# Data Mining Project 3

Olivia Hofmann and Michael Perkins

December 9, 2024

## Contents

# Data Preparation

## Define Classes

The COVID-19 pandemic highlighted the importance of preparedness for infectious disease outbreaks. Anticipating which counties are at higher risk can enable early interventions, potentially saving lives and mitigating economic impacts. This project aims to classify U.S. counties into **high**, **medium**, or **low** risk categories for future pandemics based on historical COVID-19 data and other socioeconomic factors.

The classes for COVID-19 risk levels are defined based on confirmed cases per 10,000 population per week. The following thresholds aim to categorize the severity of the pandemic into actionable categories that inform public health responses and individual precautions.

- **High Risk**: $>$ 50 cases per 10,000 population per week

  A high number of cases indicates widespread community transmission, which may overwhelm healthcare systems. This category is often used to trigger strict public health measures such as lockdowns, travel restrictions, or mass testing campaigns. The 50-case threshold for high risk captures a significant uptick in transmission, providing a signal for urgent measures.

- **Medium Risk**: 10–49 cases per 10,000 population per week

  A moderate number of cases suggests some level of community transmission. This may require targeted interventions such as localized restrictions or increased testing and vaccination efforts. The range for medium risk accommodates variability in case numbers while emphasizing the need for ongoing monitoring and targeted efforts.

- **Low Risk**: $<$ 10 cases per 10,000 population per week

  A low number of cases implies limited transmission, often seen when preventive measures are effective, or when a region is in a recovery phase. The threshold for low risk aligns with goals for maintaining control and minimizing transmission.

Examining the data helped confirm the appropriateness of these thresholds. For instance, regions with $>$ 50 cases per 10,000 showed trends of healthcare strain and higher fatality rates while regions with $<$ 10 cases were often associated with higher vaccination rates or stringent preventive measures. This classification is rooted in observed patterns and practical considerations, ensuring its relevance to real-world applications while maintaining simplicity for clear communication and policy alignment.

## Data Preparation Steps

To prepare for classification modeling, the dataset is merged, cleaned, and the data is processed to ensure that it is usable and relevant. A new column, risk_level, is created that categorizes the severity of COVID-19 cases into three levels: high, medium, and low.

These levels are defined based on the number of confirmed cases per 10,000 population per week:

- **High Risk**: $>$ 50 cases
- **Medium Risk**: 10-49 cases
- **Low Risk**: $<$ 10 cases

The risk_level column is converted to a factor, etreating it as a categorical variable in the following modeling steps. This ensures that the dataset has a clear and actionable target variable (class attribute) for classification.

Features were selected from the dataset that were likely to be predictive of the risk_level class. The selected features include:

- Mobility-related changes (**Retail Change**, **Grocery Change**, and **Workplace Change)**.
- Principal Component Analysis (PCA) components (**PC1** and **PC2**).
- The week variable, indicating the temporal context of the data.

These features are selected based on their potential to correlate with COVID-19 risk levels. PCA components are particularly useful as they reduce dimensionality while preserving variability in the data.

A preview of the processed data is displayed in Table 1 below. The data preparation steps ensure that the dataset is clean, balanced, and ready for classification modeling.

Table 1: First 10 Rows of Classification Data

| Retail Change | Grocery Change | Workplace Change | PC1 Score | PC2 Score | Week | Risk Level |
|---|---|---|---|---|---|---|
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-01-19 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-01-26 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-02-02 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-02-09 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-02-16 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-02-23 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-03-01 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-03-08 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-03-15 | low |
| 3 | 1 | -1 | -2.15 | 0.03 | 2020-03-22 | low |

# Modeling

## Model 1: Decision Tree

**Decision trees** use a tree-like structure to split data based on feature thresholds, aiming to classify samples into distinct classes.

**Advantages**:

- **Simple and Interpretable**: **Decision trees** are easy to understand and visualize, making them highly interpretable for stakeholders.
- **Fast Training and Prediction**: **Decision trees** train and predict quickly, especially for smaller datasets or datasets with few features.
- **Handles Mixed Data Types**: **Decision trees** can work with both numerical and categorical data without requiring preprocessing or scaling.
- **Captures Nonlinear Relationship**s: **Decision trees** can model complex, nonlinear decision boundaries effectively.

## Model 2: Random Forest

**Random forest** is an ensemble method that trains multiple decision trees on random subsets of the data and aggregates their predictions for classification.

**Advantages**:

- **Handles Large Datasets**: **Random forests** can efficiently handle large datasets with high feature dimensionality.
- **Robustness**: The ensemble approach reduces the risk of overfitting, providing more stable and generalized predictions.
- **Feature Importance**: **Random forests** provide a measure of feature importance, helping to identify the most influential variables in the classification task.
- **Captures Feature Interactions**: **Random forests** inherently model interactions between features due to the random splitting.

## Model 3: Support Vector Machine (SVM)

**Support vector machines** constructs a hyperplane or set of hyperplanes in high-dimensional space to separate classes with the maximum margin.

**Advantages**:

- **Effective for High-Dimensional Spaces**: **SVM** works well when the number of features is large relative to the number of samples.
- **Robust to Overfitting**: Especially effective for tasks with clear class separability in the feature space.
- **Flexibility with Kernels**: The kernel trick enables **SVM** to model nonlinear relationships by transforming data into higher-dimensional spaces.
- **Handles Smaller Subsets**: Using subsampling suits **SVM** well since it is computationally intensive on large datasets.

