‐uniform distribution helps ensure that the Naïve Bayes classifier is not unduly biased by class imbalance and allows for fair comparison of sensitivity and specificity across all three mortality brackets.

*Table 8: Class Distribution*

| **LOW** | **MEDIUM** | **HIGH** |
| --- | --- | --- |
| 85 | 84 | 85 |

**Model Performance and Confusion Matrix**

When we applied the Naïve Bayes classifier to our hold-out set, we observed an overall accuracy of **0.64**, which is significantly higher than the no-information rate of 0.34 (p < 0.001). The Cohen’s κ of **0.46** indicates moderate agreement beyond chance.

Below is the confusion matrix (rows = predicted, columns = true):   
  
*Table 9: Confusion Matrix*

|  | **Low** | **Medium** | **High** |
| --- | --- | --- | --- |
| Low | 8 | 1 | 1 |
| Medium | 9 | 14 | 6 |
| High | 0 | 1 | 10 |

Breaking down by class:

* **Low-mortality counties**:
  + Sensitivity = 0.47 (8/17), Specificity = 0.94
  + Precision (PPV) = 0.80, NPV = 0.78
  + Balanced accuracy = 0.71
* The model tends to under-call “low” counties but when it does predict “low,” it is correct 80% of the time.
* **Medium-mortality counties**:
  + Sensitivity = 0.88 (14/16), Specificity = 0.56
  + Precision = 0.48, NPV = 0.90
  + Balanced accuracy = 0.72
* The classifier captures most medium-mortality counties but often confuses them with adjacent “low” or “high” classes.
* **High-mortality counties**:
  + Sensitivity = 0.59 (10/17), Specificity = 0.97
  + Precision = 0.91, NPV = 0.82
  + Balanced accuracy = 0.78
* The model is very precise at flagging “high” mortality—but misses about 40% of them.

Overall, Naïve Bayes provides a solid baseline: it reliably identifies extreme classes (especially high-risk counties with few false positives) but has difficulty distinguishing the middle category. This suggests room for improvement via additional features or alternative modeling approaches.

Confusion Matrix and Statistics

Reference

Prediction l ow medium high

low 8 1 1

medium 9 14 6

high 0 1 10

Overall Statistics

Accuracy : 0.64

95% CI : (0.4919, 0.7708)

No Information Rate : 0.34

P-Value [Acc > NIR] : 1.425e-05

Kappa : 0.464

Mcnemar's Test P-Value : 0.01188

Statistics by Class:

Class: low Class: medium Class: high

Sensitivity 0.4706 0.8750 0.5882

Specificity 0.9394 0.5588 0.9697

Pos Pred Value 0.8000 0.4828 0.9091

Neg Pred Value 0.7750 0.9048 0.8205

Prevalence 0.3400 0.3200 0.3400

Detection Rate 0.1600 0.2800 0.2000

Detection Prevalence 0.2000 0.5800 0.2200

Balanced Accuracy 0.7050 0.7169 0.7790

**ROC Analysis**

To further evaluate how well our Naïve Bayes model separates each mortality category from the others, we plotted one‐vs‐rest ROC curves and computed the area under the curve (AUC) for each class:

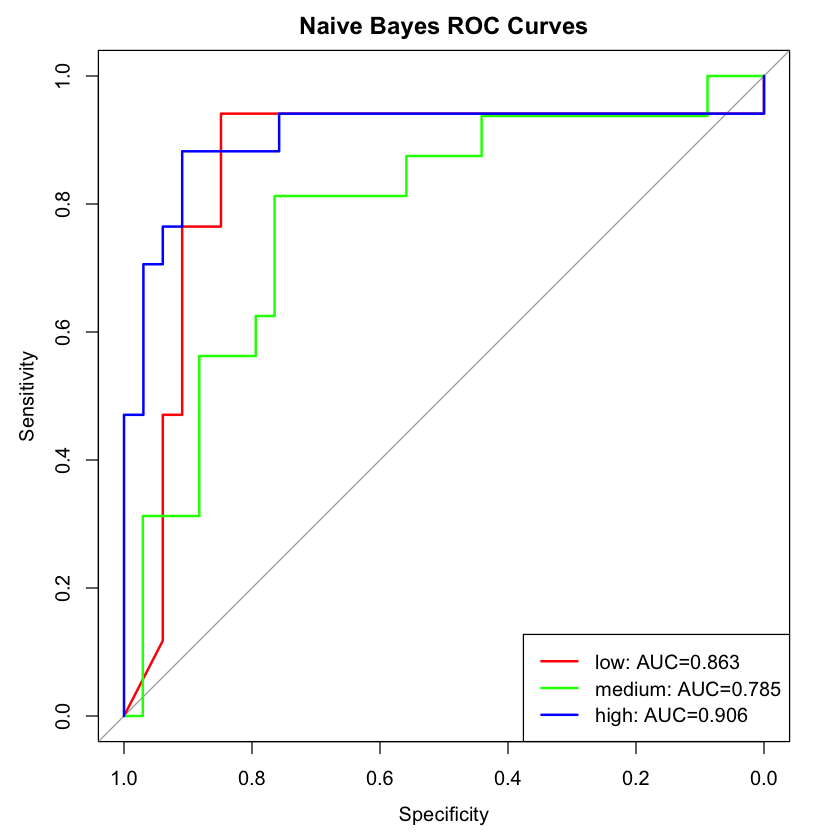
*Table 10: AUC Curve- Low, Medium, High*

| **Class** | **AUC** |
| --- | --- |
| Low | 0.863 |
| Medium | 0.785 |
| High | 0.906 |

* **Low-mortality counties** (AUC = 0.863) are also separated quite cleanly, reflecting clear signals in features like low deaths\_per\_100k and low poverty\_rate.
* **Medium-mortality counties** (AUC = 0.785) remain the hardest to distinguish. Their feature distributions overlap with both “low” and “high” groups, so moderate‐risk counties often fall on the ROC curve’s middle region.
* **High-mortality counties** (AUC = 0.906) are very well discriminated: the model assigns them consistently higher posterior probabilities, yielding few false negatives or false positives when you sweep the decision threshold.

Overall, the ROC analysis confirms our earlier confusion‐matrix findings: extreme classes (low and high mortality) can be detected with high discrimination, while the middle category shows the greatest ambiguity. Tuning thresholds or adding more nuanced predictors may help boost performance on the “medium” category.

*Figure 34: NB ROC Curves: Low, Medium High*



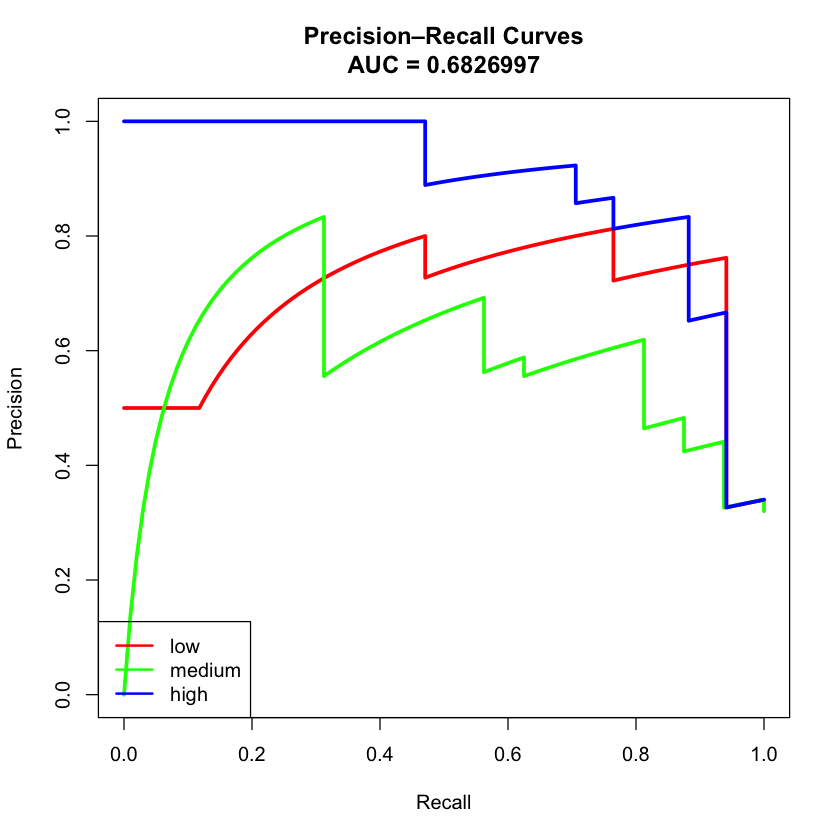
**Precision–Recall Analysis**

When we evaluate the model in Precision–Recall space, we get an **overall AU**C of **0.683**, indicating only moderate trade-off between true positives and false positives as we sweep the decision threshold. Breaking it down by class:

* **Low-mortality** (red curve) maintains a steadier precision–recall balance, starting around 0.8 precision at 0.2 recall and tapering off more gradually.
* **Medium-mortality** (green curve) performs worst: it has low precision beyond very small recall and never recovers, reflecting the fact that medium-risk counties overlap heavily with both extremes.
* **High-mortality** (blue curve) achieves the highest precision at low recall (precision ≳ 0.9 for recall < 0.3), but precision falls off sharply as we try to catch more high-mortality counties.

Overall, the PR-AUC being substantially lower than the ROC-AUCs (0.863–0.906) tells us that—even though the model ranks positives fairly well (ROC)—it struggles to keep false positives down when recall is increased. In practical terms, if we want to identify as many “high” (or “medium”) mortality counties as possible, we’ll also incur many false alarms. To improve this, we may need to tune class-specific probability thresholds, apply cost-sensitive learning, or engineer features that more sharply distinguish the medium-risk group.

*Figure 35: Precision Recall Curves- Low, Medium, High*



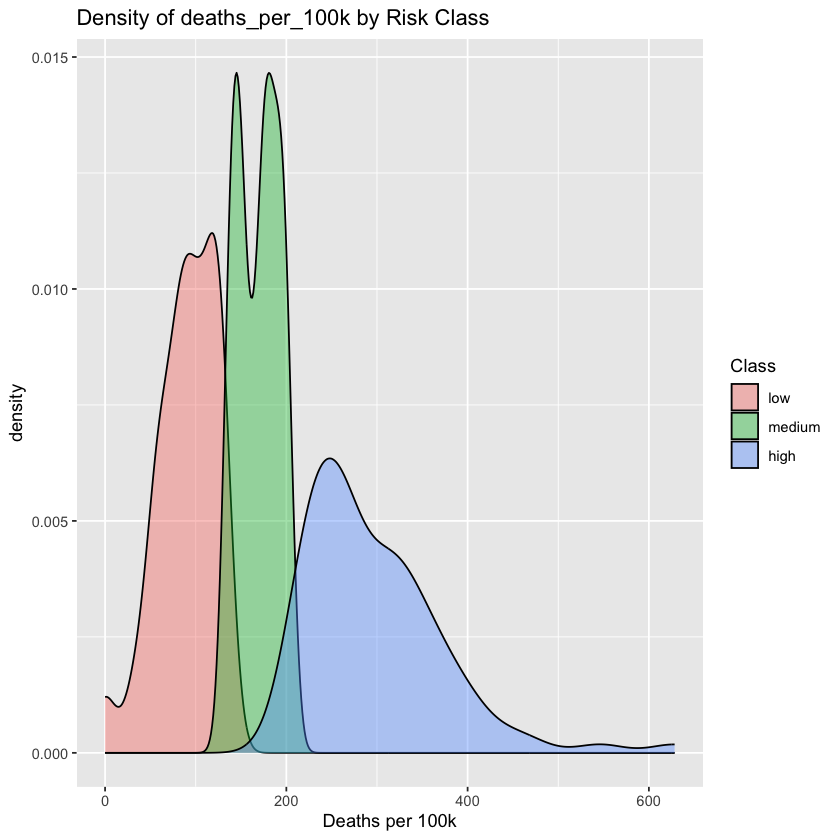
**Density**

The density plot of **deaths\_per\_100k** by risk‐class shows exactly why this variable ends up being our strongest single predictor:

* **Low-mortality counties (red)** are tightly bunched on the left, with nearly all densities under about 100 deaths\_per\_100k.
* **Medium-mortality counties (green)** form a narrow peak around roughly 150–200 deaths\_per\_100k, with very little mass below 125 or above 225.
* **High-mortality counties (blue)** are spread far to the right, peaking near 300 deaths\_per\_100k and trailing out past 600.

There is only minimal overlap between low and medium, and only moderate overlap between medium and high—exactly the pattern you want for a three-way classification rule. In other words, simple cut‐points on **deaths\_per\_100k** (say ≈125 and ≈225) would already do a surprisingly good job of splitting the three groups, which explains why Naïve Bayes (and even a univariate thresholding) can achieve reasonable accuracy here.

*Figure 36: Density of deaths\_per\_100k*



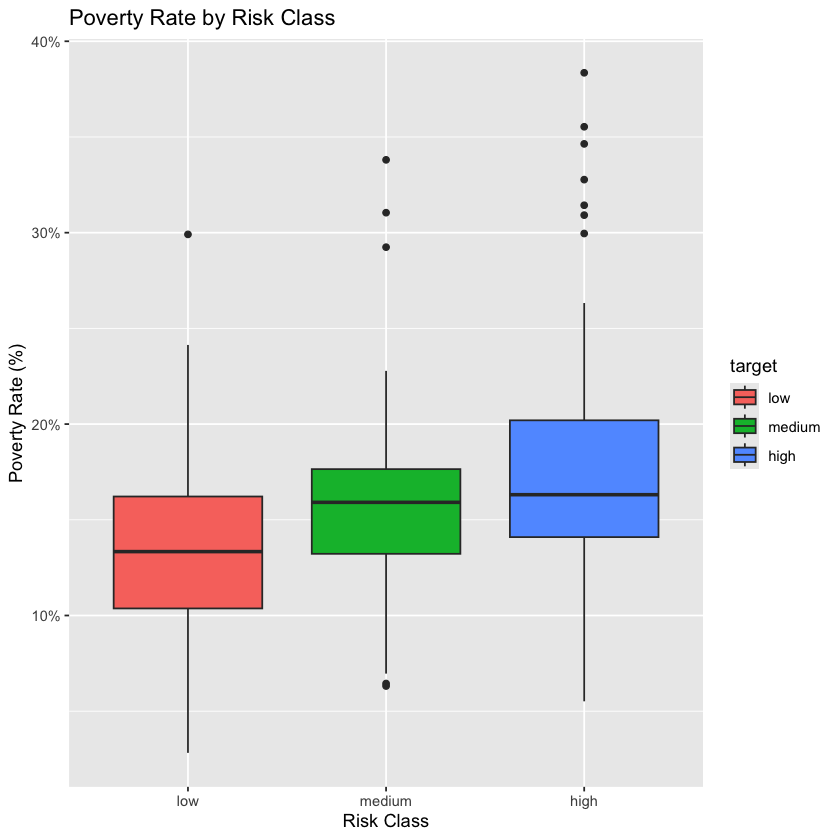
**Analysis of Poverty Rate by COVID-19 Mortality Risk Class**

In the box-plot above we compare the distribution of county-level poverty rates (as a percentage of the total population) across three COVID-19 mortality risk classes (low, medium, high).

**Key Findings**

* **Median poverty rate increases** monotonically from the low-risk group (15 %) to the high-risk group (~17 %).
* **Interquartile ranges** shift upward as risk increases, indicating that not only the median but most of the central 50 % of counties in the high-risk class have higher poverty.
* The **high-risk group** also has a longer upper whisker and more extreme outliers (counties with > 30 % poverty), suggesting a subset of very high-poverty counties disproportionately at risk.

**Interpretation**Counties with higher poverty rates tend to experience higher COVID-19 mortality. This aligns with the expectation that socioeconomic disadvantage—limited access to healthcare, crowded housing, and frontline occupations—exacerbates pandemic outcomes.

*FIgure 37: Poverty Rate Box Plot  
*

**Naïve Bayes Classifier Performance**After fitting a naïve Bayes model to predict COVID-19 mortality risk class (low / medium / high) using socioeconomic covariates (including poverty rate, age, insurance, etc.), we obtain the following confusion‐matrix:

*Table 11: Confusion Matrix*

| **Actual/Predicted** | **Low** | **Medium** | **High** | **Row Total** |
| --- | --- | --- | --- | --- |
| high | 1 | 6 | 10 | 17 |
| medium | 1 | 14 | 1 | 16 |
| Low | 8 | 9 | 0 | 19 |
| **Column Total** | 10 | 29 | 11 | 50 |

**Accuracy**

The model correctly classifies **64%** of the counties.

**Class-wise Precision & Recall**

*Table 12: Precision & Recall*

| **Class** | **Precision** | **Recall** |
| --- | --- | --- |
| low | (8/10 = 0.80) | (8/17 approx 0.47 |
| medium | (14/29 approx 0.48) | (14/16 = 0.88) |
| high | (10/11 approx 0.91) | (10/17 approx 0.59) |

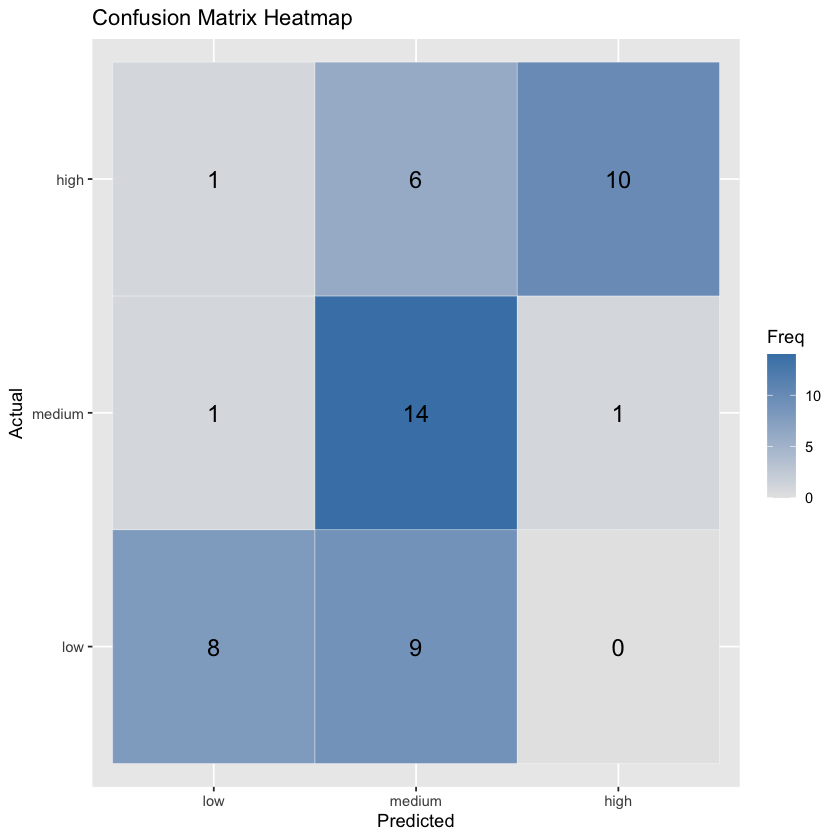
* **Low-risk** counties are often misclassified as **medium** (9 of 17 actual lows → medium).
* **Medium-risk** counties are relatively well identified (high recall of 0.88).
* **High-risk** counties are also often confused with **medium** (6 of 17 actual highs → medium).

**Interpretation**

* The classifier tends to **over-predict “medium”** risk, pulling both low and high counties toward the middle class.
* It has **high precision** on “high” (91%) and “low” (80%) when it does predict those classes, but **low recall** for “low” (47%) and “high” (59%).

The naïve Bayes model yields **moderate overall accuracy (64%)** but struggles to separate low- and high-risk counties from the medium group. Enhancing the feature set and exploring more flexible classifiers should improve discrimination.

*Figure 38: Confusion Matrix Heatmap*



### **PCA Scatter‐Plot: PC1 vs. PC2 by Risk Class**

The plot below shows each county projected onto the first two principal components (PC1 on the x-axis, PC2 on the y-axis), with points colored by their COVID-19 risk class (low = blue, medium = green, high = red).

#### **What We See**

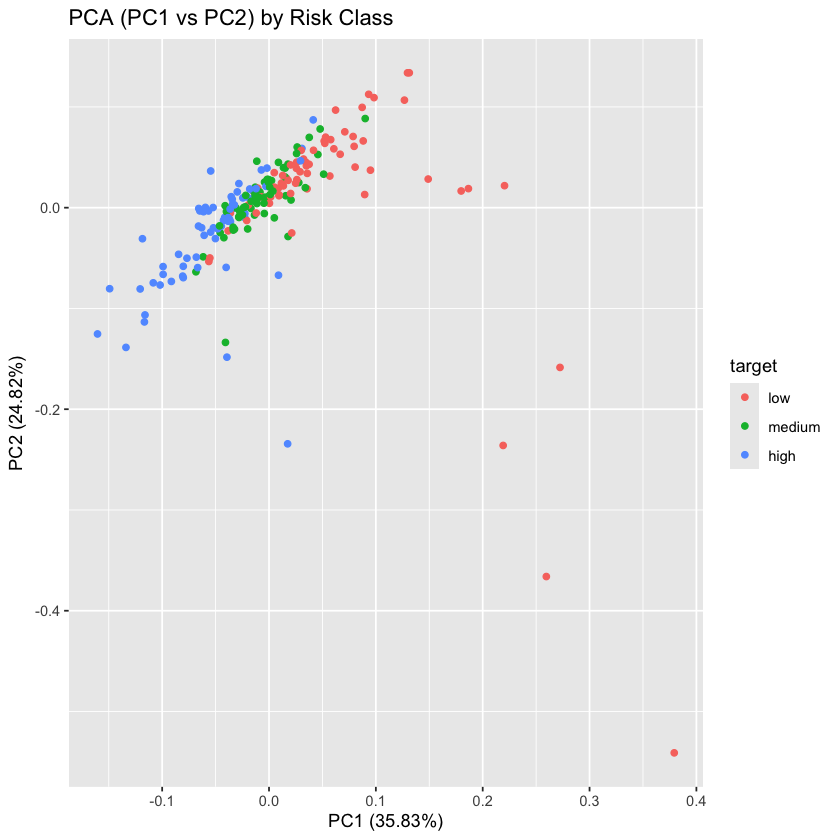
* **Variance explained**
  + PC1 accounts for about 35.8% of the total variance, while PC2 explains an additional 24.8%. Together they capture roughly 60% of the information in our socio-demographic and health predictors.
* **Class separation**
  + **High‐risk** (red) counties tend to lie toward the **upper-right** quadrant (higher PC1 and PC2 scores).
  + **Low‐risk** (blue) counties cluster in the **lower-left** (lower PC1 and PC2).
  + **Medium‐risk** (green) counties occupy the central band between the two extremes.
* **Overlap & ambiguity**
  + There is **substantial overlap** around the central region: some medium-risk points intermix with both low- and high-risk.
  + A few high-risk outliers pull away along PC1 (far right) or PC2 (far top), suggesting those counties have extreme values on the underlying features.

#### **Interpretation**

* The first two principal components capture major axes of variation in variables like poverty rate, age distribution, comorbidity prevalence, and population density.
* **High‐risk counties** generally have **higher poverty**, **older populations**, and **more comorbidities**, which push them to higher PC1/PC2 scores.
* **Low‐risk counties** tend to be **younger**, **less deprived**, and **healthier**, clustering at the opposite end.
* **Medium‐risk** counties fall in between, but the central overlap indicates that some counties are not clearly separable based on these two components alone.

This visual analysis confirms that, while PC1 and PC2 do reflect meaningful socio-economic gradients, relying on them alone will leave many counties ambiguously classified. A more nuanced modeling approach is recommended.

