Predicting Beach-Related Illness with Machine Learning: Lessons from Santa Monica Bay

By Katarina Barbosa

Every summer, millions of people visit California's beaches to swim, surf, and enjoy the ocean. But few realize that contact with contaminated water can cause serious health problems. Recreational beach water contamination is a major public health concern, as bacteria from urban runoff can lead to gastrointestinal (GI) illness among swimmers. My project tackled this issue by applying machine learning to predict the severity of GI disease using data from the *Santa Monica Bay Epidemiological Study*. The goal was to explore how different machine learning models, both linear and non-linear, could reveal patterns between bacterial water quality, human exposure, and illness outcomes.

The Data: Santa Monica Bay Beach Study

The dataset I used originates from a large epidemiological study conducted at three beaches along Santa Monica Bay: Santa Monica, Will Rogers, and Surfrider. Researchers collected detailed information including 11,686 observations with 32 variables capturing microbial water quality, swimmer demographics, and health outcomes.

Participants were included if they had immersed their heads in ocean water during the study period and had not swum elsewhere in the prior week. The data included bacterial counts for several microbial indicators such as total coliforms, fecal coliforms, *E. coli*, and enterococcus. It also recorded participant demographics such as age, sex, race, and proximity to storm drains, along with symptom variables used to calculate the Quantitative Gastrointestinal (QG) severity score.

The key outcome variable was the Quantitative Gastrointestinal (QG) score, a validated measure of GI illness severity. After cleaning and preprocessing the data, I retained 8,944 valid observations, representing about 76.5% of the original sample. This final dataset provided a rich foundation for modeling disease severity as a function of water quality, swimmer demographics, and proximity to storm drains, the major sources of bacterial runoff.

Preparing the Data

Before building any models, I focused on transforming and standardizing the data. The bacterial indicator variables were highly right skewed, some with skewness values exceeding 17, because contamination levels varied dramatically depending on location and time. To normalize these distributions, I applied a $\log_{10}(x + 1)$ transformation, which successfully reduced skewness to below 0.5 across all microbial indicators. This transformation not only stabilized variance but also made the relationships between bacterial counts and health outcomes more linear and interpretable.

Each one-unit increase in the log-transformed variable represents a tenfold increase in bacterial concentration.

The GI severity score (QG) was also highly skewed, reflecting that most swimmers reported mild symptoms while a few experienced severe illnesses. Applying the same logarithmic transformation made the QG distribution nearly symmetric, allowing the models to better capture linear relationships between exposures and outcomes.

For the classification models, I converted the continuous QG score into a binary outcome. Participants with QG scores greater than 5.18 were labeled as "High severity," and those below that threshold were labeled as "Low severity." This median split created two balanced groups, which allowed fair training and evaluation of classification models.

Demographic and exposure variables, such as age group, sex, race, and distance from the nearest storm drain, were encoded as categorical features and one-hot encoded for use in machine learning pipelines. Missing values originally coded as 99999.99 were replaced with NaN values, and incomplete cases were removed for key predictors. Outliers were flagged using the interquartile range (IQR) method and treated with an iterative imputer, preserving the full dataset while minimizing the impact of extreme values. After preprocessing, I then split the dataset into 70% training and 30% testing subsets, maintaining equal proportions of high- and low-severity cases. With preprocessing complete, the dataset was ready for modeling.

Building the Models

To capture both linear and non-linear patterns in the data, I implemented four models: Multiple Linear Regression and Random Forest Regressor for predicting continuous QG severity, and Logistic Regression and XGBoost Classifier for classifying disease severity into high versus low categories.

For regression, I used LassoCV, a form of regularized regression that automatically selects relevant features through cross-validation. Of the 27 candidate predictors, 26 were retained as significant contributors to the model. For classification, I used Random Forest feature importance scores to identify the top 15 predictors, which included all four log-transformed microbial indicators along with select symptoms such as diarrhea and fever, and demographic factors like age and race.

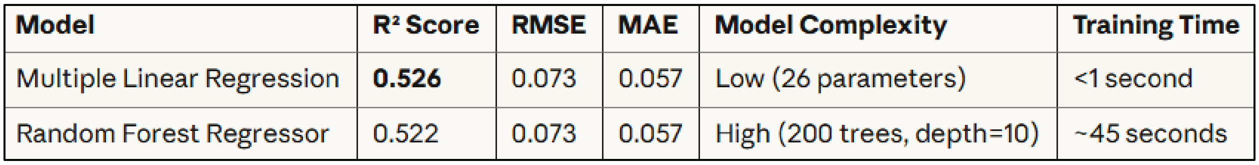
Each model underwent hyperparameter tuning using GridSearchCV with five-fold cross-validation. Although my classes were already balanced, I applied SMOTE (Synthetic Minority Oversampling Technique) to the training set to ensure robust and stable classification performance. This approach helped prevent the models from favoring one severity class over the other during training.

