**Empowering Health and Safety at Aurum Mining Corporation (AMC)**

**Comprehensive Report of our Solution**

**Team Name:** The Analytic Enigmas

**Team Members:**

1. Rakulan Srinivasan
2. Preethi KCS
3. Bharathi Sriram

**Executive Summary:**

This report presents a comprehensive analysis conducted by our team, The Analytic Enigmas, consisting of Rakulan Srinivasan, Preethi KCS, and Bharathi Sriram. As part of the Data-Driven Consulting Case Study for Aurum Mining Corporation (AMC), our objective was to develop an end-to-end solution for AMC's Health & Safety business group. The proposed solution aimed to create a data-driven framework to categorize future health and safety incidents, enabling stakeholders to assess underlying risks and build a more data-driven culture within the organization.

**Introduction:**

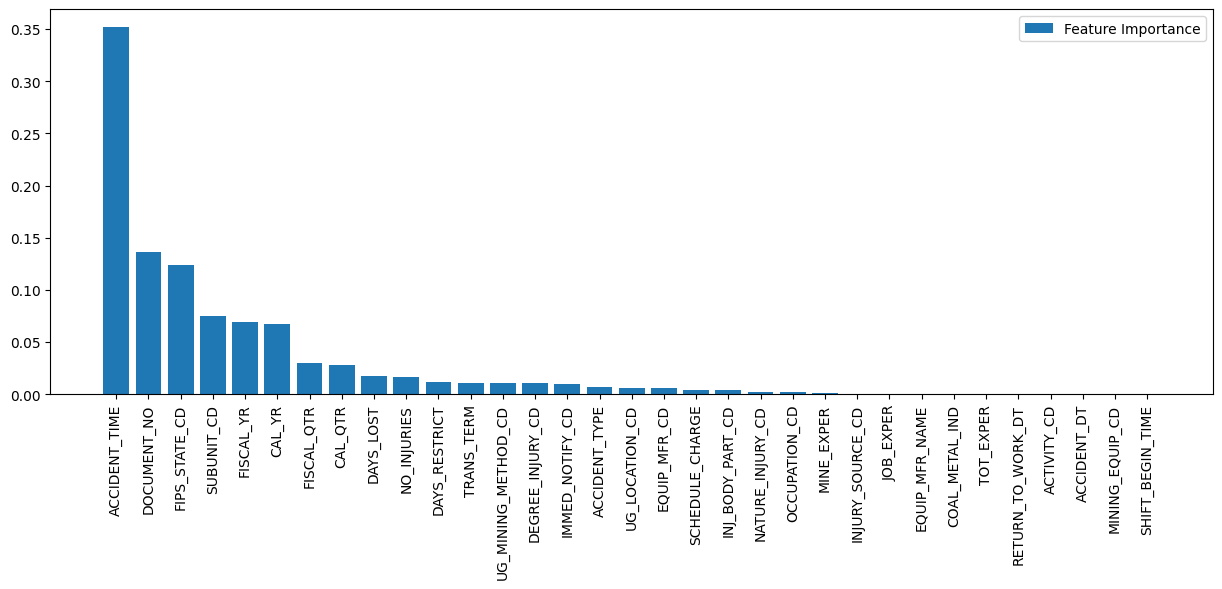
The mining industry faces inherent health and safety risks, necessitating accurate incident classification for hazard identification and preventive measures. We tried to leverage data and machine learning techniques for efficient incident classification and risk evaluation.