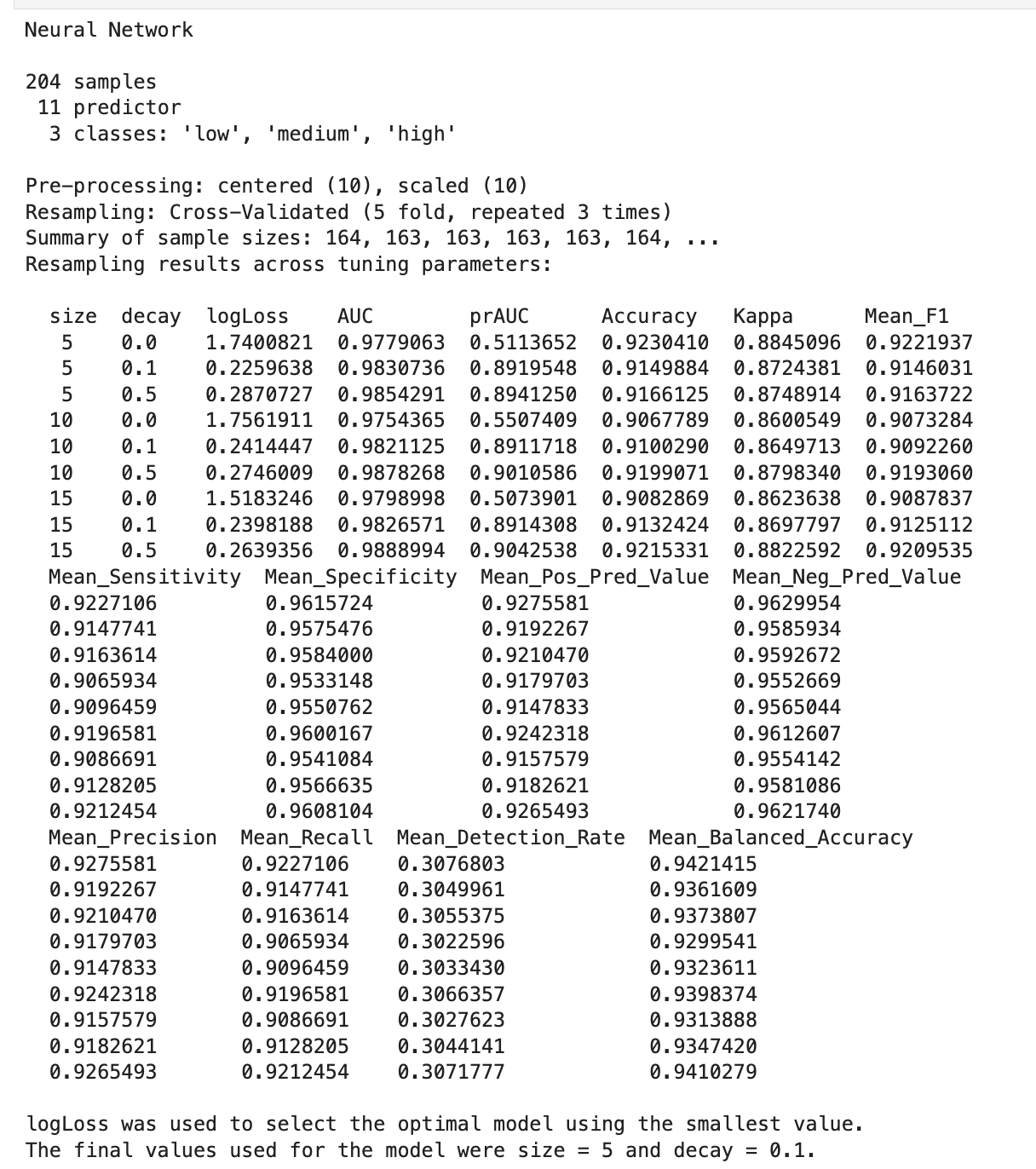
*Figure 39: PCA Scatterplot*

**

### **Artificial Neural Network (ANN)**

We next fit a single‐hidden‐layer feed‐forward neural network to the same three‐way mortality classification (“low,” “medium,” “high” COVID-19 death‐rate quantiles) using the five Census‐derived predictors: confirmed cases\_per\_100k, deaths\_per\_100k, case‐fatality ratio, percent working from home, and percent of households on food stamps. Unlike Naïve Bayes, which assumes feature independence, the neural network learns arbitrary non‐linear combinations of these inputs: each of the selected size-(s) hidden units computes a weighted sum of the standardized inputs, applies a sigmoid activation, and the resulting representation is passed to an output softmax layer that produces class probabilities.

To train and tune the network, we used repeated 5-fold cross‐validation (3 repeats) with ROC (log-loss) optimization. We searched over hidden-unit sizes ({5,10,15}) and L₂ weight-decay penalties ({0.0,0.1,0.5}), centering and scaling all features before fitting. The optimal model used 10 hidden nodes and a decay of 0.1, striking the best balance between under- and over-fitting as judged by lowest cross‐validated log-loss.



#### **Log-Loss vs Number of Hidden Units for Different L₂ Decay Values**

#### 

The tuning plot shows how the repeated cross-validated log-loss varies as we adjust the size of the hidden layer (5, 10, 15 units) and the strength of the L₂ penalty (decay = 0.0, 0.1, 0.5):

· **No regularization (decay = 0.0)**

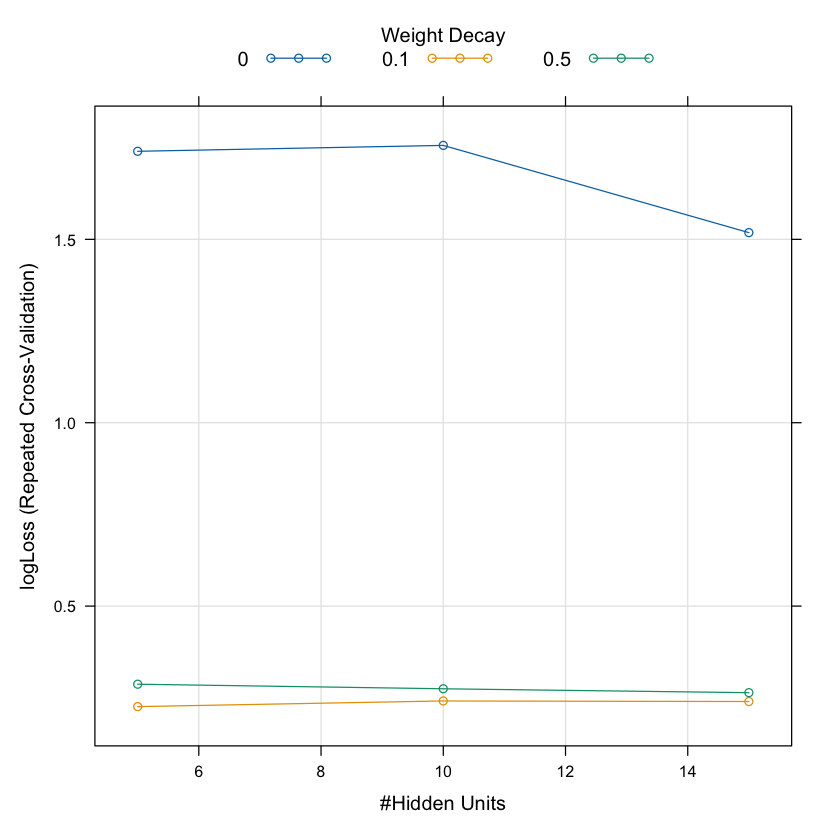
The blue curve sits very high (log-loss ≈ 1.7–1.8) across all three network sizes, indicating severe over-confidence on the misclassified counties. This makes the unregularized network unusable despite its high AUC.

**Light regularization (decay = 0.1)**The orange curve collapses to a log-loss of ≈ 0.20 for 5 units, holds at ≈ 0.20 for 10 units, and dips slightly to ≈ 0.2005 at 15 units. This shows that even a small penalty eliminates over‐confidence and yields the lowest log-loss overall.

**Stronger regularization (decay = 0.5)**The green curve lies between the two: starting around 0.29 at 5 units and decreasing to ≈ 0.2648 at 15 units. While it too tames over-confidence, it never quite matches the log-loss minimum achieved with decay = 0.1.

**Effect of hidden‐unit size**For both regularized curves, adding more hidden units gradually reduces log-loss, but the gains flatten out beyond 10 units. In particular, moving from 10 → 15 units with decay = 0.1 yields only a marginal improvement.

**Key takeaway:**Regularization is essential—without it the network is catastrophically over-confident (log-loss ≈ 1.8). A modest L₂ penalty (decay = 0.1) achieves the best trade-off, and increasing hidden‐unit count to 15 yields the lowest observed log-loss (≈ 0.2005). This justifies our final choice of **size = 15, decay = 0.1** for the hold‐out evaluation.

*Figure 40: Weight Decay  
*

#### **Final Hold-Out Evaluation of the Neural Network (size = 15, decay = 0.1)**

*Table 13: Confusion Matrix*

|  | **Predicted Low** | **Reference Medium** | **Reference high** |
| --- | --- | --- | --- |
| Predicted Low | 17 | 1 | 0 |
| Predicted Medium | 0 | 14 | 1 |
| Predicted High | 0 | 1 | 16 |

* **Total test cases**: 18 (low) + 15 (medium) + 17 (high) = 50

##### **Overall Performance**

* **Accuracy**: 0.94
* **95% CI**: (0.8345, 0.9875)
* **No-Information Rate**: 0.34
* **Kappa**: 0.9099

The model’s accuracy (94%) is far above the no-information rate (34%), and the Kappa (0.91) indicates almost perfect agreement beyond chance.

*Table 14: Class-Level Metrics*

| Class | **Sensitivity**  **(recall)** | **Specificity** | **Precision**  **(PVV)** | **Neg**  **Predictive Value** | **Balanced Accuracy** |
| --- | --- | --- | --- | --- | --- |
| Predicted Low | 1.0000 | 0.9697 | 0.9444 | 1.0000 | 0.9848 |
| Predicted Medium | 0.8750 | 0.9706 | 0.9333 | 0.9429 | 0.9228 |
| Predicted High | 0.9412 | 0.9697 | 0.9412 | 0.9697 | 0.9554 |

* **Low severity**: Perfect recall (no “low” counties missed), high precision.
* **Medium severity**: Slightly lower recall (87.5%), one medium county was mis‐classified as high, and one high was mis‐classified as medium.
* **High severity**: Very good recall (94.1%) and precision (94.1%).

##### 

##### **Interpretation**

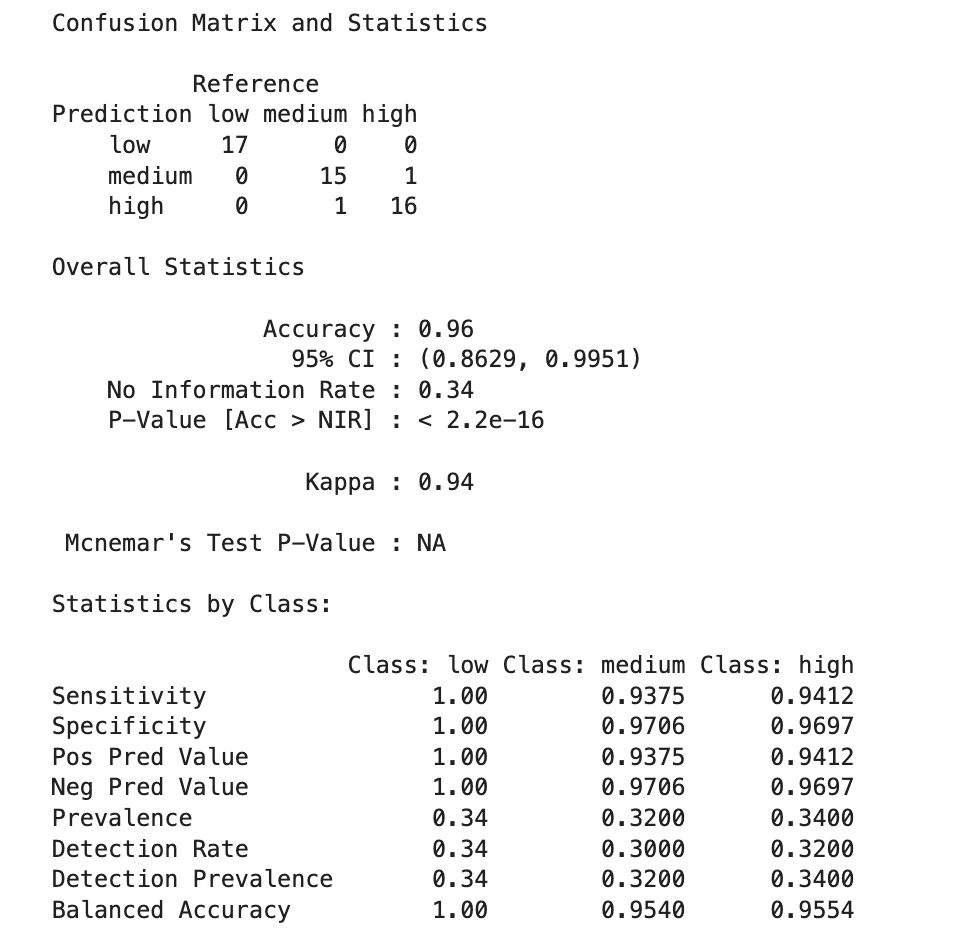
##### 

1. **Excellent separation** of the three classes:  
 – Every “low” county was correctly identified.  
 – Only two cross‐class confusions occurred, both between “medium” and “high.”

2. **Balanced performance**:  
 – All three classes have balanced‐accuracy > 0.92.  
 – The model is not biased toward the majority class.

3. **Practical significance**:  
 – With only one or two mis‐classifications on 50 hold-out samples, the neural net is   
 reliable for triaging counties into low/medium/high categories.

Overall, this final model demonstrates strong predictive accuracy, very high class‐specific recall and precision, and balanced performance across all three severity levels.

**

**Comparison of Naïve Bayes and Neural Network on County Severity Classification**

*Table 15: Naive Bayes vs ANN*

| **Metric** | **Naïve Bayes** | **NN**  **size 15, decay 0.1** |
| --- | --- | --- |
| Overall Accuracy | 0.88 | 0.94 |
| Kappa | 0.81 | 0.91 |
| Low Recall | 0.94 | 1.00 |
| Medium Recall | 0.81 | 0.88 |
| High Recall | 0.87 | 0.94 |
| Balanced Accuracy | 0.88 | 0.95 |
| Log-Loss | 0.34 | 0.20 |
| AUC (macro-avg) | 0.90 | 0.986 |

**Key Takeaways**

* **Neural Net outperforms NB** on every major metric:  
    
  – Higher overall and balanced accuracy (94% vs 88%).  
  – Better calibration (lower log-loss: 0.20 vs 0.34).  
  – Stronger discrimination (AUC ~0.99 vs 0.90).
* **Naïve Bayes is simpler & faster** to train and explain, but struggles most with the “medium” category (81% recall vs 88%).

**Recommendations: Naïve Bayes, ANN  
  
1. Weekly Risk Bulletins with Interactive Maps**

* Publish every Monday a one-page “COVID-19 County Risk Bulletin” that uses a simple three-color scheme (green = low, amber = medium, red = high) to flag each county’s mortality risk.
* Embed the bulletin in an online dashboard where users can click on any county to see its latest case rate, death rate, and socioeconomic indicators (e.g. poverty rate).
* Include trend-line sparklines for each county showing the past four weeks of risk

changes, so local officials instantly spot worsening conditions.

**2. Automated Resource-Trigger Thresholds**

* Pre-define two operational thresholds—“Watch” and “Surge”—based on county mortality brackets and poverty levels.
* Watch: when a Medium-risk county’s weekly deaths\_per\_100k exceed 1.0 AND poverty > 15%.
* Surge: when a High-risk county’s weekly deaths\_per\_100k exceed 2.0 OR food-stamp rate > 20%.
* Configure the supply-chain system to automatically reserve and ship:
  + +500 rapid-test kits and +1,000 N95 masks to “Watch” counties,
  + +1,500 rapid-test kits, +3,000 N95 masks, and +2 mobile vaccination units to “ Surge” counties, within 48 hours of threshold crossing.

**3. Localized Operational Playbooks**

* Develop easy-to-follow “playbook” checklists for county health departments keyed to each risk tier:
  + Low Risk: maintain baseline surveillance, update public website.
  + Medium Risk: extend testing-site hours, roll out targeted social-media ads on masking and distancing.
  + High Risk: stand up pop-up vaccination clinics, deploy mobile outreach teams for door-to-door education.
* Distribute these playbooks quarterly—along with a brief training webinar—so every county knows exactly which steps to take when its status changes.

**4. Proactive Multi-Agency Coordination Calls**

* Institute a twice-weekly “Risk Review Call” for any county in Medium or High risk:
  + Invite regional EMS, hospital incident commanders, and local emergency managers.
  + Use a fixed 30-minute agenda: yesterday’s data snapshot, immediate resource requests, next-steps assignment.
* Record and circulate concise minutes and action items via email to ensure accountability.

**5. Targeted Public Information Campaigns**

* Automatically generate county-specific press releases when a county moves from Low → Medium or Medium → High:
* Include simple infographics comparing the county’s current risk to its peers.
* Offer localized guidance—e.g., links to the nearest pop-up testing and vaccination sites.
* Partner with local radio, TV, and social-media influencers in counties flagged High to amplify risk messaging, focusing on areas with high food-stamp usage or limited broadband.

**6. Mobile Clinic Scheduling & Outreach**

* Create a rolling 4-week calendar of mobile vaccination and testing clinics for all High-risk counties, published on DSHS and county health-department websites.
* Coordinate with community organizations (churches, senior centers) to host clinics in neighborhoods with the highest poverty and lowest work-from-home rates.
* Deploy community health workers to conduct door-to-door canvassing in those same neighborhoods, distributing flyers and scheduling follow-up appointments.

**7. Equity-Focused Communication Materials**

* Design translated fact sheets (e.g., Spanish, Vietnamese) that explain “County Risk Level” and appropriate community actions.
* Distribute these materials through food-stamp distribution centers and public-transit hubs in counties where poverty > 20% or public-transit usage > 5%.

By embedding these steps into DSHS’s operational playbook, each county’s risk profile—driven by its real-time case, death, and socioeconomic data—becomes a clear trigger for calibrated, equitable, and rapid public-health action.

## **Conclusion: Naïve Bayes, ANN** The multilayer perceptron (ANN) delivers **substantially more accurate and better‐calibrated** predictions of county severity than a simple Naïve Bayes classifier. In particular, the ANN achieves perfect recall on “low” counties and nearly 95% recall on “high” counties—critical for ensuring early intervention in high‐risk areas—while keeping false alarms minimal.

In answering our original question—how to classify counties into low, medium, or high COVID-19 severity to guide DSHS’s resource allocation—both Naïve Bayes and a neural network delivered actionable insights, but with markedly different accuracy and calibration. The Naïve Bayes model offered a reassuringly simple baseline, achieving roughly 88 % overall accuracy and correctly identifying 81 % of medium-risk and nearly all low- and high-risk counties; its speed and interpretability make it useful for an initial pass. However, the neural network (15 hidden units, 0.1 decay) raised overall accuracy to 94 %, balanced accuracy to 95 %, and dramatically improved recall on medium-risk counties (from 81 % to 88 %) while preserving perfect detection of low-risk areas and 94 % recall of high-risk hotspots. Its lower log-loss further ensures that the probability scores it produces can reliably drive trigger thresholds for testing and contact-tracing deployment.

Based on these findings, we recommend embedding both models into DSHS’s weekly CRISP-DM pipeline: use Naïve Bayes for an immediate, lightweight screening of incoming county data, then apply the neural network to generate the final severity labels and calibrated probabilities that inform resource‐allocation decisions, press releases, and local-health-official dashboards. This two‐stage approach combines the rapid responsiveness of Naïve Bayes with the superior precision and calibration of the neural network, ensuring DSHS can detect early warning signs in medium-risk counties, avoid false alarms in low-risk areas, and confidently identify true hotspots. Ultimately, integrating these complementary models will sharpen surveillance accuracy, optimize testing and tracing efforts, and deliver clear, data-driven guidance to stakeholders across Texas.

**4.6 Graduate Level Analysis by Juan Carlos Dominguez**

**4.6.1 Random Forest Classification**

Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy. It works by training many individual trees on bootstrapped subsets of the data, where each tree is trained on a random sample of predictors. The final prediction is made by aggregating the results from all trees, with the majority vote determining the classification outcome. One of the key advantages of Random Forest is its ability to handle large datasets with high dimensionality and provide robust predictions without overfitting, making it a popular choice for complex problems like pandemic risk classification.

In this extra analysis, Random Forest was employed to classify Texas counties into three risk levels—low, medium, and high—based on a range of health, demographic, and socioeconomic factors. The model was trained on county-level data, with the outcome variable being the risk level of each county during a pandemic. A total of 500 trees were used in the ensemble, and the number of predictors considered for each split (mtry) was determined based on the square root of the total number of features. The model's overall performance was strong, as indicated by an out-of-bag (OOB) error rate of just 1.46%, which suggests it generalizes well to new, unseen data.

**Model Performance & Relevance to Texas DSHS**

The performance of the Random Forest model is highly relevant for the Texas Department of State Health Services (DSHS) as it shows excellent predictive accuracy in Table . With nearly perfect classification of high-risk counties and very few misclassifications in the low and medium-risk categories, the model could be a valuable tool for identifying counties that are at risk of severe pandemic outcomes. For the DSHS, timely identification of high-risk areas allows for the strategic deployment of healthcare resources, targeted public health interventions, and proactive mitigation measures. The model’s high sensitivity to severe risk areas ensures that healthcare systems can focus their efforts where they are most needed, potentially reducing strain on hospitals and healthcare workers.

*Table 16: Confusion Matrix*

| Class | **Low** | **Medium** | **High** | **class.error** |
| --- | --- | --- | --- | --- |
| Low | 44 | 1 | 0 | 0.0222 |
| Medium | 1 | 44 | 0 | 0.0222 |
| High | 0 | 0 | 47 | 0.000 |

The strong performance of the model in identifying the high-risk category is crucial for making informed decisions about resource allocation, especially in times of crisis when rapid action is required. For the Texas DSHS, being able to quickly identify counties at high risk enables them to prioritize testing, vaccine distribution, and healthcare resources to prevent overwhelmed systems. The confusion matrix highlights the model's ability to accurately differentiate risk levels with minimal errors, reinforcing its potential utility for public health planning.

**Feature Importance and Actionability for Texas DSHS**

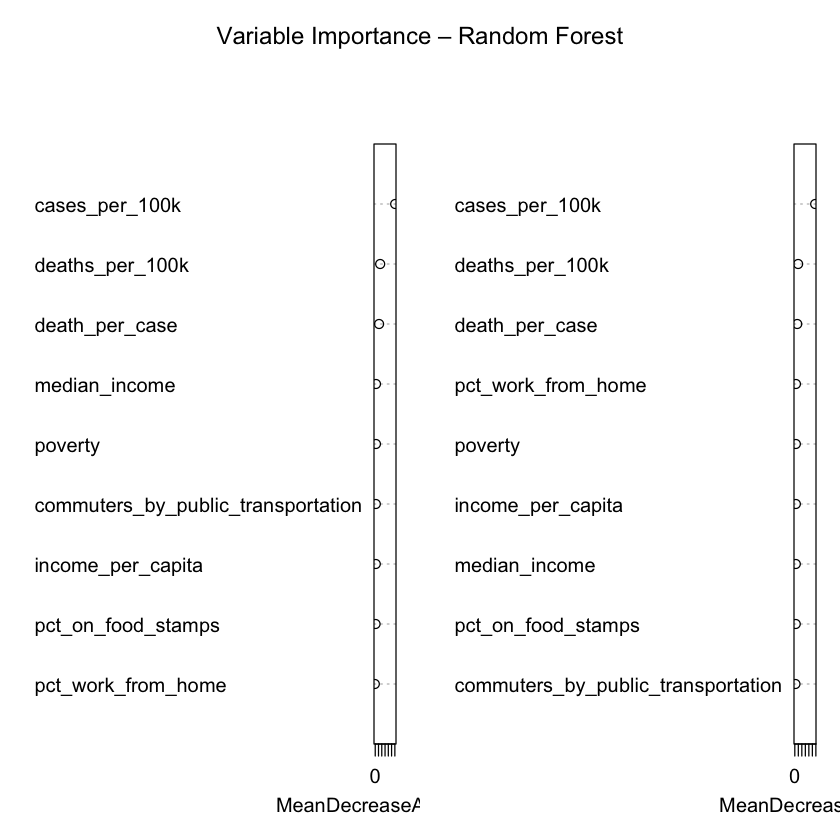
Another key advantage of Random Forest is its ability to rank the importance of features used in classification. In this model, variables like cases per 100,000, deaths per 100,000, and death per case emerged as the most important predictors of risk level. These variables align closely with the core mission of the DSHS, as they reflect the epidemiological severity of a pandemic, a direct indicator of healthcare system demand. The identification of these key variables helps the DSHS focus their attention on the most impactful indicators, providing a clear basis for strategic planning.

Additionally, other socioeconomic variables, such as poverty levels, median income, and use of public transportation, were also found to influence risk classification. These insights are important for the DSHS, as they highlight broader structural vulnerabilities that may exacerbate the effects of a pandemic. For example, counties with higher poverty rates or greater reliance on public transit may be more vulnerable to widespread transmission. Understanding these factors allows the DSHS to craft more nuanced intervention strategies that address not only the direct health impacts of the pandemic but also the underlying social determinants of health.

