Final Paper:
Predicting Critical Health Violations in King County Restaurants

Caroline Amsbary [1] **Emily Friedman** [1] **Suhanee Singh** [1]

Abstract

Food safety inspections play a crucial role in protecting public health, yet limited resources require counties to prioritize establishments most likely to exhibit critical violations. This project develops a machine learning framework to predict whether a restaurant's *next* inspection in King County, Washington will contain at least one critical ("red") violation. Using more than 277,000 inspection records from 2006 to the present, we clean, aggregate, and transform violation-level data into a restaurant-level prediction dataset capturing historical performance, violation severity, and inspection frequency. We engineer features such as average past inspection score, proportion of red violations, mean violation points, number of prior inspections, and temporal coverage of inspection history, and use these predictors to train logistic regression, k-nearest neighbors, decision tree, and random forest classifiers

To prevent information leakage, we split the data at the restaurant level so that all inspections from the same place stayed in either the training or test set. We evaluated each model using accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices, focusing especially on recall because missing a high-risk restaurant (a false negative) is more serious for public health. Overall, the models performed very well, with accuracy between 95–96% and ROC–AUC scores between 0.96–0.99. Logistic regression, k-NN, and random forest all reached recall values around 0.96, meaning they correctly identified most restaurants that later received a red violation. When looking at feature importance, we found that a restaurant's history, especially the proportion of past red violations and average violation points, was the strongest predictor of future problems. These results show that inspection history provides meaningful information and that machine learning can help support risk-based scheduling for county health departments.

Overall, our results suggest that relatively simple classification models can accurately anticipate which establishments are most likely to incur critical violations in subsequent inspections. This provides a data-driven foundation for targeted inspections and resource allocation, and lays the groundwork for future extensions incorporating temporal dynamics, spatial modeling, and richer representations of violation patterns.

1. Data

We use the publicly available King County Food Establishment Inspection dataset (County).

1.1. Dataset Description

The dataset includes all food service establishment inspections conducted by the King County Health Department from 2006 to the present. Each row represents a single violation identified during an inspection, meaning the raw data contains multiple entries per inspection. The dataset reports detailed information including:

- Violation type (red or blue)
- Violation description and severity points
- Inspection date and inspection result
- Risk category (I–IV) based on food safety risk
- Inspection score (0 is ideal; higher values indicate worse performance)
- Business identifiers and restaurant-level metadata
- Geographic coordinates (latitude and longitude)

[1] Department of Data Science, University of Virginia, Charlottesville, VA, USA. Correspondence to: Suhanee Singh <email@virginia.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

King County is an ideal case study because it is the most populated county in Washington State and includes over 13,000 unique establishments, providing substantial variation in inspection behavior and outcomes.

1.2. Data Preparation and Cleaning

Several steps were taken to convert the raw violation-level dataset into a usable modeling dataset:

- Date and numeric standardization: Inspection dates were converted to datetime objects; inspection scores, violation points, and grades were cast to numeric types.
- Handling missing values: Missing violation descriptions or points usually indicate no violation was issued. These were filled with neutral defaults (e.g., "None," violation points $= 0$).
- Risk category extraction: Risk Category (I–IV) was extracted from free-text descriptions using regular expressions. Establishments without an explicit category were labeled "Unknown."
- Deduplication: Rows missing core identifiers (Inspection Date, Business ID) were removed.
- Aggregation to inspection-level: Since multiple violations can occur in one inspection, all violations associated with a unique inspection were collapsed into a single row, summarizing total violation points and whether any red violations occurred.
- Creation of a binary indicator: A key variable, has_red_violation, indicates whether an inspection included at least one critical violation.

This produced a clean inspection-level dataset ready for exploratory analysis.

