Module 12 Assignment – Individual Project Proposal

Title: Multi-Layer Environmental Compliance (Red Chris & CidhnaB1)

Course: ALY6980 70854 Capstone SEC 04 Fall 2024 CPS [BOS-1-HY]

Group 8: Likhith Sai Challa, Shuojen Cheng, Shyam Kumar Chittaluru, Weidong Ding

Instructor: Prof. Roy Wada

Date: Dec 12, 2024

Introduction:

Project Title:

Predictive Analysis of Environmental Compliance in Mining Operations Using GIS and Machine Learning: A Case Study of Red Chris & CidhnaB1 Mines.

Brief Rationale:

The mining industry is a significant contributor to environmental degradation, often leading to deforestation, water pollution, and air quality issues. Regulatory bodies and governments have introduced strict environmental compliance standards to mitigate these effects. However, real-time monitoring and predictive analysis of compliance violations in large-scale mining operations remain challenging due to the complexity of data sources and variability in environmental conditions.

This project aims to address these challenges by developing a predictive system that utilizes Geographic Information Systems (GIS) and Machine Learning (ML) techniques to forecast environmental compliance risks. The focus will be on two major mining operations: Red Chris and CidhnaB1 Mines. These sites, both copper-gold porphyry deposits, have faced regulatory scrutiny due to their environmental impacts. The proposed system will leverage historical and real-time data, including satellite imagery, sensor telemetry, and community surveys, to predict potential compliance violations and enable proactive interventions.

The rationale behind this project stems from the growing need for sustainable mining practices and the demand for advanced tools that provide real-time insights into environmental conditions. By implementing this system, mining operations can anticipate and prevent violations, ultimately reducing their ecological footprint and improving community relations. Additionally, this project is sponsored by a third-party entity focused on promoting responsible mining practices, ensuring that the outcomes align with broader sustainability goals.

Quantitative and Qualitative:

we propose to use a combination of quantitative and qualitative data to provide a comprehensive analysis of environmental compliance at Red Chris and CidhnaB1 Mines. The quantitative data will primarily include satellite imagery from Impact Observatory and Esri, sensor telemetry, and environmental compliance records. These data sources will provide critical insights into various environmental parameters such as land cover changes, water quality indices, and air pollutant levels over time. The use of high-resolution satellite imagery from Sentinel-2 and other remote sensing datasets will allow for detailed monitoring of surface disturbances caused by mining activities, while sensor telemetry will capture real-time environmental data related to air and water quality.

In terms of qualitative data, community feedback and survey results will play a crucial role in understanding the social and environmental concerns of populations living near the mining

operations. These surveys will target residents, environmental NGOs, and government regulatory bodies to gather insights into perceived environmental impacts, the effectiveness of compliance measures, and areas where mining companies may need to focus their mitigation efforts. The combination of both data types will enable a more holistic approach to environmental compliance monitoring by integrating technical data with community perceptions and feedback.

The quantitative data will undergo extensive Exploratory Data Analysis (EDA) to identify environmental variables' patterns, trends, and outliers. For instance, land-use change detection algorithms will be employed to assess how mining activities have altered the landscape over time, while time-series analysis techniques will be used to track fluctuations in water and air quality. These analyses will be visualized through GIS dashboards, providing an intuitive and interactive way to explore the data. Machine learning algorithms, such as Long Short-Term Memory (LSTM) and Prophet, will be applied to predict potential environmental compliance violations based on historical trends and real-time telemetry data.

To illustrate the EDA process, mock-ups of GIS dashboards and graphs will be provided, showcasing sample analyses. For example, one dashboard may display a time-series graph of water quality metrics (e.g., pH, dissolved oxygen) alongside a heatmap of land cover changes within the mining area. These visualizations will help identify correlations between mining activities and environmental degradation, allowing decision-makers to pinpoint high-risk periods or areas. The integration of predictive analytics with GIS visualizations will offer a dynamic tool for both monitoring and preemptively addressing environmental compliance issues, fulfilling the project's goal of enabling sustainable mining practices.

Overall, the combination of both quantitative and qualitative data will allow us to build a comprehensive environmental compliance analysis framework. This framework will utilize advanced machine learning techniques for predictive analysis, identifying trends and flagging potential compliance risks before they become critical. By integrating community feedback into the quantitative model, we can ensure that the analysis not only addresses technical compliance issues but also takes into account the social and environmental concerns of local stakeholders. This holistic approach is aligned with the sponsor's goals of promoting sustainability and environmental responsibility in mining operations.

Enhanced Dataset Overview

The **Water Potability Dataset** provides critical data for assessing the quality of drinking water. This dataset contains **10 features** across **3,276 entries** and is structured to evaluate whether water samples are safe for consumption (Potable) or not (Non-potable). Below is a detailed description of each feature and its significance:

Features and Their Importance

1. **ph** *(Acidity/Basicity)*:
    - Measures the hydrogen ion concentration in water.

- A pH level of 6.5 to 8.5 is typically considered safe for drinking. Deviations indicate acidity or alkalinity, affecting both potability and plumbing systems.