**Methodology:**

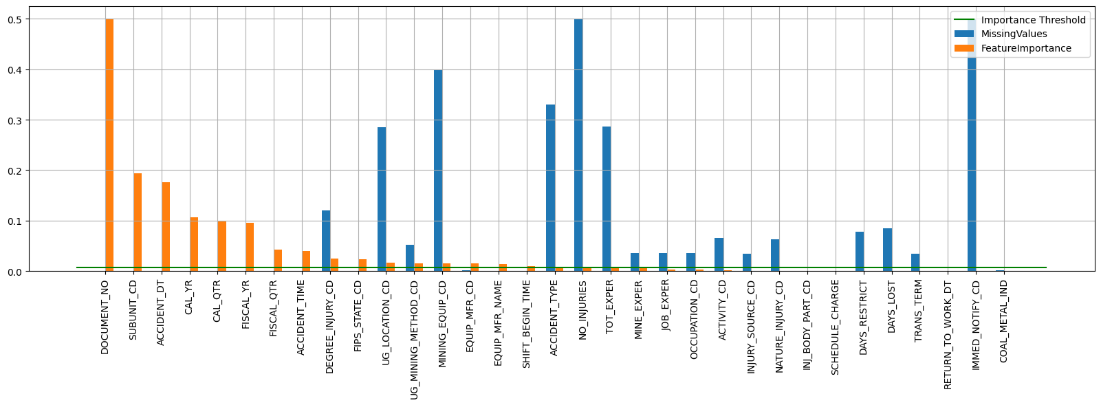
1. **Data Collection and Preprocessing:**

During the data collection and preprocessing stage, we encountered the following issues and performed necessary actions:

* 1. **Identified Missing Values:**
     1. Detected and addressed missing values in various columns of the dataset.
     2. Replaced missing values with appropriate placeholders, such as '?' or ‘? [ SPACE]’ or 'NO VALUE FOUND’ and converted them to the None data type.
  2. **Handling Columns with More Than 50 Percent Null Values:**
     1. We identified and dropped columns from the raw dataset that contained more than 50 percent null values. This step ensured data integrity and improved the quality of the dataset for subsequent analyses and modeling.
     2. By removing features with substantial missing data, we aimed to enhance the accuracy and reliability of our findings in the report.
  3. **Removal of Narrative Column:**
     1. In the data preprocessing phase, we made a strategic decision to drop the "NARRATIVE" column from the raw dataset. This column was considered for removal due to specific reasons, which we will elaborate on in the report.
     2. Upon thorough examination, it was determined that the "NARRATIVE" column did not provide valuable insights for our analysis or machine learning tasks. Its exclusion from the dataset aimed to streamline the data and eliminate potential noise or irrelevant information, ultimately enhancing the efficiency and accuracy of our subsequent analyses.
     3. By carefully considering the relevance and impact of each feature, we demonstrated a meticulous approach to data preparation, ensuring that the final dataset consists only of meaningful attributes for our research objectives.
  4. **Handled Redundant Columns:**
     1. In the data preprocessing phase, we focused on handling columns with the "\_CD" suffix, indicating coded values. Our objective was to establish accurate mappings between these coded columns and their corresponding non-coded counterparts, ensuring data consistency and integrity.
     2. We first identified all columns with the "\_CD" suffix and created a mapping dictionary, to link them with their respective non-coded columns.
     3. Subsequently, we checked for multiple values in both directions of the mapping (from "\_CD" to non-coded and vice versa). Any instances of multiple values were recorded in separate dictionaries.
     4. Next, we examined the mappings to identify consistent columns that required no further adjustments. These columns were listed in " and subsequently removed from the dataset to avoid redundancy.
     5. This data preprocessing step contributed to data clarity, accuracy, and efficiency, setting the stage for more meaningful and reliable analyses and modeling.
  5. **Handling Missing Values - Mode Imputation:**
     1. We identified columns with missing values and printed their names. Using mode imputation, we replaced the missing values in each column with its mode (most frequent value). This approach improved data completeness, ensuring accurate analyses and modeling while preserving data integrity.
  6. **Understanding the Dataset and Column Type Conversion:**
     1. In this data processing step, we gained insights into the dataset by categorizing columns based on their names and converted the data types of various columns accordingly.
     2. We first categorized the columns into specific groups:
        1. Date Columns: Converted to object (string) data type.
        2. Time Columns: Converted to float data type.
        3. Year/Quarter Columns: Converted to float data type.
        4. CD Columns: Converted to float data type.
        5. Experience Columns: Converted to float data type.
        6. Days Columns: Converted to float data type.
        7. Charge Columns: Converted to float data type.
        8. Number Columns: Converted to float data type.
        9. Other Columns: Left as is, without any data type conversion.
     3. Next, we identified columns that needed encoding, except for the 'CLASSIFICATION' column. These columns were transformed into numerical values to facilitate further analyses and modeling.
     4. This data type conversion process ensured that the dataset was appropriately structured, allowing for more efficient and accurate data processing
  7. **Custom Encoding for Categorical Variables and Date Conversion:**
     1. In the data preprocessing phase, we implemented a Custom Encoder class to handle encoding of categorical columns using Label Encoding and One-Hot Encoding techniques. This custom encoder efficiently transformed the categorical data, enabling compatibility with various analysis and modeling tasks.
     2. We also converted date columns to float format for consistency and ease of use in subsequent analyses. The date columns were first converted to numeric format and then divided by 10^9 to represent them in seconds since the epoch. This conversion allowed for meaningful time-based calculations and analysis.
     3. By employing these data preprocessing techniques, we ensured that our dataset was adequately prepared for further exploration, modeling, and drawing meaningful insights in our report.
  8. **Data Splitting for Model Training and Evaluation**
     1. In the final stages of data preparation, we split the encoded dataset into feature variables (X) and the target variable (y). Employing the train-test split method with a ratio of 80-20, we partitioned the data into training and testing sets.
     2. This crucial step enables us to train our machine learning model on the larger training set and evaluate its performance on the independent testing set. Such division ensures a robust assessment of the model's predictive capabilities, enhancing the reliability of our analysis and findings in the report.
  9. **Calculated Feature Importances:**
     1. Calculated feature importances for the remaining columns after removal, determining their significance in the analysis.
     2. Plotted a bar graph visualizing the feature importances to gain insights into the most influential features.