1.3. Exploratory Data Analysis

Initial visualizations revealed several important trends:

- Inspection scores are right-skewed. Most inspections have low scores, with a long tail of more severe outcomes.
- Risk Category strongly correlates with red violations. Higher-risk categories show substantially higher rates of critical violations.
- Geography matters. Red violations cluster more densely in central Seattle and other commercial areas, motivating the inclusion of spatial features.

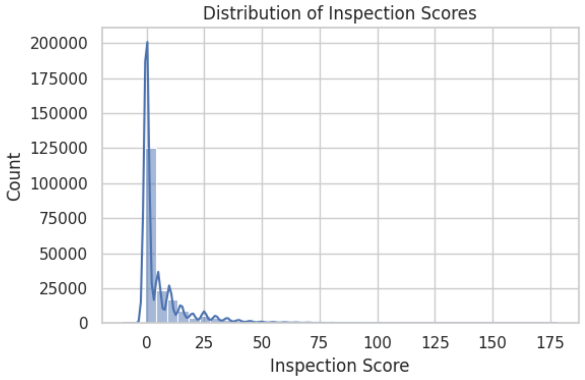

Figure 1. Distribution of inspection scores across all inspections.

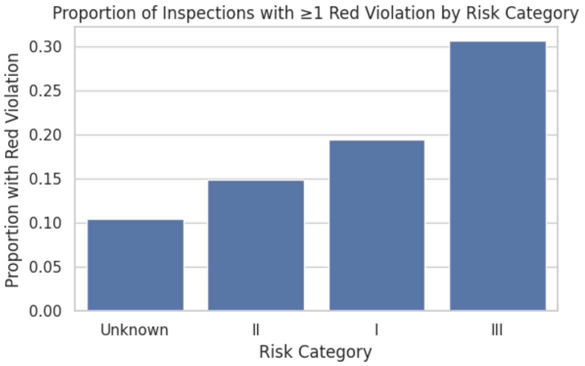

Figure 2. Proportion of inspections with at least one red violation by official risk category.

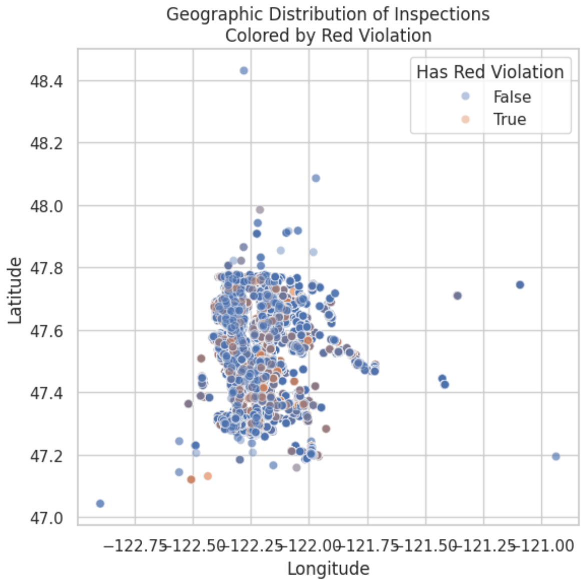

Figure 3. Geographic distribution of restaurant inspections in King County, colored by presence of a red violation.

These patterns informed our feature engineering choices and guided our modeling approach.

2. Methods and Results

2.1. Problem Definition

We define a binary classification target:

- **1** = the restaurant's next inspection contains at least one red violation
- **0** = no red violations in the next inspection

The goal is to predict this outcome using only information available from past inspections.

2.2. Feature Engineering

To build a restaurant-level prediction dataset, we aggregated all historical inspections for each establishment and constructed a set of summary features designed to capture past performance and violation patterns. For each restaurant, we computed: (1) the average past inspection score (`avg_past_score`), (2) the number of previous inspections (`num_prev_inspections`), (3) the proportion of inspections with at least one red violation (`prop_red_violations`), (4) the average violation points per inspection (`avg_violation_points`), and (5) the total timespan between the first and most recent inspection (`time_since_last_insp`). These features summarize inspection frequency, typical violation severity, and the historical consistency of critical violations.

