Predicting Contaminated Lakes
AI6102: Machine Learning Final Project

Tricoire Timour (*G2304288L*), Joe Tan (*G2302851L*), Joseph UWINEZA (*G2303477F*)

26 April 2024

1 Introduction

Predicting the presence of human waste in lakes is an important problem with implications for water quality and public health. This report presents a machine learning approach to address this challenge as part of an ongoing Kaggle competition titled "Predicting Contaminated Lakes" (https://www.kaggle.com/competitions/predicting-contaminated-lakes). The competition, ending on May 24, 2024, aims to assess the biogeochemical health of lakes within the US using data from the National Lakes Assessment (NLA) reports conducted by the Environmental Protection Agency of the US government in 2007 and 2012.

The goal of this project is to develop an accurate predictive model that can classify lakes as containing human waste or not based on various physical and chemical measurements collected in the NLA survey. The evaluation metric for the competition is the F1-Score, which considers both precision and recall.

2 Challenges

The main challenges associated with this problem include the complexity of the data and its particular nonlinearity; the unbalanced distribution of the target variable also may affect the model's performance negatively; the lack of domain knowledge on lake ecology; finally, the proper selection and tuning of machine learning models to meet the selected task and to achieve high predictive accuracy.

3 Data Loading and Pre-processing

Based on the physical distribution of the data as shown in Figure 1, the data was collected from the contiguous United States. The huge geographical distribution results in a large variation within the input data for certain chemical and physical measurements.

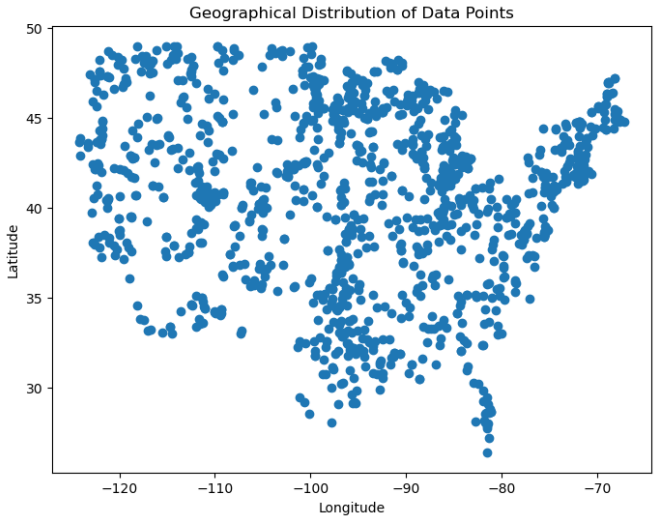

Figure 1: Geographical Distribution of Data Points

To pre-process the data, an initial check was conducted and returned with no missing values. Subsequently, the data was plotted in histograms to assist in data visualisation (The histograms being too large to readably fit in this report, they are available upon request). The visualisation of the data indicated that the general chemical and physical measurements followed a roughly normal distribution curve, however, certain outliers exist with certain measurements skewed. The notable outliers includes the size of the lakes which accounts for smaller lakes and the great lakes. Additionally, certain skewed measurements can be attributed to the different climate faced across the contiguous United States to affect the snowfall along with the different amount of agriculture uses such as the Midwest region.

The correlation matrix was constructed using the Pearson's Correlation Coefficient for each variable pair. The resulting symmetric metric with values ranging from -1 to +1 represents the strength and direction of correlation.