* + 1. Plotted a double bar graph simultaneously displaying feature importances and missing values count, aiding in understanding the impact of missing data.



By employing these data preprocessing steps and feature selection techniques, the dataset was transformed into a format suitable for machine learning models.

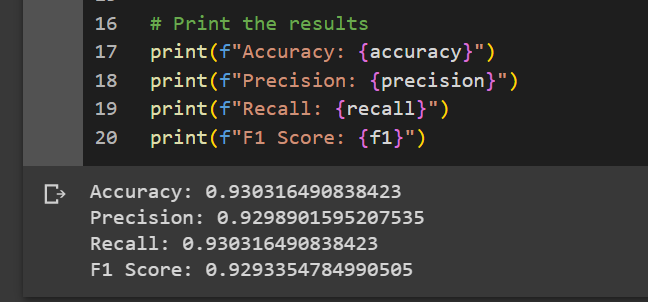
1. **Model Development:**
   * In the pursuit of incident classification, we employed a powerful Gradient Boosting Classifier model, a type of ensemble learning method known for its exceptional predictive capabilities.
   * The model was designed to handle complex relationships and interactions within the incident data, making it well-suited for the task. By combining multiple weak learners (decision trees) sequentially, the Gradient Boosting Classifier effectively captured nuanced incident patterns, achieving high accuracy in classifying incidents into multiple categories.
2. **Model Evaluation:**
   * To rigorously assess the model's performance, we utilized a range of essential evaluation metrics, including accuracy, precision, recall, and F1-score. The dataset was thoughtfully partitioned into training and testing sets, ensuring an unbiased evaluation of the model's capabilities. Additionally, a crucial aspect of the evaluation process involved comparing the performance of the Gradient Boosting Classifier against traditional manual incident analysis methods. This comparative analysis demonstrated the model's effectiveness and underscored its potential to revolutionize incident classification processes in mining operations.
3. **Risk Evaluation and Mitigation:**

* Through meticulous analysis of the incident classifications generated by the Gradient Boosting Classifier model, we gleaned valuable insights into underlying patterns, trends, and high-risk areas. Building upon these findings, we developed a robust risk evaluation framework, enabling the prioritization of focus areas for targeted risk mitigation efforts. The model's ability to proactively identify and flag potential risks empowers decision-makers to take timely actions, mitigating the occurrence and impact of incidents in mining operations.