We also defined the target variable `Next Red Violation`, which indicates whether the restaurant's next recorded inspection (after its history) contained at least one red violation. This transforms the problem into a binary classification task at the restaurant level, where each row represents the restaurant's entire inspection history up to but not including its next inspection.

After constructing the feature matrix X and target vector y, we handled any remaining missing values using mean imputation. We then performed a 70/30 train–test split to evaluate model performance on unseen restaurants. Finally, features were standardized using a `StandardScaler` for models sensitive to feature magnitude (e.g., logistic regression and k-NN).

2.3. Modeling Approach

To establish a clear benchmark for comparison, we trained four standard supervised classification models using commonly accepted default settings as a starting point. Logistic Regression served as the primary baseline because it is interpretable and performs well on linearly separable problems. The benchmark for "good performance" was defined before model training as:

- Accuracy ≥ 0.90
- Recall ≥ 0.90 for the positive class (red violation next inspection)
- ROC–AUC ≥ 0.90

These thresholds reflect a practical standard for public-health decision-making: the model must correctly identify most high-risk restaurants while maintaining overall reliability. We then trained the following models:

- Logistic Regression
  - Serves as the baseline benchmark model
- k-Nearest Neighbors (kNN)
  - k tuned from 3, 5, 7, 9, 11
  - Optimal k=7 based on cross-validation
- Decision Tree
  - Max depth tuned from 3, 4, 5
  - Best depth = 5
- Random Forest
  - 100 trees
  - Max depth = 5

Hyperparameters for all models were selected using 5-fold cross-validation on the training set to ensure fair benchmarking across methods. Figures demonstrating how optimal values were determined can be found in Appendix A.

For models sensitive to feature scale (logistic regression and k-NN), we trained and evaluated using standardized versions of the features. Decision trees and random forests were trained on unscaled data because they are invariant to monotonic transformations. Each model was trained on 70% of the restaurants and evaluated on the remaining 30%.

In addition to the supervised models, we also explored unsupervised structure in the data using a 2-cluster k-means algorithm. The resulting clusters aligned loosely with inspection history severity. This shows that restaurants naturally group into lower-risk and higher-risk profiles even without labels. This supports the idea that historical violation patterns contain meaningful predictive signal.

2.4. Performance Metrics

Models were evaluated using accuracy, precision, recall, F1-score, ROC–AUC, and confusion matrices.

Because our application concerns public health, recall is the most important metric: a false negative (failing to flag a high-risk restaurant) is far more costly than a false positive. F1-score helps balance precision and recall, while ROC–AUC measures the model's ability to discriminate between high- and low-risk restaurants across all classification thresholds, not just the 0.5 cutoff. Confusion matrices allow us to directly inspect how often each model misclassifies each category.

Together, these metrics provide a complete view of both overall accuracy and public-health-specific performance.

2.5. Empirical Results

All models performed strongly, with accuracy above 95% and high recall, indicating that the models were effective at identifying restaurants that later received at least one red violation. The decision tree and random forest models performed particularly well in terms of F1-score and ROC-AUC, suggesting that even shallow tree-based structures capture meaningful nonlinear patterns in inspection history.

Table 1. Comparison of model performance on the test set.

Model	**Accuracy**	**Recall**	**F1**	**ROC-AUC**
Logistic Regression	0.952	0.968	0.969	0.969
k-NN (best: $k = 9$)	0.950	**0.969**	0.968	0.967
Decision Tree (depth=3)	**0.959**	0.967	**0.973**	0.971
Random Forest	0.952	0.967	0.969	**0.984**

Key takeaways:

- All models successfully identified high-risk restaurants with very few false negatives.
- Random Forest achieved the highest ROC-AUC, indicating the best overall discriminative power.
- Precision for low-risk restaurants (class 0) was lower, resulting in some false positives, which is acceptable in public health contexts.