Regression Results: Predicting Continuous Illness Severity

The regression models sought to predict the continuous log-transformed QG severity score. Interestingly, both the Multiple Linear Regression and the Random Forest Regressor performed almost identically. The Linear Regression model achieved an R^2 value of 0.526, meaning it explained about 52.6% of the variance in GI disease severity. It also produced a Root Mean Squared Error (RMSE) of 0.073 and a Mean Absolute Error (MAE) of 0.057 on the test data.

The Random Forest Regressor, optimized through grid search with 200 trees and a maximum depth of 10, achieved nearly identical results: an R^2 of 0.522 with the same RMSE and MAE values. This finding was both surprising and informative: the more complex ensemble method offered no real performance advantage over a simple linear model. The reason lies in the data transformation; by applying logarithmic scaling, I effectively linearized the underlying relationships between bacterial contamination and illness severity.

Table 1 Regression Model Performance Comparison



Model	R^2 Score	RMSE	MAE	Model Complexity	Training Time
Multiple Linear Regression	**0.526**	0.073	0.057	Low (26 parameters)	<1 second
Random Forest Regressor	0.522	0.073	0.057	High (200 trees, depth=10)	~45 seconds

In this case, feature engineering mattered more than algorithmic complexity. The linear model, while simple and interpretable, captured nearly all the predictive information that the non-linear model could.

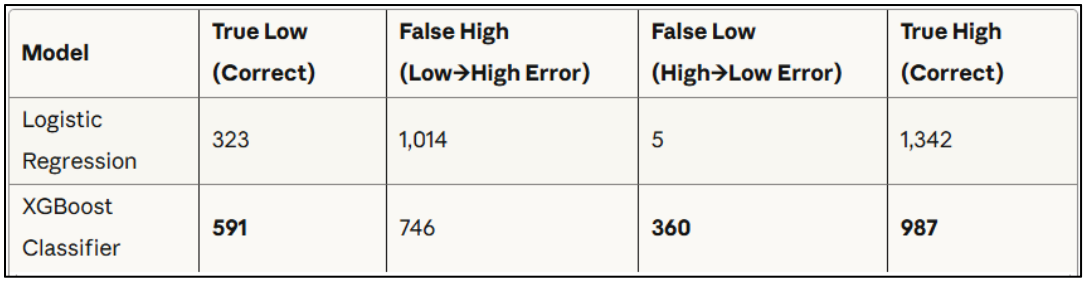
Classification Results: Identifying High vs. Low Severity

The classification task introduced more complexity. Here, the goal was to correctly classify beachgoers as having high or low GI illness severity based on bacterial exposure and other predictors.

The Logistic Regression model with L1 regularization achieved an accuracy of 62.0% and an F1-score of 0.556 after optimizing its threshold at 0.518. However, the model's confusion matrix revealed a major issue: it was heavily biased toward predicting the "High severity" class. It correctly identified nearly all low-severity cases but missed most of the truly high-severity ones; a critical flaw in a public health context where failing to identify high-risk individuals can have serious consequences.

The XGBoost Classifier, on the other hand, achieved slightly lower accuracy (58.8%) but a higher F1-score (0.579) and a more balanced classification performance. It correctly identified 44% of high-severity cases and 73% of low-severity cases. Both models achieved identical ROC-AUC values of 0.611, meaning their overall ranking ability was similar. Yet, XGBoost's ability to balance sensitivity and specificity made it the more practical choice for health risk prediction.