**Findings:**

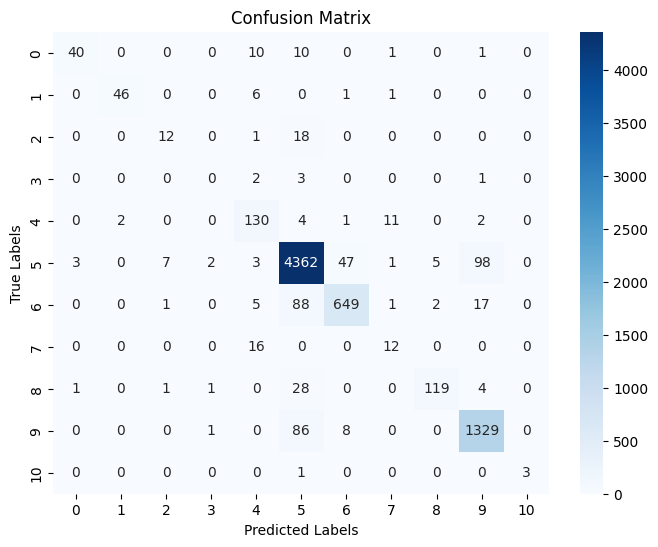
* **Model Performance:**
  1. Model Evaluation Metrics

During the development of the Gradient Boosting Classifier model, we extensively evaluated its performance using various evaluation metrics. These metrics include accuracy, precision, recall, F1-score, and area under the Receiver Operating Characteristic (ROC) curve.