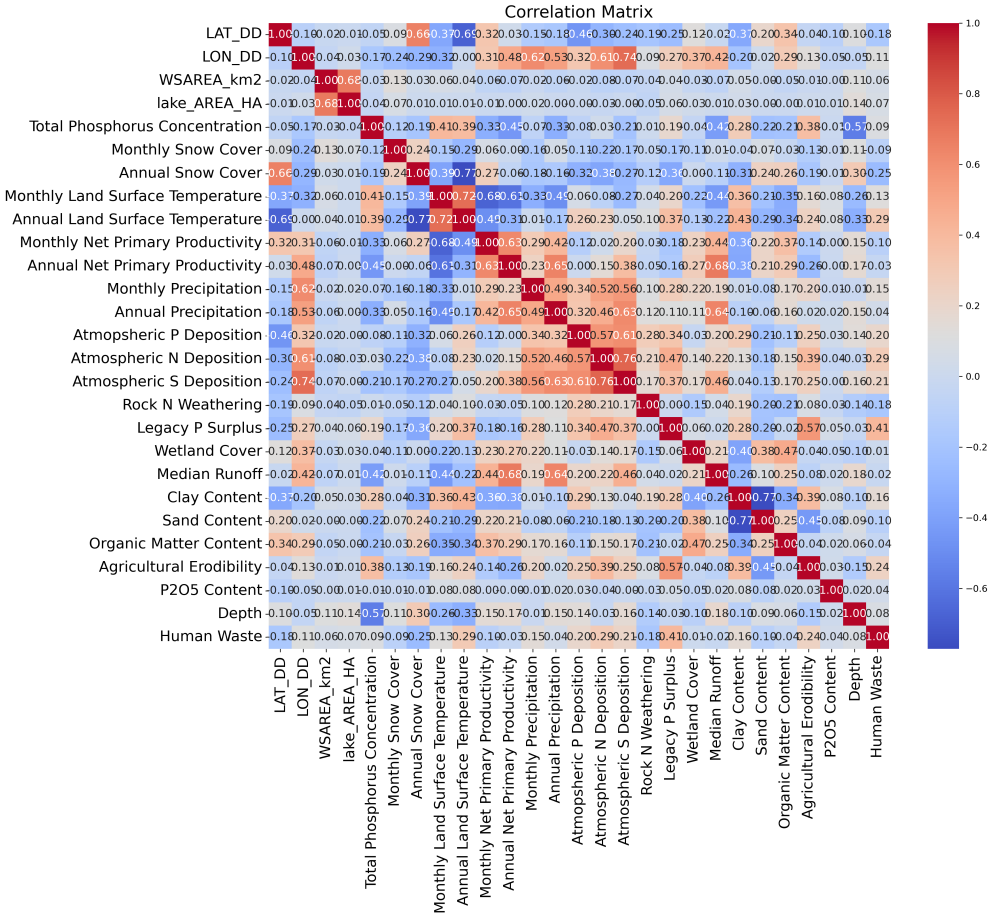

Figure 2: Correlation Matrix

The correlation matrix revealed several notable patterns and clusters of strongly correlated variables. Latitude, longitude, and atmospheric deposition variables (P, N, S) exhibited high positive correlations (> 0.9), suggesting strong regional patterns in these factors. Temperature variables, both monthly and annual, formed another cluster with very strong positive correlations (0.8 to 1.0), reflecting the inherent relationships between these measures. Precipitation and snow cover variables also displayed strong positive correlations within their respective clusters.

Latitude showed negative correlations (-0.6 to -0.9) with most temperature variables, aligning with the trend of decreasing temperatures at higher latitudes. Atmospheric deposition variables had moderate negative correlations (-0.3 to -0.6) with snow cover, potentially indicating the role of snow in limiting atmospheric fallout.

A few variables, such as lake area, depth, wetland cover, and human waste, exhibited weak correlations with most other factors, suggesting their relative independence in the dataset.

Without expert knowledge in the domain, a PCA was conducted to shows the explained variance ratio for each principal component, ordered from highest to lowest. As shown in 3 a clear inflection point at the second principal component is visible, indicating that two principal components capture the majority of the meaningful information in the dataset. The remaining components likely represent residual noise or less important dimensions. Based on the scree plot, the first two principal components would preserve the most variance and would thus be the minimal

components needed during the experiment.

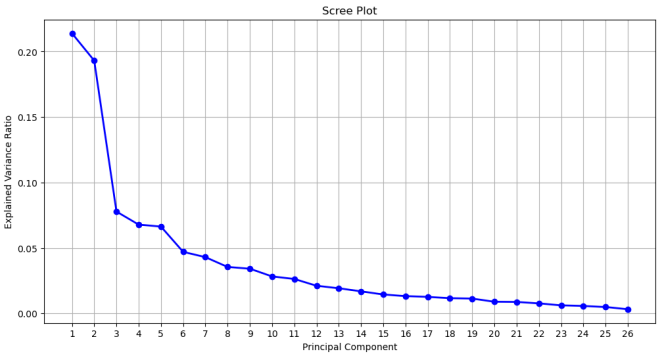

Figure 3: Scree Plot

4 Proposed Solution

4.1 Approach

The proposed solution involves comparing various machine learning models to find the best approach for predicting human waste in lakes. The models considered include Logistic Regression, Random Forest, Feedforward Neural Networks (FNNs) with varying depths, Gaussian Naive Bayes, CART, Kernel SVM, Bagging, and Quadratic Discriminant Analysis (QDA). The motivation is to explore different model architectures and algorithms to capture the underlying patterns in the data.