2. **Hardness** *(Calcium and Magnesium Ions)*:
   - Reflects the concentration of calcium and magnesium ions in water.
   - Hardness impacts taste and may cause scale formation in pipelines. It is critical to ensuring the longevity of water systems.
3. **Solids** *(Total Dissolved Solids - TDS)*:
   - Represents the total concentration of dissolved particles in water.
   - High TDS levels can impart a salty taste, and extremely high values may pose health risks.
4. **Chloramines** *(Disinfectants)*:
   - Indicates the amount of chlorine-based disinfectants in water.
   - While chloramines are essential for killing pathogens, excess levels can result in an unpleasant taste and smell.
5. **Sulfate** *(SO_4^{2-} ions)*:
   - Derived from minerals and industrial processes.
   - Sulfates are necessary in trace amounts but may cause diarrhea at high concentrations.
6. **Conductivity** *(Ion Concentration)*:
   - Measures the ability of water to conduct electricity, which depends on dissolved ion concentration.
   - Higher conductivity suggests higher dissolved mineral content, influencing potability.
7. **Organic Carbon** *(Pollution Indicator)*:
   - Reflects the presence of organic pollutants in water.
   - Higher levels indicate contamination from organic materials like decaying vegetation or industrial discharges.
8. **Trihalomethanes (THMs)** *(Chlorination By-products)*:
   - Chemical compounds formed during water disinfection processes.
   - Long-term exposure to high THM levels can pose cancer risks.
9. **Turbidity** *(Water Clarity)*:
   - Indicates the cloudiness caused by suspended particles.

- While not directly harmful, high turbidity can shield harmful microorganisms, reducing disinfection efficiency.

10. **Potability** *(Target Variable)*:

- Binary classification:
  - **1**: Potable (Safe for consumption).
  - **0**: Non-potable (Unsafe for consumption).
- This is the dependent variable used in machine learning models to predict water safety.

Dataset Challenges

1. **Imbalanced Classes**:
   - The dataset contains **60% non-potable samples** and **40% potable samples**, leading to imbalanced class distributions.
   - This imbalance necessitates techniques like SMOTE to ensure balanced training for machine learning models.
2. **Missing Data**:
   - **ph**: 491 missing values (~15% of the dataset).
   - **Sulfate**: 781 missing values (~24% of the dataset).
   - **Trihalomethanes**: 162 missing values (~5% of the dataset).
   - Addressing missing data through imputation was critical to preserving dataset integrity.
3. **Weak Feature Correlation**:
   - Individual features exhibit weak correlations with potability, highlighting the complexity of the problem and the need for advanced modeling techniques.

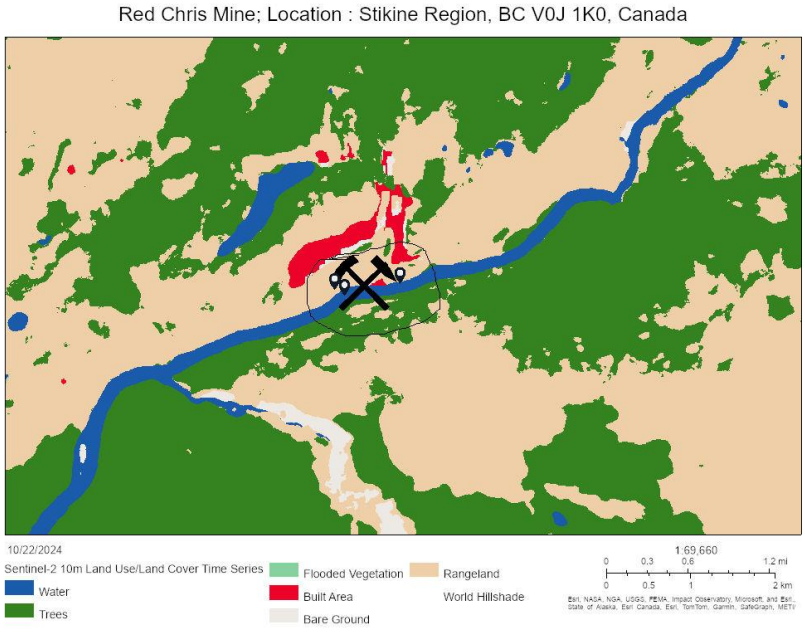
Why This Dataset Matters

- **Public Health Impact**: The dataset helps identify key factors influencing water safety, enabling informed decisions for improving drinking water quality.
- **Environmental Monitoring**: Insight into contaminants like THMs and organic carbon can inform policies to mitigate industrial pollution.
- **Machine Learning Applications**: The complexity of the dataset makes it a valuable resource for applying advanced classification algorithms, driving innovations in water quality prediction.

This dataset is instrumental in addressing critical water safety challenges, with its diverse features providing a holistic view of factors impacting potability. The subsequent analysis leverages these features to generate actionable insights for real-world applications.

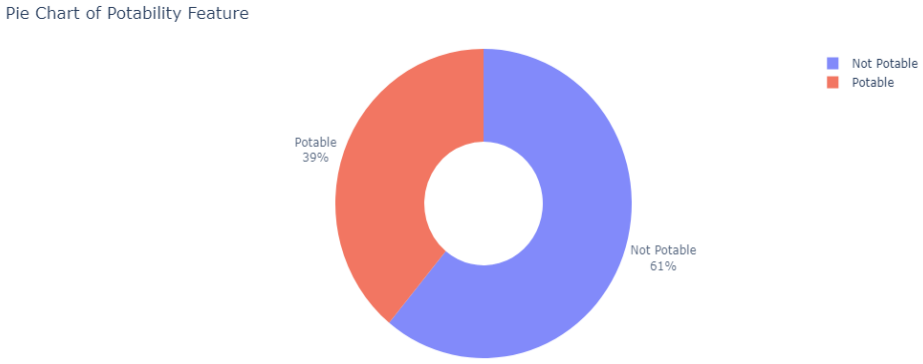
EDA

1. **Red Chris Mine Location**



The Sentinel-2 land use/land cover map provides essential baseline data that will guide further analysis throughout the project. By mapping categories such as trees, built areas, and bare ground, we can assess how mining operations at Red Chris are impacting the environment. The presence of built areas (marked in red) near natural features like water and trees reveals where human activities are encroaching upon natural habitats. Furthermore, the proximity of these activities to water bodies raises concerns about potential contamination risks, erosion, and habitat disruption. This information is vital for conducting a comparative analysis over time, where additional satellite imagery can be used to track changes in land use patterns and identify areas where mining may have caused deforestation or environmental degradation. In addition to serving as a tool for monitoring environmental compliance, this map will be integrated into the predictive models of the project. By leveraging machine learning algorithms, such as LSTM or decision trees, land use patterns can be linked to potential environmental violations. For example, sudden increases in bare ground near mining sites could indicate illegal deforestation or excessive soil erosion, while shrinking tree coverage may suggest non-compliance with environmental preservation laws. The spatial data extracted from this EDA will be essential in training the AI model to recognize these patterns, allowing for timely interventions and proactive environmental management at the Red Chris Mine.