2.6. Confidence in Assessments

We carried out a validation step to better understand how the trained models behave on the held-out test set and to check that their predictions are reasonable in the context of public health risk.

First, for each tuned model in our `trained_models` set (logistic regression, k-NN, decision tree, and random forest), we generated confusion matrices on the test data. The confusion matrices confirmed the earlier summary metrics: all models correctly identified the vast majority of restaurants whose next inspection contained a red violation. As expected given the class imbalance, most errors were false positives (flagging a restaurant that did not receive a red violation), which is a safer failure mode in this application. Confusion matrices confirm that false negatives are rare across models, which is critical for identifying high-risk establishments. Below is the Confusion Matrix of the Random Forest model:

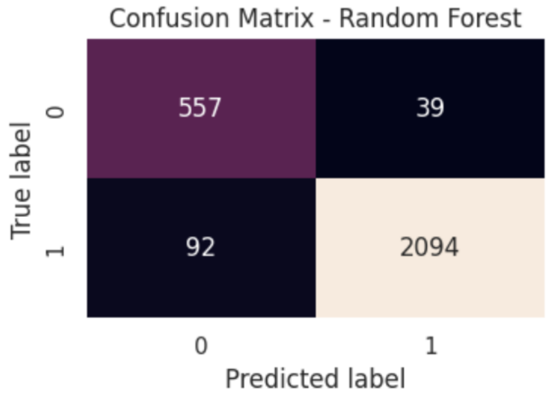

Figure 4. Random Forest Confusion Matrix

For models that provide class probabilities, we also recomputed ROC–AUC scores using the predicted probabilities on the test set. These values were consistently high (approximately 0.96–0.98), indicating that the models are able to rank higher- and lower-risk restaurants effectively. The random forest obtained the highest ROC–AUC, reinforcing its strong overall performance.

To interpret the models, we examined feature contributions where possible. For logistic regression, we analyzed the learned coefficients; features such as the proportion of past red violations and average violation points had large positive coefficients, indicating that worse historical behavior is strongly associated with a higher probability of a future red violation. For the decision tree and random forest, we inspected the feature importance values. These again showed that historical severity (proportion of red violations, average violation points) and temporal coverage (length of inspection history) were the strongest predictors. These patterns are intuitive and suggest that our engineered features capture meaningful aspects of restaurant risk.

We also plotted ROC curves for all four models to evaluate their ability to distinguish high-risk restaurants (those likely to have a red violation) from low-risk ones. All models perform well, with curves rising steeply above the diagonal, indicating strong discriminative power. The Random Forest achieves the highest AUC (0.984), followed by the Decision Tree (0.971), Logistic Regression (0.969), and k-NN (0.967), suggesting that Random Forest is the most effective overall. Using a logarithmic scale for the False Positive Rate emphasizes performance at very low FPR values, which is important when minimizing false positives. In this region, Random Forest maintains higher true positive rates than the other models, reliably identifying high-risk restaurants with fewer false alarms. Overall, while all models perform strongly, Random Forest shows a slight but meaningful advantage, particularly in the low-FPR range most relevant for practice.

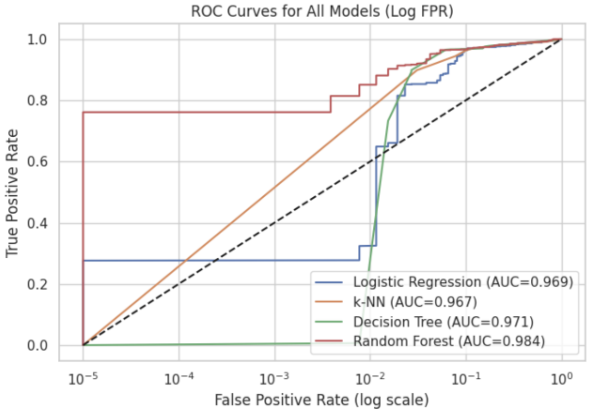

Figure 5. ROC Curves

Overall, the validation confirms that the models are not only performing well in terms of standard metrics, but are also aligned with domain expectations: restaurants with frequent or severe past violations, and longer or more irregular inspection histories, are more likely to experience critical violations in future inspections.