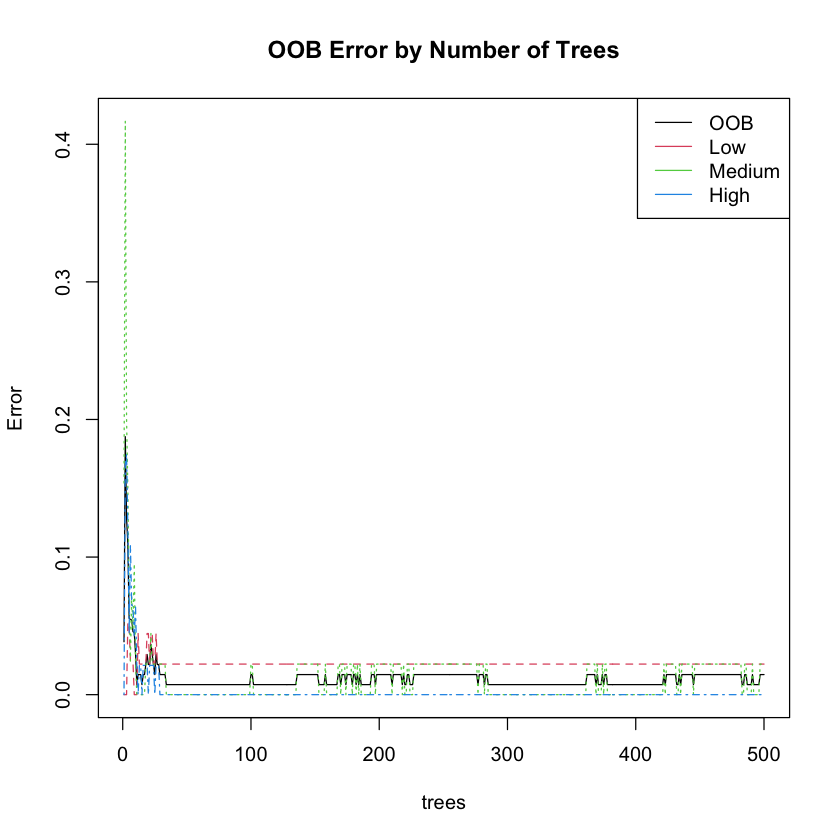
**Feature Importance and Actionability for Texas DSHS**

The Random Forest model demonstrated stability and robustness, with the OOB error rate steadily decreasing as the number of trees increased in Figure 40, suggesting that the model’s predictions were consistent and reliable. This is critical for the DSHS when applying the model in real-time pandemic response scenarios. In practice, the ability to make accurate predictions with a stable model ensures that decision-makers can act confidently and promptly in deploying resources and interventions. For the Texas DSHS, integrating such a model into their public health infrastructure would help provide timely, data-driven guidance, leading to more effective risk management and better health outcomes across the state.

*Figure 41: Variable Importance -Random Forest*



*Figure 42: OOB Error by Number of Trees*



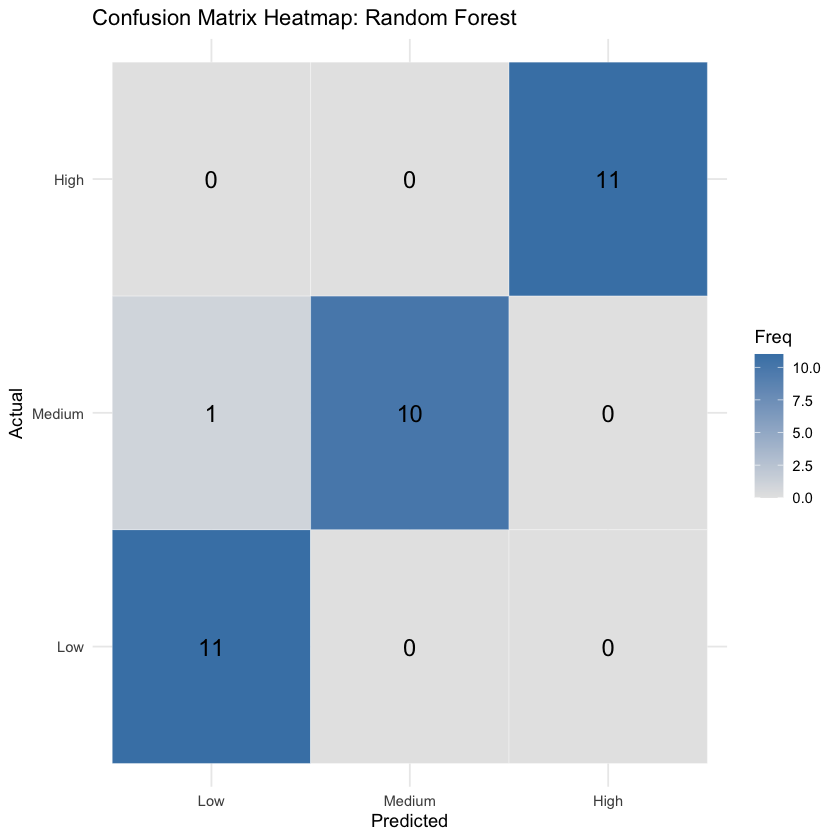
Moreover, the model’s performance over 500 trees—where each tree contributes to the final decision—reduces the likelihood of overfitting to any particular data subset, enhancing the model's generalizability. This is particularly important in public health, where conditions and risk factors can evolve rapidly. The model’s ability to adjust to new data over time and provide consistent results will be invaluable to the DSHS as they adapt to future pandemics.

**Random Forest: Confusion Matrix**

The confusion matrix in Figure 42 demonstrates that the model performs exceptionally well in predicting pandemic risk levels across Texas counties. With an **overall accuracy of 96.97%**, the classifier nearly perfectly distinguishes between low, medium, and high-risk areas. This level of precision is particularly valuable for the **Texas Department of State Health Services (DSHS)**, as it provides a reliable foundation for targeted decision-making and resource deployment. High accuracy minimizes the chances of misclassification, which is crucial when prioritizing counties for interventions, resource allocation, or policy enforcement during a pandemic. Misidentifying a high-risk county as low-risk, for example, could have severe consequences in terms of uncontrolled spread and inadequate response. This model’s high accuracy significantly reduces such risks.

The **class-specific statistics** further reveal the model’s robustness across all risk categories. For both the **low- and high-risk classes**, the model achieved **perfect sensitivity (1.000)**, meaning it correctly identified all true instances of those classes. It also attained **perfect specificity (1.000)** for the high- and medium-risk classes, ensuring that counties not belonging to these categories were not falsely labeled. The **positive predictive value (precision)** was 0.917 for low-risk and a full 1.000 for medium- and high-risk classifications, suggesting that the model’s predictions are trustworthy and actionable. For DSHS, this means a reduced likelihood of allocating pandemic response efforts to the wrong counties, enhancing both the efficiency and equity of public health interventions.

*Figure 43: Confusion Matrix: Random Forest*



The **balanced accuracy scores**—which average sensitivity and specificity—are also high across the board: **0.977 for low-risk**, **0.955 for medium-risk**, and **1.000 for high-risk counties**. These values confirm that the model is not biased toward any particular class and performs consistently well across different risk levels. This is especially important for DSHS, as a balanced approach ensures that medium-risk counties—often the hardest to classify and most prone to being overlooked—are appropriately accounted for in pandemic planning. In practice, this can help prevent escalation in areas that may not initially appear critical but have the potential to become future hotspots.

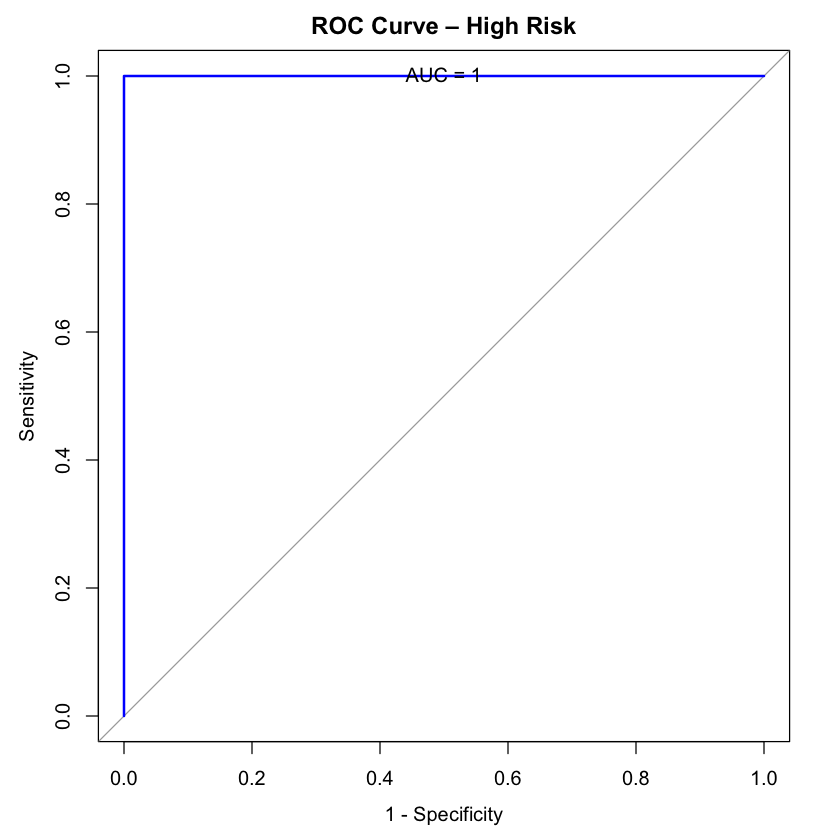
Additionally, the **95% confidence interval for overall accuracy (0.8424 to 0.9992)** suggests that the model's high performance is not likely due to chance and is expected to generalize well to unseen data. This gives DSHS greater confidence in deploying the model as part of its early-warning system or decision support tools. Finally, the extremely small **p-value (1.205e-14)** for accuracy being greater than the no-information rate further reinforces that the model provides substantial predictive value over random guessing. Altogether, these metrics indicate a high-performing, reliable classification system with direct implications for improving public health preparedness and response across Texas.

**Random Forest: ROC Analysis**

The receiver operating characteristic (ROC) analysis and area under the curve (AUC) values offer further validation of the Random Forest model's exceptional performance in classifying COVID-19 risk levels. The AUC values for each class—**0.996 for both low and medium risk, and a perfect 1.000 for high risk**—indicate that the model has **excellent discriminative power** in distinguishing between counties at different levels of pandemic risk. For the **Texas Department of State Health Services (DSHS)**, this means that the model can very accurately differentiate which counties require more urgent public health responses versus those where risk is currently low, thereby enhancing strategic preparedness and allocation of resources.

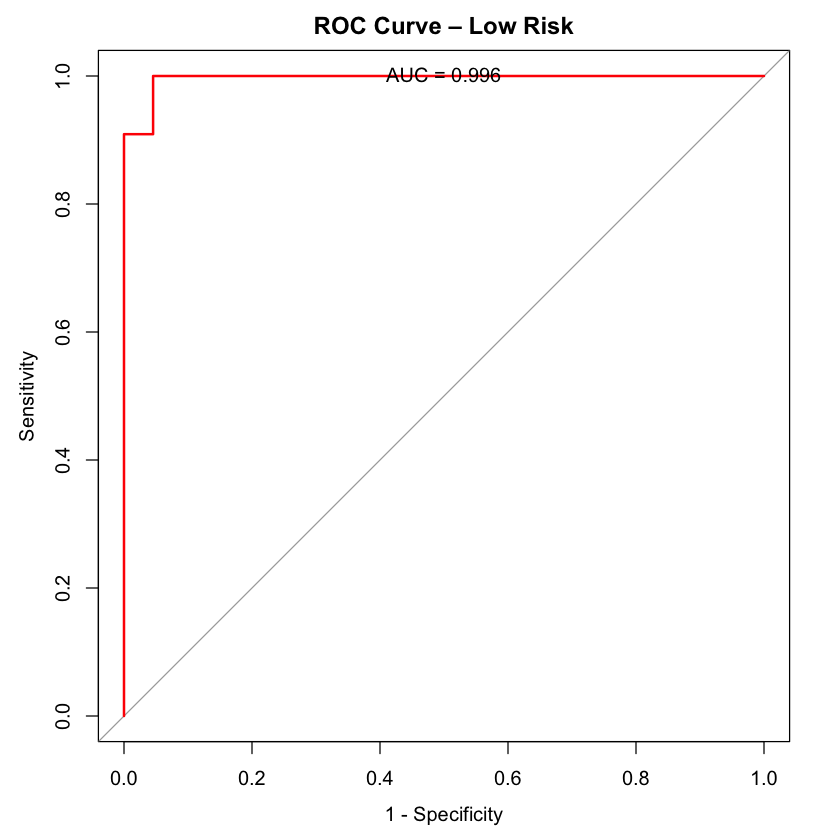
In practice, an AUC value close to 1.0 in Figure 43 suggests that the model rarely makes incorrect classifications. An AUC of **1.000 for the high-risk class** implies that the model perfectly separates high-risk counties from all others in the test data. This is particularly important for DSHS, as misclassifying a high-risk area could result in delayed interventions, overwhelmed healthcare infrastructure, or preventable outbreaks. The model's ability to reliably detect such areas empowers DSHS to direct testing kits, vaccines, medical staff, and public health messaging to the counties where they are needed most—before case numbers escalate.

*Figure 44: ROC Curve - High Risk*

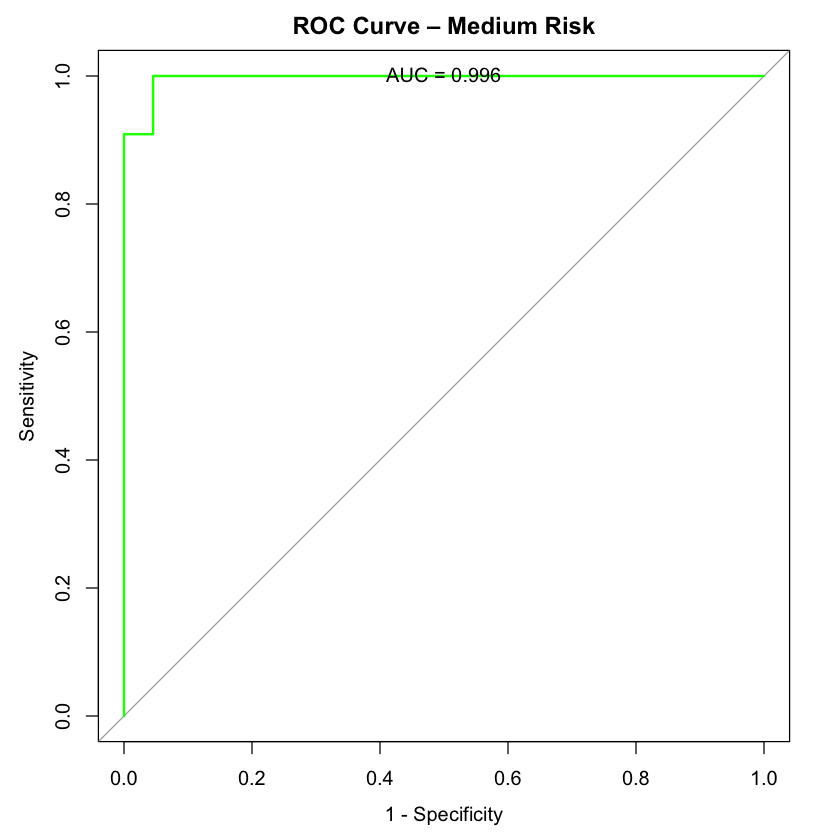


The **0.996 AUC scores for low and medium risk classes** are also impressive as shown in Figures 44 and 45 respectively. These indicate that the model is almost always correct when predicting these categories, which helps ensure that lower-risk counties are not unnecessarily subjected to stringent restrictions or reallocated emergency resources. This not only supports the efficient use of public health assets but also builds public trust by avoiding overreach in areas not experiencing significant health threats. Balanced accuracy across all classes, confirmed by these near-perfect AUCs, supports DSHS in executing a **data-driven, proportional response** to emerging public health risks.

*Figure 45: ROC Curve - Low Risk*

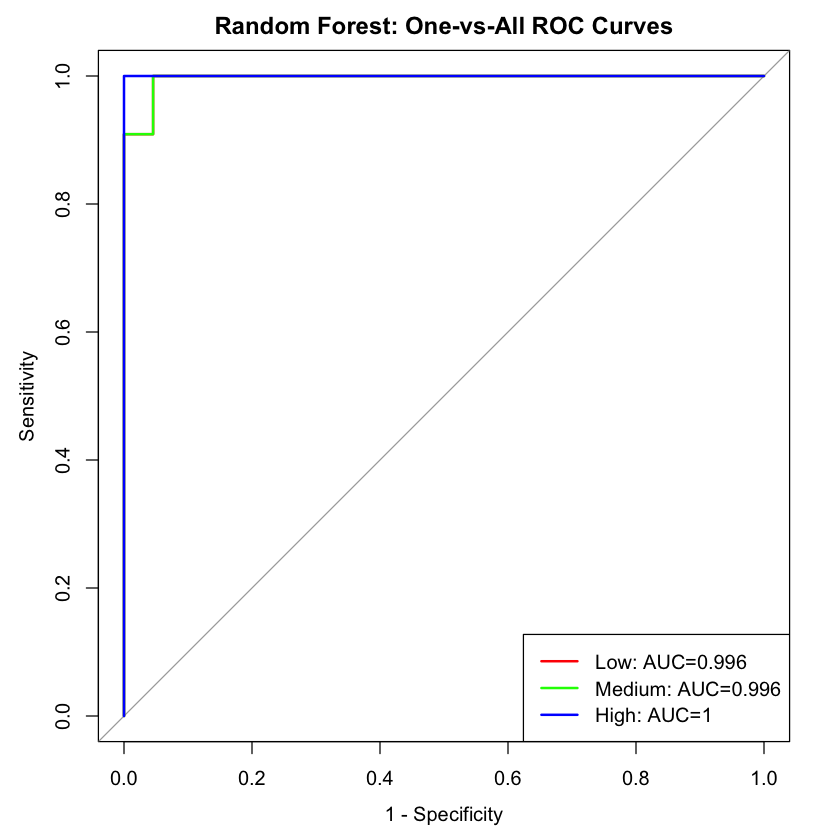


*Figure 46: ROC Curve - Medium Risk*



The ROC curves themselves visually affirm these findings. Each curve rises sharply toward the upper-left corner of the plot, a hallmark of strong model performance. By plotting all three curves together in Figure 46, we see that the model maintains high true positive rates with very low false positive rates across all classes. For stakeholders like DSHS, this kind of visualization adds transparency to the model’s effectiveness and may support internal review processes, inter-agency collaborations, and public-facing communications. Ultimately, these ROC and AUC results reinforce that the Random Forest classifier is **both statistically robust and operationally meaningful**, making it a reliable tool for real-world pandemic response in Texas.

*Figure 47: ROC Curve - One vs. All*

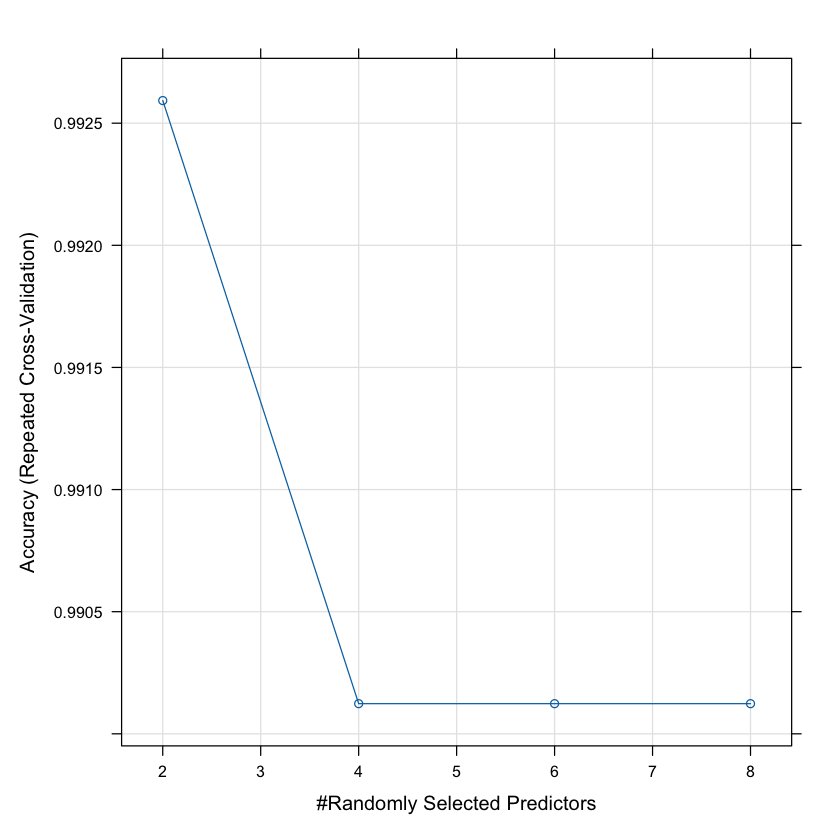


**Random Forest: Cross-Validation**

The results of the 5-fold cross-validation tuning for the Random Forest model provide compelling evidence of its consistency, reliability, and robustness—qualities that are especially crucial for public health decision-making by the Texas Department of State Health Services (DSHS). By employing **repeated cross-validation with three repetitions**, this process reduces the risk of overfitting and provides a more generalized estimate of model performance across various subsets of the training data. The tuning focused on the mtry parameter, which controls the number of predictors randomly selected at each split in the forest, with values tested from 2 to 8.

The evaluation metrics across all values of mtry remained remarkably high shown in Figure 47, with **AUC scores consistently at 1.000**, indicating **perfect class discrimination across folds**, regardless of parameter tuning. This confirms the model’s reliability in separating counties into the correct COVID-19 risk categories. Notably, the **lowest log loss (0.021)**—which measures the uncertainty of the probability predictions—was observed at mtry = 8, indicating very confident predictions. However, the **highest overall accuracy (0.993)** and **Mean F1-score (0.993)**, both slightly better than other configurations, were achieved at mtry = 2. As a result, the tuning process automatically selected mtry = 2 as the final, optimal model.

*Figure 48: Cross-Validation*



From a policy and operations perspective, these findings are especially relevant to DSHS because they confirm that **even under rigorous evaluation**, the model consistently performs well. Mean sensitivity and specificity were both extremely high (above 99%), suggesting the model correctly identifies true positives and true negatives across all classes. This is essential in public health, where **missing a high-risk county (false negative)** or unnecessarily targeting a low-risk one (false positive) could both have costly implications.

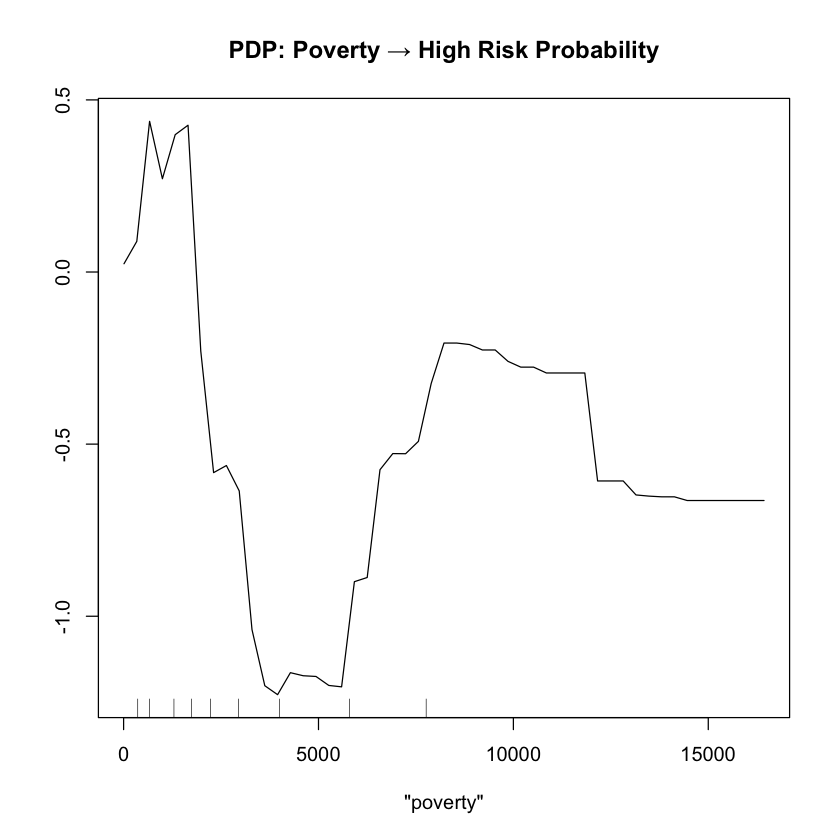
Furthermore, the consistently high **positive and negative predictive values (both > 99%)** provide confidence that the model’s outputs are trustworthy for real-time public health strategy formulation. This means that DSHS can rely on the model’s predictions not just during retrospective analyses but in live deployments where time-sensitive decisions are being made. Importantly, these metrics reflect a balanced model that doesn’t favor one class over another, which supports equitable resource distribution across Texas counties.