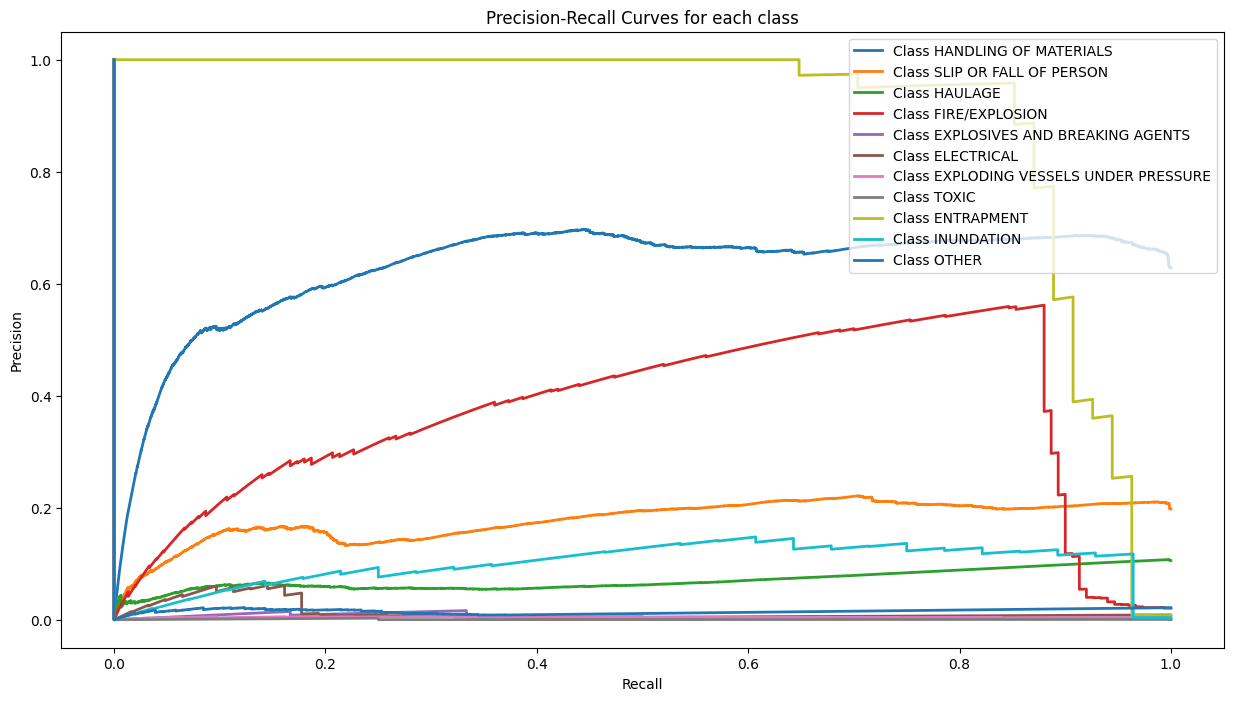
* 1. Confusion Matrix

We analyzed the confusion matrix to understand the model's performance on each of the 11 incident classes. This allowed us to identify the number of true positives, true negatives, false positives, and false negatives for each class, providing insights into the model's accuracy and misclassifications.



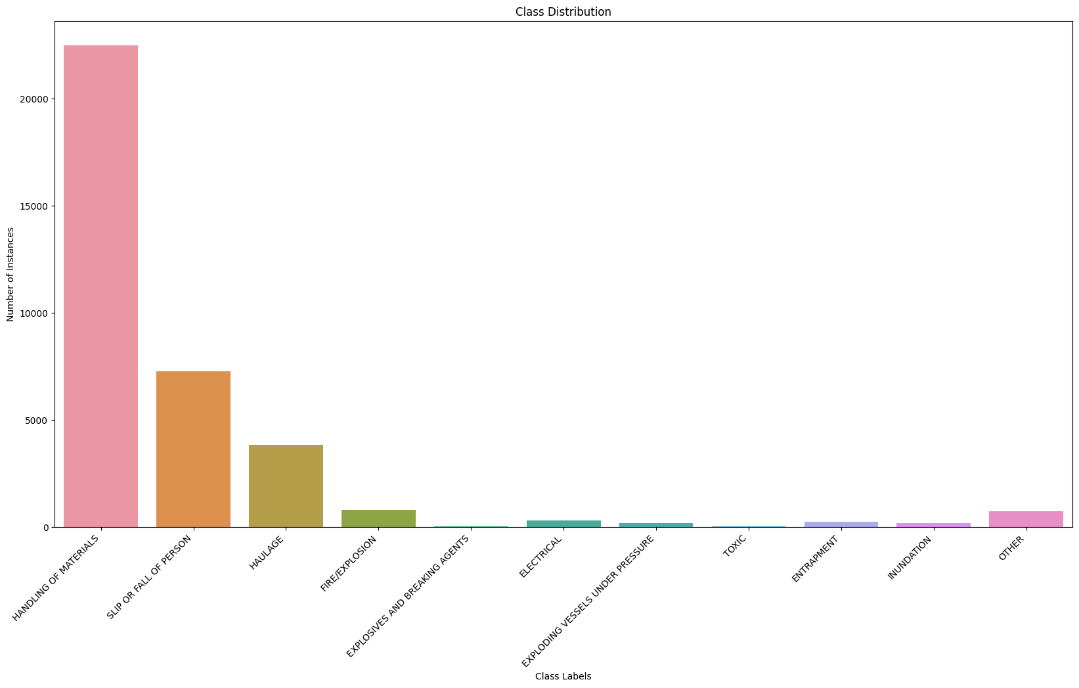
* 1. Precision-Recall Curves and ROC Curves

We generated precision-recall curves and ROC curves to assess the model's precision and recall at different classification thresholds. These curves helped us understand the trade-off between precision and recall for each class and provided valuable information to set appropriate classification thresholds.