3. Conclusion

The project that we worked on demonstrates that machine learning models can accurately predict whether a restaurant's next inspection will include a critical food safety violation. By transforming raw violation-level data into meaningful restaurant-level features, we are able to capture key patterns in inspection history, severity, and risk classification.

Our modeling design intentionally focused on a small but diverse set of standard classification methods. Logistic regression provided an interpretable linear baseline, making it straightforward to relate each engineered feature to changes in predicted risk. k-nearest neighbors served as a simple instance-based comparator that predicts risk by analogy to similar restaurants in feature space. Decision trees allowed us to capture nonlinear threshold effects and interactions in inspection histories, while random forests leveraged ensembles of shallow trees to improve stability and generalization. This collection of models enabled a systematic comparison of performance and interpretability across different modeling paradigms, all applied to the same risk-prediction task.

Across our experiments, decision trees and random forests consistently outperformed simpler baselines such as logistic regression and k-nearest neighbors. These models achieved higher accuracy and recall and also offered interpretable measures of feature importance, which made it possible to identify which aspects of inspection history mattered most. The strongest predictors included the average severity of past red violations, the proportion of previous inspections that involved serious issues, and the number of prior inspections. These findings align with the intuitive expectation that restaurants which repeatedly display unsafe food-handling practices or receive many serious violations over time are significantly more likely to incur critical violations in following inspections. The consistency of this pattern across model classes strengthens confidence in our conclusions.

Beyond the technical evaluation, the implications of our findings highlight several opportunities for practical application. Predictive models of this kind could help health departments more efficiently allocate limited inspection resources by identifying establishments that present elevated risk. Rather than relying solely on fixed inspection schedules, agencies could use data-driven prioritization to intervene earlier at restaurants that show signs of minimal compliance. These targeted strategies have the potential to reduce the incidence of foodborne illness, improve inspection productivity, and promote equitable regulatory enforcement.

However, our work also revealed important limitations in the available data. The dataset does not include operational characteristics such as staffing levels, management turnover, restaurant size, or training practices. These are all factors that may influence violation patterns but remain unobserved in our models. Inspection frequency varies widely between establishments, which may introduce subtle forms of bias. This bias is that restaurants with more frequent inspections naturally accumulate more violation records, while those inspected less often may appear artificially compliant. In addition, our models do not currently account for external influences such as policy changes, countywide inspection backlogs, or variability in inspector behavior. These factors may introduce noise that reduces model generalizability across time.

These limitations motivate several directions for future work. Incorporating temporal modeling, such as rolling-window summaries or recurrent architectures, may allow us to capture changes in compliance behavior more precisely. Testing additional ensemble models, such as gradient boosting machines, could further improve predictive performance and stability. We also plan to evaluate fairness across geographic, demographic, and socioeconomic subpopulations to ensure that our predictions do not inadvertently reinforce disparities in regulatory enforcement.

Another direction for future work would be to move beyond purely predictive modeling toward causal and intervention analysis. Instead of only asking “Which restaurants are likely to have critical violations?”, we could also ask “Which targeted interventions (educational visits, warning letters, or follow-up inspections) are most effective at reducing future violations for specific types of establishments?”. This would combine our risk predictions with quasi-experimental methods or policy simulations to allow health departments not only to identify high-risk locations but also to evaluate which actions produce the largest improvements in food safety outcomes.

Overall, this project demonstrates the strong potential of machine learning for enhancing food safety oversight and provides a foundation for more advanced modeling in the future. By integrating richer features, improving temporal sensitivity, and expanding model evaluation, we hope to advance toward a scalable and equitable predictive system capable of supporting real-world public health decision-making at the county level.