## Model 4: Gradient Boosting

**Gradient boosting** trains sequential decision trees, where each tree corrects the errors of the previous one by minimizing a specified loss function.

**Advantages**:

- **Highly Accurate**: **Gradient boosting** often achieves state-of-the-art performance for classification tasks.

- **Customizable**: The learning rate, tree depth, and number of iterations can be tuned for optimal performance.
- **Handles Missing Data**: **Gradient boosting** models handle missing values effectively.
- **Feature Importance**: Similar to random forest, **gradient boosting** provides insights into feature importance.
- **Handles Multiclass Classification**: The model can output class probabilities for each class, aiding in more nuanced decision-making.

## Model 5: Logistic Regression

**Logistic regression** models the probability of class membership using a logistic function and assumes a linear relationship between features and the log-odds of the outcome.

**Advantages**:

- **Simplicity**: **Logistic regression** is easy to implement and computationally efficient, even for large datasets.
- **Interpretable Coefficients**: The coefficients represent the strength and direction of the association between features and the outcome, providing clear interpretability.
- **Works Well for Linearly Separable Data**: It performs best when classes are linearly separable in the feature space.
- **Baseline Model**: **Logistic regression** serves as a reliable baseline to compare against more complex models.
- **Probabilistic Predictions**: It provides probabilities for class membership, allowing for more informed decision-making thresholds.

## Model Analysis

### Confusion Matrices

**Confusion matrix** heatmaps visually represent the classification performance of each model across the three classes: **low**, **medium**, and **high**. Each heatmap, shown in Figure 1, displays the counts of true positives, false positives, and false negatives. The color intensity indicates the magnitude of the counts, with darker colors representing higher values.

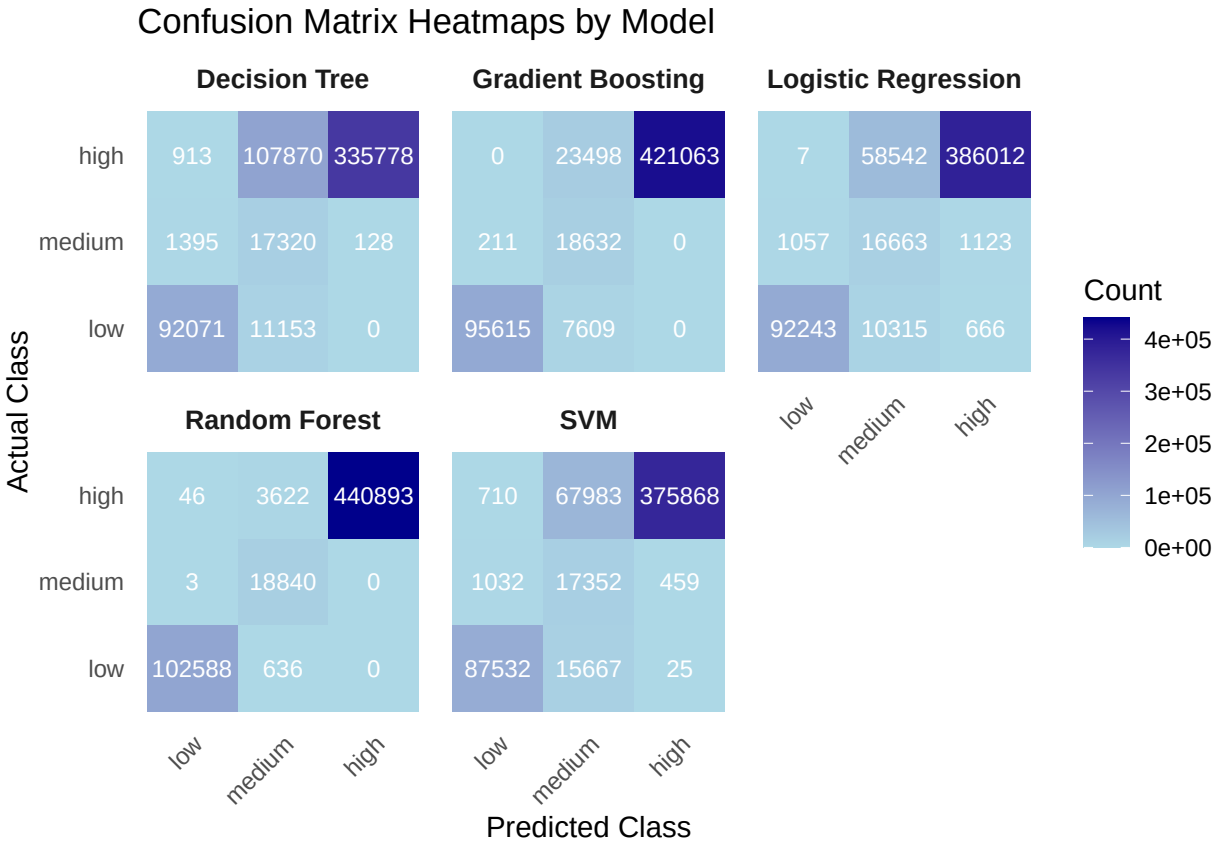


Figure 1: Confusion Matrix Heatmaps by Model

- **Decision Tree**: Performs well for the **high** class with many correct predictions, but struggles with misclassifications for **medium** and **low** classes. The model misclassifies a large number of **low** instances as **high**.
- **Random Forest**: Significantly improves accuracy across all classes. The model excels in **high** and **low** class predictions, with minimal false negatives and false positives compared to other models.
- **SVM**: Effective in classifying the **low** class, but struggles with **medium** and **high** classes. The model has a substantial number of misclassifications occur for **medium**, indicating difficulty in separating this class in the feature space.
- **Gradient Boosting**: Handles **high** and **low** classes with high precision and recall, as evident by the strong diagonal counts in the heatmap. This model struggles with **medium**, likely due to overlapping characteristics in the feature space.
- **Logistic Regression**: Performs well for **low** and **high**. The high false positives and false negatives for **medium** suggest a linear decision boundary may not fully capture the complexity of the data.