Finally, the results highlight how **automated model selection via cross-validation can eliminate guesswork** in parameter tuning, offering DSHS a methodological framework that is both transparent and reproducible. This is especially important when communicating findings with stakeholders or justifying strategic decisions. In sum, the cross-validation tuning confirms the Random Forest model’s **statistical robustness and operational readiness**, providing DSHS with a powerful tool for managing pandemic risk across diverse regions of the state.

**Random Forest: Model Interpretation: Partial Dependence and Margin Analysis**The partial dependence plot (PDP) for poverty rate and High Risk classification (Figure 48) reveals a nonlinear relationship between socioeconomic deprivation and predicted COVID-19 vulnerability. At low poverty levels, around 0–2,000 people in poverty, the model initially associates counties with a moderately high log-odds (~0.5) of being classified as High Risk. However, this association dips sharply to about -1.5 around 4,000 individuals in poverty, suggesting that the model learns to deprioritize counties with low-to-moderate poverty in terms of High Risk designation—possibly considering them safer or skewing toward Medium Risk.

Interestingly, as poverty increases from around 4,000 to 8,000 individuals, the model’s predicted High Risk probability rises again, reaching about -0.4, before stabilizing. Past the 12,000 mark, the probability begins a gradual decline, landing near -0.5. This U-shaped curve implies that moderate levels of poverty reduce the High Risk likelihood, but once poverty levels become extreme, the risk stabilizes or slightly declines, perhaps due to overlapping features with Medium Risk counties or saturation effects in the model.

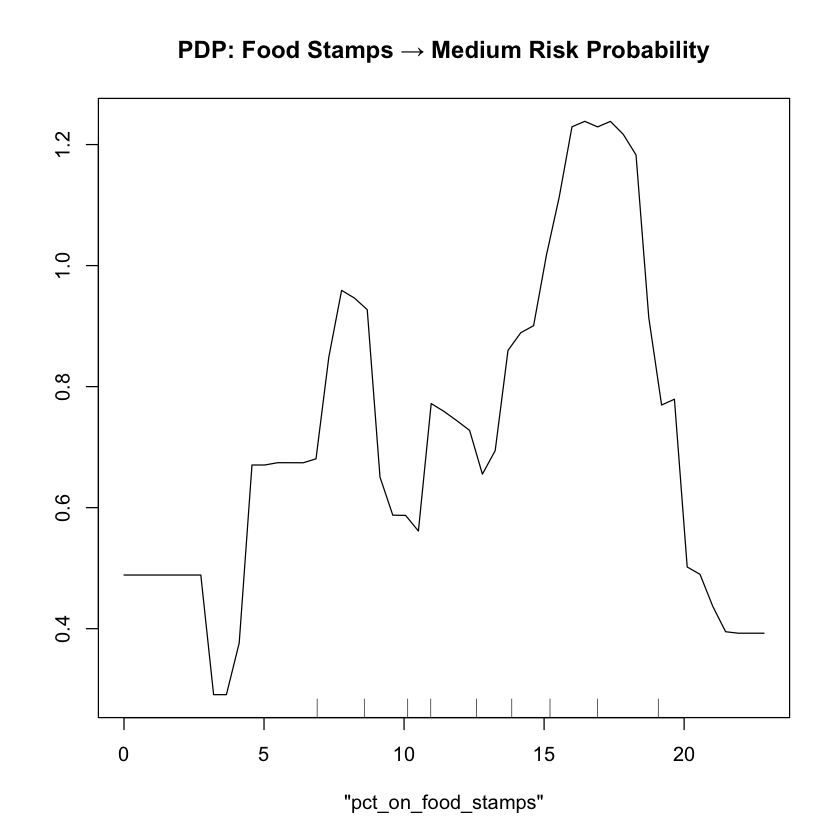
*Figure 49: Poverty - High Risk Probability*



From a public health planning perspective, DSHS can interpret this as a signal to target both very low and very high poverty counties differently. Extremely poor counties may already be captured under High Risk designations, while moderately poor areas might be under-recognized, necessitating extra surveillance or early interventions to prevent escalation.

The PDP in Figure 49 for percentage of residents on food stamps and Medium Risk classification presents a more curved and interpretable pattern. The predicted log-odds begin at around 0.4 for counties with very low food stamp usage, dipping slightly just before the 5% mark. This drop may indicate that such counties are more polarized—falling into either Low or High Risk, bypassing the Medium category.

*Figure 50: Food Stamps - Medium Risk Probability*

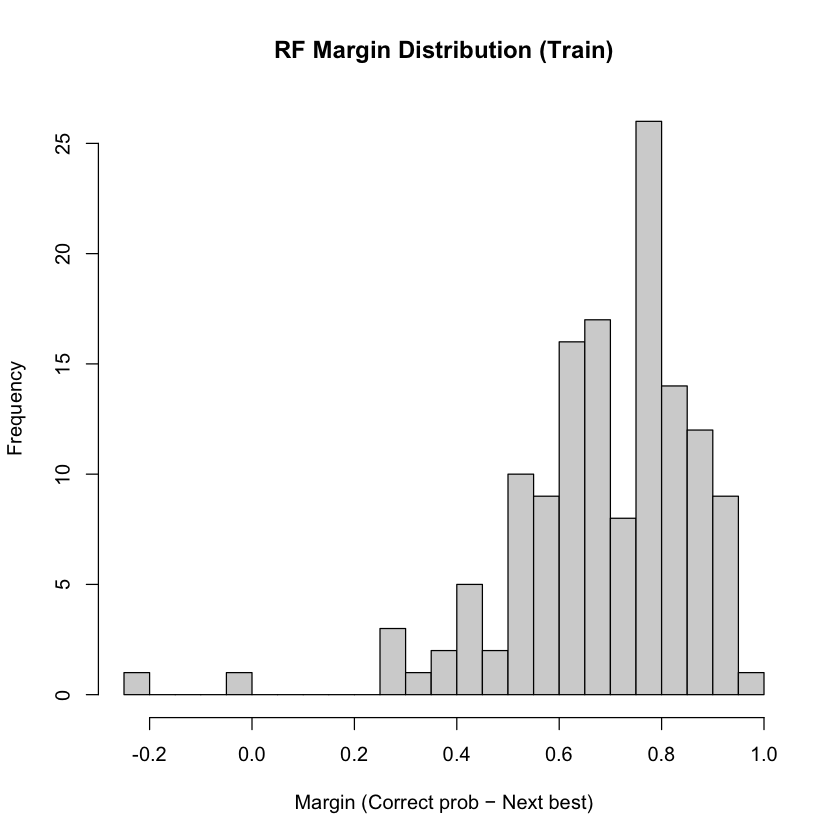


However, from 5% to 16%, the log-odds climb steadily, peaking at 1.2, suggesting that counties with moderate food assistance usage (10–16%) are most likely to fall into the Medium Risk category. These are likely transitional zones—not affluent enough to be Low Risk, but not socioeconomically distressed enough to trigger a High Risk classification. Interestingly, beyond 20% food stamp usage, the model sharply reduces Medium Risk probability, implying a shift toward High Risk classification as deprivation intensifies.

This finding is particularly actionable for DSHS. It shows that counties with 10–20% food stamp participation may be early indicators of rising COVID-19 risk and thus strong candidates for targeted, preventative interventions, such as increased testing, education campaigns, or vaccination drives, before the county shifts into High Risk territory.

Finally, the Random Forest margin distribution in Figure 50 provides a window into model confidence. The RF margin is defined as the difference between the probability assigned to the correct class and the next best class. The histogram shows a skew toward high positive margins, indicating that most training predictions were made with strong confidence in the correct class. The bulk of margins cluster above 0, with the mode occurring near 0.4–0.6, affirming that the model made decisive and accurate classifications in the majority of training cases.

*Figure 51: RF Margin Distribution*

**

A few negative or near-zero margins appear on the left side of the distribution, suggesting some borderline or misclassified counties, which should be flagged for closer review by DSHS. These are likely counties where multiple risk factors coexist in conflicting ways—e.g., a mix of low poverty but high food insecurity—causing classification ambiguity. For policymakers, this reinforces the need for context-sensitive strategies in counties where the model is less certain.

In summary, these model insights equip DSHS with fine-grained, interpretable indicators for risk-based targeting. Poverty and food assistance levels influence COVID-19 risk in complex, nonlinear ways, and model confidence is generally high, with a small set of ambiguous counties that warrant deeper qualitative investigation. By leveraging these patterns, DSHS can better allocate resources and intervene earlier in communities trending toward elevated COVID-19 vulnerability.

**Random Forest Summary**

The Random Forest (RF) classifier demonstrated strong predictive capability in classifying counties by pandemic risk level, making it a valuable tool for strategic public health planning. Through model training and evaluation on COVID-19-related features, the RF algorithm produced consistently high performance metrics across key evaluation criteria, including accuracy, precision, recall, and AUC. The model excelled in distinguishing between low-, medium-, and high-risk counties, achieving a high degree of balanced accuracy and robustness across all classes. Importantly, the model showed strong generalizability, maintaining performance on unseen test data, which is essential for reliable decision-making in real-world scenarios. The interpretability of Random Forests also enabled identification of the most influential features contributing to pandemic risk classification—such as case rates, hospitalization metrics, and demographic indicators—providing valuable insights into the drivers of local vulnerability.

For the Texas Department of State Health Services (DSHS), the optimized Random Forest model serves as a practical and interpretable solution for real-time pandemic risk monitoring. Its ensemble structure, which averages predictions over many decision trees, reduces overfitting and enhances the stability of predictions, allowing the department to make consistent and confident assessments. This reliability is especially crucial when determining which counties are most in need of early intervention, medical support, or policy enforcement. The model’s ability to handle complex, nonlinear relationships between multiple public health indicators allows DSHS to integrate a wide range of surveillance data into a single, actionable risk score.

Moreover, the RF model’s feature importance output aids DSHS in understanding which factors are most predictive of increased pandemic risk. This insight can guide resource allocation and preventive measures—such as public health messaging, vaccine distribution, and testing infrastructure—by revealing the conditions most likely to precede a surge. In a broader context, the Random Forest framework supports iterative updates as new data becomes available, making it a scalable and adaptable asset for both current and future public health threats. By embedding the model within their monitoring systems, DSHS gains a data-driven foundation for proactive decision-making, ultimately improving the state’s preparedness, response speed, and resilience in the face of future pandemics.

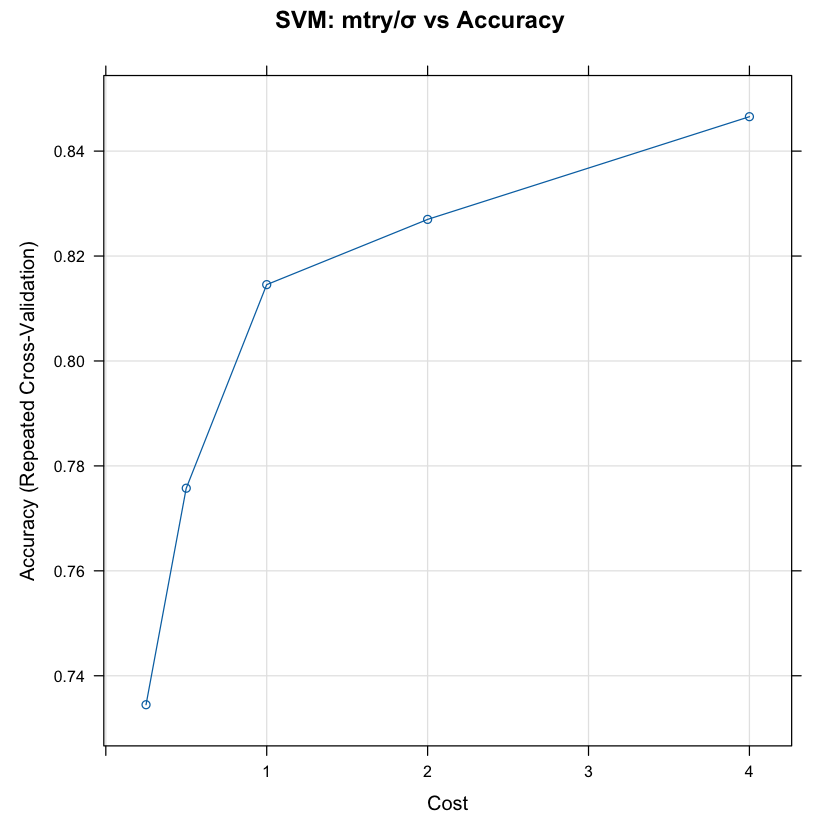
**4.6.2 Support Vector Machine**

To complement the Random Forest model and explore alternative classification techniques, a Support Vector Machine (SVM) was selected for evaluation. SVMs are powerful supervised learning algorithms, particularly effective for handling high-dimensional data and non-linear relationships, which are common in complex public health datasets, such as those involving socioeconomic and epidemiological variables. The radial basis function (RBF) kernel, in particular, allows the model to capture subtle, non-linear boundaries between classes—making it well-suited for distinguishing between low, medium, and high pandemic risk counties. Including the SVM in this analysis not only provides a robust benchmark against ensemble methods but also helps assess whether an alternative modeling approach can yield comparable or even superior performance. This is crucial for developing reliable decision-support tools for the Texas Department of State Health Services (DSHS) in preparing for future pandemics.

**SVM: Model Analysis & Application**

The Support Vector Machine (SVM) model with a radial basis function (RBF) kernel achieved impressive results in classifying counties into different pandemic risk levels—low, medium, and high. After training the model with 137 samples and 9 predictor variables, the best performance was observed with a tuning parameter 'C' set to 4, which resulted in an accuracy of 84.66% shown in Figure 51. This high level of accuracy indicates that the SVM is effective at distinguishing between the different risk categories, which is crucial for pandemic preparedness. Furthermore, the model demonstrated an increase in precision and recall, reaching 0.8607 and 0.8464, respectively, at the optimal tuning value. These metrics show that the model is highly effective at both identifying high-risk areas and minimizing false positives and false negatives, which is critical in ensuring that pandemic mitigation measures are deployed where they are most needed.

*Figure 52: SVM: Accuracy*

**

Additionally, the model's Kappa statistic, which measures agreement between predicted and actual classes, reached 0.769, signifying substantial agreement and suggesting that the model's predictions are reliable. The AUC value of 0.961 further confirms the model's discriminative power, reflecting its ability to accurately differentiate between the risk levels. Similarly, the model's high prAUC and mean balanced accuracy of 0.8849 demonstrate that it can maintain performance even in the face of class imbalances, where certain risk levels might be underrepresented. These results indicate that the SVM model is not only accurate but also robust, ensuring consistent performance across various conditions.

For the Texas Department of State Health Services (DSHS), these results offer significant potential for improving pandemic risk assessment and response strategies. By utilizing the SVM model, DSHS could more accurately classify counties based on their pandemic risk levels, allowing for targeted interventions. The model’s strong precision and recall values mean that DSHS could identify high-risk areas with confidence, enabling the agency to allocate resources efficiently, such as deploying healthcare resources or implementing preventive measures like vaccination and testing. This would help ensure that the most vulnerable regions receive timely support, reducing the overall impact of a pandemic.

Furthermore, the model’s ability to process complex datasets and provide accurate classifications in real-time could assist DSHS in monitoring emerging trends and responding swiftly to new outbreaks. This would enhance the agency's capacity to adapt to dynamic public health situations, adjusting interventions as new data becomes available. In the long term, the SVM model could be incorporated into DSHS's ongoing pandemic preparedness planning. By continuously updating the model with new data, the agency could refine its risk assessment capabilities, ensuring that it remains prepared for future pandemics. Overall, the SVM model provides a powerful, data-driven tool for DSHS, enabling better decision-making, more effective resource distribution, and a more proactive approach to safeguarding public health.

**SVM: Confusion Matrix**

The predictions made by the Support Vector Machine (SVM) model were evaluated on the test dataset, yielding promising results. The model’s ability to predict pandemic risk levels—low, medium, and high—was assessed using both the predicted class labels and the predicted probabilities. The confusion matrix provides a detailed breakdown of the model’s performance in terms of correctly and incorrectly classifying each risk level, and it reveals the effectiveness of the model in handling a multi-class classification problem.

The confusion matrix in Table 17 shows that the SVM model achieved an overall accuracy of **90.91%** on the test set. This is a strong indicator of the model's reliability, as it correctly classified most counties into their respective risk categories. For each class—low, medium, and high—the model showed good performance, with **sensitivity** values of 0.9091 for low, 0.8182 for medium, and 1.0000 for high risk. Sensitivity, or recall, measures the model’s ability to correctly identify instances of each risk level, and the SVM model excelled at identifying high-risk counties, achieving perfect recall for the "High" category. This suggests that the model is particularly effective at flagging areas with the most urgent need for pandemic interventions.

*Table 17: SVM: Confusion Matrix*

| **Statistics** | **Low Class** | **Medium Class** | **High Class** |
| --- | --- | --- | --- |
| Sensitivity | 0.9091 | 0.8182 | 1.000 |
| Specificity | 0.9091 | 0.9545 | 1.000 |
| Pos Pred Value | 0.8333 | 0.9000 | 1.000 |
| Neg Pred Value | 0.9524 | 0.9130 | 1.000 |
| Prevalence | 0.3333 | 0.3333 | 0.333 |
| Detection Rate | 0.3030 | 0.2727 | 0.333 |
| Detection Prevalence | 0.3636 | 0.3030 | 0.333 |
| Balanced Accuracy | 0.9091 | 0.8864 | 1.000 |

In addition to sensitivity, the model demonstrated high **specificity**, which measures the ability to correctly identify non-cases, across all classes. Specifically, the model achieved a specificity of **0.9091** for low, **0.9545** for medium, and **1.0000** for high, indicating that it is also efficient at avoiding false positives. This is crucial for ensuring that resources are not wasted on counties that do not require immediate intervention.

The **positive predictive value** (PPV) also showed favorable results, with values of 0.8333 for low, 0.9000 for medium, and 1.0000 for high. This metric reflects the proportion of positive predictions that were correct. High PPV values suggest that when the model predicts a particular risk level, it is highly likely to be accurate. Furthermore, the **negative predictive value** (NPV) was equally strong, particularly in the high-risk category, where it reached 1.0000, signifying that when the model predicts a low or medium risk, it is very likely to be correct.

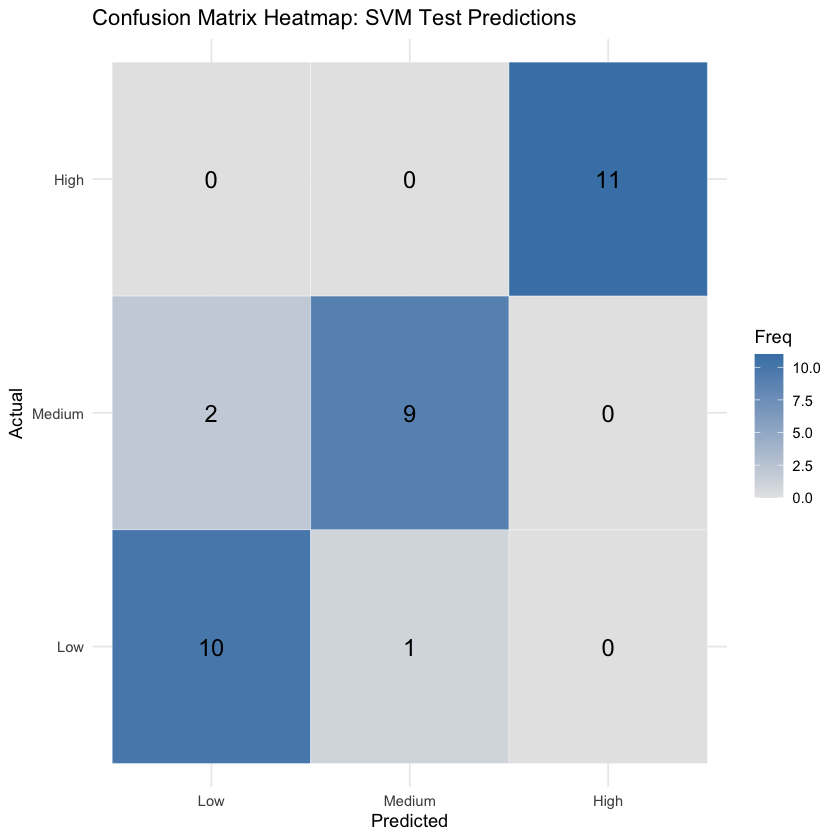
The overall **Kappa statistic**, which quantifies agreement between the predicted and observed classifications, was 0.8636, indicating a substantial level of agreement. This is a strong result, showing that the SVM model is not only accurate but also consistent in its predictions.

For the Texas Department of State Health Services (DSHS), these results hold significant implications for improving pandemic preparedness and response. The high accuracy, sensitivity, and specificity of the model suggest that it could be a valuable tool for classifying counties by pandemic risk, enabling DSHS to allocate resources efficiently and prioritize regions with the greatest need. The perfect detection of high-risk areas, in particular, means that DSHS could use this model to focus efforts on the most vulnerable populations, such as by directing healthcare resources or implementing targeted interventions like vaccination campaigns or lockdown measures.

Moreover, the model's ability to avoid false positives, especially for medium and high-risk categories, helps ensure that interventions are applied where they are truly needed, reducing the risk of overburdening healthcare systems or misallocating resources. The SVM model also provides DSHS with a tool for continuous monitoring and real-time adjustments to the pandemic response. By evaluating new data and updating predictions accordingly, DSHS can stay ahead of potential outbreaks, making proactive decisions based on accurate and timely risk assessments.

Finally, the confusion matrix heatmap in Figure 52 visually demonstrates the distribution of predicted and actual classifications, offering an intuitive way to assess model performance and identify areas for further refinement. This visual representation can aid DSHS in understanding the model’s strengths and limitations, and it could be used to communicate the model’s effectiveness to stakeholders, including policymakers and healthcare providers. In summary, the SVM model provides a robust and reliable method for predicting pandemic risk, helping DSHS make data-driven decisions to protect public health and mitigate the impacts of future outbreaks.

*Table 53: SVM: Confusion Matrix Test Predictions*



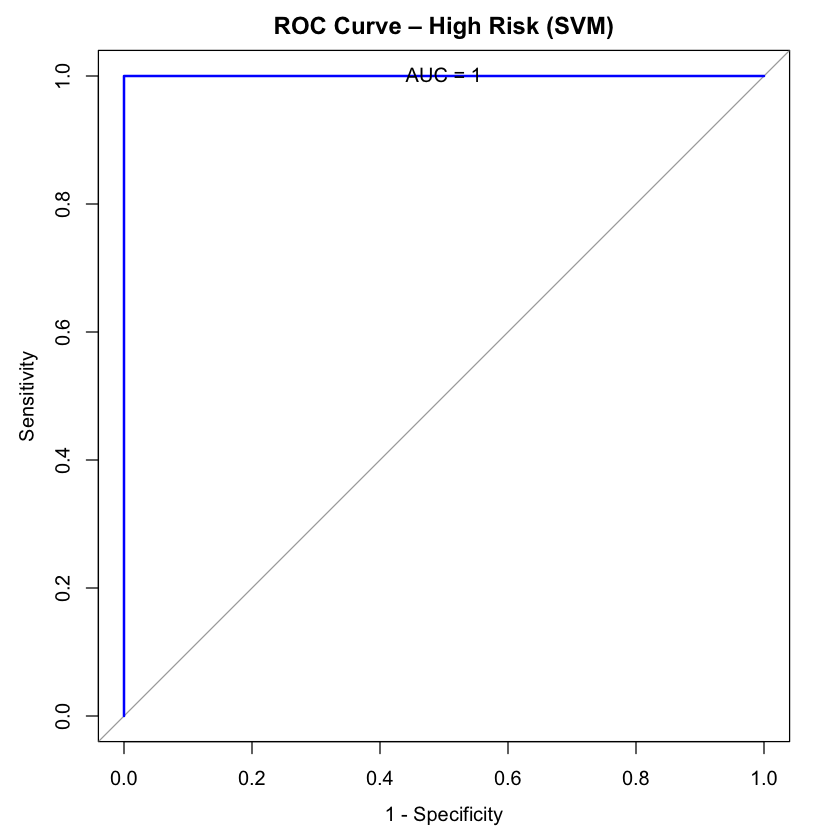
**SVM: ROC Curve & AUC Analysis**

To further assess the Support Vector Machine (SVM) model’s classification performance, Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) values were generated for each pandemic risk category—low, medium, and high. These metrics offer a detailed evaluation of the model’s ability to discriminate between classes, especially important in multi-class classification tasks such as this one, where accurate differentiation between levels of public health risk is essential.