References

County, K. Food establishment inspection data. https://data.kingcounty.gov/d/f29f-zza5. Accessed: December 7, 2025.

Contributions

- Emily Friedman completed Abstract and 1. Data
- Caroline Amsbary completed 2. Methods and Results
- Suhanee Singh completed 3. Conclusion, References, and Appendices

A. Appendix A

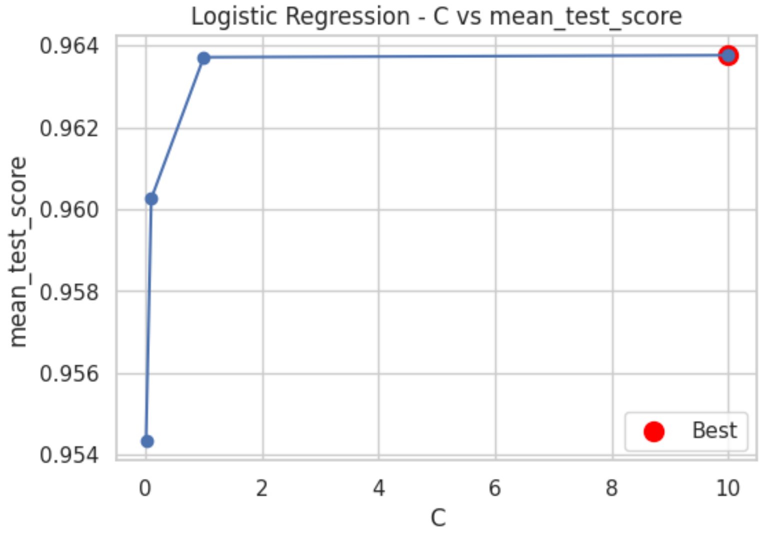

Figure 6. Appendix A1.

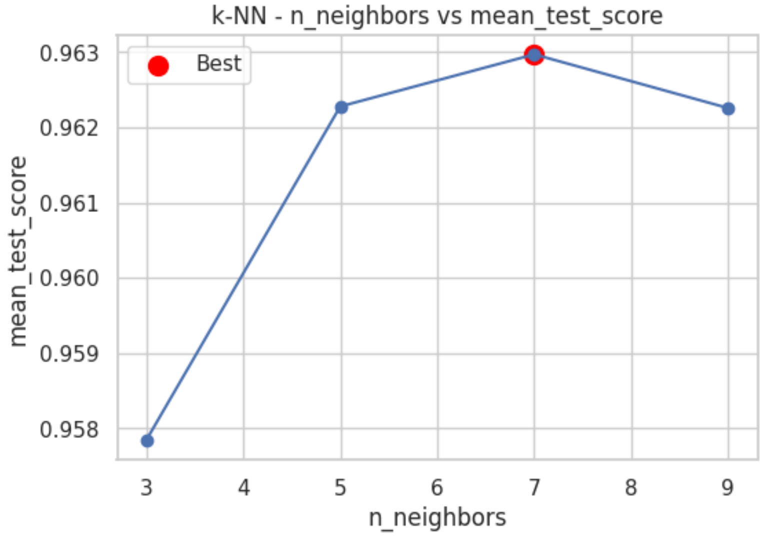

Figure 7. Appendix A2.