The **confusion matrices** metrics table, Table 2, provides detailed numerical insights into the performance of each model across the three classes. It includes:

- **True Positives**: Correctly predicted instances of a class.

- **True Negatives**: Instances correctly classified as not belonging to the class.
- **False Positives**: Instances incorrectly predicted as belonging to the class.
- **False Negatives**: Instances of the class incorrectly classified as another class.

Table 2: Confusion Matrix Metrics for Each Model

| Model | Class | True Positives | True Negatives | False Positives | False Negatives |
|---|---|---|---|---|---|
| **Decision Tree** | | | | | |
| Decision Tree | low | 92,071 | 461,096 | 11,153 | 2,308 |
| Decision Tree | medium | 17,320 | 428,762 | 1,523 | 119,023 |
| Decision Tree | high | 335,778 | 121,939 | 108,783 | 128 |
| **Random Forest** | | | | | |
| Random Forest | low | 102,588 | 463,355 | 636 | 49 |
| Random Forest | medium | 18,840 | 543,527 | 3 | 4,258 |
| Random Forest | high | 440,893 | 122,067 | 3,668 | 0 |
| **SVM** | | | | | |
| SVM | low | 87,532 | 461,662 | 15,692 | 1,742 |
| SVM | medium | 17,352 | 464,135 | 1,491 | 83,650 |
| SVM | high | 375,868 | 121,583 | 68,693 | 484 |
| **Gradient Boosting** | | | | | |
| Gradient Boosting | low | 95,615 | 463,193 | 7,609 | 211 |
| Gradient Boosting | medium | 18,632 | 516,678 | 211 | 31,107 |
| Gradient Boosting | high | 421,063 | 122,067 | 23,498 | 0 |
| **Logistic Regression** | | | | | |
| Logistic Regression | low | 92,243 | 462,340 | 10,981 | 1,064 |
| Logistic Regression | medium | 16,663 | 478,928 | 2,180 | 68,857 |
| Logistic Regression | high | 386,012 | 120,278 | 58,549 | 1,789 |

- **Decision Tree**: High false positives for the **high** class (108,783) indicate over-prediction of this class. Low false negatives for the **low** class show that most **low** instances are correctly identified.
- **Random Forest**: Lowest false positives and false negatives across all models for the **high** class. Demonstrates a balanced performance, excelling in correctly identifying all three classes.
- **SVM**: High false negatives for **medium** (83,650) indicate poor recall for this class. Strong performance for the **low** class but inconsistent for the others.
- **Gradient Boosting**: Achieves low false negatives for **high** (0) and **low** (211), reflecting strong recall. High false positives for **medium** (516,678), highlighting challenges in distinguishing this class.
- **Logistic Regression**: Performs adequately for **low** and **high**, with low false negatives. Struggles with **medium**, resulting in high false positives (478,928) and false negatives (68,857).

The **Random Forest** model demonstrates the most balanced performance across all classes, with minimal misclassifications and strong metrics for true positives and true negatives.The **medium** class consistently shows the highest misclassification rates across all models, indicating that it shares overlapping features with **low** and **high**. Both **Random Forest** and **Gradient Boosting** outperform simpler models (e.g., **Decision Tree**, **Logistic Regression**) due to their ability to capture complex patterns and interactions. The combination of heatmaps and metrics tables provides a comprehensive understanding of model strengths and weaknesses, helping identify areas for improvement.

### Precision, Recall, and F1-Scores

Figure 2 shows the **Precision**, **Recall**, and **F1-Score** for each class across the chosen models. These metrics provide insight into how well each model performs in predicting the different classes.

Metrics Definitions:

- **Precision**: The proportion of correctly predicted instances of a class out of all instances predicted as that class. High **precision** indicates fewer false positives.
- **Recall**: The proportion of correctly predicted instances of a class out of all actual instances of that class. High **recall** indicates fewer false negatives.

- **F1-Score**: The harmonic mean of **precision** and **recall**, balancing the two metrics.