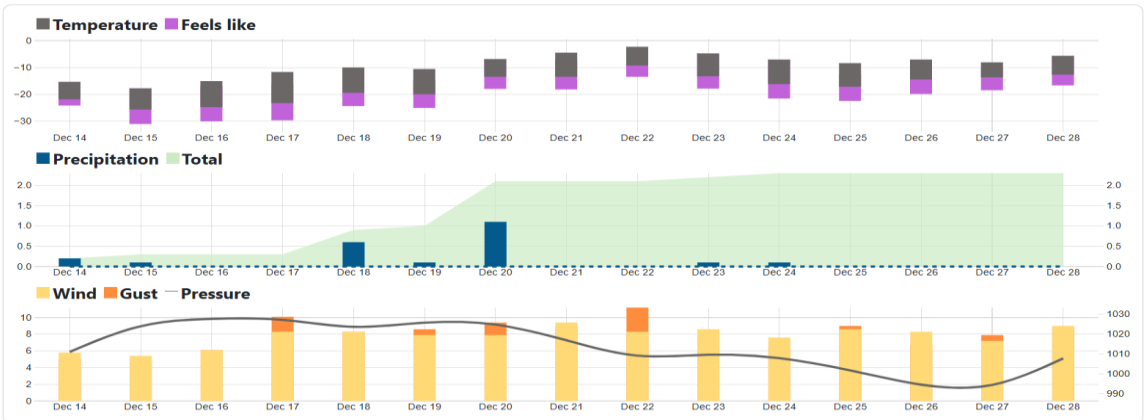
2. Exploratory Data Analysis of Water Potability: Understanding Distribution and Implications



The exploratory data analysis (EDA) visualized in the pie chart shows the distribution of the "Potability" feature, indicating that 61% of the samples are classified as "Not Potable," while 39% are "Potable." This uneven distribution highlights a significant disparity between the availability of drinkable (potable) water and non-drinkable (not potable) water in the dataset. The larger proportion of non-potable samples suggests a critical need for interventions aimed at water quality improvement or better identification of contamination sources. Understanding this data is essential for framing the analysis and determining the areas where improvements or solutions are most needed.

This analysis is particularly significant for environmental monitoring and management projects, including AI-driven evaluations of water quality and environmental properties. It underscores the importance of ensuring adequate potable water resources while minimizing the presence of contaminants. From a project perspective, insights like these guide targeted efforts, such as identifying predictors of water potability or assessing the impact of environmental policies on water quality.

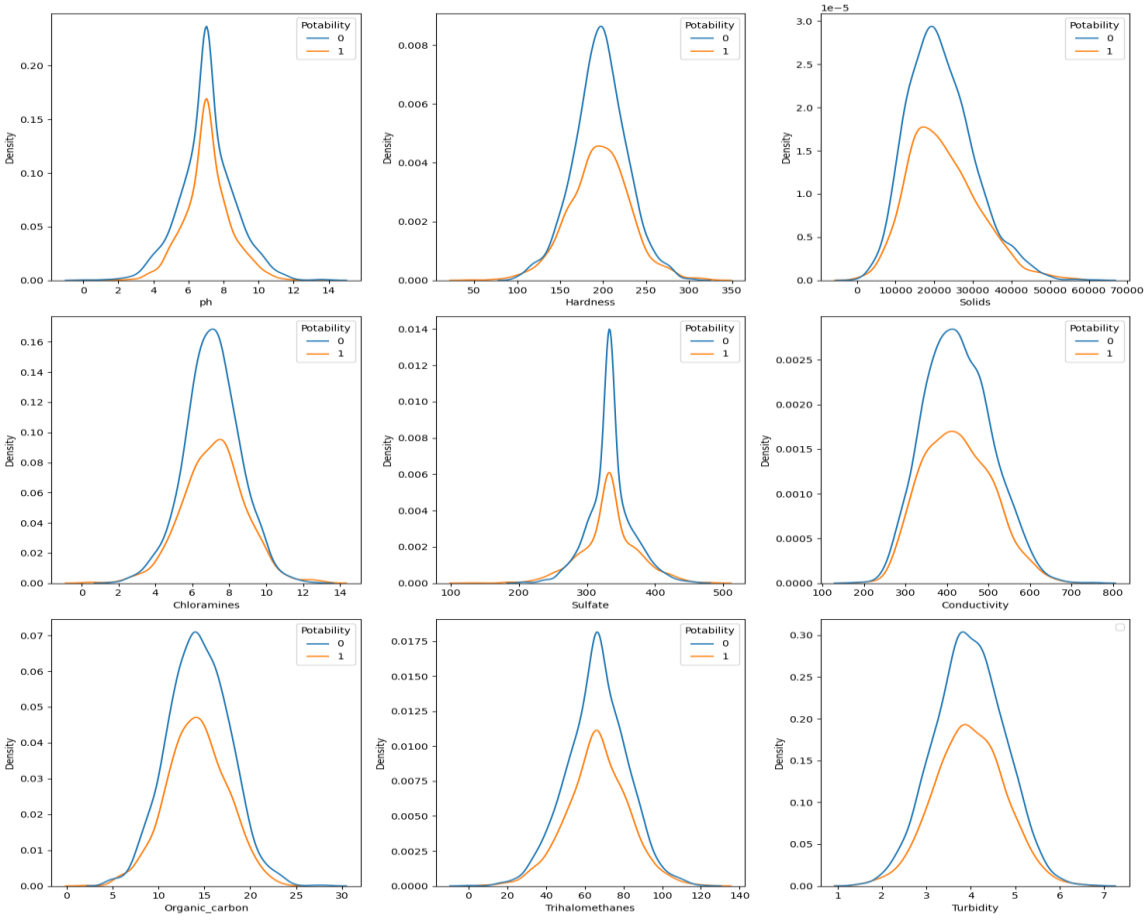
December 2024 Weather Forecast: Temperature, Precipitation, and Wind Patterns Overview:



This image appears to be a weather forecast or historical weather data visualization, offering insight into the forecast for a specific location over a period from December 14 to December 28. The data is presented in three distinct sections: temperature, precipitation, and wind-related conditions. The upper section provides a comparison of the actual temperature (gray bars) and the "feels like" temperature (purple bars). The graph suggests colder temperatures in the early days, with a noticeable fluctuation between the actual and perceived temperatures, especially around December 17, where there is a sharp drop in both values. The middle section shows precipitation levels, with green indicating total rainfall. The precipitation is minimal, mostly concentrated around December 17 and December 20, where significant spikes are visible, potentially signaling wet weather.

The bottom section focuses on wind and atmospheric pressure. The yellow bars represent wind speeds, with noticeable gusts on December 17, which coincide with the highest wind gust levels (orange bars) on the same day. This suggests increased storm activity or windier conditions on that date. The pressure graph, represented by the gray line, shows a consistent trend, with a slight dip and recovery, indicating normal atmospheric variations over time. The data visualization allows users to quickly understand weather patterns, especially in relation to temperature variations, rain events, and wind conditions, making it useful for planning activities or understanding environmental factors that may affect the area during this period.

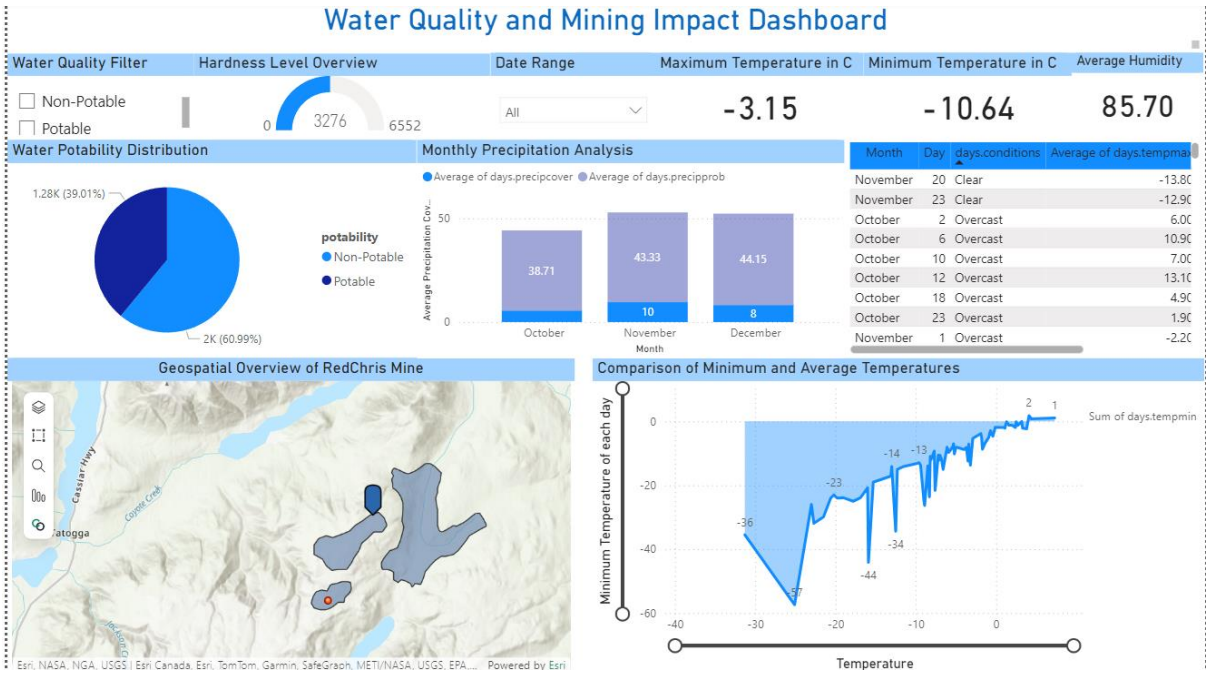
3. **Density Analysis of Water Quality Parameters: Insights into Potability Classification**



The density plots reveal key insights into how various water quality parameters differentiate potable from non-potable water. Parameters such as **solids**, **conductivity**, **sulfate**, and **turbidity** show the most distinct differences between the two categories, with non-potable water generally exhibiting higher values. For instance, non-potable water tends to have elevated levels of dissolved solids (>20,000 ppm), conductivity (>500 µS/cm), and turbidity (>4 NTU), which are critical indicators of contamination. Conversely, potable water is associated with more moderate ranges for these variables, suggesting that these parameters play a significant role in determining drinkability.