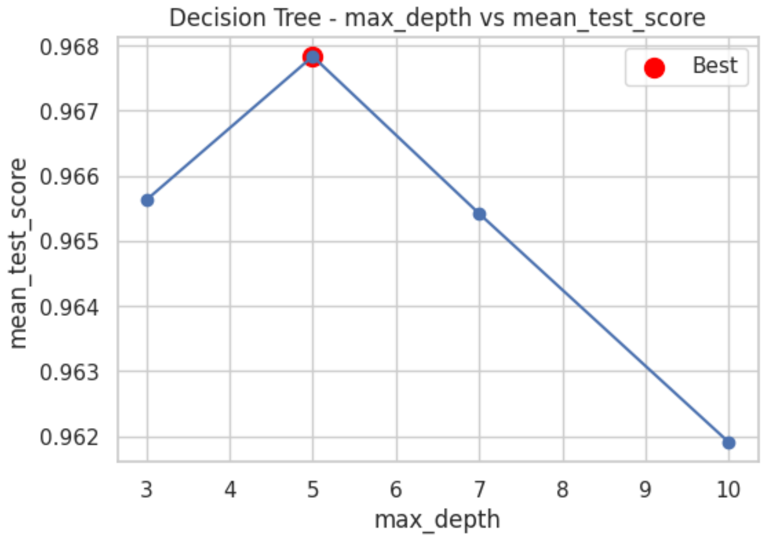

Figure 8. Appendix A3.

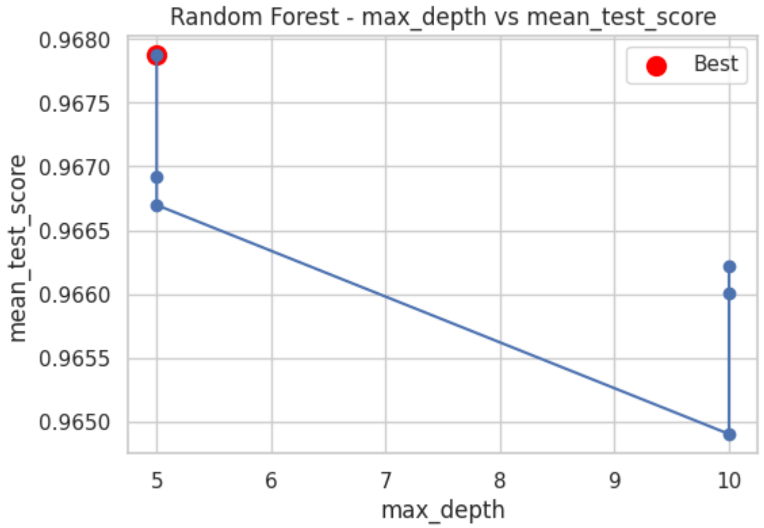

Figure 9. Appendix A4.

B. Appendix B

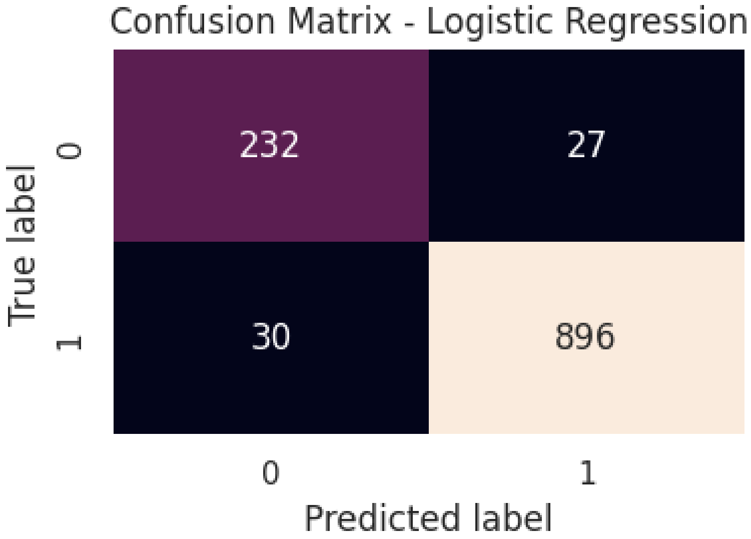

Figure 10. Appendix B1.

