Data Mining Project 3

Olivia Hofmann and Michael Perkins

December 9, 2024

Contents

Introduction

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.

Data Preparation

Define Classes

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. These thresholds align with public health standards observed in similar epidemiological studies and guidelines from health authorities such as the CDC or WHO.

- **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 and 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, ensuring that it is treated 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 the table 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 Relationships: 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

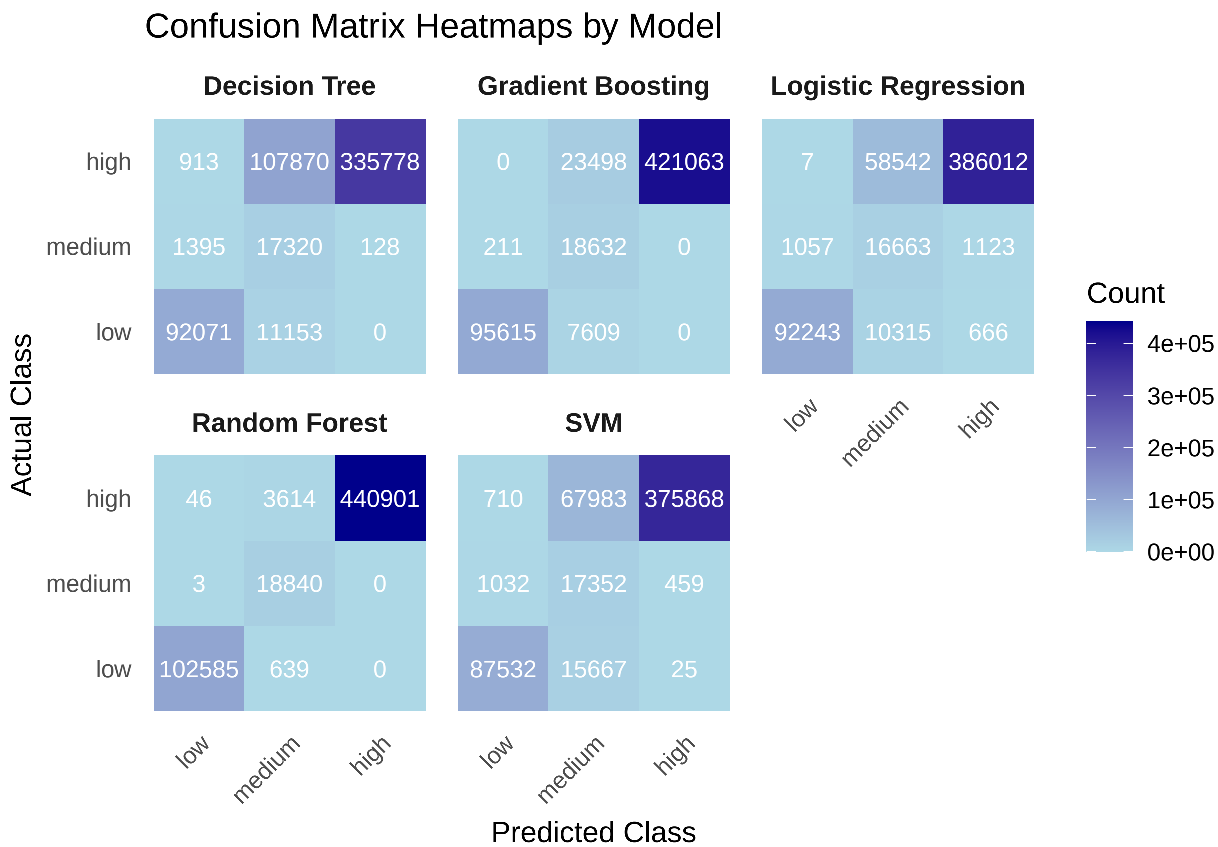


Table 2: Confusion Matrix Metrics for Each Model

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	102,585	463,355	639	49
Random Forest	medium	18,840	543,532	3	4,253
Random Forest	high	440,901	122,067	3,660	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