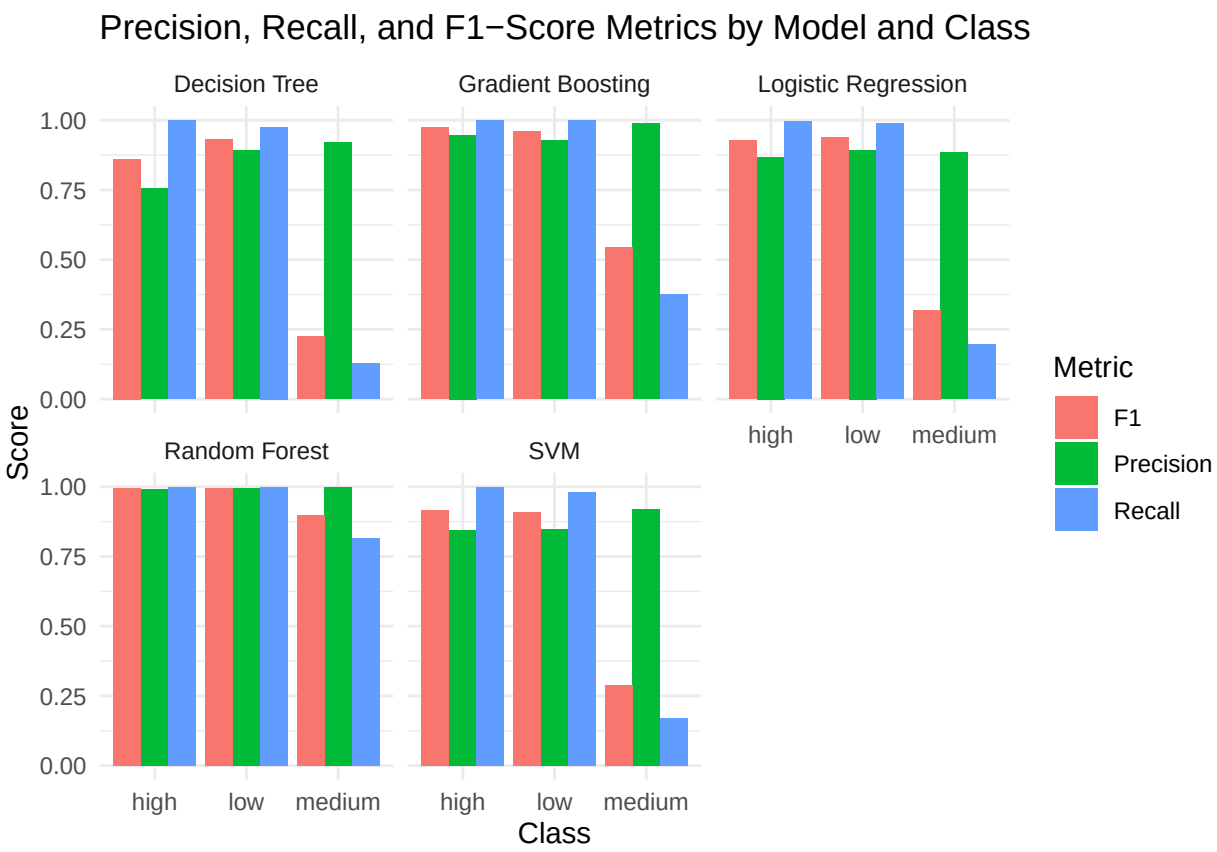


Figure 2: Precision, Recall, and F1-Score Metrics by Model and Class

- **Decision Tree**: High **recall** for the **high** class indicates most **high**-risk levels are correctly identified. Low **recall** for the **medium** class highlights challenges in detecting **medium**-risk cases accurately, leading to a low **F1-score** for this class.
- **Random Forest**: Consistent high **precision** and **recall** across all classes. This is reflected in perfect **F1-scores** (1.00) for the **low** and **high** classes and a slightly reduced score for **medium**.
- **SVM**: High **precision** for **medium** and **high** classes, but low **recall** for **medium**, indicating it struggles to capture all **medium** cases. Balanced performance for the **low** class, with reasonable **F1-score**.
- **Gradient Boosting**: Strong performance for the **low** and **high** classes, with near-perfect **precision**, **recall**, and **F1-scores**. Moderate recall for the **medium** class, which reduces its **F1-score**.
- **Logistic Regression**: Good performance for **low** and **high** classes, with balanced metrics. Significant drop in **recall** for **medium**, leading to a low **F1-score** for this class.

Table 3 presents a numerical summary of the metrics for each model and class. It complements the graph by providing exact values, making it easier to compare models quantitatively.

Table 3: Precision, Recall, and F1-Score for Each Model

| Model | Class | Precision | Recall | F1 Score |
|---|---|---|---|---|
| **Decision Tree** | | | | |
| Decision Tree | low | 0.89 | 0.98 | 0.93 |
| Decision Tree | medium | 0.92 | 0.13 | 0.22 |
| Decision Tree | high | 0.76 | 1.00 | 0.86 |

| | | | | |
|---|---|---|---|---|
| **Random Forest** | | | | |
| Random Forest | low | 0.99 | 1.00 | 1.00 |
| Random Forest | medium | 1.00 | 0.82 | 0.90 |
| Random Forest | high | 0.99 | 1.00 | 1.00 |
| **SVM** | | | | |
| SVM | low | 0.85 | 0.98 | 0.91 |
| SVM | medium | 0.92 | 0.17 | 0.29 |
| SVM | high | 0.85 | 1.00 | 0.92 |
| **Gradient Boosting** | | | | |
| Gradient Boosting | low | 0.93 | 1.00 | 0.96 |
| Gradient Boosting | medium | 0.99 | 0.37 | 0.54 |
| Gradient Boosting | high | 0.95 | 1.00 | 0.97 |
| **Logistic Regression** | | | | |
| Logistic Regression | low | 0.89 | 0.99 | 0.94 |
| Logistic Regression | medium | 0.88 | 0.19 | 0.32 |
| Logistic Regression | high | 0.87 | 1.00 | 0.93 |

- **Decision Tree**: **Precision** performs reasonably for all classes. **Recall** struggles with **medium** class (0.13), leading to a very low **F1-score** (0.22).
- **Random Forest**: Perfect or near-perfect **precision** and **recall** for all classes, making it the best-performing model overall. Slight drop in **F1-score** for the **medium** class (0.90).
- **SVM**: Strong **precision** across classes, but **recall** issues with **medium** (0.17) result in a low **F1-score** (0.29).
- **Gradient Boosting**: Near-perfect metrics for **low** and **high** classes. Reduced performance for **medium** class due to **recall** challenges, reflected in the **F1-score** (0.54).
- **Logistic Regression**: Balanced **precision** and **recall** for **low** and **high**. Struggles significantly with **medium** class (**recall**: 0.19, **F1-score*:** 0.32).