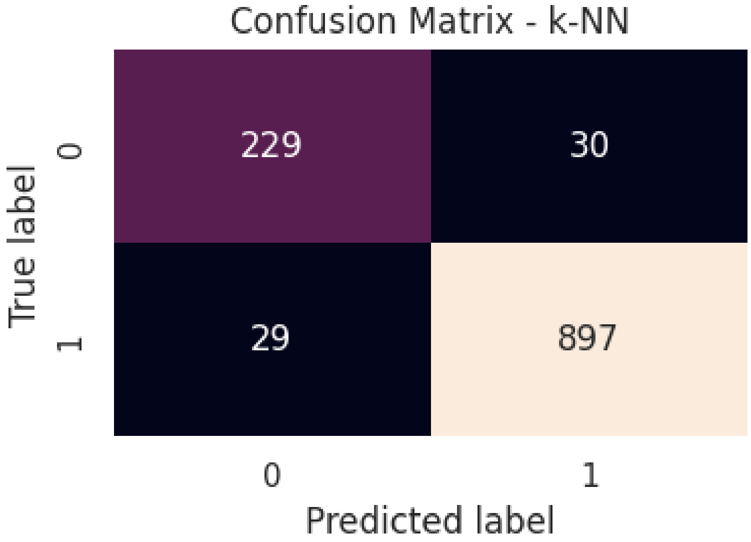

Figure 11. Appendix B2.

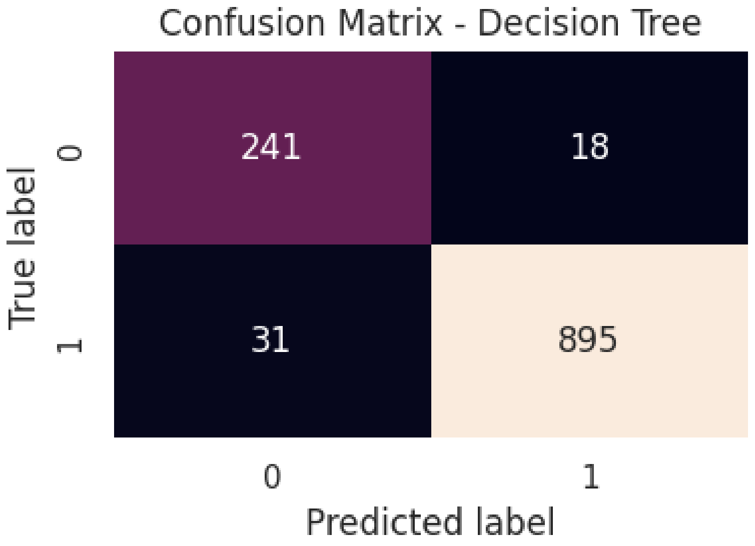

Figure 12. Appendix B3.

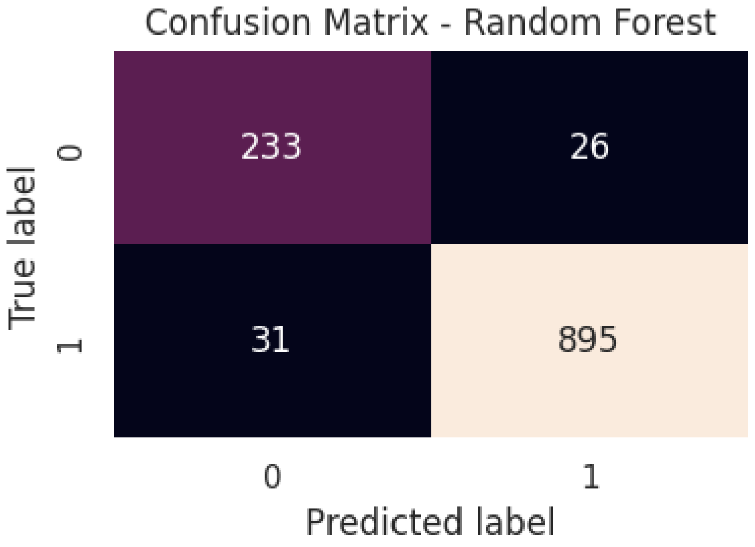

Figure 13. Appendix B4.