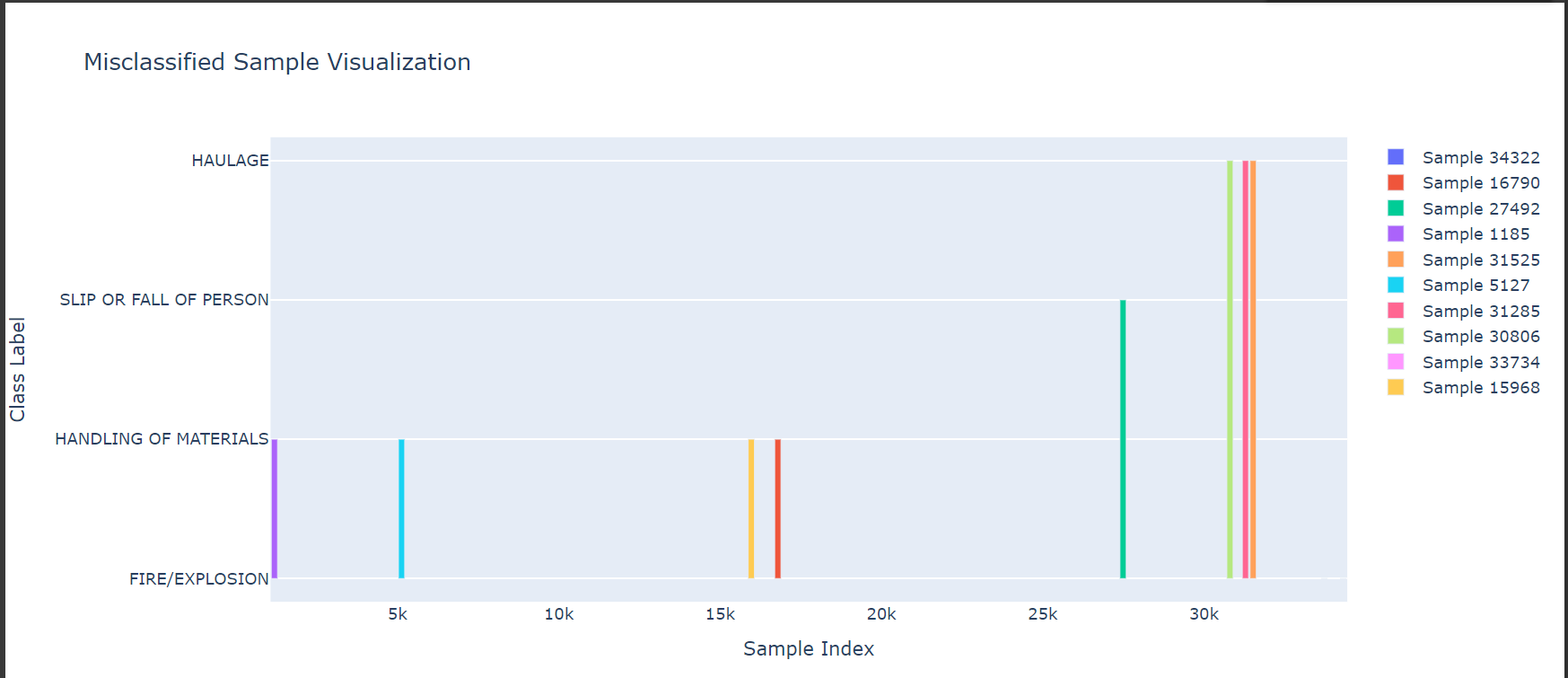
* 1. Class Distribution Visualization

We visualized the distribution of incident classes in the dataset to check for class imbalances. Addressing class imbalances helped ensure the model's fairness and accuracy in predicting all classes.