The **Random Forest** model emerges as the most robust model, with consistently high metrics across all classes. It effectively captures class boundaries and performs well even for challenging classes like **medium**. The **medium** class consistently shows lower metrics across all models, suggesting that this class has overlapping features with **low** and **high**, making it harder to distinguish.

### ROC Curves

The **Receiver Operating Characteristic (ROC)** curves measure the performance of classification models by comparing sensitivity (True Positive Rate) and 1-specificity (False Positive Rate). Each curve, in Figure 3, represents the performance of a specific model for predicting whether an instance belongs to a particular class (**high**, **medium**, or **low**) in a One-vs-All classification setup.

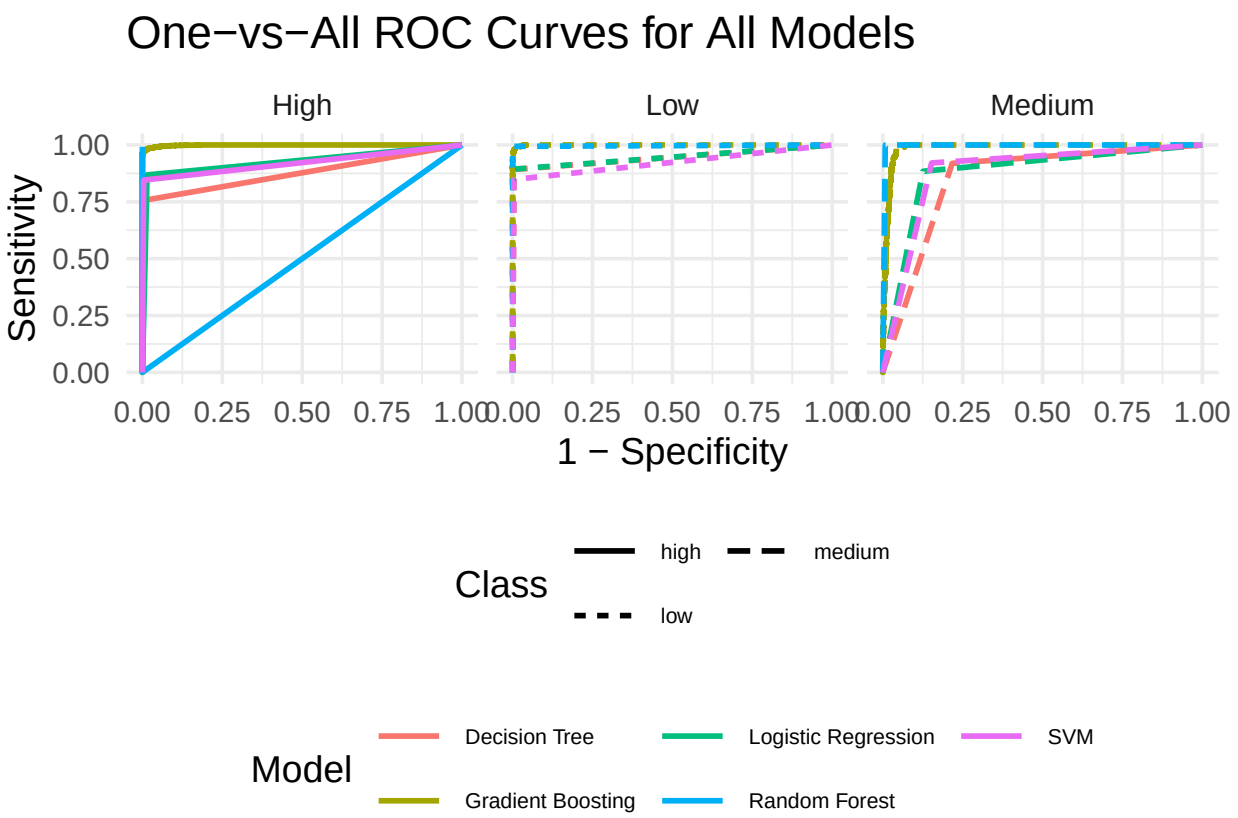


Figure 3: One-vs-All ROC Curves for All Models

- **High Class**: **Gradient Boosting** and **Random Forest** display the highest sensitivity across the full range of specificity. They indicate excellent performance in distinguishing **high** instances. **Decision Tree** and **SVM** show a drop in sensitivity at higher false positive rates. The **Decision Tree** model's curve is linear, which suggests a less nuanced separation of the **high** class.
- **Medium Class**: **Random Forest** shows the most balanced performance for the **medium** class, with consistently high sensitivity and specificity. **SVM** and **Logistic Regression** lag behind slightly, especially at intermediate false positive rates, indicating challenges in distinguishing the **medium** class. **Gradient Boosting** performs well but dips slightly compared to its performance for the **high** and **low** classes.
- **Low Class**: All models perform well, with nearly overlapping **ROC** curves close to the top-left corner, indicating high sensitivity and specificity. **Gradient Boosting** and **Random Forest** marginally outperform the other models by maintaining a steeper curve near the origin, indicating minimal false positives.