4.2 Detailed Steps

The process of the experiment is first to load the data. The scikit-learn StandardScaler is then used to ensure that different features do not have greater effects depending on their magnitude. Next, data augmentation is performed using specified parameters such as noise standard deviation (0.1) and scaling ranges ([(0.8, 1.2), (0.9, 1.1), (0.95, 1.05)]). The addition of random Gaussian noise allows for the models to simulate real-world scenarios and to prevent overfitting of the models on the training data. The scaling ranges were used to scale the data within different intervals, allowing the model to learn from different variation in magnitude. The code iterates over different scaling ranges and for each combination, and performs a stratified 5-fold cross-validation on the augmented data. PCA was also incorporated into the pipeline in an attempt to reduce the dimensionality of the augmented data before training the models. The number of principal components to retain was also iterated over to determine the optimal number of components that yields the best performance. For each combination of scaling range and number of principal components, the F1 score was calculated for each fold, and the mean F1 score across all folds is computed. The combination of scaling range and number of principal components that yields the best mean F1 score was selected. This allows for the identification of the most effective data augmentation strategy and the optimal number of principal components to use for dimensionality reduction. Finally, the model with the highest F1 score is trained on the full training set using the selected scaling range and number of principal components. Predictions are then made on the test set based on that model. This approach ensures that the model's performance is optimized by considering different data augmentation techniques, dimensionality reduction using PCA, and selecting the best combination of parameters that maximizes the F1 score.

5 Experiments

5.1 Verification of Motivations

The cross-validation results verify the motivations behind using different models. The Random Forest model achieved the highest cross-validated F1 score of 0.909, outperforming other models. This suggests that the Random Forest's ensemble approach and ability to capture non-linear relationships are effective for this problem. Meanwhile, QDA, which assumes that the decision boundary can be expressed as a quadratic form – its poor results simply show how inaccurate this assumption is in this case. Similarly, the Gaussian Naïve Bayes classifier makes incorrect

assumptions, key amongst which the independence assumption after which it is named. The other classifiers tend to behave similarly, with a score around 0.87.

5.2 Performance Evaluation

The performance of each model is evaluated using 5-fold stratified cross-validation and the F1 score metric. The cross-validated F1 scores for the models are as follows:

Method	F1 Scores
Logistic Regression	0.832
Random Forest	**0.909**
2-layer FNN	0.864
3-layer FNN	0.873
4-layer FNN	0.876
Gaussian Naïve Bayes	0.601
CART	0.867
Kernel SVM	0.856
Bagging	0.895
QDA	0.574

At time of submission, our team (under the name Joe Tan) is at the top of the leaderboard, as shown in Figure 4, with a Random Forest solution that also makes uses of data preprocessing, performing similarly on the test set than during cross-validation (0.901). For the competition, the leaderboard is calculated with approximately 43% of the test data with the final results will be based on the other 57%, so the final standings may be different. This preliminary result demonstrates the superiority of ensemble models, and the strength of decision trees in situations where features are meaningful.

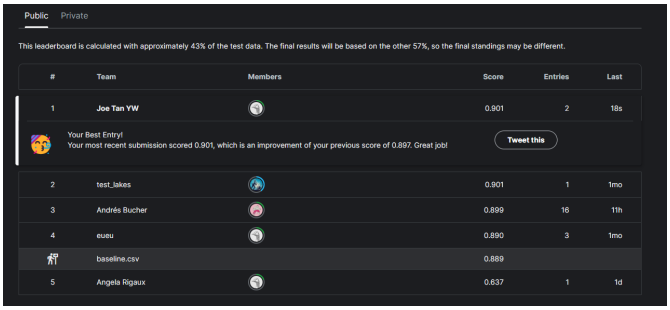

Figure 4: Kaggle Leaderboard

6 Conclusion

Based on the cross-validation results, the Random Forest model emerged as the best performing model with an F1 score of 0.909. The Random Forest's ability to handle complex relationships and its robustness to overfitting contributed to its superior performance compared to other models.

The project demonstrates the application of machine learning techniques to predict the presence of human waste in lakes as part of the Kaggle competition "Predicting Contaminated Lakes". The Random Forest model, trained on the full training set, is used to make predictions on the test set. The predictions are saved to a CSV file in the required submission format.

This project highlights the importance of model selection, evaluation, and comparison to identify the most suitable approach for a given problem. The insights gained from this analysis can be valuable for decision-making and further research in the domain of water quality monitoring and management.

References

[1] Casimir Fisch, geobiology. (2024). Predicting Contaminated Lakes. Kaggle. https://kaggle.com/competitions/predicting-contaminated-lakes