Precision, Recall, and F1-Scores

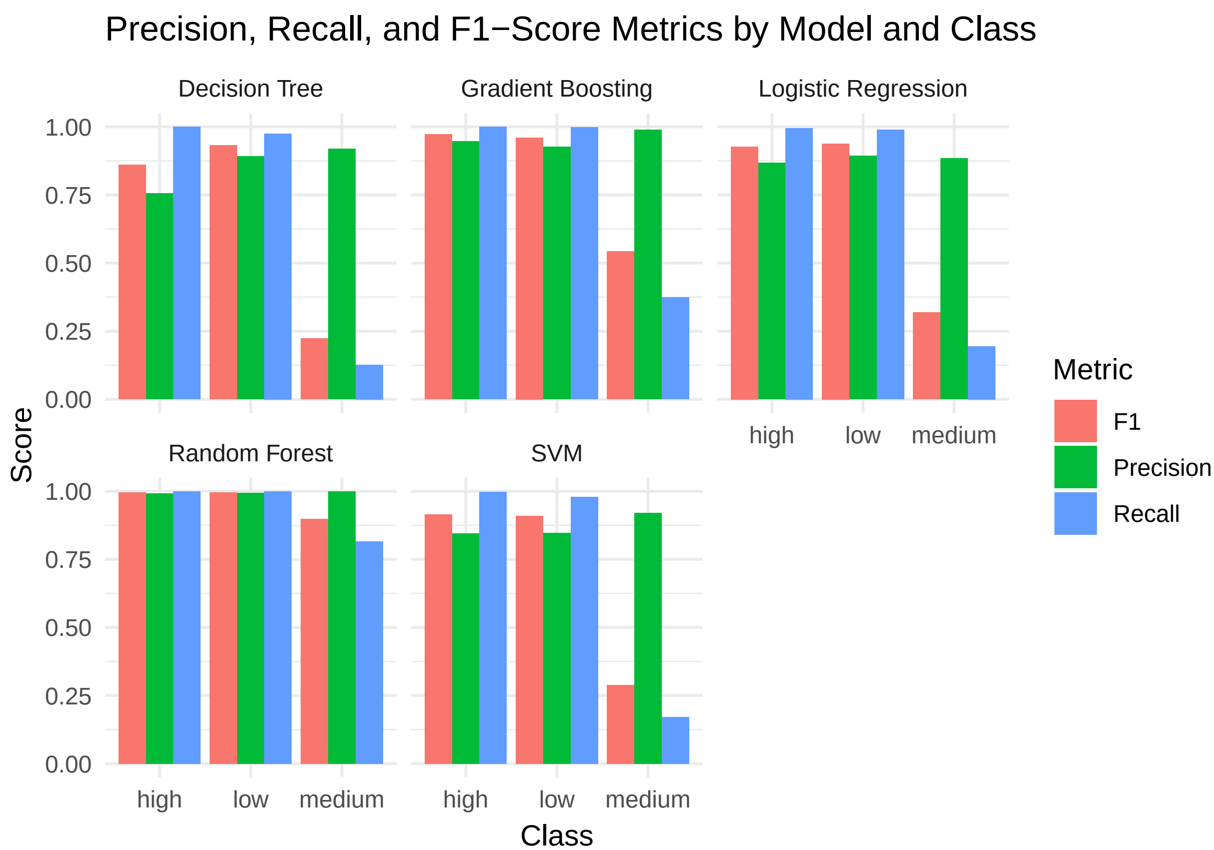


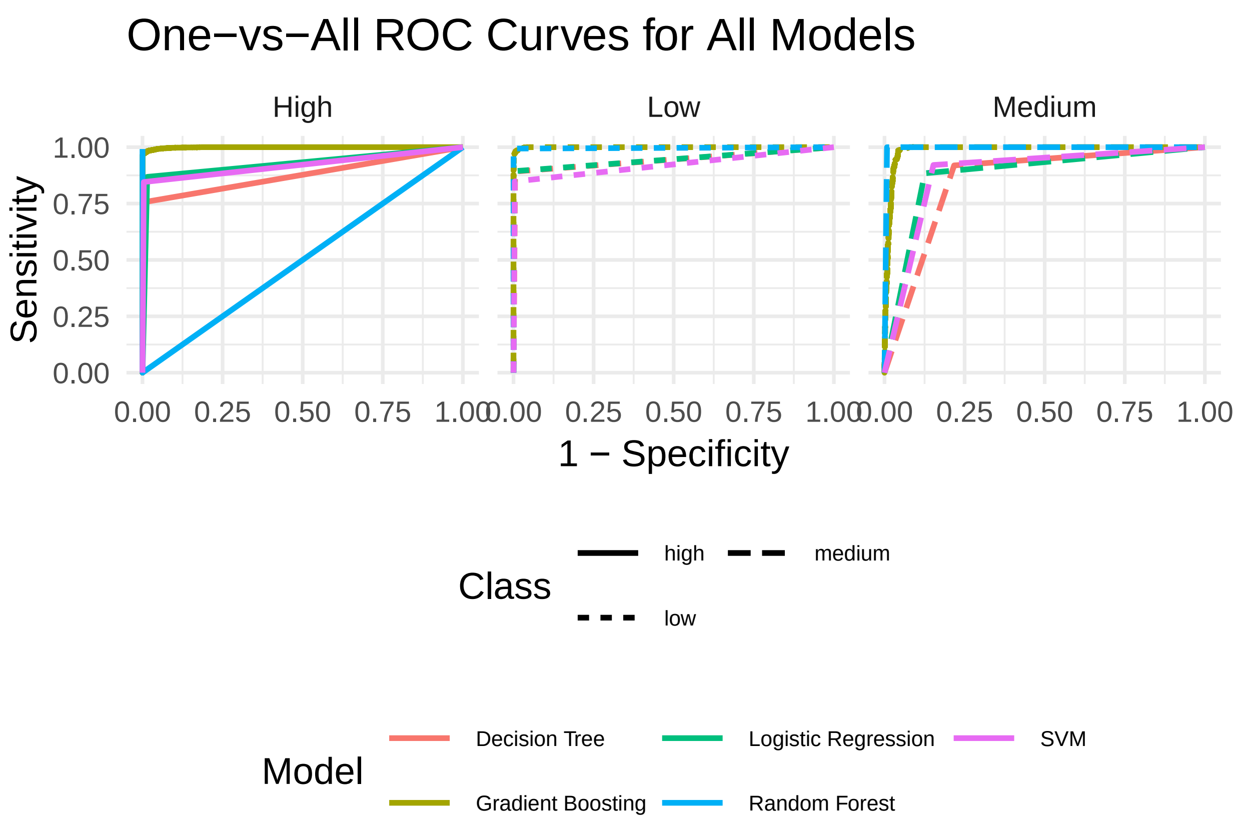
Table 3: Precision, Recall, and F1-Score for Each Model