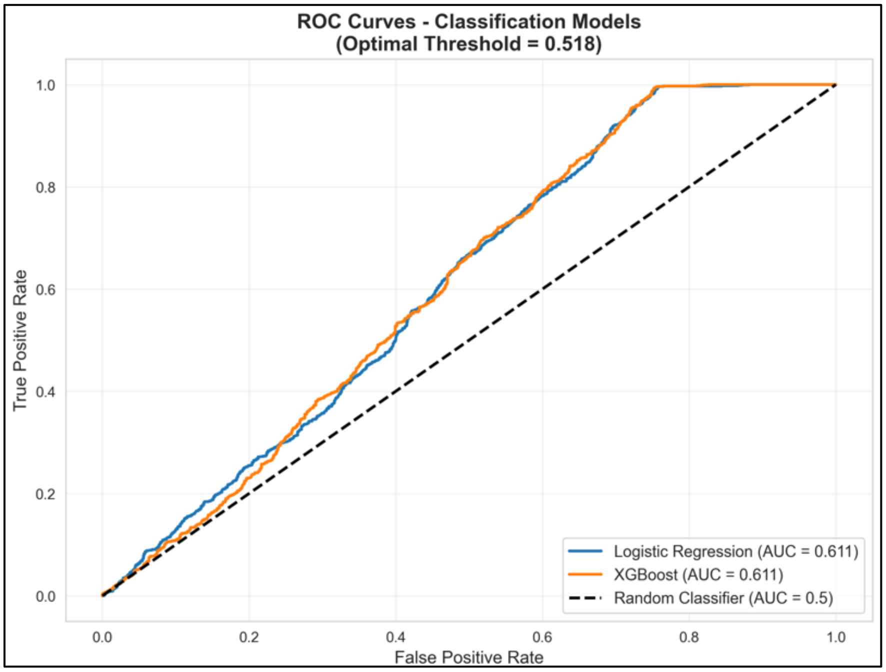
Table 2 Confusion Matrix Analysis



Model	**True Low (Correct)**	**False High (Low→High Error)**	**False Low (High→Low Error)**	**True High (Correct)**
Logistic Regression	323	1,014	5	1,342
XGBoost Classifier	**591**	746	**360**	**987**

The ROC curves (Figure 1) provide additional insight into model discrimination capability. Both Logistic Regression and XGBoost achieved identical Area Under the Curve (AUC) values of 0.611, indicating moderate but clearly superior performance compared to random classification (AUC = 0.5). The overlapping ROC curves demonstrate that both models possess similar discriminatory power in terms of their ability to separate high and low severity cases across all possible threshold settings. However, the critical difference lies in their performance at the specific operating point (threshold = 0.518) selected for deployment. While Logistic Regression achieves higher specificity at this threshold (correctly rejecting low-severity cases), it does so at an unacceptable cost to sensitivity (failing to identify most true high-severity cases). XGBoost, in contrast, operates at a more balanced point on the ROC curve, trading modest decreases in specificity for substantial gains in sensitivity.

Figure 1 ROC Curves- Classification Models



In public health modeling, a balanced classifier is far more valuable than one that maximizes overall accuracy at the expense of missing true positive cases. For this reason, XGBoost emerged as the preferred classification model despite its marginally lower accuracy score.

Comparing Linear and Non-Linear Approaches

This project underscored a key principle in applied machine learning: model simplicity can be just as powerful as complexity when data are properly prepared. For the regression task, the linear model performed as well as the non-linear Random Forest, proving that thoughtful feature transformations can linearize relationships and make simple models highly effective.

However, for classification, the non-linear XGBoost clearly outperformed Logistic Regression in terms of balanced sensitivity and specificity. While both achieved identical AUC values, XGBoost was better at adjusting to subtle, localized patterns in the data that a single linear decision boundary could not capture. This flexibility is critical when working with biological or environmental systems, where relationships are rarely perfectly linear.

Feature Importance and Practical Implications

An important part of the analysis was examining which features mattered most. In the regression models, the strongest predictors were symptom-related variables, particularly diarrhea, nausea, vomiting, and stomach pain, alongside the log-transformed microbial indicators. This makes biological sense, as those symptoms directly reflect illness severity. However, these variables are not useful for early prediction, since they occur after exposure.

In contrast, for the classification models, the feature importance rankings told a different story. Here, the log-transformed bacterial indicators dominated the list, with total coliform emerging as the single most predictive variable, followed by fecal coliform, *E. coli*, and enterococcus. This shift indicates that exposure-based variables, those measurable before symptoms occur, are most useful for predicting who might develop severe illness.

These results reinforce the importance of combining routine environmental monitoring with data-driven modeling. By translating raw water quality data into illness risk predictions, such models can support early-warning systems, improve public health advisories, and ultimately make beach recreation safer.

Conclusion

Through this project, I developed and compared four machine learning models to predict gastrointestinal disease severity among beachgoers exposed to contaminated ocean water. After rigorous preprocessing, transformation, and model optimization, the results were clear: Linear

Regression provided a transparent and efficient approach for continuous predictions, while XGBoost offered the most balanced and actionable results for classification tasks.

Ultimately, this study supports the hypothesis that higher bacterial contamination and closer proximity to storm drains increase the risk and severity of gastrointestinal illness. Beyond that, it demonstrates how data science can transform environmental monitoring data into actionable insights. With thoughtful preprocessing and model selection, machine learning can help inform public health policy, real-time beach closures, and early warning systems that protect millions of swimmers every year.

References

Gold, M., Bartlett, M., et al. (1991). *Santa Monica Bay Restoration Project.*
Santa Monica Bay Restoration Project. (1995). *State of the Bay.*