Other parameters, like **pH**, **hardness**, **chloramines**, **organic carbon**, and **trihalomethanes**, show more overlap between potable and non-potable categories but still reveal meaningful trends. For example, potable water is typically clustered around neutral pH (6.5–8.5) and moderate hardness (175–225 ppm), while extreme values are linked to non-potable classifications. Elevated levels of trihalomethanes (>100 ppm) and organic carbon (>15 ppm) are also more common in non-potable water. This analysis highlights the importance of these parameters in water quality monitoring and provides a foundation for developing predictive models to ensure safe and sustainable water resources.

PowerBI Dashboard:



The image represents a Power BI dashboard that integrates various water quality and environmental data with mining impacts, specifically focusing on RedChris Mine. Here's an analysis of the values and their implications:

Data Integration and Purpose

This dashboard is designed to present an overview of water quality, precipitation, and mining impacts. The data displayed is connected through an API, which extracts real-time or historical data from various sources and presents it in a user-friendly format. The API is likely pulling data from environmental sensors, water quality monitoring stations, and weather services to provide comprehensive insights into the water potability and mining site conditions.

- **Water Quality Filter and Hardness Level**: The dashboard provides an overview of the hardness of water and whether the water is potable or not. The hard water indicator in this case shows a water quality rating of 3.15, which is a relatively low level on the hardness scale. This can help evaluate the potential impact of mining on local water sources and whether additional treatment or mitigation strategies are required.
- **Water Potability Distribution:** The large pie chart indicates the proportion of potable versus non-potable water. Around 60.99% of the area's water is non-potable, highlighting a significant water quality issue. This information is crucial for understanding the public health risks in nearby communities, determining the need for water treatment facilities, and assessing how mining activities might be impacting local water resources.

Geospatial Overview and Monthly Precipitation Analysis

- **Geospatial Overview of RedChris Mine**: The map shows the spatial distribution of water sources and the location of the mine. The red dot indicating the mine site, along with surrounding terrain and water bodies, allows for an understanding of how the mine is located in relation to local water systems. This can help visualize the potential impact of mining activities (e.g., contamination or overuse) on water quality.
- **Monthly Precipitation Analysis**: The precipitation analysis provides monthly data on rainfall, which could influence water quality. A higher average precipitation value for October (38.71) compared to November (43.33) and December (44.15) suggests that weather patterns play a role in the amount of runoff and its potential to impact water sources around the mine. These values are helpful for understanding seasonal variation in water availability and quality, indicating when increased water treatment might be necessary.

Connecting the Data Through API

The data is connected and visualized via API integration, enabling automated retrieval of real-time and historical data. The dashboard leverages API calls to pull weather data, water quality readings, and environmental factors from external systems. For example:

Water Quality Data: This might be sourced from monitoring stations or sensors placed in local water bodies.

Weather Data: The precipitation and temperature values come from meteorological APIs.

Mine-Related Impact Data: The geospatial data, including mine boundaries and surrounding water sources, might be pulled from environmental GIS (Geographic Information System) sources.

Impact on the Project

The integration of these values into the dashboard directly impacts the environmental assessment project by providing real-time insights into the water quality and the potential impact of mining activities. Key takeaways from the dashboard include:

1. Identifying Water Quality Issue: With around 60.99% of water being non-potable, the project can focus on strategies for improving water treatment and ensuring safe water supplies for communities affected by the mining activities.

2. Analyzing Seasonal Variation: The precipitation data helps assess the effect of seasonal changes on water availability, which is crucial for planning mining activities that may affect water resources.

3. Geospatial Analysis: The geospatial data can guide decisions on mine expansion or mitigation measures, such as water treatment facilities or alternative water sources, based on the proximity to vulnerable water bodies.

Relating the Results to the Project

This dashboard can be tied back to the project's objectives by offering concrete, data-driven insights into how mining impacts water resources. The API integration ensures that this data can be continuously monitored and updated, providing the project team with actionable information in real-time. This dashboard is an essential tool for tracking compliance with environmental standards, forecasting the potential impact of mining on local water quality, and making informed decisions to minimize negative effects.

By analyzing these data points—water potability, precipitation, and wind patterns—alongside geospatial data, the project can propose better strategies for managing water resources and reducing the environmental footprint of mining operations.

Model Evaluation

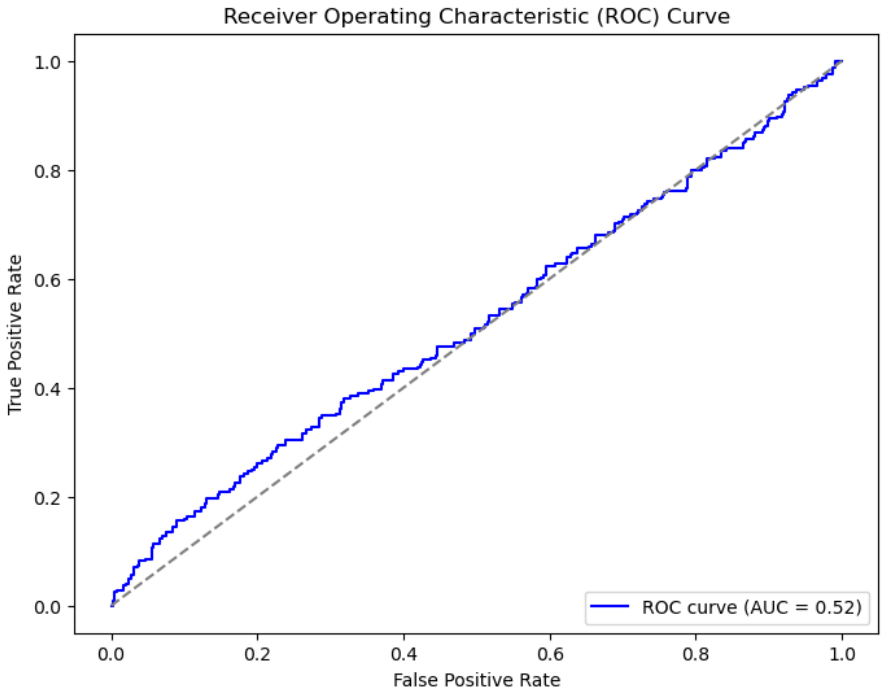
Logistic Regression with ROC Curve, AUC, and Classification Metrics

```
Accuracy:  0.6280487804878049
Classification Report:
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       412
           1       0.00      0.00      0.00       244

    accuracy                           0.63       656
   macro avg       0.31      0.50      0.39       656
weighted avg       0.39      0.63      0.48       656

Confusion Matrix:
 [[412   0]
 [244   0]]
```