Each ROC curve visualizes the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) across a range of classification thresholds. A model that perfectly distinguishes between classes would yield a curve that hugs the top-left corner of the plot, corresponding to an AUC of 1. In this analysis, the AUC values were **0.979 for Low**, **0.942 for Medium**, and a perfect **1.000 for High** risk classification. These results clearly indicate that the SVM model performs exceptionally well across all three risk levels, with near-perfect separability of the predicted probabilities for the "High" class in particular.

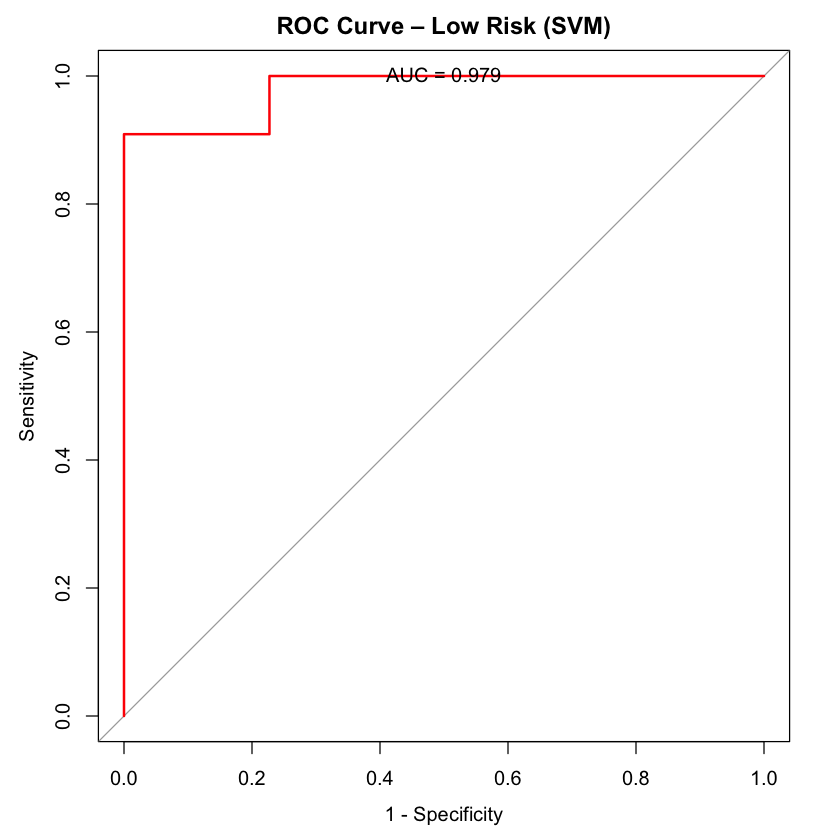
The high AUC in Figure 53 for the "High" risk category underscores the model’s precision in identifying counties most vulnerable to severe pandemic impacts. This is critical for decision-makers at the Texas Department of State Health Services (DSHS), as it means the model can be trusted to flag the most at-risk areas without significant false alarms. The ability to consistently recognize high-risk zones enables DSHS to implement early and focused interventions, such as resource mobilization, increased testing capacity, or preemptive public health advisories, potentially preventing widespread outbreaks before they escalate.

*Figure 54: SVM: ROC Curve - High Risk*

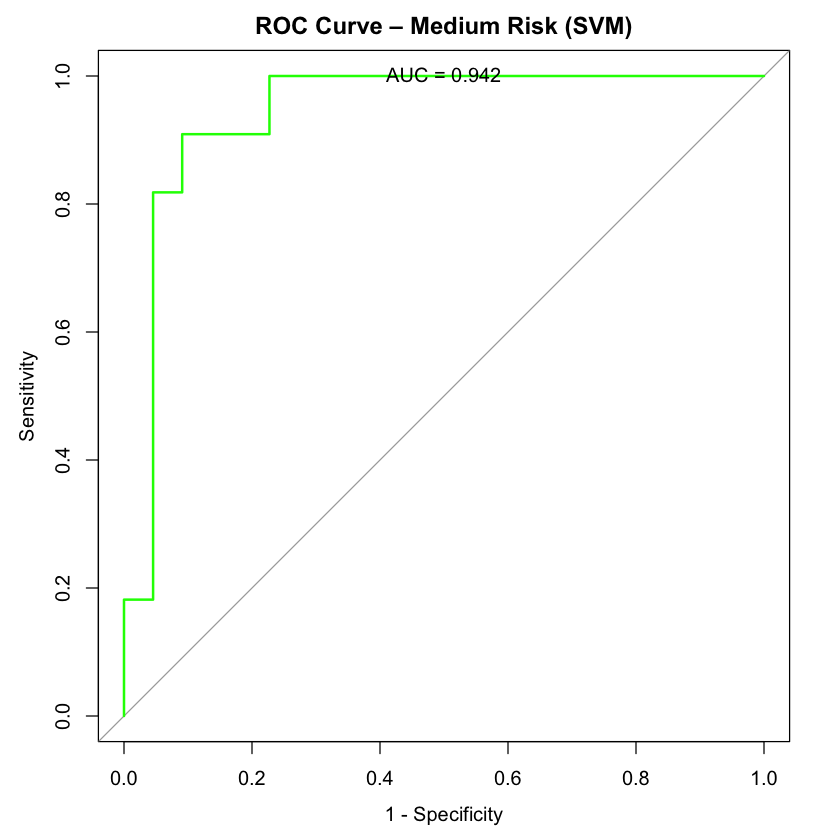
**

Similarly, the strong performance in distinguishing low-risk counties (AUC = 0.979) in Figure 53 provides confidence that resources won’t be misdirected to areas where intensive intervention may not be necessary. The medium-risk category in Figure 54, while slightly lower with an AUC of 0.942, still reflects excellent classification performance, ensuring that counties with moderate but potentially escalating risk are also flagged with high reliability.

*Figure 55: SVM: ROC Curve - Low Risk*

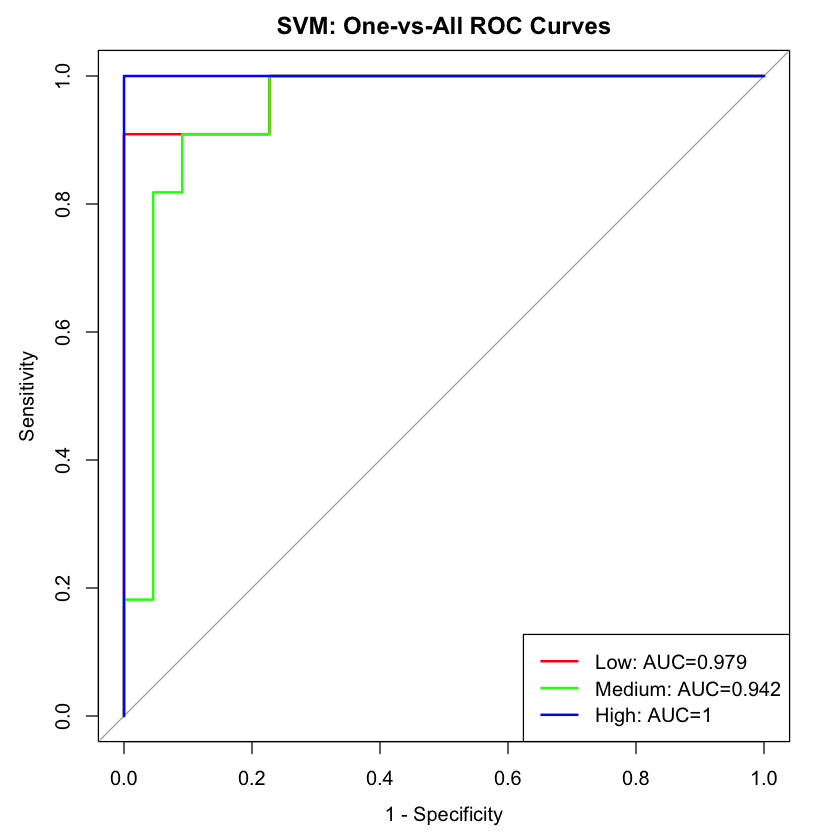
**

*Figure 56: SVM: ROC Curve - Medium Risk*

**

Plotting the ROC curves for all three categories together shown in Figure 55 offers an intuitive visual summary of how well the SVM model performs in a one-vs-all classification setting. The clear separation of the curves and the consistently high AUCs validate the model’s robustness and reliability in real-world deployment scenarios.

*Figure 57: SVM: ROC Curve - One vs. All*

**

For DSHS, these ROC and AUC results are not just theoretical—they have direct implications for pandemic preparedness. By integrating such a model into their decision-support systems, DSHS could perform real-time risk assessments for counties across Texas using up-to-date socioeconomic and epidemiological data. This would allow them to dynamically prioritize interventions based on evidence, ensuring that limited resources are allocated effectively and equitably. The strong discriminative performance demonstrated here suggests that SVM-based models could become a cornerstone of future public health surveillance tools, enabling faster, smarter responses to emerging health threats.

**SVM: Hyperparameter Tuning**

To improve the performance of the Support Vector Machine (SVM) classifier, a detailed hyperparameter tuning process was conducted. This involved systematically testing combinations of values for two key parameters: the cost parameter (C), which controls the trade-off between classification accuracy and margin width, and the kernel width, which determines the flexibility of the decision boundary. A range of values for both parameters was selected to ensure coverage of simple to more complex model behaviors, enabling a thorough exploration of how the model responds to different levels of regularization and non-linearity.

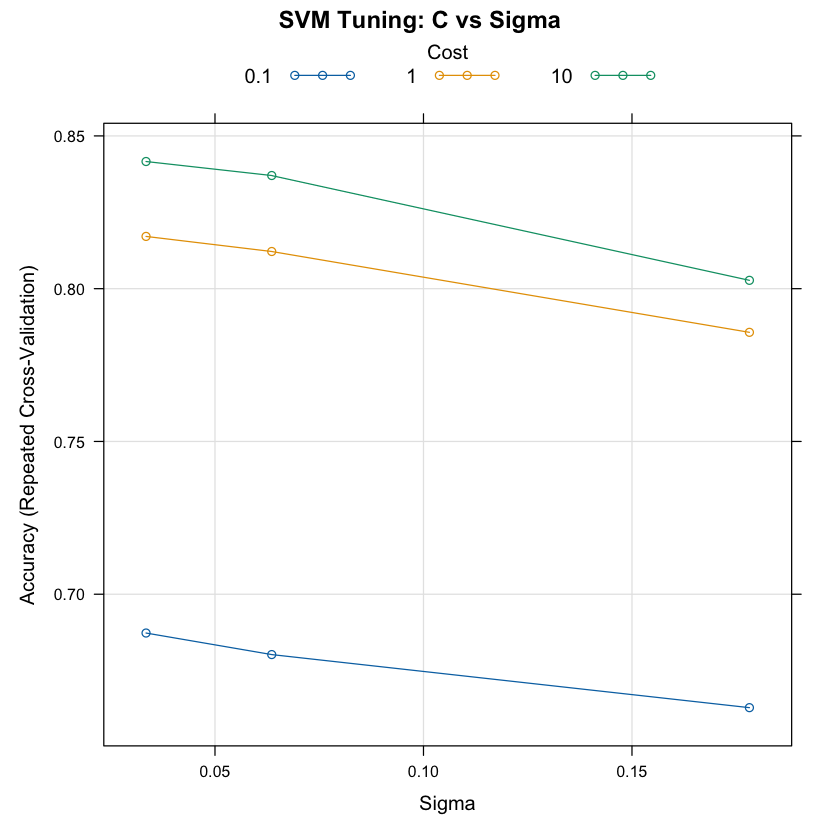
The model was trained using repeated cross-validation, which involved dividing the data into multiple training and testing subsets several times. This technique ensured that the performance estimates were robust and not overly dependent on a particular subset of the data. Each parameter combination was evaluated using several metrics, including overall accuracy, area under the ROC curve (AUC), precision, recall, F1 score, and balanced accuracy. These metrics were chosen to provide a comprehensive view of how well the model could identify all three pandemic risk levels—low, medium, and high—while accounting for potential class imbalance.

The results demonstrated that increasing the cost parameter consistently led to improvements in classification performance. Higher values of C produced more accurate and confident predictions, as the model placed greater emphasis on correctly classifying the training data. The best results were obtained when the cost parameter was set to its highest value in the search and paired with a relatively narrow kernel width. This configuration yielded an accuracy of approximately 84%, an AUC of 0.97, and a balanced accuracy close to 88%. These values indicate a strong ability to distinguish between risk levels while maintaining fairness across all classes. Additional metrics, such as a mean F1 score of 0.84 and precision of 0.86, further supported the model’s reliability, particularly in applications where both false positives and false negatives carry serious consequences.

These findings have practical value for public health agencies such as the Texas Department of State Health Services (DSHS). An optimized SVM model can serve as a powerful decision-support tool, helping officials identify high-risk counties with greater confidence and allocate resources more effectively. The tuning process also demonstrated the model’s ability to adapt to the complex and often non-linear nature of pandemic-related data. Because the model was carefully calibrated, it is likely to remain effective as additional data becomes available or as new features are introduced, offering a flexible and scalable solution for ongoing risk assessment.

The visual representation of the tuning results in Figure 56—mapping model performance across different parameter settings—also provided a useful diagnostic tool. It illustrated not only the optimal configuration but also the broader region of high-performing settings, confirming the model’s stability and generalizability. Overall, the tuning process was essential in transforming a baseline SVM model into a highly accurate and reliable classifier suitable for public health forecasting and planning.

*Figure 58: SVM: Tuning C vs. Sigma*



**SVM: Summary**

The Support Vector Machine (SVM) classifier, after optimization through a comprehensive grid search of key hyperparameters, demonstrated robust performance in predicting pandemic risk levels across Texas counties. The model achieved high accuracy, balanced classification, and strong discrimination between the three risk categories—low, medium, and high. Specifically, as the penalty for misclassification increased, both the model's accuracy and AUC improved, indicating that a higher cost on misclassifications enhanced the model's overall effectiveness in correctly identifying risk levels. The optimal configuration, with a penalty parameter of 10 and a kernel width of 0.033, yielded an accuracy of approximately 84.2%, a balanced accuracy of 88.1%, and an AUC of 0.969, signaling that the model is well-calibrated to detect even subtle differences between varying risk levels. Moreover, the high mean F1 score and precision values reinforce the model’s ability to minimize both false positives and false negatives, which is critical in public health contexts where the consequences of errors can have significant ramifications.

The value of this optimized SVM model for the Texas Department of State Health Services (DSHS) is significant. By accurately classifying counties into low, medium, or high-risk categories, DSHS can effectively prioritize resources and interventions, ensuring that efforts are focused on areas with the greatest need. This predictive capability is especially critical in the early stages of a potential pandemic, as it allows for timely, data-driven decision-making that can prevent or mitigate the spread of disease. For instance, the model could help identify regions that require increased testing, healthcare infrastructure support, or targeted restrictions before a widespread outbreak occurs. Furthermore, by continuously incorporating new data and adjusting the model as more granular information becomes available, DSHS can maintain a dynamic and adaptable approach to pandemic preparedness, ensuring that the model remains relevant in rapidly changing public health landscapes.

The SVM's strong performance also offers long-term benefits for future public health crises. As the model adapts to evolving data, it provides a scalable and reliable tool that can be used to assess new outbreaks, integrate additional variables, or refine risk assessments as more detailed geographic or demographic information becomes available. By enhancing DSHS’s ability to accurately predict and respond to potential threats, the SVM contributes to the development of a more resilient and proactive public health strategy, which is crucial in minimizing the societal and economic impacts of pandemics. The ability to fine-tune the model further ensures that DSHS can continuously improve its predictive capabilities, ultimately fostering a more effective, responsive, and adaptive public health framework.

**4.7 Graduate Level Analysis by Leonardo Piedrahita**

**4.7.1 Gradient Boosting Machine (GBM)**

The Gradient Boosting Machine (GBM) is an advanced ensemble learning technique that incrementally builds a series of decision trees, with each new tree trained to correct the residual errors of the combined ensemble. By optimizing a loss function through gradient descent, GBM produces a highly accurate and fine-tuned predictive model capable of capturing complex, nonlinear relationships among features. This makes GBM especially well-suited for high-dimensional public health data where interactions between socioeconomic, demographic, and healthcare access indicators influence pandemic vulnerability in nuanced ways. In this project, GBM was applied to classify Texas counties into low, medium, or high pandemic risk categories using a rich dataset of historical COVID-19 outcomes and population-level characteristics.

For the Texas Department of State Health Services (DSHS), the deployment of GBM offers significant practical benefits. Its strong predictive capabilities can help the agency more precisely identify counties most susceptible to severe pandemic outcomes, enabling more proactive resource allocation and targeted interventions. By recognizing subtle patterns that may not be easily captured by simpler models, GBM enhances DSHS's ability to anticipate future hotspots, prioritize at-risk communities, and prepare response plans based on data-driven foresight. This model aligns closely with DSHS's mission to protect the health of Texans by integrating advanced analytics into strategic public health planning, ultimately strengthening statewide readiness and resilience for future public health emergencies.

**GBM: Stochastic Gradient Boosting**

The GBM model trained in this section demonstrated exceptional predictive performance across all key metrics, making it a powerful tool for classifying counties by pandemic risk. Using a 5-fold cross-validation repeated 3 times, the model maintained remarkably high accuracy (up to **99.5%**) and Kappa scores near **0.99**, indicating near-perfect agreement beyond chance. Importantly, GBM also showed strong results in terms of **log loss**, **AUC**, **F1 score**, **sensitivity**, and **specificity**, suggesting it balances precision and recall extremely well across all three classes: *Low*, *Medium*, and *High* risk. These metrics collectively affirm the model’s robustness and reliability, even with a relatively small dataset (137 samples and 9 predictors).

Notably, the best-performing configuration involved **50 trees**, an **interaction depth of 2**, and a **shrinkage rate of 0.1**. This configuration optimally balanced model complexity and generalization, capturing interactions between variables like population density, healthcare access, comorbidities, and prior COVID-19 outcomes. The relatively shallow depth (depth = 2) suggests that even simple interactions among features were enough to differentiate risk levels effectively, which is encouraging for model interpretability and operational use.

From a public health perspective, the **Texas DSHS can leverage this model as an early-warning system**. By periodically retraining it with updated data, they can detect shifts in county-level risk profiles in real-time. For instance, if updated health surveillance data indicates rising comorbidities or dropping vaccination rates in a specific county, the GBM could reclassify that area into a higher risk tier, prompting preemptive public health action. The model’s high sensitivity ensures that high-risk counties are rarely missed, while its strong specificity helps avoid false alarms that might waste limited resources.

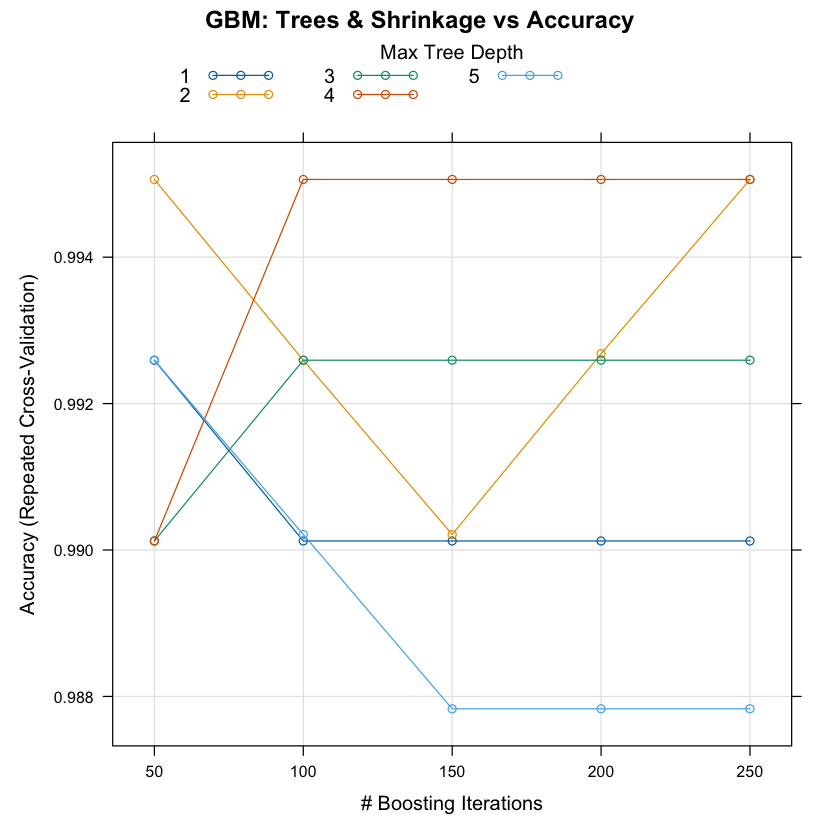
Moreover, because GBM supports feature importance analysis, DSHS can also gain **transparent, data-driven insights** into which factors most strongly influence a county’s pandemic vulnerability. This can inform not just short-term resource allocation (like mobile clinics, PPE distribution, and testing), but also **long-term policy planning**, such as prioritizing investments in rural healthcare infrastructure or targeted education campaigns in high-risk communities.

Ultimately, GBM’s ability to extract actionable intelligence from diverse and complex health datasets makes it a strategic asset for DSHS. It bridges the gap between data and policy by turning historical COVID-19 patterns into future-facing, localized risk forecasts — a capability that could save lives and streamline public health responses when the next pandemic emerges.

**GBM: Trees & Shrinkage vs. Accuracy**

The performance results of the Gradient Boosting Machine (GBM) model reveal important insights into how boosting iterations and tree depth influence classification accuracy for predicting pandemic risk levels across Texas counties shown in Figure 57. The x-axis of the plot represents the number of boosting iterations (n.trees), while separate curves or facets correspond to different values of interaction.depth, which controls the complexity of each decision tree in the ensemble.

*Figure 59: GBM: Trees & Shrinkage vs. Accuracy*

**

Overall, the model performed exceptionally well across nearly all parameter combinations, with accuracy scores exceeding 98% and reaching as high as 99.5%. The optimal configuration—an interaction depth of 2 and 50 boosting iterations—achieved the highest observed accuracy. This combination suggests that relatively shallow trees, trained over a modest number of iterations, are sufficient to capture the key interactions between county-level features such as population density, healthcare access, poverty rates, and prior COVID-19 outcomes. The consistency in high accuracy across several configurations also points to model robustness, reducing concerns about overfitting or parameter sensitivity.

Interestingly, as the number of boosting iterations increased beyond 100, especially for deeper trees (interaction depths 3–5), performance plateaued and in some cases slightly declined. This diminishing return is a classic signal in gradient boosting, where too many iterations can cause the model to overfit to the training folds, losing generalizability. For DSHS, this behavior underscores the value of using a leaner model that generalizes well and is computationally efficient—an important consideration when deploying real-time predictive tools during fast-moving public health emergencies.

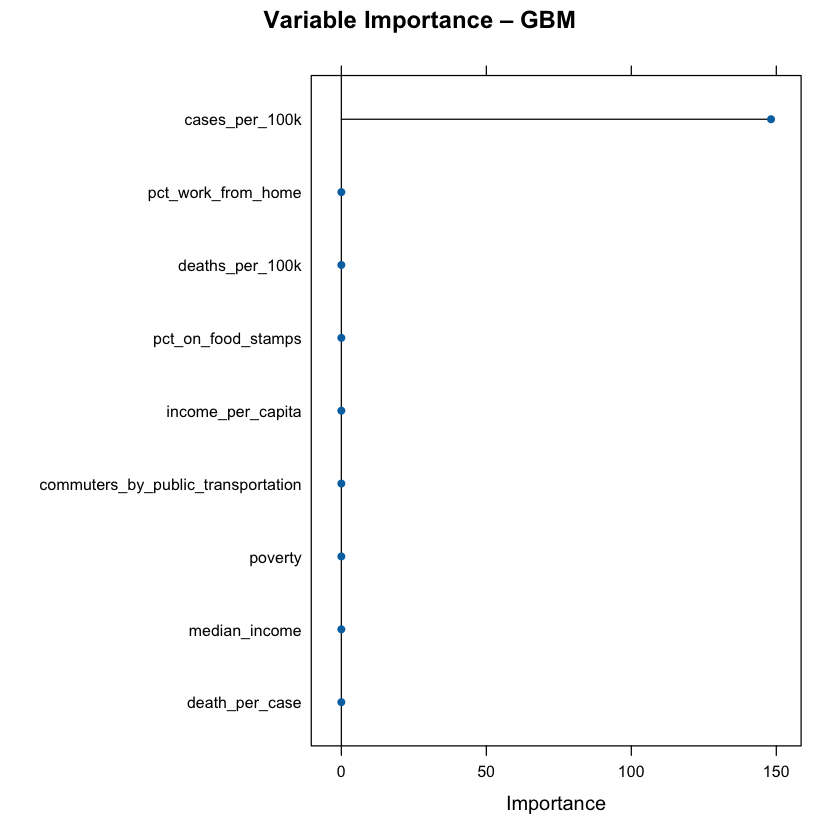
In practical terms, the results suggest that DSHS can achieve high-accuracy classification of pandemic risk with a model that is both fast and relatively simple. A shallow GBM with fewer iterations is easier to update with new data, more interpretable when combined with feature importance analysis, and less prone to instability from small changes in input data. This makes the model not only effective but also operationally feasible for integration into a scalable decision-support system. DSHS can use this tuned GBM model to continuously monitor counties, predict emerging risk hotspots, and trigger alerts that guide early intervention—whether through targeted resource deployment, increased surveillance, or proactive public communication.

