D682 Task 1

Dennis J Pothier

Student ID: 011690596

22Jul2025

B. Three Researched Algorithms and Analysis

1. Three Potential Algorithms

Naïve Bayes Algorithm: This algorithm assumes that features are independent given the labeled data, hence the Naïve part of the algorithm. “The Naive Bayes classifier does this by making a conditional independence assumption that dramatically reduces the number of parameters to be estimated.” (Mitchell, 2019) It is a probabilistic machine learning algorithm based on Bayes’ theorem of statistics. It calculates the probability of each class (health risk level) based on the input features (air quality and weather data) and selects the class with the highest probability. It is simplistic and fast for computations, leading to speedier model training and predictions. It suffers from the innate nature of assuming all features are independent, which may not fit this application best. Exploration of the dataset and subsequent training would determine whether the model is accurate enough to apply to this scenario.

Decision Tree: This algorithm assumes that features are not independent and splits the data based on the features in the dataset into a binary tree. By a series of yes/no questions, the data is classified into a series of nodes, creating the decision tree part of the model. The best features and split points are determined through training to make the most accurate model with the given dataset. This algorithm does tend to suffer from overfitting, which can lead to gross over-specification of the features, causing the model to be unable to generalize during future predictions.

Random Forest Regression: “Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently.” (Breiman, 2001) This algorithm combines the choices from multiple decision trees and creates a more robust and accurate model. It uses bagging and random feature selection to reduce overfitting, thereby improving generalization. By selecting various random samples from the data (with replacement), a decision tree for each subset of the data can be created by choosing random features for each split. The trade-off is that there is a longer computational time for training and predictions.

2. Chosen Algorithm: Random Forest Regression

a. Relation to Specified Optimization Problem

Ultimately, it was determined that the Random Forest Regression technique would be the best option to implement out of these three algorithms. It also applies to the EPA’s requirements by modeling the complex relationships between air quality, weather conditions, and environmental interactions. The ensemble approach captures linear and non-linear interactions and ensures all ecological factors are considered in its predictive modeling. The problem requires capturing nonlinear relationships among pollutants, meteorology, and health risk. RFs model complex interactions without heavy feature engineering while still providing global importance measures; “it is versatile enough to be applied to large-scale problems, is easily adapted to various ad-hoc learning tasks, and returns measures of variable importance.” (Biau & Scornet, 2016). It was chosen over the Naïve Bayes algorithm because it is highly improbable that weather features are genuinely independent of one another in actual practice. For example, humidity affects temperatures and vice versa. It was chosen over the decision tree to combat potential overfitting issues. The random forest provides the ideal balance of predictive accuracy, interpretability, and robustness for real-world EPA deployment. Unlike simpler algorithms that may lack sufficient accuracy for critical health decisions or more complex methods that sacrifice interpretability, it delivers the high performance and the transparency required for the public health sector to make informed decisions.

b. Two Strengths

Strength 1: Policy-Relevant Interpretability

Random Forest provides transparent, quantitative information about variable impact in the form of feature-importance rankings, allowing EPA officials to recognize leading causes of pollution and substantiate regulatory emphases. Global interpretability is furthered by partial dependence and interaction plots, which outline how environmental conditions help mold health outcomes. For example, if PM2.5 has an importance rating of 0.35 and a temperature rating of 0.22, monitoring and intervention efforts can effectively target those metrics. Ultimately, the model transforms data into actions and talking points for stakeholders.

Strength 2: Resiliency to Raw Data

Environmental data streams typically mix many variable types and are riddled with outliers, missing data, and sensor noise. Random Forests alleviate these challenges natively: tree splits require no scaling, are resistant to extreme observations, and ensemble averaging dampens random measurement error. Bootstrap sampling and surrogate split also preserve predictive performance when sensors fail or are offline for maintenance. This robustness enables continuous, large-scale monitoring with little preprocessing or fragile data-cleaning pipelines.

c. Two Potential Limitations

Limitation 1: Poor Temporal Dependency Modeling

As Random Forest predicts each instance independently, it does not naturally capture temporal patterns such as cumulative exposure, seasonal trends, or lagged effects. These patterns are essential for environmental health uses, where risk often depends on sequences of past data rather than individual snapshots. As a result, essential dynamics such as prolonged pollution events can be underemphasized unless additional time-derived features or auxiliary time-series models are included.

Limitation 2: Case-by-Case Explainability is Limited

Though Random Forests provide good global interpretability, the ensemble structure of the model makes it challenging to explain individual predictions. The process of averaging over numerous trees can obscure the rationale for a single risk prediction, which becomes problematic when policymakers must justify particular choices or answer public queries. The collective prediction can temper signals that single trees identify as being high risk, and confidence communication, along with instance-level explanation, becomes arduous without auxiliary explanation methods.

D. Test and Validate

1. Two metrics for evaluation

Evaluation utilizes Root Mean Squared Error (RMSE) and Coefficient of Determination (R²) with Cross Validation on the training data to check for overfitting. The RMSE quantifies the average magnitude of errors in predictions, while R² indicates the proportion of variance in the dependent variables explained by the model. RMSE describes how close predictions are to the actual values, while R² describes how well the model captures the overall variability in the data. RMSE and R² are used because they directly express error magnitude and variance explained, which are crucial for judging public‑health impact.

3. Results Interpretation

The model achieved Holdout RMSE ≈ 0.113 and R² ≈ 0.972, with CV RMSE ≈ 0.112 ± 0.017 and CV R² ≈ 0.973 ± 0.013—strong performance, but high accuracy warrants leakage checks and time-aware validation (Kohavi, 1995). The high accuracy is probably because this is likely engineered data for academia's sake, and a real-world evaluation would not yield as accurate a prediction.

4. Area for Improvement

The next step is hyperparameter tuning, customary for Random Forest regressors. This helps to lower error while minimizing overfitting risk. Tuning systematically explores the model’s configurations to locate settings that better balance bias and variance rather than relying on the defaults. As Bergstra and Bengio observe, “Randomly chosen trials are more efficient for hyper-parameter optimization than trials on a grid” (Bergstra & Bengio, 2012), which leads to the conclusion that the chosen method will likely be that of a Random Search instead of a Grid Search.

E. References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Biau, G., & Scornet, E. (2016). A random forest guided tour. TEST, 25(2), 197–227. https://link.springer.com/article/10.1007/s11749-016-0481-7

Mitchell, T. M. (2019). Naive Bayes and logistic regression (Lecture notes). Carnegie Mellon University. https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305. https://www.jmlr.org/papers/v13/bergstra12a.html