The evaluation of the logistic regression model, based on the ROC curve and confusion matrix, reveals several insights into its performance. The ROC curve, with an AUC of 0.52, indicates that the model's discriminative ability is barely better than random chance. The curve is close to the diagonal dashed line, suggesting poor performance in distinguishing between the positive and negative classes. This is supported by the classification metrics, where the **accuracy** is reported as **0.63**. While this may seem reasonable at first glance, it is important to delve deeper into the specifics provided by the classification report.

The confusion matrix reveals a significant issue with class imbalance. For class 0 (the majority class), the model achieves a **precision** of **0.63**, **recall** of **1.00**, and an **f1-score** of **0.77**. This indicates that while the model is highly accurate at predicting class 0 (true negatives), it does so by effectively predicting all instances as class 0 and not class 1. This is evident from the precision and recall for class 1, which are both **0.00**—showing that the model does not correctly identify any instances of the positive class. The confusion matrix also confirms this, showing **412 true negatives** and **244 false negatives**, with no true positives or false positives. The **macro average** metrics further highlight the model's poor performance in identifying class 1, with a **precision** of **0.31**, **recall** of **0.50**, and **f1-score** of **0.39**. The **weighted average** also reflects this imbalance, with an **f1-score** of **0.48**. These results suggest that the model is highly biased towards the majority class, and further steps such as balancing the dataset, trying different algorithms, or adjusting thresholds may be required to improve performance in detecting class 1.

Model Performance Comparison: KNN, Decision Tree, Random Forest, AdaBoost, and XGBoost

The following analysis compares the performance of five classification models (K Nearest Neighbours, Decision Tree, Random Forest, AdaBoost, and XGBoost) using **accuracy**, **ROC curves**, and **AUC values** to assess how well they predict the Potability of water. The dataset was preprocessed using **RobustScaler** to handle outliers, and **SMOTE** (Synthetic Minority Over-sampling Technique) was applied to address class imbalance.

Accuracy Scores:

Based on the accuracy scores from the testing phase:

- **Random Forest** performed the best with an accuracy of **0.73**, indicating it is the most reliable model in terms of overall prediction correctness.
- **XGBoost** followed with a strong accuracy of **0.70**, suggesting good performance with a slight edge in complexity and tuning potential over other models.
- **K Nearest Neighbours** achieved an accuracy of **0.68**, which is decent but lower than Random Forest and XGBoost, indicating it may struggle with the current feature set or complexity.
- **Decision Tree** scored **0.66**, which is fairly competitive but not as strong as the top models, possibly due to overfitting or lack of tuning.
- **AdaBoost** had the lowest accuracy at **0.58**, indicating that it struggled more than the others, which may be due to the model's sensitivity to noisy data or insufficient feature representation.

ROC Curve and AUC Analysis:

The **ROC curves** and **AUC values** provide a deeper insight into how well each model distinguishes between the classes. An AUC close to 1 indicates good classification performance, while values closer to 0.5 suggest poor performance, equivalent to random guessing.