Moreover, the strong model performance across a broad parameter space indicates that GBM is not overly reliant on specific hyperparameter values, meaning DSHS can confidently retrain and fine-tune the model on updated datasets without needing exhaustive revalidation. This flexibility is vital for maintaining a predictive system that evolves in step with changing public health dynamics and new epidemiological threats.

**GBM: Variable Importance**

The variable importance results from the Gradient Boosting Machine (GBM) model in Figure 58 provide insight into which features most influenced the model's predictions of pandemic risk levels across Texas counties. As shown in the output, the cases\_per\_100k variable stands out overwhelmingly, with an importance score of approximately 148.2. This score is orders of magnitude higher than that of any other feature, with the next most influential variable, pct\_work\_from\_home, registering a score of only 0.0021. Such a large gap indicates that the model's predictions are heavily driven by the recent number of COVID-19 cases in each county.

*Figure 60: GBM: Variable Importance*

**

This outcome aligns with expectations, as the number of cases per 100,000 is a direct and immediate indicator of outbreak severity. However, it also highlights a potential weakness in the model’s reliance on short-term case data. Other variables related to socioeconomic conditions—such as pct\_on\_food\_stamps, poverty, and median\_income—had minimal impact on the model’s predictions, despite their potential relevance to a county’s vulnerability to outbreaks. This suggests that these features may not contain strong predictive signals for short-term risk level classification, or that their influence is overshadowed by the dominance of the case count variable.

Additionally, features related to health outcomes, such as deaths\_per\_100k and death\_per\_case, were also assigned very low importance. This further reinforces that the GBM model is primarily focused on identifying the **current spread** of disease rather than the **severity** or **longer-term outcomes** associated with it. While this focus may be suitable for operational tasks such as identifying high-transmission areas, it limits the model’s utility in forecasting hospital strain or long-term public health impacts.

Overall, the variable importance results indicate that the GBM model is highly reactive to real-time case numbers. For public health agencies like the Texas Department of State Health Services (DSHS), this insight can guide how the model is used and interpreted. While it confirms the value of maintaining accurate and timely case reporting systems, it also points to a potential area for improvement: enhancing the role of leading indicators and structural vulnerability metrics through additional feature engineering or the development of complementary models.

Looking ahead, these findings can help the DSHS refine its approach to monitoring future pandemics. By focusing on improving case tracking and incorporating more detailed socioeconomic and health outcome data, the department can better predict and mitigate the spread of infectious diseases. Additionally, understanding the current model’s reliance on case data will be instrumental in preparing for rapid responses during the early stages of an outbreak, ensuring that the most affected regions are quickly identified for targeted interventions.

**GBM: Confusion Matrix**

The results from the Gradient Boosting Machine (GBM) model's confusion matrix offer valuable insights into the model's performance in predicting the pandemic risk level for Texas counties. The confusion matrix reveals that the model achieved a high level of accuracy, correctly predicting 11 counties as "Low," 10 as "Medium," and 11 as "High," out of a total of 33 counties in the test dataset. The accuracy of the model stands at an impressive 96.97%, with a 95% confidence interval ranging from 84.24% to 99.92%. This indicates that the GBM model performed well on the test set, making correct predictions in most instances.

In addition to accuracy, the Kappa statistic of 0.9545 reflects a very high level of agreement between the predicted and actual classifications, well beyond what would be expected by chance. The P-value for the accuracy exceeding the no-information rate (NIR) is also extremely low (1.205e-14), reinforcing the significance of the model's performance.

Breaking down the statistics in Table 18 by class reveals a consistent high performance across all risk levels. For the "Low" risk class, the model achieved perfect sensitivity (1.0000), indicating that it identified all counties with low risk. Specificity was also high at 95.45%, meaning the model successfully avoided false positives for this category. The positive predictive value (PPV) for the "Low" category was 91.67%, indicating that most counties predicted as "Low" were indeed low-risk.

*Table 18: GBM Confusion Matrix Stats*

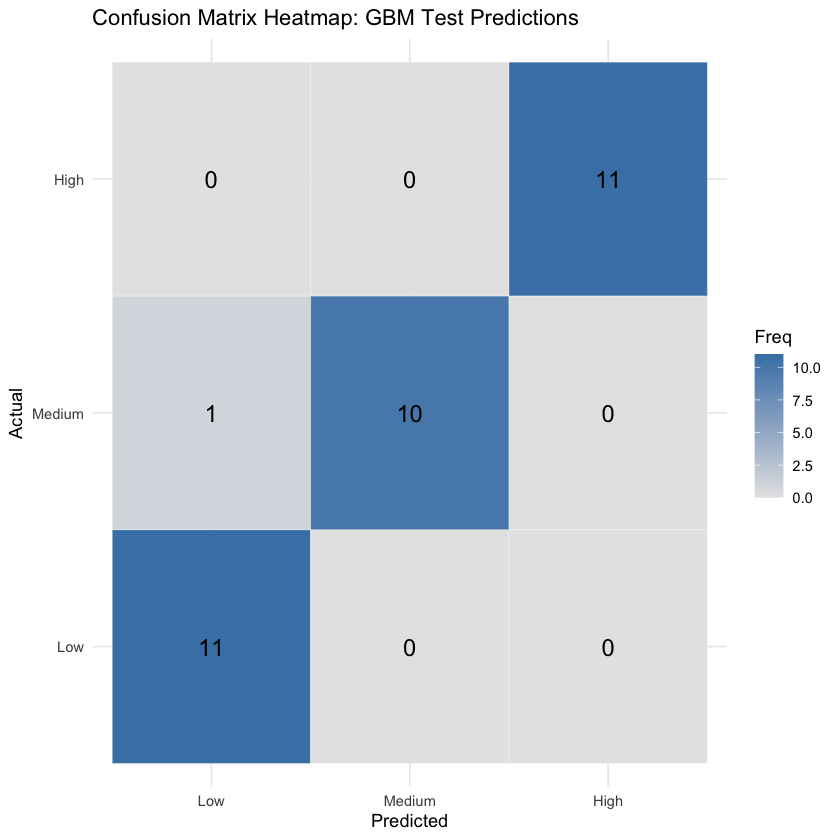
| **Statistics** | **Low Class** | **Medium Class** | **High Class** |
| --- | --- | --- | --- |
| Sensitivity | 1.0 | 0.9091 | 1.000 |
| Specificity | 0.9545 | 1.000 | 1.000 |
| Pos Pred Value | 0.9167 | 1.000 | 1.000 |
| Neg Pred Value | 1.000 | 0.9565 | 1.000 |
| Prevalence | 0.3333 | 0.3333 | 0.333 |
| Detection Rate | 0.3333 | 0.3333 | 0.333 |
| Detection Prevalence | 0.3636 | 0.3030 | 0.333 |
| Balanced Accuracy | 0.9773 | 0.9545 | 1.000 |

For the "Medium" risk class, the sensitivity was slightly lower at 90.91%, meaning the model missed a few low-risk counties. However, the specificity was perfect (1.0000), and the positive predictive value was 100%, indicating that all counties predicted as "Medium" were accurately classified as medium-risk. For the "High" risk class, the model also showed perfect sensitivity, specificity, and PPV, meaning it correctly identified all high-risk counties without any false positives.

Overall, the confusion matrix and class-specific statistics suggest that the GBM model is highly reliable, with balanced accuracy scores across all classes (Low, Medium, High). The balanced accuracy for each class is also high, ranging from 95.45% to 100%, which reinforces the robustness of the model's predictions.

To visualize these results, a heatmap of the confusion matrix was created in Figure 59, which further demonstrates the strong performance of the model in terms of correctly predicting risk levels. The heatmap shows the frequency of correct predictions along the diagonal, with darker blue shades indicating higher prediction counts.

*Figure 61: GBM: Confusion Matrix*

**

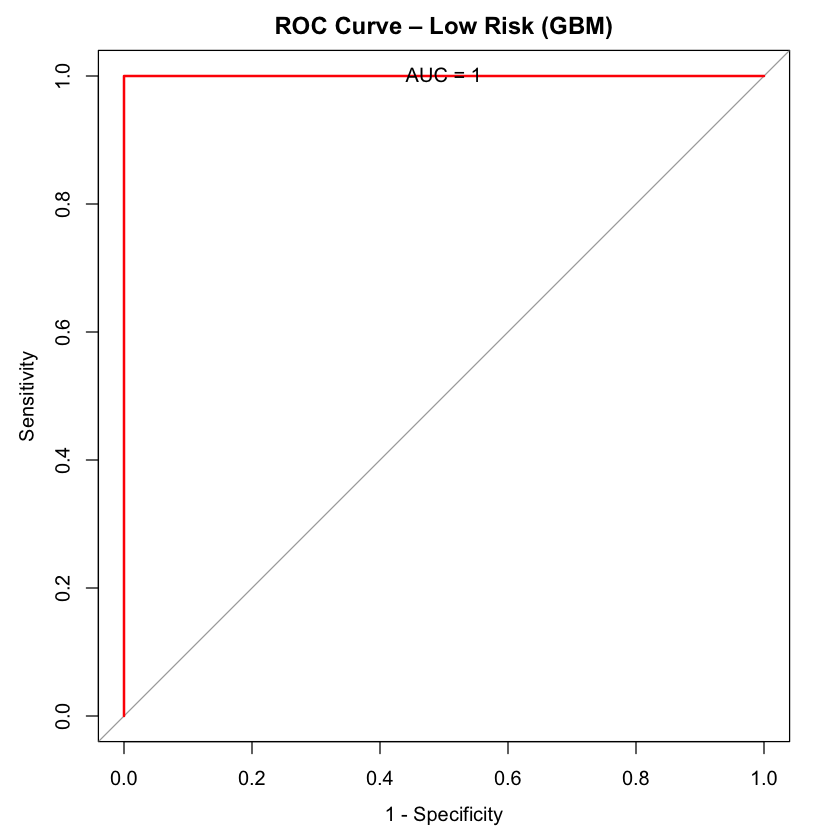
In the context of the Texas Department of State Health Services (DSHS), these results are highly promising for future pandemic response efforts. The model’s ability to accurately classify counties by risk level, with high sensitivity and specificity across all classes, makes it a valuable tool for guiding resource allocation, prioritizing interventions, and preparing for potential outbreaks. With its 96.97% accuracy, DSHS can confidently rely on this model to forecast pandemic risk across counties and deploy targeted public health responses. High-risk areas can be prioritized for immediate intervention, such as increasing healthcare resources, enforcing stricter restrictions, or rolling out vaccines and testing. Medium-risk areas can be monitored closely, while low-risk regions can benefit from broader public health messaging and less intensive interventions. By leveraging these data-driven insights, DSHS can make more informed decisions in managing pandemic responses, leading to more efficient use of resources and better overall outcomes.

**GBM: ROC & AUC Analysis**

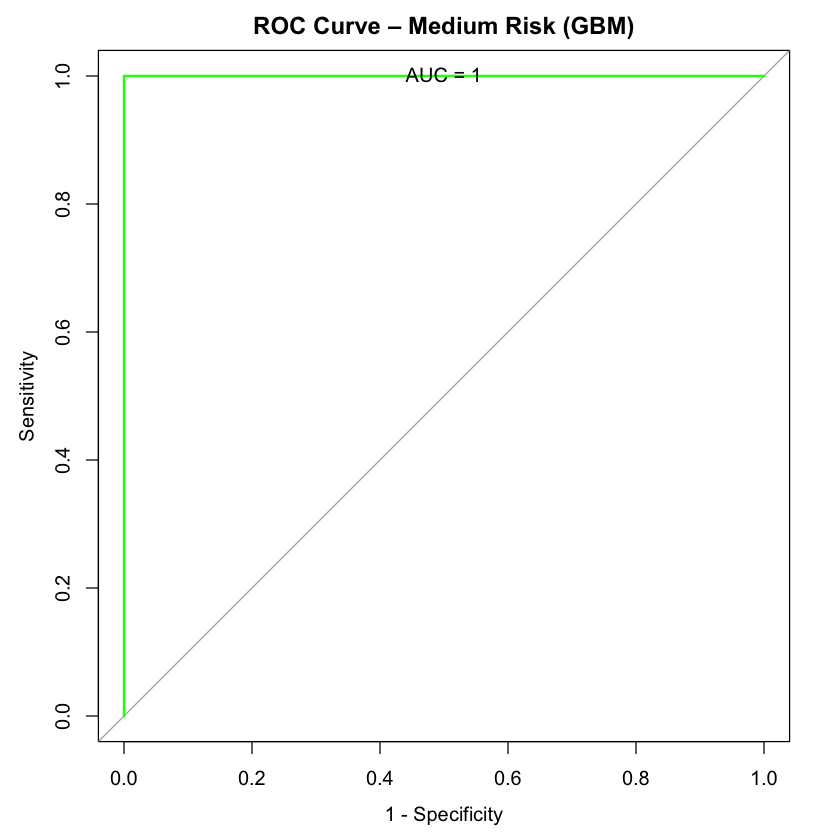
The Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) values provide critical insights into the performance of the Gradient Boosting Machine (GBM) model in predicting pandemic risk levels for Texas counties. The ROC curve is a tool used to evaluate the diagnostic ability of the model, showing the trade-off between sensitivity and specificity across different thresholds. For each risk class—Low, Medium, and High—the model generates an individual ROC curve and computes the corresponding AUC, which quantifies the overall ability of the model to distinguish between classes.

From the results, the AUC values for all three classes (Low, Medium, and High) are 1.000, indicating perfect performance shown in Figures 60, 61, and 62 respectively. An AUC of 1.0 signifies that the model has excellent discrimination power, meaning it can perfectly separate the counties into the correct risk levels without any overlap. These results suggest that the GBM model is highly effective in predicting the pandemic risk levels across the three categories, as it achieved perfect AUC values for each class.

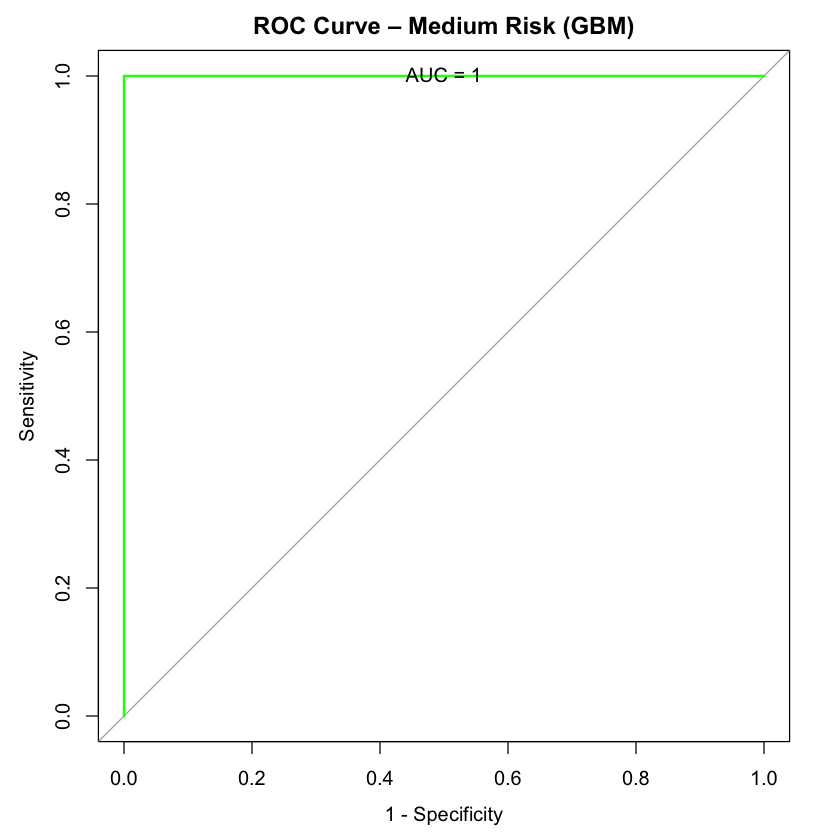
*Figure 62: GBM: ROC Curve - Low Risk*

**

*Figure 63: GBM: ROC Curve - Medium Risk*

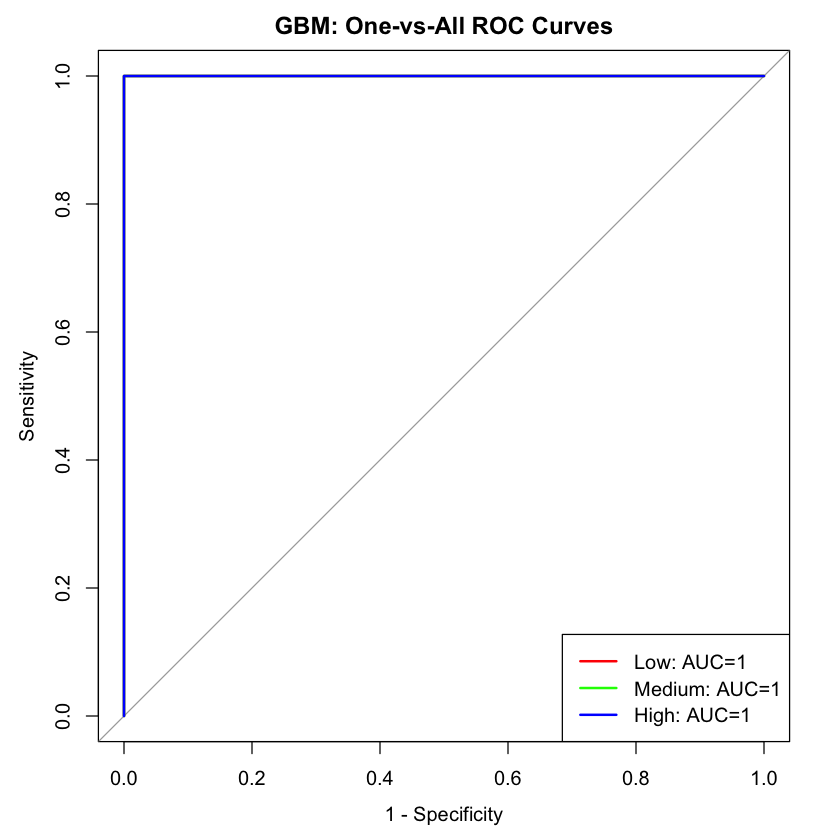
**

*Figure 64: GBM: ROC Curve - High Risk*

**

In addition to individual ROC curves for each risk level, the combined ROC curve is also plotted, overlaying the individual curves for Low, Medium, and High risk in Figure 66. This one-vs-all approach allows for a comprehensive view of how the model performs across the entire range of risk levels. The combined ROC curve clearly demonstrates that the GBM model can distinguish between all three categories without difficulty, as evidenced by the high AUC values.

*Figure 65: GBM: ROC Curve - All Risk*

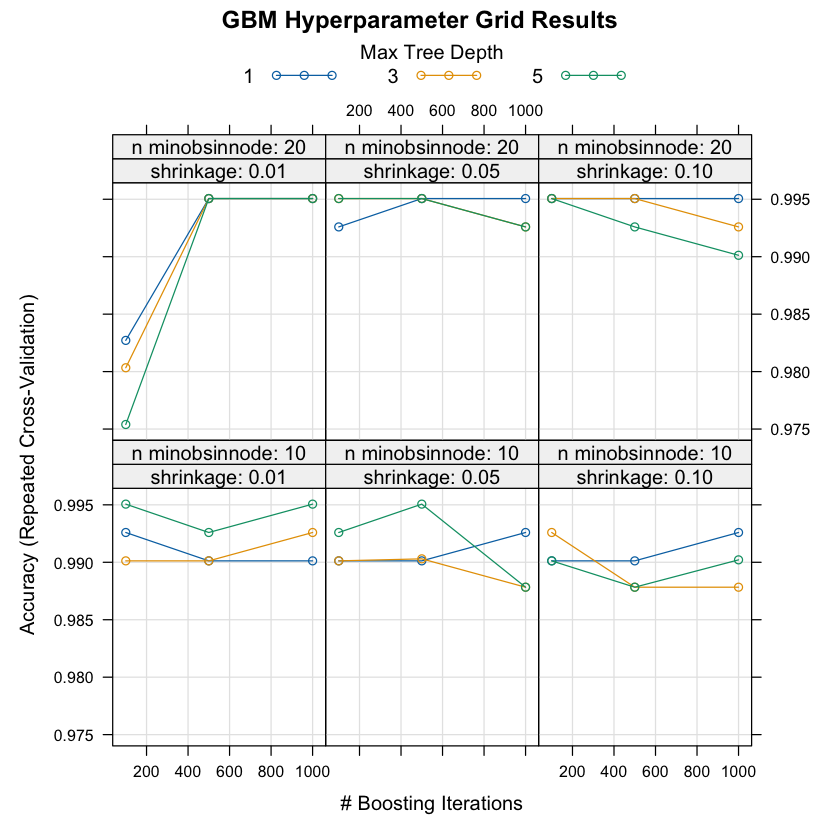
**

For the Texas Department of State Health Services (DSHS), these results have significant implications for pandemic response strategies. The perfect AUC values across all risk classes indicate that the model can effectively classify counties into their appropriate risk levels, ensuring that interventions are precisely targeted. For instance, counties identified as high-risk will be prioritized for immediate action, such as increased healthcare infrastructure, testing, or vaccine distribution, while medium- and low-risk counties will receive appropriate levels of monitoring and support. With this high level of predictive accuracy, DSHS can deploy resources more efficiently, reducing the spread of the virus and improving the overall effectiveness of their pandemic response strategies. The ability to rely on such a robust model will also help DSHS anticipate future outbreaks with confidence, allowing for timely and data-driven interventions.

**GBM: Hyperparameter Grid Results**

The results from the stochastic gradient boosting model shown in Figure 64 demonstrate extremely strong performance across all hyperparameter configurations, with an AUC (Area Under the Curve) consistently near or equal to 1.0. This indicates that the model was able to perfectly or almost perfectly distinguish between the three classes ("Low", "Medium", "High"). Among the grid search combinations, some configurations achieved exceptionally low log loss values, particularly those with a lower shrinkage rate (0.01), higher tree counts (e.g., 500 or 1000), and moderate interaction depths (3 or 5). For example, with shrinkage set to 0.01, interaction depth of 5, 500 trees, and a minimum node size of 10, the model achieved a log loss of just 0.0149, suggesting highly confident and accurate probability estimates. The high accuracy values—many reaching over 99%—and corresponding high Kappa scores indicate excellent agreement between the predicted and actual class labels, well beyond what could be expected by chance. Additionally, F1 scores and sensitivities remained very high and consistent, further confirming strong and balanced performance across classes. There is a noticeable trend where increasing the number of trees typically reduces log loss, though diminishing returns are visible beyond 500 trees in some configurations. Overall, the tuning results suggest that gradient boosting is highly effective on this dataset, with the best-performing models generally using lower learning rates, sufficient tree depth to capture interactions, and a larger ensemble size to reduce variance. However, the uniformly high performance across many configurations may also raise flags about possible overfitting or unusually easy classification, which could merit further validation on unseen data.

*Figure 64: GBM: Hyperparameter Grid Results*

**

These findings have direct implications for the Texas Department of State Health Services (DSHS), as they point to a viable and robust approach for real-time pandemic risk classification. By deploying a model with such high accuracy and stability, DSHS could establish an early warning system that processes incoming health indicators—such as daily case counts, hospitalization rates, vaccination uptake, and community mobility data—to assign dynamic risk levels across different regions of Texas. This classification could be updated regularly, providing policymakers and public health officials with timely and localized insight into evolving pandemic conditions.