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

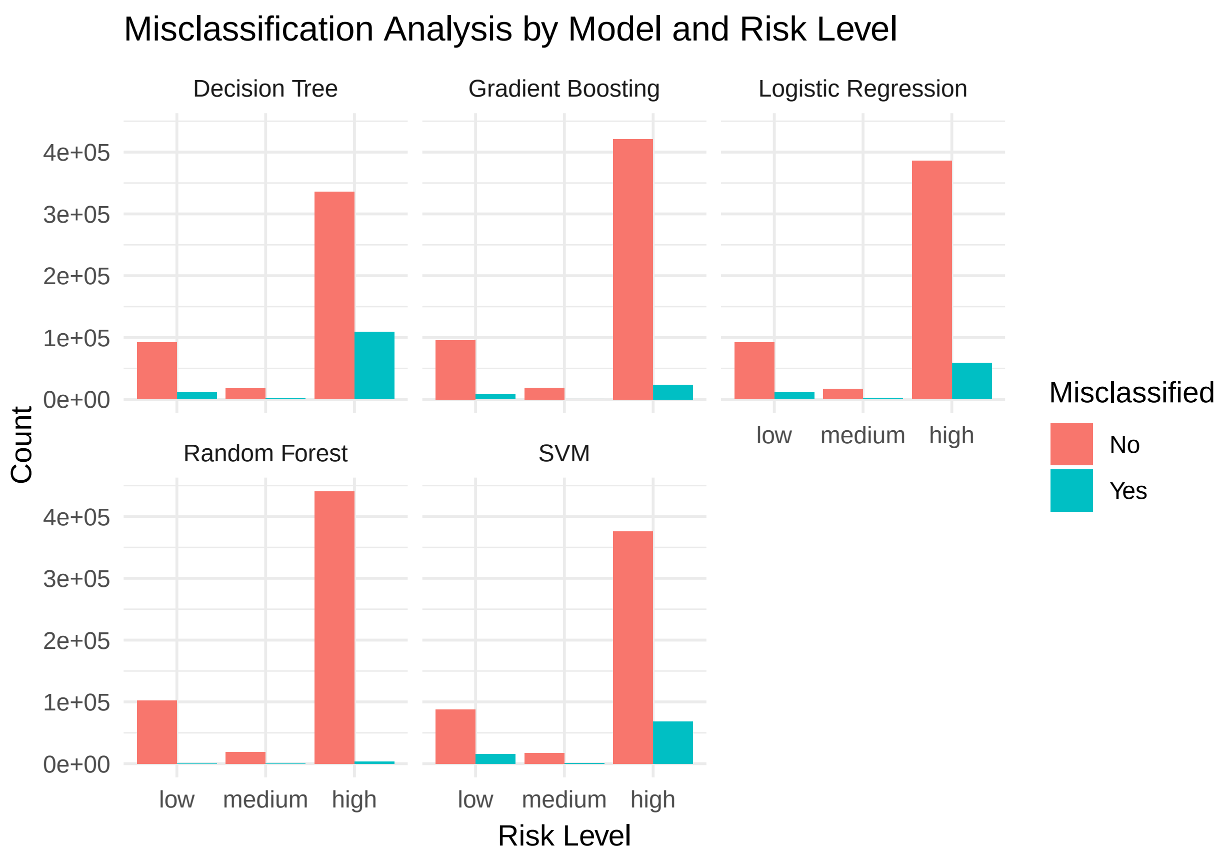
Logistic Regression	high	0.87	1.00	0.93

Precision-Recall Curves

ROC Curves

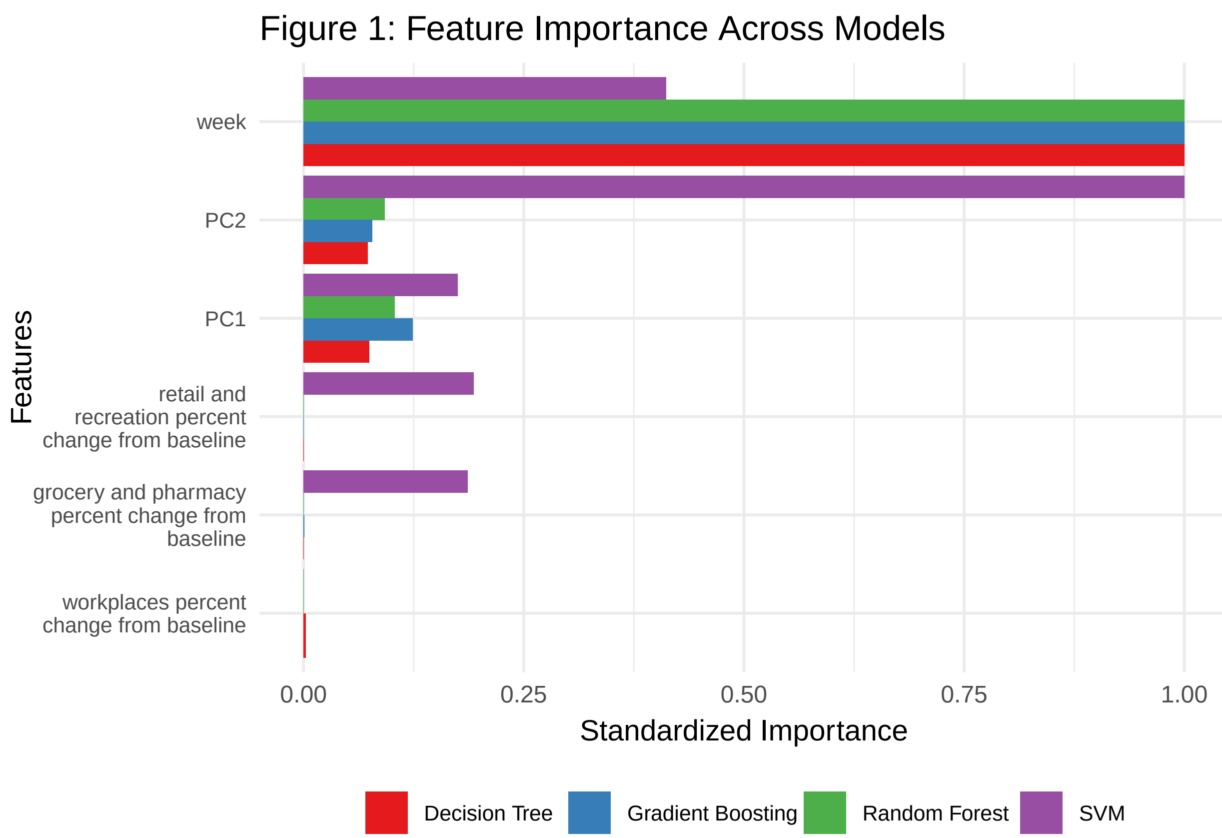


Misclassification Analysis



Feature Importance

Understanding which features significantly influence the classification outcomes is vital for both model interpretability and actionable insights. In this section, we present a consolidated feature importance analysis for four models: **Decision Tree**, **Random Forest**, **SVM**, and **Gradient Boosting**.



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.

Final Evaluation

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

Conclusion

These visuals provide insights into the models' performance and help stakeholders understand the trade-offs between accuracy, precision, and recall for each classification method. They also highlight areas for improvement, such as addressing classifications.

Deployment

- **Practical Use**: The model can guide early interventions (e.g., mask mandates, closures).
- **Update Frequency**: Weekly updates based on new data.
- **Integration**: Stakeholders can incorporate model predictions into decision-making frameworks.

Appendix

Team Contributions

- 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.