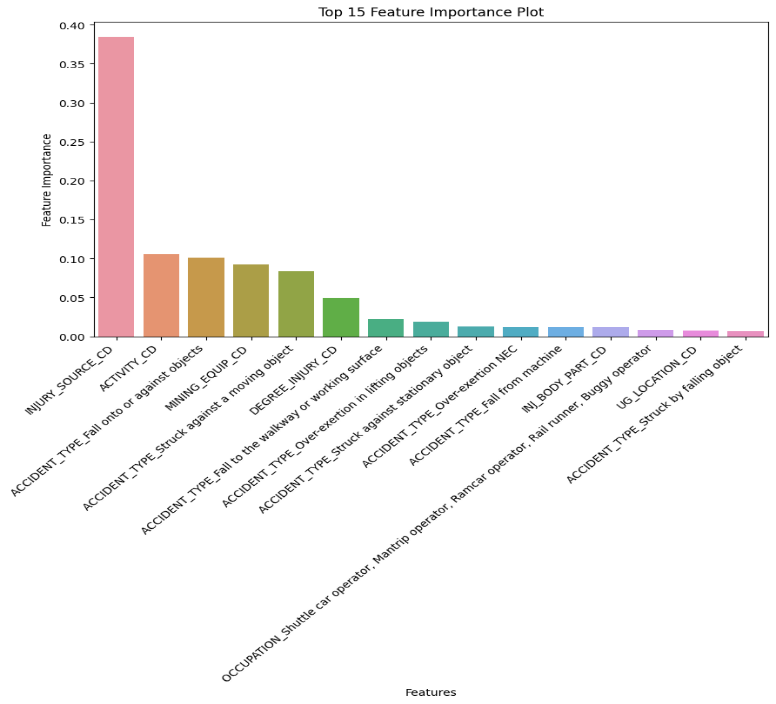
****

**Risk Evaluation and Mitigation:**

* **Analyzing Misclassifications and Identifying Risk Areas**
  1. By analyzing misclassified samples, we identified common patterns and characteristics among misclassifications. This enabled us to focus on classes with higher misclassification rates as potential high-risk areas.
     1. **Frequent Misclassifications**: First, we observed classes that are frequently misclassified by the model. From the confusion matrix analysis, we identified classes with a high number of false positives and false negatives. These classes are potential candidates for further investigation, as their misclassifications may indicate areas of difficulty for the model.
     2. **Confused Classes**: Next, we looked for pairs of classes that are often confused with each other. For example, we noticed that Class A is frequently misclassified as Class B and vice versa. This observation suggests that there might be similarities or overlaps between the incident characteristics of these two classes, leading to confusion during classification.
     3. **Characteristics of Misclassified Samples**: We delved deeper into the misclassified samples and analyzed their characteristics. By examining these samples, we sought to identify common patterns or attributes that consistently lead to misclassifications. This analysis provided us with insights into the factors contributing to misclassifications.
     4. **Comparison of Incident Characteristics:** To gain a comprehensive understanding, we compared the incident characteristics of misclassified samples with those of correctly classified samples. This comparison allowed us to identify any discernible differences or similarities that might be influencing the misclassification patterns.



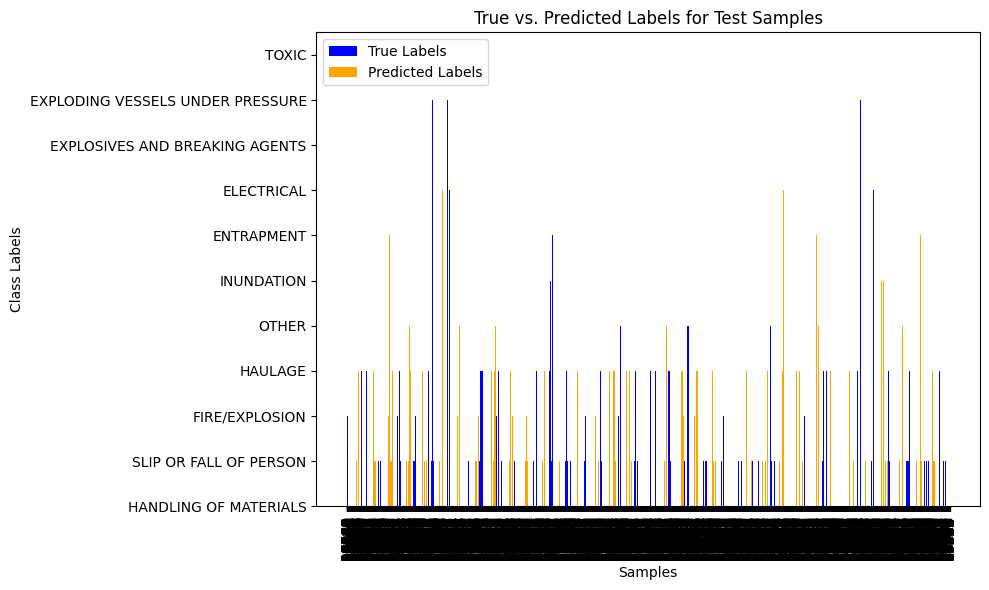
* **Importance of Features and Decision Boundaries**
  1. Analyzing feature importance plots and decision boundaries allowed us to understand which features influenced the model's predictions the most. Identifying critical factors that impact incident classification helps in understanding risk factors.