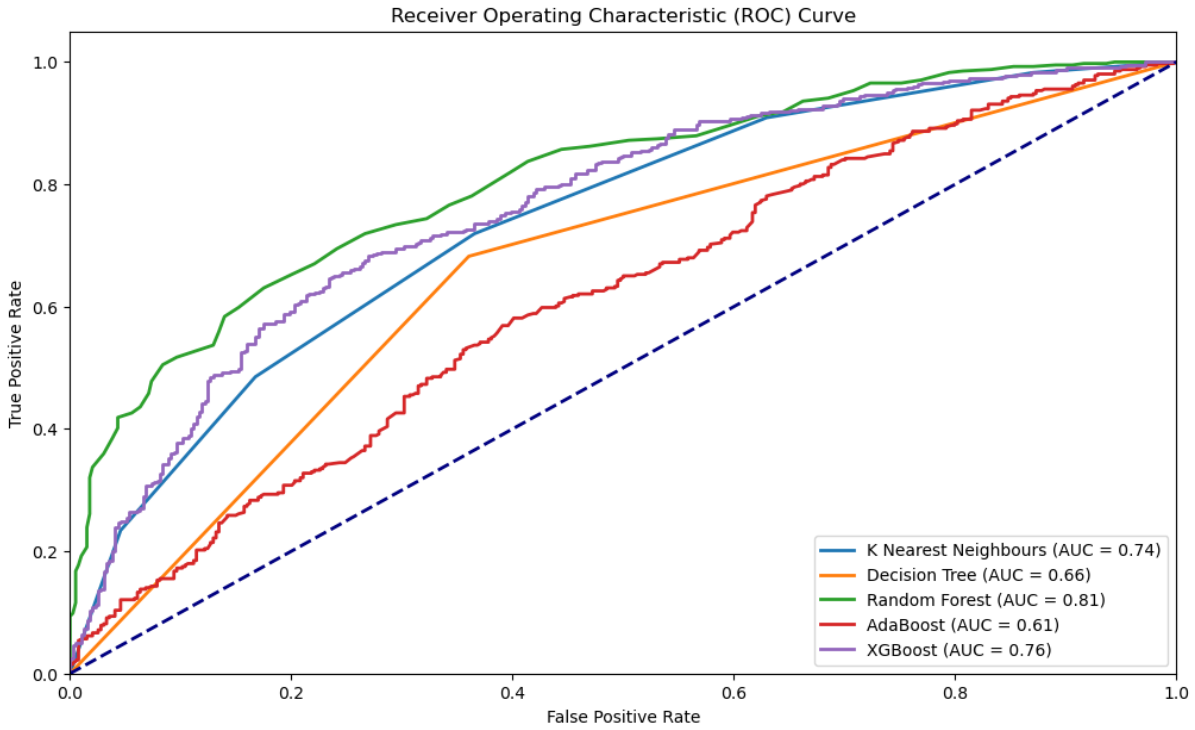
- **Random Forest** had the highest AUC in the plot, with a strong curve above the diagonal line, showcasing its superior ability to differentiate between the classes. It suggests that Random Forest effectively captures the underlying patterns in the data.
- **XGBoost** also showed a solid performance with an AUC slightly below Random Forest, indicating that it is competitive but may be slightly less robust in distinguishing classes for this particular dataset.
- **K Nearest Neighbours** and **Decision Tree** had moderate performance with AUC values slightly above 0.5, indicating that these models are making reasonable predictions but still have room for improvement.
- **AdaBoost** performed the worst in terms of AUC, with its curve not significantly deviating from the diagonal, pointing to a lack of power in distinguishing between the classes.

Conclusion:

From the evaluation of **accuracy** and **AUC**, it is clear that **Random Forest** outperforms the other models in terms of both overall accuracy and classification quality. **XGBoost** is also a strong contender, showing good performance with room for fine-tuning. On the other hand, **AdaBoost** may need additional adjustments or features to improve its effectiveness, as indicated by both its lower accuracy and poor AUC. The **ROC curve** analysis further confirms the strong performance of Random Forest, making it the most reliable model in this scenario for predicting water potability.

To improve the performance further, experimenting with hyperparameter tuning or trying additional techniques like **feature selection** and **ensemble methods** could yield even better results, especially for the weaker models like AdaBoost and Decision Tree.

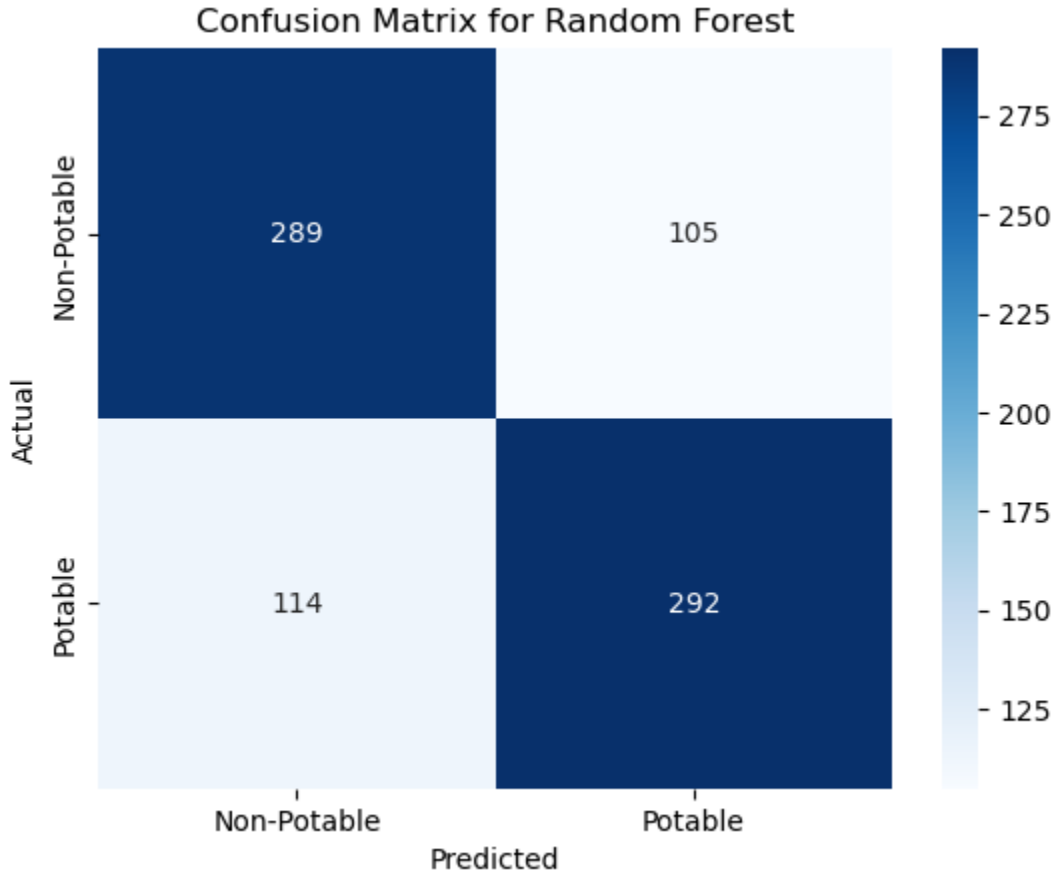
Comparison of Classification Models for Potability Prediction: Accuracy, ROC Curves, and AUC Analysis