**Gradient Boosting** and **Random Forest** demonstrate the best trade-off between sensitivity and specificity for all classes. Their steeper curves and proximity to the top-left corner of the plots confirm their ability to minimize false positives while maintaining high true positive rates. All models, including the top performers, show relatively lower sensitivity for the **medium** class. This indicates that the **medium** class shares overlapping features with the **low** and **high** classes, making it harder to distinguish.

### Misclassification Analysis

The misclassification analysis, Figure 4, provides insights into the prediction errors for each model across the three risk levels. It quantifies the frequency of correct predictions (no misclassification) versus incorrect predictions (yes misclassification) for each combination of model and class. This allows for evaluating which risk levels are most challenging to predict and comparing the relative performance of models in reducing misclassifications.

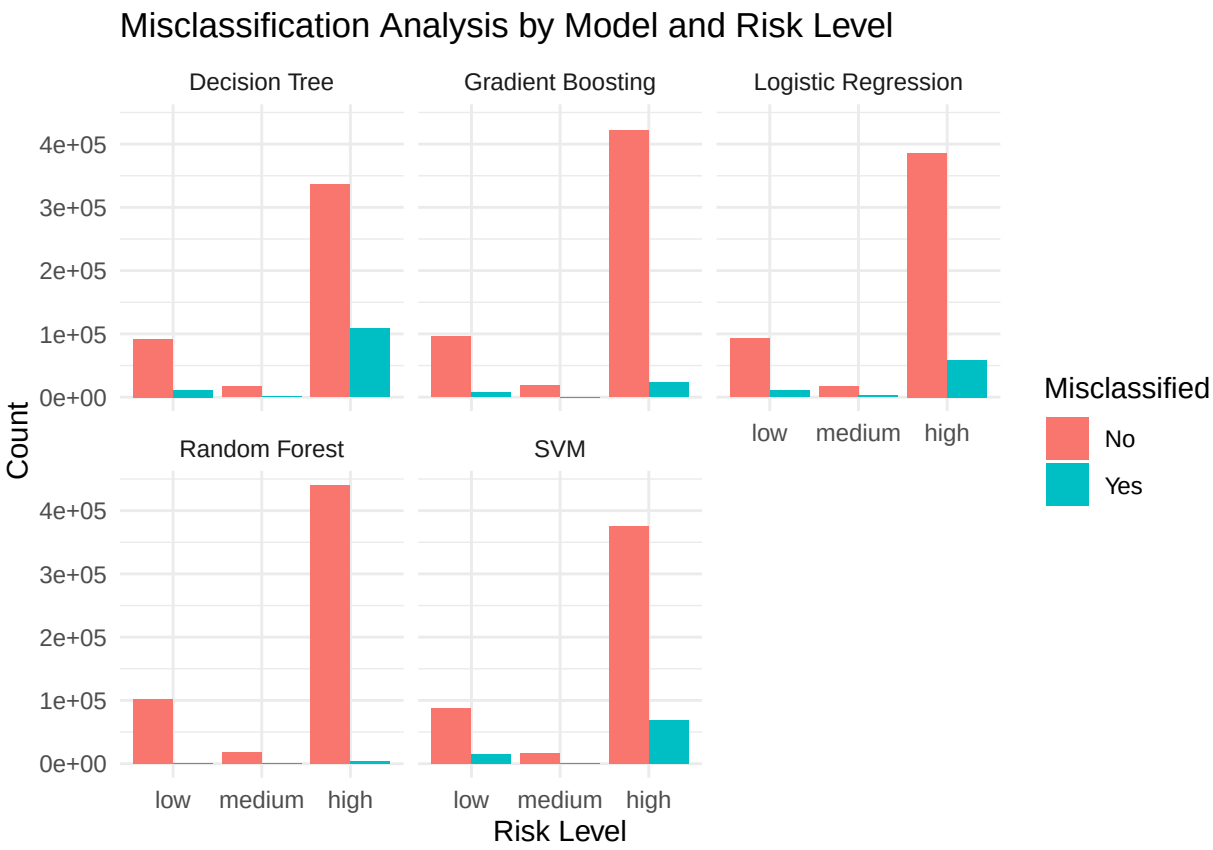


Figure 4: Misclassification Analysis by Model and Risk Level

- **Decision Tree**: High misclassification count for the **medium** class, indicating poor predictive ability for this risk level. Moderate misclassification for the **low** class, but **high** class predictions are relatively accurate. Struggles to handle the **medium** class, likely due to simplistic decision boundaries inherent to decision trees.
- **Random Forest**: Strong performance across all classes, particularly for the **high** class, with minimal misclassification. The **low** class is also well-classified, but there are moderate misclassification counts for the **medium class**. As an ensemble model, **Random Forest** captures complex interactions between features, reducing errors.
- **SVM**: Moderate misclassification counts for the **medium** and **high** classes. Performs well for the **low** class.
- **Gradient Boosting**: High accuracy for the **low** class, with almost negligible misclassification. Some challenges with the **medium** class. The **Gradient Boosting** model's iterative approach improves its ability to capture nuances in the data, reducing errors for most classes.
- **Logistic Regression**: Performs relatively well for the **low** and **high** classes, but struggles significantly with the **medium** class. High misclassification for the **medium** class may stem from its linear decision boundaries, which are less suited for complex data patterns.