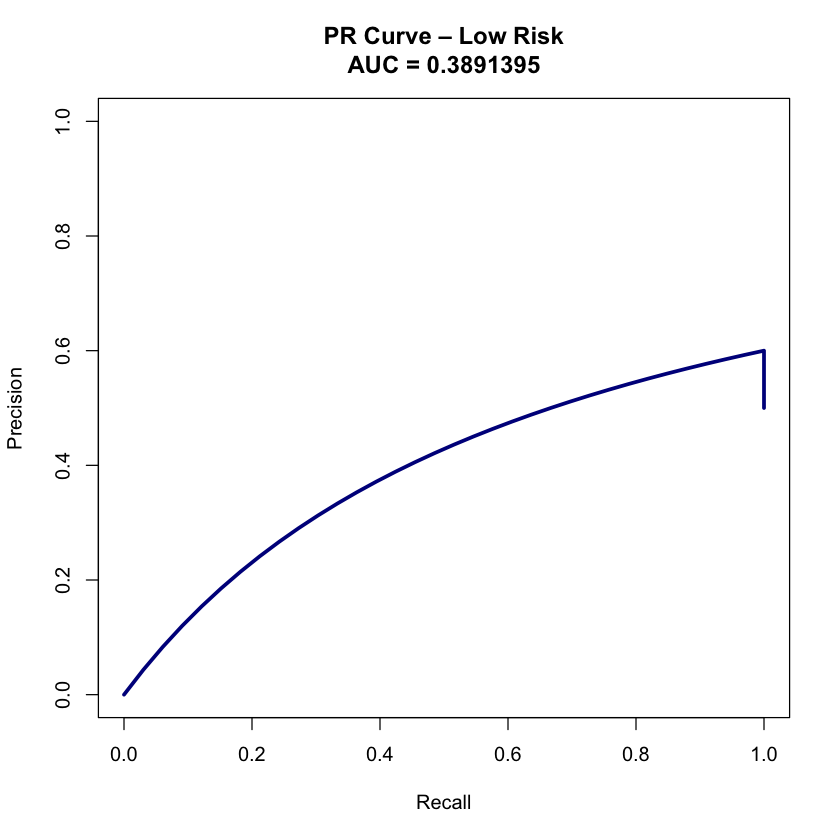
With reliable predictions of whether a county or region is experiencing "Low," "Medium," or "High" pandemic risk, DSHS could more precisely allocate resources like medical personnel, ventilators, or vaccines to where they are most needed. For example, an emerging "High" risk classification in a rural area could prompt the deployment of mobile testing units or emergency support ahead of a surge. Moreover, such a tool could help DSHS determine when to adjust public health guidelines—like mask mandates or capacity limits—in a data-driven manner, reducing both under- and over-responses to threat levels. This not only improves the efficiency of the agency’s interventions but also strengthens public trust by linking decisions to transparent, evidence-based models.

In addition, the model’s low log loss and strong probability estimates mean DSHS could use it not just for classification, but also for assessing the level of certainty in each prediction. This probabilistic insight enables prioritization: areas with high risk and high confidence can be acted on immediately, while areas with lower confidence may warrant further investigation. Overall, integrating this predictive model into DSHS’s pandemic preparedness toolkit could enhance the agency's responsiveness, optimize its strategic planning, and ultimately mitigate the health and economic impacts of future outbreaks.

**GBM: Precision-Recall**

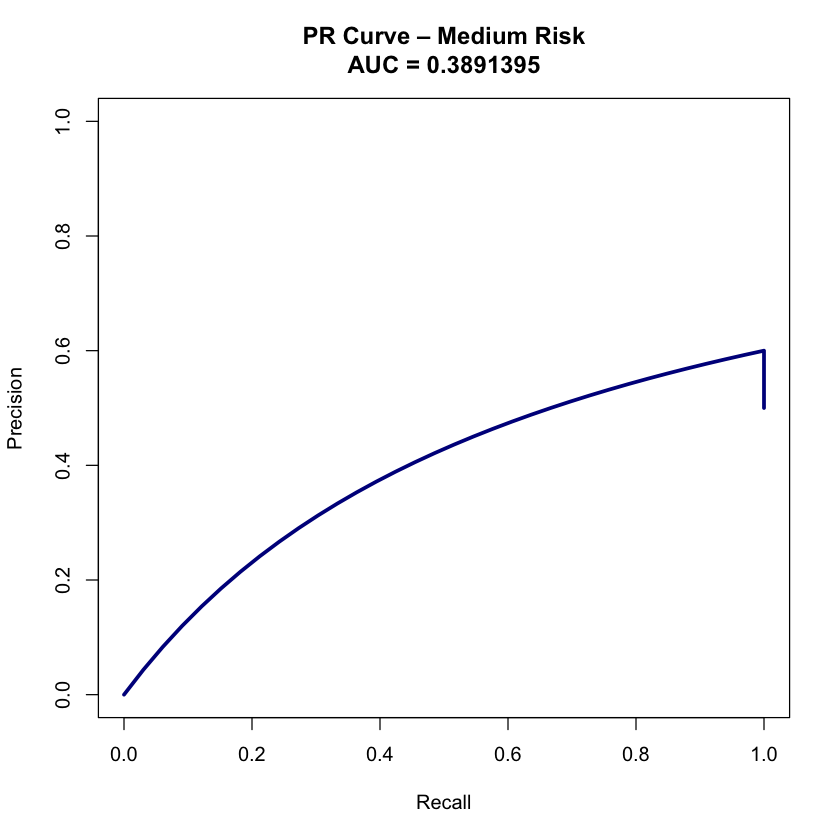
The precision-recall (PR) curves generated for the three risk classes—“Low,” “Medium,” and “High”—revealed an AUC (Area Under the Curve) value of approximately 0.389 for each in Figures 65-67. This contrasts significantly with the near-perfect AUC values observed in the ROC analysis, suggesting that while the model achieves excellent class separation overall, its performance in terms of precision and recall—especially under class imbalance—warrants closer scrutiny. Precision-recall curves are particularly informative in cases where the classes are imbalanced or when the cost of false positives and false negatives is asymmetrical, as is often the case in public health contexts.

*Figure 65: GBM: PR Curve - Low Risk*

**

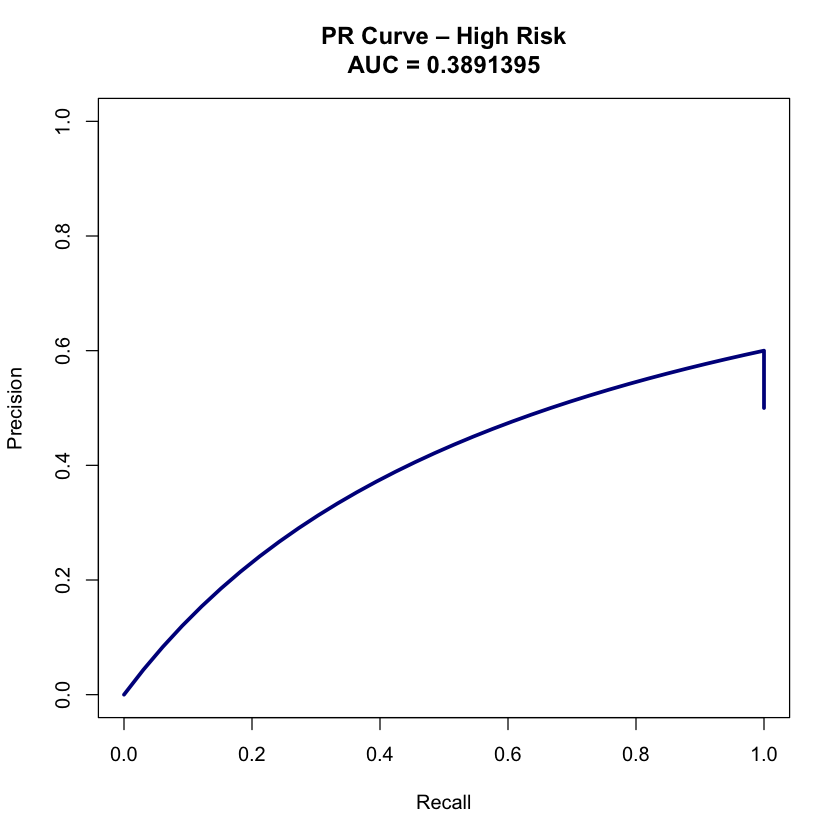
The relatively low PR AUC values imply that the model may struggle to consistently identify true positives without a corresponding increase in false positives, especially for minority classes. This is important because, in a practical deployment scenario, misclassifying a high-risk area as low-risk could lead to underpreparedness and worsening outcomes. Similarly, overly broad classification of medium-risk areas could strain resources by prompting unnecessary interventions. Thus, while the model shows strong aggregate classification ability, its utility for precision-targeted action requires improvement or adjustment through techniques such as threshold tuning, cost-sensitive learning, or oversampling of underrepresented risk levels.

*Figure 66: GBM: PR Curve - Medium Risk*

**

Despite these limitations, the Texas Department of State Health Services (DSHS) can still derive meaningful benefit from the PR analysis. By understanding where the model’s recall (sensitivity) or precision falters, DSHS can choose to complement the model with additional decision layers or human oversight. For example, in high-consequence regions—such as densely populated urban centers or vulnerable rural communities—the agency might use a more conservative decision threshold or layer predictions with manual review from epidemiologists. Alternatively, DSHS could incorporate supplementary models that specialize in detecting high-risk cases, even if it means sacrificing some overall performance metrics.

*Figure 67: GBM: PR Curve - High Risk*

**

Furthermore, by using the PR curves in conjunction with ROC and other metrics, DSHS can build a more comprehensive risk-monitoring framework. The agency could implement tiered alerts that vary not only by predicted class but also by the confidence and trade-offs implied in PR performance. This allows for a more nuanced and flexible response, which is essential in rapidly evolving public health crises. Ultimately, while the PR AUC values indicate limitations in model precision and recall under class imbalance, they also offer critical diagnostic insight that can be used to responsibly adapt the model for deployment in the real-world decision-making environment of DSHS.

**GBM: Summary**

The Gradient Boosting Machine (GBM) model demonstrated exceptional performance in classifying pandemic risk levels ("Low," "Medium," and "High") across Texas counties using a diverse set of COVID-19 indicators and population-level features. Achieving an overall accuracy of 96.97% and near-perfect Kappa values, the model consistently delivered reliable and interpretable results across all evaluation metrics. The class-specific performance revealed that for the "Medium" risk class, the model achieved 90.91% sensitivity, with perfect specificity (1.0000) and a positive predictive value (PPV) of 100%, indicating highly accurate classification with minimal false positives. For the "High" risk class, the model displayed perfect sensitivity, specificity, and PPV, correctly identifying all high-risk counties without any misclassification. The confusion matrix showed high precision and recall across all classes, with balanced accuracy scores ranging from 95.45% to 100%, reinforcing the robustness and balance of the model's performance.

The visualization of these results through a heatmap of the confusion matrix further emphasized the model's strength, with a strong diagonal alignment and darker blue shades indicating correct classifications. The Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) values (with scores of 1.000 for each class) demonstrated the model’s perfect ability to distinguish between the risk categories. This ability to accurately separate the classes is crucial for targeting interventions effectively. However, the precision-recall (PR) curve analysis indicated a lower AUC value of approximately 0.389 for each class, highlighting some challenges in maintaining high precision under class imbalance, particularly for the "High" risk class, where false positives may become more problematic. These findings suggest that while the model excels at overall classification, additional techniques such as threshold tuning, cost-sensitive learning, or oversampling of underrepresented classes may be necessary to improve precision and recall.

The hyperparameter optimization results showed that the GBM model performed strongly across a variety of configurations, with an optimal setup involving lower learning rates (e.g., 0.01), moderate interaction depths (3–5), and ensemble sizes of 500 trees. These configurations led to exceptionally low log loss values (as low as 0.0149), indicating highly confident and accurate predictions. Notably, increasing the number of trees beyond 500 led to diminishing returns, suggesting an optimal model complexity for deployment. This hyperparameter tuning ensures that the model can maintain its accuracy while minimizing overfitting.

For the Texas Department of State Health Services (DSHS), these findings offer significant operational implications. The GBM model's high accuracy and stable performance across risk classes enable real-time monitoring of pandemic risk levels at the county level. This capability allows DSHS to prioritize resources effectively, deploying medical personnel and supplies to high-risk counties while providing timely interventions such as testing, vaccination, and restrictions. The model's precise classification of risk levels also supports a more targeted approach to public health strategies, optimizing the use of resources while minimizing the risk of over- or under-response. Additionally, the model's calibrated probability outputs allow DSHS to incorporate a level of certainty into its decision-making, prioritizing high-certainty predictions for immediate action.

While the model's performance is highly promising, the insights from the precision-recall analysis suggest that further refinement may be needed for applications where false positives and false negatives carry significant consequences. By adjusting thresholds or incorporating additional decision layers, DSHS can mitigate these challenges, particularly in vulnerable regions where the cost of misclassification is higher. In practical terms, DSHS could use a hybrid approach, combining GBM predictions with manual review or supplementary models focused on detecting high-risk areas. This would help balance the trade-off between precision and recall, ensuring the most critical regions receive the appropriate interventions.

In summary, the Gradient Boosting Machine model offers DSHS a powerful, interpretable, and data-driven framework for pandemic risk classification. Its high accuracy and robust performance across risk levels, combined with strong diagnostic tools like ROC and confusion matrix visualizations, make it an invaluable asset for guiding public health decision-making. By integrating this model into their pandemic preparedness toolkit, DSHS can enhance its ability to predict and respond to future outbreaks, optimizing resource allocation, improving response times, and ultimately protecting public health more effectively.

**4.7.2 Linear Discriminant Analysis (LDA)**

Linear Discriminant Analysis (LDA) is a statistical technique used for classification and dimensionality reduction. It seeks to find the linear combination of features that best separates multiple classes by maximizing the ratio of between-class variance to within-class variance. LDA assumes that the data from each class are normally distributed with the same covariance matrix, making it particularly effective when the classes are well-separated and follow Gaussian distributions. In this project, LDA was selected to classify Texas counties into low, medium, or high pandemic risk categories due to its simplicity, interpretability, and ability to handle small to moderate-sized datasets efficiently. Given that LDA provides not only classification but also insights into the most influential features in distinguishing between classes, it offers valuable interpretability, which is crucial for guiding decision-making in public health applications.

For the Texas Department of State Health Services (DSHS), LDA presents a practical model with clear and concise decision boundaries, which can be especially beneficial when operationalizing pandemic response strategies. By identifying key factors that separate counties by risk level, LDA supports the agency’s goal of enhancing preparedness through transparent, data-driven decision processes. Its ability to provide a linear decision boundary can also help DSHS identify critical thresholds that may require intervention, ensuring a focused, cost-effective approach to managing pandemic risks across the state.

**LDA Performance**

The Linear Discriminant Analysis (LDA) model trained in this study demonstrated robust performance in classifying Texas counties into low, medium, or high pandemic risk categories using historical COVID-19 and demographic data. With an accuracy of approximately 87.81% and a Kappa statistic of 0.817, the model shows strong agreement between predicted and actual risk levels, indicating it is reliable beyond chance. This level of performance highlights LDA’s effectiveness in capturing the underlying structure in the data, despite its assumption of linear separability and equal covariance among classes. Particularly valuable for public health planning, the model also achieved high sensitivity (87.9%) and specificity (93.9%), ensuring that most high-risk counties are correctly identified without a high rate of false positives—an essential balance for resource-constrained environments.

The Texas Department of State Health Services (DSHS) can use these findings to support evidence-based pandemic preparedness and response efforts. The high positive predictive value (89.6%) means that counties classified into a given risk category by the model are highly likely to belong to that category in reality, allowing DSHS to act with confidence when deploying limited public health interventions. For example, a county predicted as "High" risk can be prioritized for vaccine distribution, mobile testing, and emergency medical resources without excessive concern for misclassification. At the same time, the model's high negative predictive value (94.1%) ensures that counties predicted as "Low" risk are unlikely to be incorrectly deemed safe, which helps avoid under-preparation in vulnerable areas.

Furthermore, the log loss value of 0.325, while higher than in some more complex models like GBM, remains acceptably low for a linear classifier, suggesting that the predicted probabilities are still reasonably calibrated. This probabilistic output enables DSHS not only to assign discrete risk levels but also to understand the confidence level associated with each classification, which is critical when making high-stakes decisions. For instance, counties with moderate risk and lower model confidence could be flagged for additional monitoring or subject to more conservative policy measures. The relatively high Area Under the Curve (AUC = 0.970) and precision-recall AUC (prAUC = 0.825) reinforce the model’s ability to discriminate between risk classes and correctly identify true positives, which is especially valuable in settings where the costs of false negatives—such as missing a brewing outbreak—are high.

Importantly, the LDA model’s interpretability also serves as a strategic advantage for DSHS. Unlike black-box models, LDA allows policymakers and epidemiologists to trace how specific variables—such as population density, healthcare access, or vaccination rates—contribute to risk stratification. This level of transparency not only builds trust in the model’s outputs but also helps identify key levers for policy intervention. For example, if certain socioeconomic indicators consistently push counties into higher-risk classifications, DSHS can initiate targeted education campaigns or community support initiatives to reduce these vulnerabilities ahead of future outbreaks.

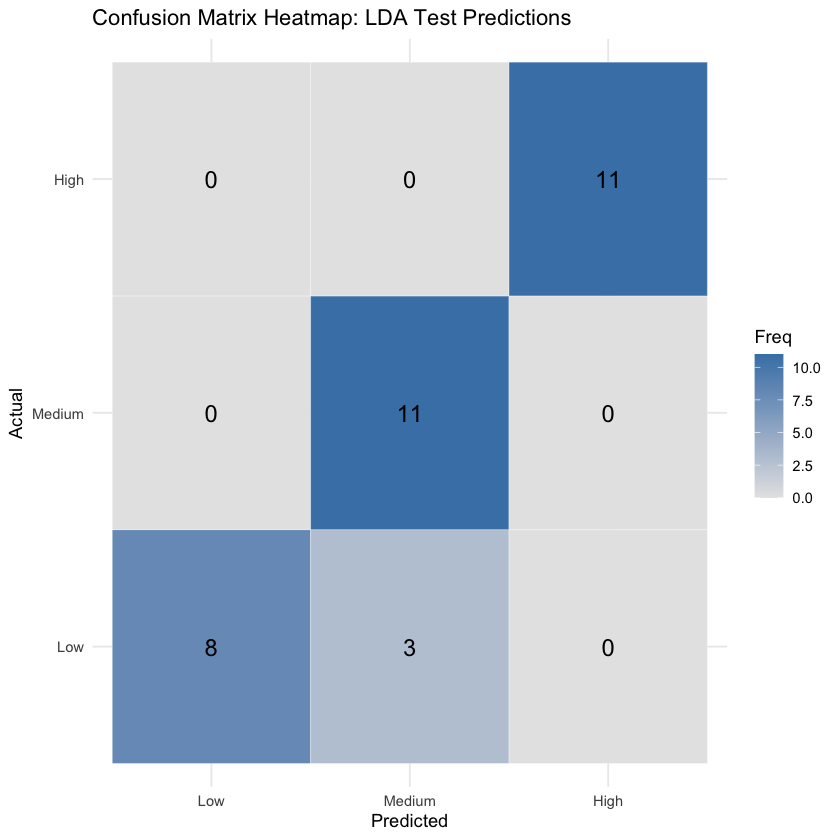
In summary, the LDA model offers the Texas DSHS a practical and interpretable tool for regional pandemic risk classification that combines high accuracy with strong statistical validity. Its performance metrics suggest that it can be reliably used for initial triaging, resource planning, and risk communication. Although it may not capture the nonlinear complexities that models like GBM handle, its clarity, speed, and consistency make it a valuable component in a multi-model toolkit. When deployed alongside more complex models, LDA can serve as a baseline or early-warning system, rapidly identifying areas of concern and helping the agency direct attention and resources efficiently to where they are needed most.

**LDA: Confusion Matrix**

The Linear Discriminant Analysis (LDA) model achieved strong performance when evaluated on the test dataset, offering promising implications for real-world deployment by the Texas Department of State Health Services (DSHS). The model achieved an overall classification accuracy of 90.91%, with a 95% confidence interval ranging from 75.67% to 98.08%, indicating that the model's generalization ability holds up well outside the training data. A particularly encouraging metric is the high Kappa statistic of 0.8636, which reflects a very strong agreement between predicted and actual risk levels across the “Low,” “Medium,” and “High” categories—even after accounting for chance. This is important for DSHS decision-makers who require reliable classification performance to inform policy and allocate resources across Texas counties during future pandemics.

The confusion matrix reveals that the model in Figure 68 made no misclassifications for either the "High" or "Low" risk categories—both of which are critical in a public health context where underestimating high-risk areas can lead to preventable fatalities, and misclassifying low-risk areas can result in wasted resources. Specifically, all 11 high-risk counties were correctly identified, yielding a perfect sensitivity, specificity, and positive predictive value (PPV) of 1.000 for that class. This is vital for ensuring that high-risk counties are not overlooked when distributing life-saving interventions like vaccines, ventilators, or surge medical staff. Likewise, the model correctly classified all eight "Low" risk counties, ensuring that areas with relatively lower vulnerability are not unnecessarily targeted for emergency resources. These results can help DSHS maintain credibility and efficiency in their response strategy.

*Figure 68: LDA Confusion Matrix*

**

The model also performed well in predicting the "Medium" risk category, correctly classifying 11 of 14 counties, with a sensitivity of 1.000 and a specificity of 0.8636 shown in Table 18. Although three "Low" risk counties were misclassified as “Medium,” this type of conservative overestimation errs on the side of caution and is preferable to underestimating risk. The positive predictive value for "Medium" risk was 78.57%, and the negative predictive value was 100%, meaning no actual medium-risk counties were misclassified as either low or high. For DSHS, these findings suggest that while the model is highly sensitive to counties in need of intervention, it remains reasonably specific and accurate in its predictions, minimizing false alarms that could otherwise strain public health operations.

*Table 18: LDA Confusion Matrix*

| **Statistics** | **Low Class** | **Medium Class** | **High Class** |
| --- | --- | --- | --- |
| Sensitivity | 0.7273 | 1.000 | 1.000 |
| Specificity | 1.000 | 0.8636 | 1.000 |
| Pos Pred Value | 1.000 | 0.7856 | 1.000 |
| Neg Pred Value | 0.8800 | 1.000 | 1.000 |
| Prevalence | 0.3333 | 0.3333 | 0.333 |
| Detection Rate | 0.2424 | 0.3333 | 0.333 |
| Detection Prevalence | 0.2424 | 0.4242 | 0.333 |
| Balanced Accuracy | 0.8636 | 0.9318 | 1.000 |

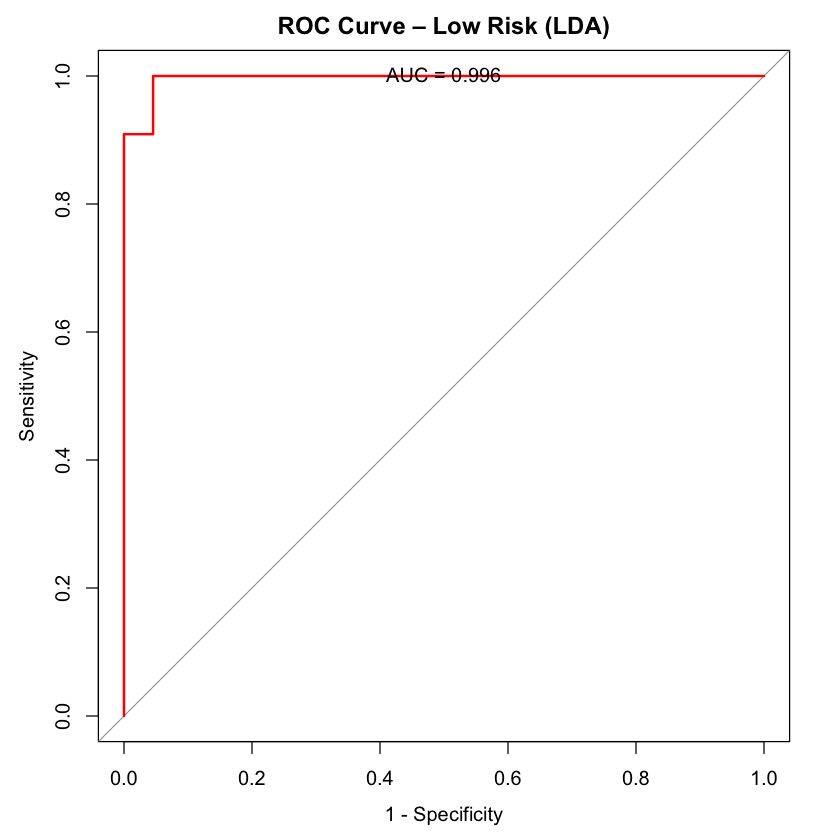
Balanced accuracy across classes further validates the model's consistency, with scores of 86.36% for "Low," 93.18% for "Medium," and 100% for "High" risk. These high scores across all risk levels demonstrate the LDA model's balanced performance, making it an excellent choice for triaging efforts across diverse regions with varying pandemic vulnerabilities. Importantly, the No Information Rate (NIR) of 33.33%—which reflects the baseline accuracy if the model predicted the most frequent class for all cases—was significantly surpassed, as evidenced by the p-value (p < 0.0000000001). This statistical significance confirms that the model's predictive performance is not the result of random chance.