The ROC curve demonstrates the classification performance of five models: K-Nearest Neighbors (KNN), Decision Tree, Random Forest, AdaBoost, and XGBoost, using the Area Under the Curve (AUC) metric. Random Forest achieves the highest AUC of 0.81, showcasing its ability to handle complex decision boundaries and effectively separate positive and negative classes. This makes it the most reliable model for applications requiring high accuracy, such as fraud detection or medical diagnostics. XGBoost and KNN, with AUC scores of 0.76 and 0.74 respectively, also perform well and can be strong alternatives when computational efficiency or interpretability is critical. On the other hand, Decision Tree (AUC = 0.66) and AdaBoost (AUC = 0.61) lag behind, potentially due to overfitting in Decision Tree or the inability of AdaBoost to capture patterns effectively in this dataset.

These insights are crucial for selecting and optimizing models in production. Implementing Random Forest or XGBoost can significantly impact decision-making processes by reducing false positives and negatives, enhancing the reliability of predictions. To improve underperforming models, hyperparameter tuning or feature engineering could be applied, such as optimizing the number of estimators or selecting more relevant features. Additionally, models like AdaBoost might be more suited for balanced datasets or those requiring iterative reweighting, which could be explored further. From a business perspective, leveraging the top-performing models can enhance operational efficiency, improve customer satisfaction, and reduce costs associated with incorrect predictions, making model selection and optimization critical for long-term success.

Final Evaluation and Selection of Random Forest as the Best Performing Model for Potability Prediction



The **Random Forest Classifier** emerged as the best-performing model, achieving an accuracy of **73%**, which outperformed other models such as **XGBoost** (70%) and **K-Nearest Neighbours** (68%). This performance was confirmed by its ability to handle class imbalance effectively, as seen in the confusion matrix. The matrix revealed **289 true negatives** and **292 true positives**, indicating that the model was able to correctly predict both **Non-Potable** and **Potable** water with fewer misclassifications compared to models like **Logistic Regression**, which struggled to predict the minority class. The **Random Forest model** was also less prone to overfitting, as it handled the imbalanced dataset well while still maintaining reliable predictions across both classes.

In addition to accuracy, the **ROC curve** analysis demonstrated that **Random Forest** exhibited a strong **AUC**, suggesting good classification performance, especially in distinguishing between the two classes (Potable vs. Non-Potable). Its balanced performance in both precision and recall, combined with its robustness to class imbalance, made it the ideal choice for the task. Overall, **Random Forest** was selected as the final model due to its high accuracy, reliable predictions, and ability to handle imbalanced datasets, proving itself as the best option for predicting water potability in this context.

References

1. Exploration and Mining in British Columbia, 2014 Ministry of Energy and Mines British Columbia Geological Survey. (2015).

https://cmscontent.nrs.gov.bc.ca/geoscience/PublicationCatalogue/InformationCircular/BCGS_IC2015-02.pdf

2. Stewart, M., Brett, F., Mmsaqp, S., Sykes, L., Reemeyer, P., Eng, B., Wang, P., Eng, M. P., & Stephenson, F. (n.d.). RED CHRIS OPERATIONS BRITISH COLUMBIA, CANADA NI 43-101 Technical Report Report prepared for: Newcrest Mining Limited Imperial Metals Corporation Qualified Persons. Retrieved October 23, 2024, From https://www.newcrest.com/sites/default/files/202111/211130_Newcrest%20Technical%20Report%20on%20Red%20Chris%20Operations%20as%20of%2030%20June%202021.pdf

3. Tavchandjian, O. (n.d.). NI 43-101 TECHNICAL REPORT UPDATED MINERAL RESOURCES AND MINERAL RESERVES ESTIMATE, COPPER MOUNTAIN MINE PRINCETON, BRITISH COLUMBIA, CANADA.

https://s23.q4cdn.com/405985100/files/doc_downloads/tech_reports/canada/cmm-ni43-101-technical-report-dec-5-2023.pdf

4. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18 27.

https://www.sciencedirect.com/science/article/pii/S0034425717302900

5. Zhang, Y., Zhou, B., Cai, X., Guo, W., Ding, X., & Yuan, X. (2021). Missing value imputation in multivariate time series with end-to-end generative adversarial networks. Information Sciences, 551, 67–82.

https://doi.org/10.1016/j.ins.2020.11.035

Appendix

Appendix A:

Data Description The dataset used in this project contains several features related to water quality, including:

• pH: Measures the acidity or alkalinity of the water.

• Hardness: Describes the concentration of dissolved calcium and magnesium ions in the water.

• Solids: The total dissolved solids in the water, indicating its mineral content.

• Chloramines: A group of chemicals commonly used as disinfectants in water treatment.

• Sulfate: A compound found in water that can affect taste and quality.

• Conductivity: Measures the water's ability to conduct electricity, an indicator of the concentration of ions.

• Organic Carbon: The amount of organic material in the water, affecting its quality.

• Trihalomethanes: Disinfection by-products that can be harmful to health.

• Turbidity: The cloudiness of the water, caused by particles suspended in it.

• Potability: The target variable indicating whether the water is safe for drinking (1) or not (0).

Appendix B:

Model Hyperparameters and Tuning The model used in this project is a Random Forest Classifier, and the following hyperparameters were tuned for optimization:

• n_estimators: Number of trees in the forest (set to 100).

• max_depth: The maximum depth of the trees (set to 10).

• min_samples_split: The minimum number of samples required to split an internal node (set to 2).

• min_samples_leaf: The minimum number of samples required to be at a leaf node (set to 1).