The **Random Forest** and **Gradient Boosting** demonstrate superior performance, with low misclassification rates across all classes. They are particularly effective for the **low** and **high** risk levels. All models struggle with the **medium** class, likely due to overlapping feature distributions.

### Feature Importance

Feature importance analysis is crucial to identify which features contribute the most to the predictive performance of each model, understand how different algorithms interpret the dataset, and provide actionable insights for decision-making and further data engineering. This analysis, seen in Figure 5, compares feature

importance across four models: **Decision Tree**, **Random Forest**, **SVM**, and **Gradient Boosting**. The importance scores are standardized for consistency and plotted for comparison.

**Note**: **Logistic Regression** was excluded from the combined Feature Importance visualization because it primarily relies on the intercept and provides non-informative coefficients for features in this context. Including it alongside other models with meaningful feature importance metrics would distort the comparative analysis, leading to misleading interpretations.

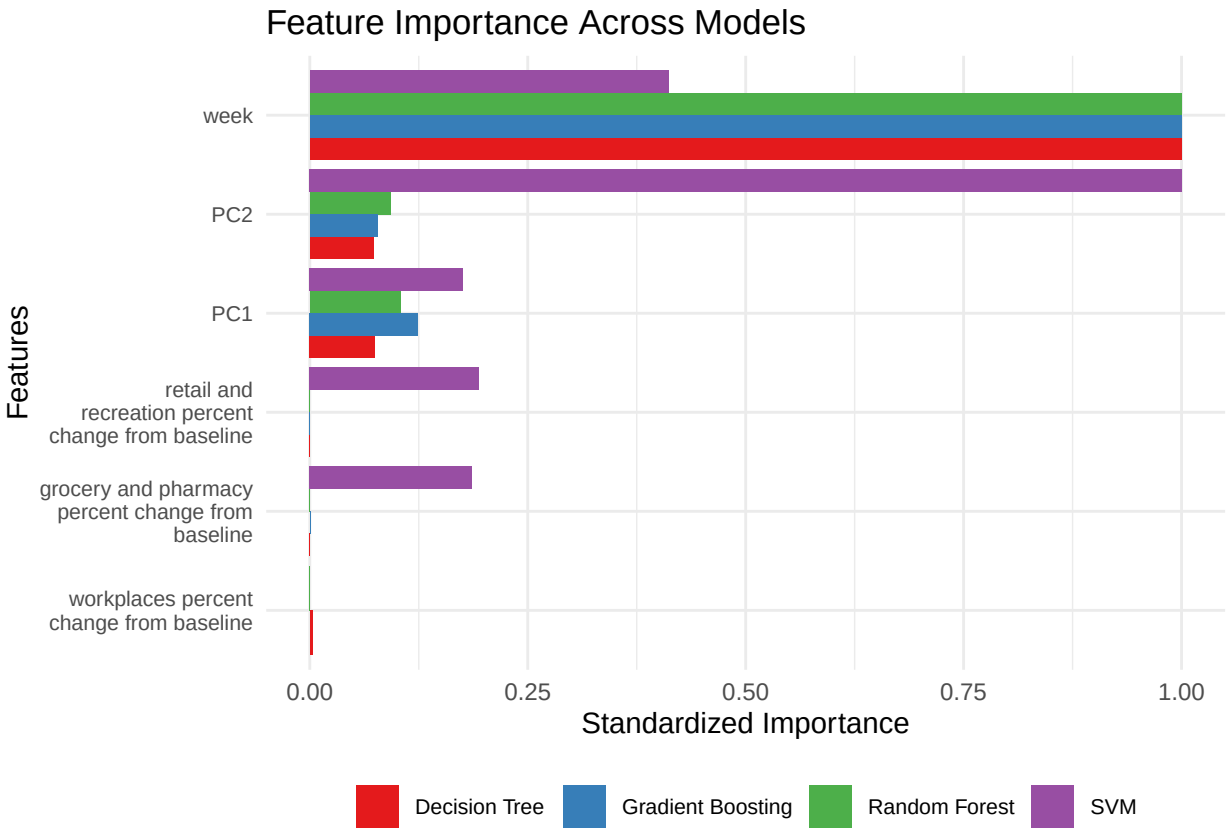


Figure 5: Feature Importance Across Models

- **week**: This feature is consistently rated as highly important across all models. The feature **week** captures temporal trends or seasonality that strongly influence the classification task. **Gradient Boosting** and **Random Forest** show the highest reliance on this feature, reflecting their strength in handling temporal patterns.
- **PC2 and PC1**: These features rank high for **SVM** and **Gradient Boosting**. **Principal components** represent linear combinations of original features and help reduce dimensionality. Their importance indicates that aggregated patterns captured by PCA strongly influence predictions.
- **retail and recreation percent change from baseline**: Important for **Gradient Boosting** and **SVM**. Changes in **retail and recreation activities** may correlate with specific risk levels in the classification task, making this feature valuable.
- **grocery and pharmacy percent change from baseline**: Important for **Random Forest** and **SVM**. Behavioral changes captured by this feature may reflect risk level distinctions.
- **workplaces percent change from baseline**: Found to have very low or negligible importance across models, suggesting limited contribution to classification.