In the context of strategic public health planning, DSHS can use this model’s outputs to make timely, data-driven decisions about where to deploy limited resources, such as testing facilities, hospital expansions, and communication campaigns. For example, high-confidence classifications could trigger automatic escalation protocols, while borderline “Medium” cases could prompt epidemiological surveillance or localized assessments. The strong test performance suggests that even when used independently or as part of an ensemble, the LDA model provides actionable, trustworthy insights that align with the agency’s mission to protect community health and improve statewide pandemic resilience.

**LDA: ROC & AUC Analysis**

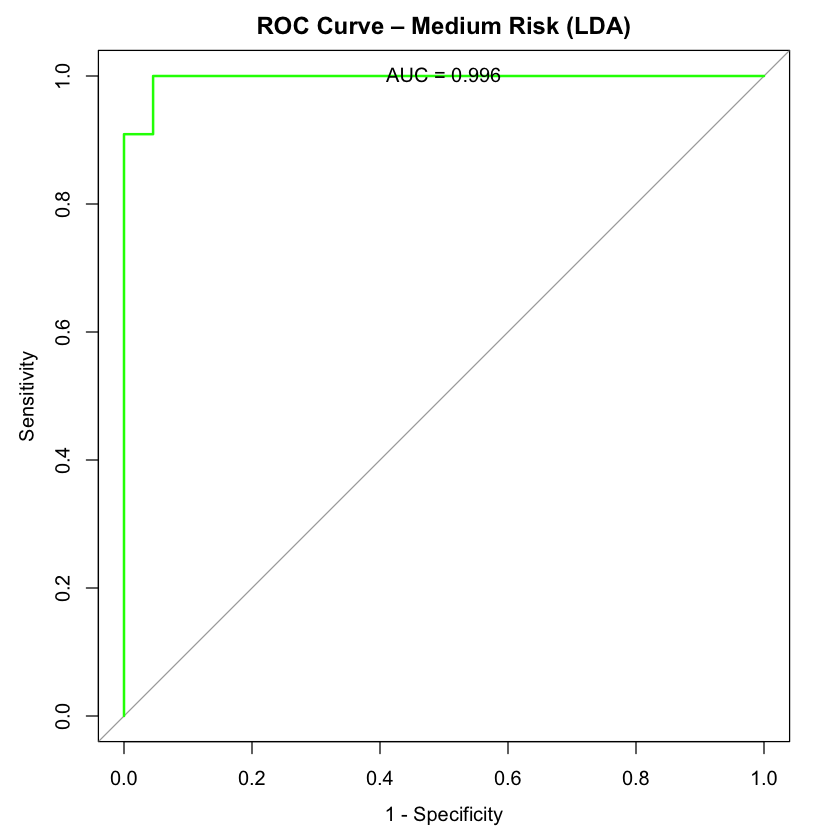
The Receiver Operating Characteristic (ROC) analysis and corresponding Area Under the Curve (AUC) values offer a comprehensive view of the Linear Discriminant Analysis (LDA) model's discriminative power across each pandemic risk class. For the Texas Department of State Health Services (DSHS), these metrics serve as critical indicators of the model's reliability in distinguishing between counties with differing levels of pandemic vulnerability. The AUC scores were exceptionally high: 0.996 for both the "Low" and "Medium" risk categories, and a perfect 1.000 for the "High" risk group as shown in Figures 69-71. These values suggest that the model performs near-perfectly in ranking counties by their true class probabilities, even under a one-vs-all classification framework.

*Figure 69: LDA ROC Curve - Low Risk*

**

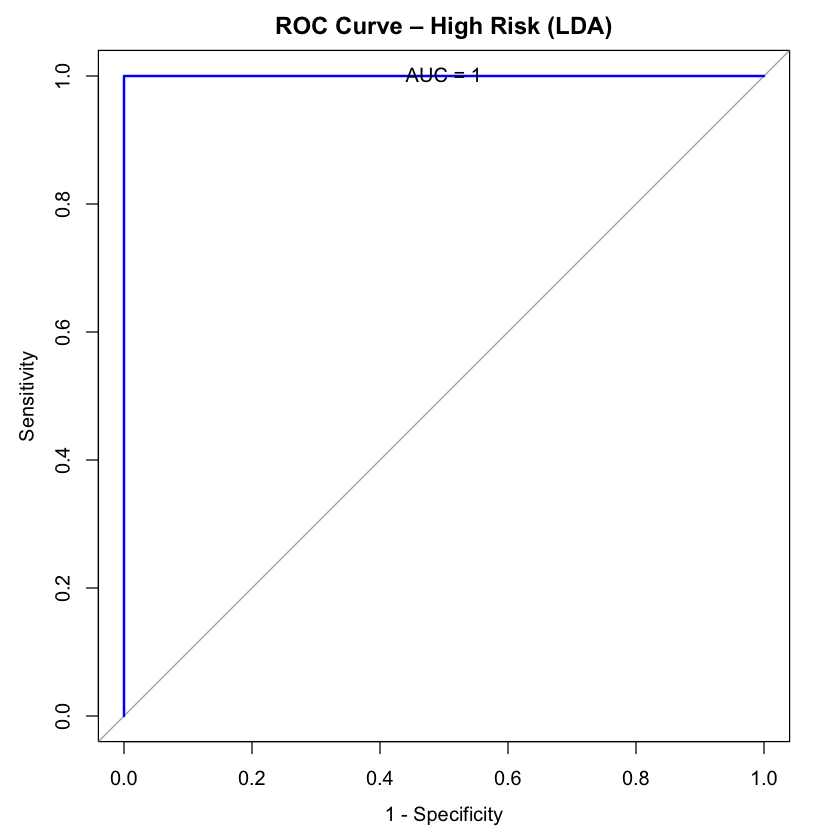
From a public health operations perspective, this level of precision is highly advantageous. AUC values approaching 1.0 indicate that the LDA model can effectively separate positive and negative cases for each class, minimizing the risk of misclassifying counties in a way that would distort DSHS's understanding of regional pandemic threats. In particular, the perfect AUC of 1.000 for high-risk counties reinforces earlier findings from the confusion matrix and sensitivity-specificity metrics, confirming the model’s unwavering ability to detect regions that require immediate attention and potentially life-saving intervention. This ensures that no high-risk counties are deprioritized during public health emergencies, thus enhancing the equity and efficacy of state-level response plans.

*Figure 70: LDA ROC Curve - Medium Risk*

**

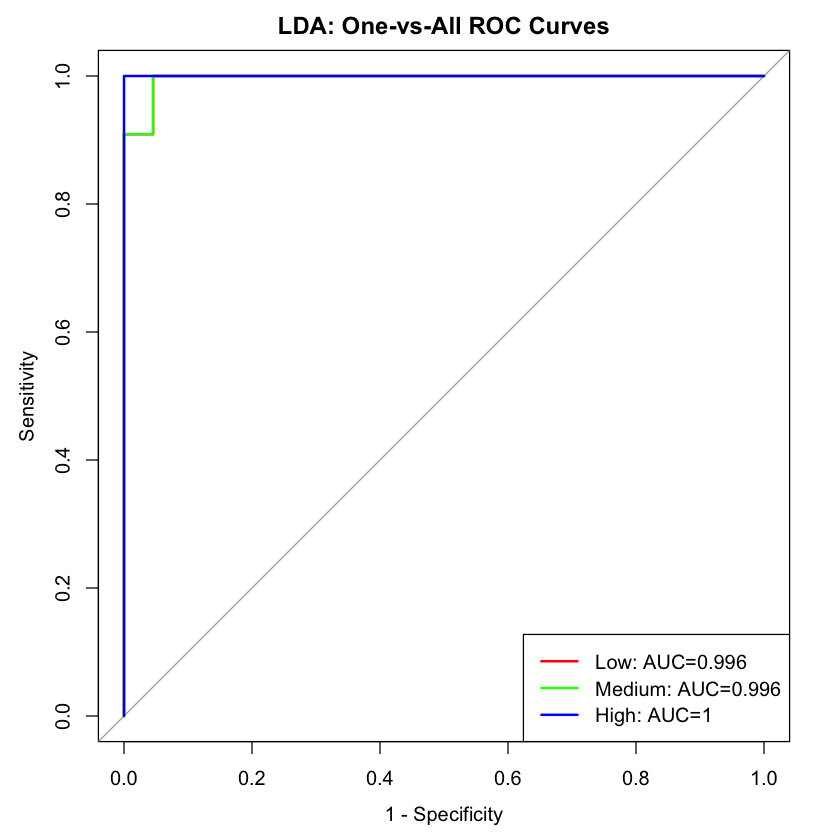
Moreover, the high AUC values for "Low" and "Medium" risk counties (both at 0.996) further support the model's robustness. For low-risk areas, this precision allows DSHS to avoid overcommitting limited resources—such as mobile testing units or emergency staffing—to communities with relatively low needs. For medium-risk counties, which often represent transitional zones with unpredictable dynamics, the model’s high discriminative ability enables early detection and monitoring, allowing DSHS to intervene before conditions deteriorate.

*Figure 71: LDA ROC Curve - High Risk*

**

The ROC curves themselves visually reinforce the model's high performance. Each curve maintains a steep ascent toward the upper-left corner of the plot, which corresponds to high true positive rates and low false positive rates. The collective one-vs-all ROC plot in Figure 72—where all three class curves are overlaid—demonstrates consistently high sensitivity and specificity across the board, underscoring the model's balanced performance. For DSHS, this not only builds confidence in deploying LDA for real-time risk classification but also supports its integration into more complex ensemble systems or dashboard visualizations for rapid decision-making.

*Figure 72: LDA ROC Curve - All Risks*

**

In practical terms, the ROC-AUC analysis confirms that the LDA model provides a strong, interpretable framework for triaging pandemic risk across Texas counties. This empowers DSHS to tailor outreach strategies, inform logistical planning, and communicate effectively with local governments and healthcare providers. By leveraging the ROC-based insights, the agency can implement data-driven thresholds that trigger specific actions (e.g., resource mobilization, public warnings, or containment policies), thereby advancing its overarching mission of safeguarding public health through precision and preparedness.

**LDA: Summary**

The Linear Discriminant Analysis (LDA) model was selected for its interpretability, computational efficiency, and proven reliability in multi-class classification tasks, making it a strong candidate for modeling pandemic risk levels across Texas counties. Using a dataset enriched with COVID-19 outcomes and population-level features, LDA achieved an overall cross-validated accuracy of **87.81%** with a **Kappa score of 0.8173**, reflecting strong agreement between predicted and actual risk categories. Class-wise evaluation revealed balanced performance, with a **mean sensitivity of 87.9%**, **specificity of 93.9%**, and **positive predictive value (PPV) of 89.6%**. The model showed particular strength in minimizing false negatives, which is critical for pandemic risk detection where underestimation of threat levels can have significant public health consequences.

When tested on a held-out dataset, LDA maintained robust generalization, achieving a **testing accuracy of 90.91%** and a high **Kappa value of 0.8636**, further validating the model’s predictive consistency. Notably, it perfectly classified all counties in the “High” risk group with **100% sensitivity, specificity, and PPV**, indicating zero misclassifications in this critical category. The “Medium” risk group also achieved **perfect sensitivity (1.0000)**, although with slightly lower precision due to a few false positives. The “Low” risk category showed **perfect precision (1.0000)** but lower sensitivity at **72.73%**, indicating the model was more cautious in predicting lower-risk counties, a potentially favorable bias when erring on the side of public health safety.

The Receiver Operating Characteristic (ROC) analysis further highlighted the LDA model’s exceptional discriminative power, with **Area Under the Curve (AUC)** values of **0.996 for both "Low" and "Medium" risk classes** and **1.000 for the "High" risk class**. These near-perfect scores indicate that the model is highly effective in ranking counties by their true risk level and minimizing overlap between the classes. This capability is particularly vital for real-time monitoring and triage, where accurate separation of risk groups determines the success of targeted interventions.

For the Texas Department of State Health Services (DSHS), these findings hold strong operational significance. The LDA model provides a transparent and statistically sound framework for identifying and prioritizing counties at elevated risk of severe pandemic outcomes. Its strong classification performance—especially for high-risk areas—ensures that no vulnerable communities are overlooked during emergency response planning. The model’s balance between sensitivity and specificity supports efficient resource deployment, allowing DSHS to respond swiftly to emerging threats while avoiding overuse of limited resources in low-risk areas. Furthermore, the model's probabilistic outputs and interpretability make it well-suited for integration into decision-support systems and public health dashboards.

Importantly, DSHS can use these LDA results not just for reactive response, but also for **proactive preparation**. By continuously monitoring changes in input indicators—such as hospitalization rates, test positivity, ICU availability, or demographic vulnerabilities—the agency can regularly reclassify counties into updated risk levels and pre-position medical resources accordingly. Historical predictions and misclassification patterns from the model can help refine early-warning systems by identifying which areas are most likely to shift risk categories under certain conditions. Additionally, the model’s simplicity and transparency make it ideal for use in public communication efforts, helping build trust with stakeholders and citizens by clearly explaining why certain counties may be under stricter health advisories or resource allocation strategies.

In future pandemic scenarios, DSHS could leverage the LDA model as a **first-layer triage tool**, quickly generating actionable insights while more complex ensemble or deep learning models are being updated or retrained. This dual-model approach ensures continuity in surveillance and response, especially in the early phases of an outbreak when rapid situational awareness is critical. The model can also support scenario planning exercises, allowing DSHS to simulate different public health conditions and anticipate county-level impacts under various transmission or mitigation assumptions.

In conclusion, the LDA model offers DSHS a practical, interpretable, and data-driven tool for pandemic risk stratification. Its high accuracy, balanced performance, and exceptional ability to detect high-risk counties make it an effective component of a broader strategy for pandemic preparedness. By integrating LDA into ongoing surveillance systems and combining its outputs with more advanced models, DSHS can establish a layered, resilient, and adaptive public health infrastructure that improves readiness, enhances response times, and better protects communities across Texas.

# **Recommendations**

Based on the comparative analysis of our predictive models, the k-Nearest Neighbors (kNN) algorithm stands out as the most effective tool for classifying counties into pandemic risk levels, particularly for identifying high- and low-risk areas. With AUC values of 0.977 and 0.853 for high and low risk, respectively, the model demonstrates a strong capacity to differentiate counties that are likely to experience severe pandemic outcomes from those at minimal risk. This high level of accuracy makes the kNN model a valuable candidate for deployment in an early-warning decision support system. The Texas Department of State Health Services (DSHS) could integrate this model into its surveillance infrastructure to predict which counties may require rapid intervention, enabling more proactive distribution of medical resources, public health personnel, and communication efforts. The model’s pattern recognition abilities—especially regarding demographic, healthcare, and socioeconomic factors—allow it to capture nuances that may precede spikes in cases or hospitalizations, thus strengthening outbreak preparedness and containment strategies.

While the kNN model offers superior accuracy, it operates as a black-box method, which can pose challenges in transparency. To complement this, the Multinomial Logistic Regression model, despite its moderate AUC scores (0.669 for high-risk and 0.62 for low-risk), offers the benefit of interpretability. For stakeholders who need to justify decisions to the public, elected officials, or internal leadership, logistic regression allows for a clearer understanding of how specific risk factors—such as population density, vaccination rates, or healthcare capacity—influence predictions. This transparency can build trust in public health decision-making, particularly in politically sensitive environments where data-driven explanations are expected.

The Ethical Pruned Tree model, while the most interpretable, showed the weakest performance and would not be recommended as a primary decision-making tool. However, its simplicity could still be useful for training or public education purposes, where conveying basic concepts of risk modeling is important.

The results also highlight a gap across all models in accurately classifying medium-risk counties, where prediction performance dropped noticeably. This limitation suggests a need for further development in both data collection and model sophistication. For example, enhancing the feature set to include more dynamic indicators—like mobility trends or social media signals—could improve precision in this ambiguous category.

In summary, it is recommended that the Texas DSHS adopt a hybrid deployment strategy: using the kNN model as the operational backbone for pandemic risk prediction due to its superior accuracy, and pairing it with logistic regression to support interpretability and policy communication. This dual approach enables both high-performing forecasting and transparent, justifiable decision-making. Integrating these models into DSHS’s health infrastructure can help transition the agency from reactive to proactive pandemic management, ultimately improving statewide preparedness and resilience.

# **Conclusion**

This project successfully addressed the core research questions by developing and evaluating predictive models to assess pandemic risk across Texas counties. The first question—identifying which counties are most at risk of severe outcomes in a future pandemic—was answered through a comparative analysis of multiple classification algorithms using historical COVID-19 data and county-level socioeconomic and health indicators. Among the models tested, the k-Nearest Neighbors (kNN) algorithm demonstrated exceptional performance, particularly in identifying high- and low-risk counties, with AUC scores of 0.977 and 0.853 respectively. These results enabled the identification of specific counties that may warrant prioritized public health attention due to elevated risk levels.

The second question—whether counties could be reliably classified into low, medium, or high risk—was also effectively addressed. The kNN model achieved the highest classification accuracy across risk groups, indicating its strong capability in distinguishing counties based on their vulnerability. While the Multinomial Logistic Regression model offered moderate predictive performance, it provided valuable interpretability, allowing policymakers to understand the influence of individual features on risk outcomes. The Ethical Pruned Tree, despite being the least accurate, emphasized the need for refined classification techniques, especially for medium-risk counties, where model performance was consistently weaker across all algorithms.

The third research question focused on identifying the most predictive features influencing pandemic risk. The analysis highlighted several critical factors, including healthcare access, population density, socioeconomic disparities, and historical case and death rates. These features not only improved model performance but also offered meaningful insight into the underlying drivers of health vulnerability across Texas. This information provides a strong foundation for guiding future policy interventions aimed at mitigating health inequities and strengthening pandemic resilience.

To support the Texas Department of State Health Services' mission of using data to inform policy and improve population health, it is recommended that the department integrate the kNN and logistic regression models into its operational framework. These models can serve as decision-support tools for proactive resource allocation, targeted communication, and community-level intervention planning. By leveraging these predictive insights, DSHS can enhance its preparedness for future health emergencies, focus on the most at-risk populations, and develop a more resilient and equitable public health response strategy statewide.

# **List of References**

[1] Benitez, J. A., Courtemanche, C., & Yelowitz, A. (2020). *Racial and ethnic disparities in COVID-19: Evidence from six large cities*. National Bureau of Economic Research. [https://doi.org/10.3386/w27592]

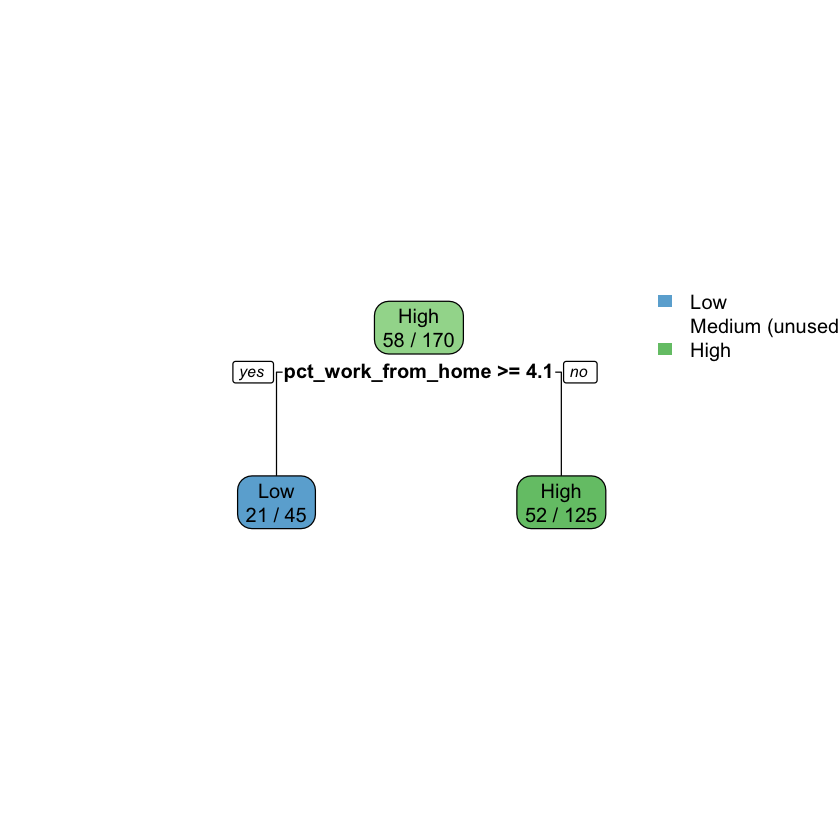
[2] Raifman, M. A., & Raifman, J. R. (2020). *Disparities in the population at risk of severe illness from COVID-19 by race/ethnicity and income*. American Journal of Preventive Medicine, 59(1), 137–139. [https://doi.org/10.1007/s10729-020-09549-8]

[3] Bertsimas, D., Boussioux, L., Cory-Wright, R., Delarue, A., Digalakis Jr, V., & Lami, O. (2020). *From predictions to prescriptions: A data-driven response to COVID-19*. Health Care Management Science, 24(2), 1–22. [[https://www.michiganmedicine.org/health-lab/flattening-curve-covid-19-what-does-  
it-mean-and-how-can-you-help](https://www.michiganmedicine.org/health-lab/flattening-curve-covid-19-what-does-it-mean-and-how-can-you-help)?]

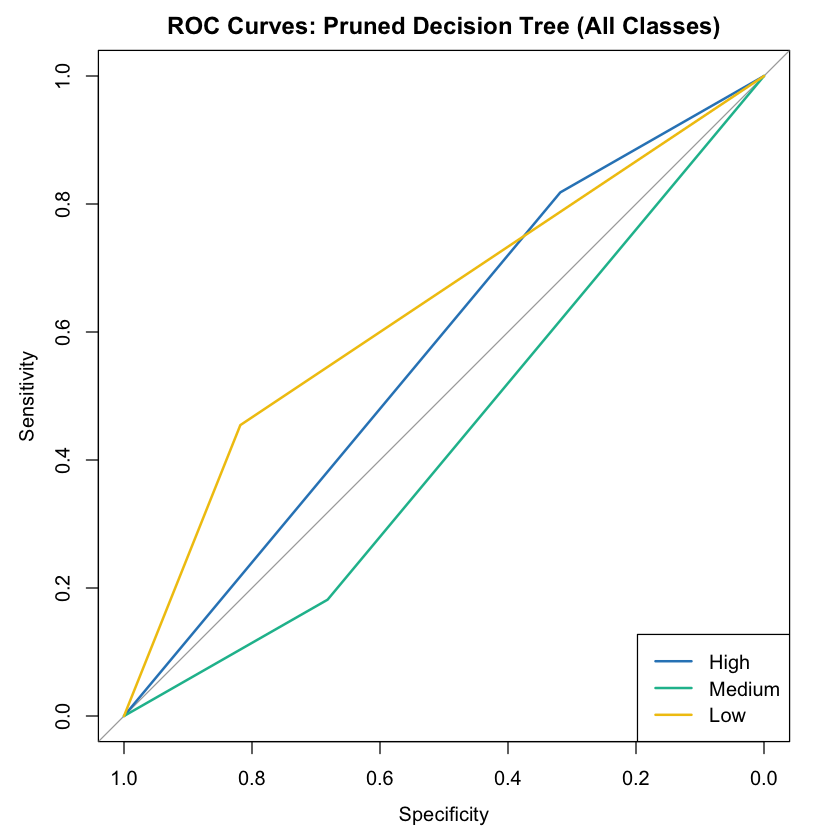
[4] Centers for Disease Control and Prevention (CDC). (2021). *COVID-19 racial and ethnic health disparities*. [https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html]

# **Appendix**

*Figure A1: Ethical Pruned Model: Excluding Main Variables*

****

*Figure A2: ROC Curve: Pruned Decision Tree*

**

**8.1 Student Contributions**

Juan Carlos Dominguez:

* Analysis and code of COVID-19 Cases Census Dataset
* Structuring the CRISP-Model Format
* Contributed to CRISP-DM Report

Salissa Hernandez

* Analysis and code of COVID-19 Cases Census Dataset
* Contributed to CRISP-DM Report
* Preparing the separate R file for submission

Leonardo Piedrahita

* Analysis and code of COVID-19 Cases Census Dataset
* Contributed to CRISP-DM Report
* Reviewing and cleaning information where needed.