Features like **week** and **principal components (PC1 and PC2)** consistently rank high, suggesting they capture the most predictive signal across the dataset. Behavioral features (e.g., **grocery and pharmacy**, **retail and recreation**) are also critical for certain models. **Decision Tree** and **Random Forest** place

greater emphasis on raw features like **week**. **SVM** and **Gradient Boosting** effectively leverage transformed features (e.g., principal components), indicating their flexibility in high-dimensional spaces. Features like **workplaces percent change from baseline** have low importance across models, suggesting they do not significantly contribute to risk-level predictions.

# Evaluation

Table 4 provides a quantitative comparison of the five models based on key metrics: **Accuracy**, **Precision**, **Recall**, and **F1-Score**. These metrics highlight the strengths and weaknesses of each model and help identify the most suitable model for deployment in real-world scenarios.

Table 4: Comparative Model Performance Metrics

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Decision Tree | 0.79 | 0.86 | 0.70 | 0.67 |
| Random Forest | 0.99 | 1.00 | 0.94 | 0.96 |
| SVM | 0.85 | 0.87 | 0.72 | 0.70 |
| Gradient Boosting | 0.94 | 0.95 | 0.79 | 0.83 |
| Logistic Regression | 0.87 | 0.88 | 0.73 | 0.73 |

The primary goal of the classification model is to predict COVID-19 risk levels for counties (**low**, **medium**, **high**) and provide actionable insights for early interventions. The evaluation demonstrates the following:

**Random Forest**

- **Utility**: This model achieves near-perfect **accuracy** (0.99) and perfect **precision** (1.00), ensuring minimal false positives. Its high **recall** (0.94) also indicates the ability to capture most **high**-risk cases. This balance is crucial for stakeholders aiming to allocate resources efficiently without missing critical areas requiring intervention.
- **Real-Life Impact**: With its strong performance, this model is ideal for guiding decisions like prioritizing vaccination campaigns, allocating medical supplies, and implementing lockdowns in **high**-risk areas. Its low misclassification rate across all classes ensures reliable predictions that build trust among stakeholders.

**Gradient Boosting**

- **Utility**: **Gradient Boosting** achieves excellent performance across all metrics, particularly in **accuracy** (0.94) and **F1-score** (0.83). However, its **recall** (0.79) slightly lags behind **Random Forest**, meaning it could miss some **high**-risk areas.
- **Real-Life Impact**: This model is a strong alternative, particularly in scenarios where interpretability or computational efficiency is a priority. Stakeholders can use it to refine decisions in **medium-**risk areas, where it provides slightly better precision than **Random Forest**.

**Logistic Regression**

- **Utility**: **Logistic Regression** provides moderate performance, with an **accuracy** of 0.87 and balanced **precision** and **recall** (0.88 and 0.73). While it is less effective at distinguishing **medium**-risk areas, it serves as a reliable baseline for comparison.
- **Real-Life Impact**: This model is suitable for stakeholders needing a simpler, computationally efficient option. However, it may require manual adjustment or supplementary models for improved performance in **medium**-risk predictions.

**SVM**

- **Utility**: **SVM** performs well for **low** and **high**-risk categories, but its recall for the **medium**-risk class is limited (0.72), leading to a lower **F1-score** (0.70).
- **Real-Life Impact**: Due to its computational intensity and moderate performance, **SVM** may not be ideal for large-scale deployment. However, it can still provide value in small datasets or as a supplementary model for **high**-precision tasks.

**Decision Tree**

- **Utility**: The **Decision Tree** model is the most interpretable, with an **accuracy** of 0.79 and a relatively low **F1-score** of 0.67. It struggles with capturing **medium**-risk cases due to its simplicity.
- **Real-Life Impact**: Its transparency makes it suitable for stakeholders who prioritize interpretability over performance, such as policy advisors needing clear decision-making rules. However, it is less effective in high-stakes scenarios requiring high **recall**.

**Medium**-risk misclassification is a recurring issue. Stakeholders should interpret **medium**-risk predictions with caution and supplement the model with domain expertise.

The **Random Forest** model is the most robust option for deployment, offering excellent **accuracy** and **precision** while minimizing false negatives. This ensures stakeholders can trust its predictions to make impactful, data-driven decisions. For example, it can prioritize counties for vaccination campaigns or testing efforts, minimizing wasted resources. **Gradient Boosting** serves as a strong backup, particularly in scenarios requiring better performance for **low** and **medium**-risk predictions. Models like **Logistic Regression** and **Decision Tree** provide value as interpretable baselines but are outperformed by ensemble methods for this task. By leveraging the insights provided, stakeholders can allocate resources effectively, implement timely interventions, and mitigate the societal and economic impacts of future pandemics.

# Deployment

The models can predict **high**-risk counties early, enabling stakeholders to deploy healthcare resources effectively, implement containment measures such as testing and quarantine, and mobilize vaccination efforts to high-priority areas. The high **precision** of models like **Random Forest** ensures resources are not wasted on false alarms, particularly when distributing medical supplies or setting up treatment facilities. **Low**-risk predictions can guide policymakers in lifting restrictions to support economic recovery. **Medium**-risk predictions, while challenging for all models, can help refine decisions where uncertainty exists.

The model should be updated weekly to reflect new COVID-19 data, ensuring its predictions remain relevant. Weekly updates allow it to capture dynamic changes in case counts and population behavior, critical for accurate risk categorization.

# Appendix

## Team Contributions

Olivia Hofmann and Michael Perkins split the work equally, using an iterative method.

- Olivia Hofmann: Lead on data preparation and feature engineering.
- Michael Perkins: Lead on modeling and evaluation.

## Graduate Work

- **Additional Models**: Gradient Boosting and Logistic Regression.