**Visualizations and Interpretations:**

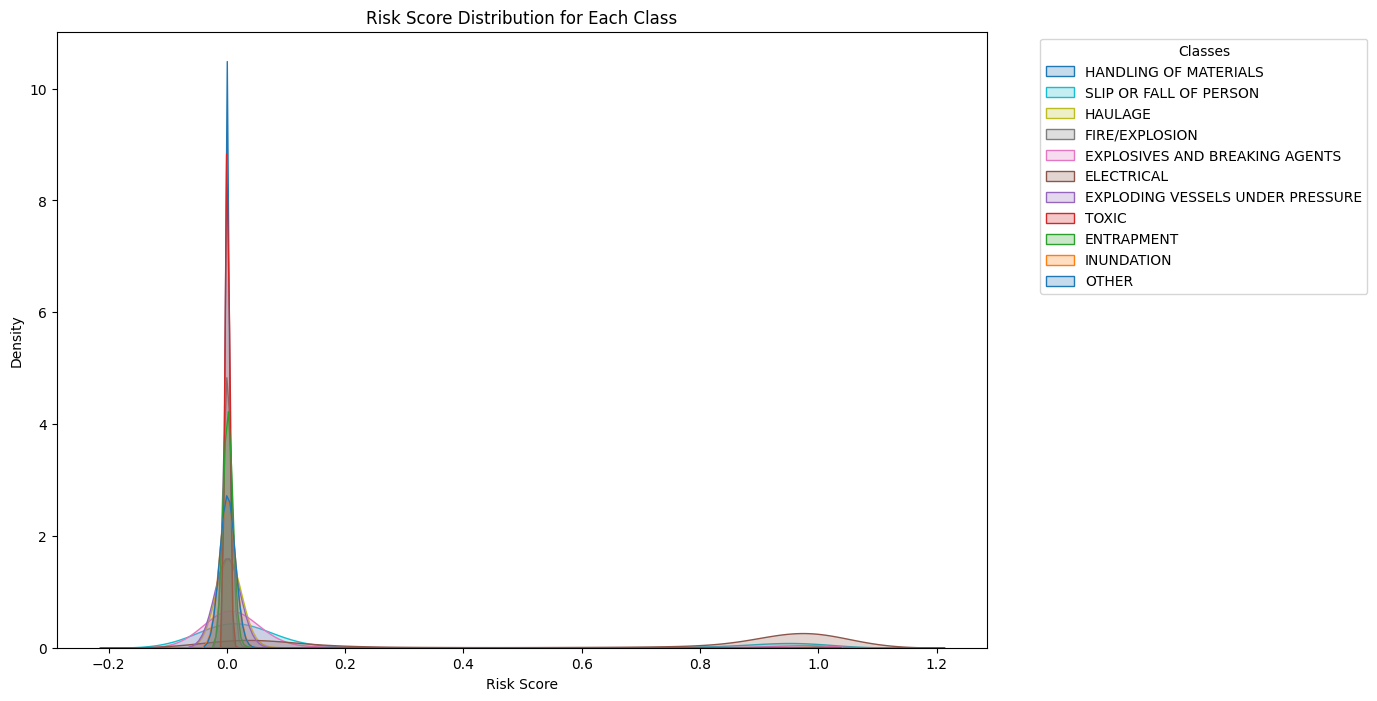
1. **Incident Classification Results:**

* A bar plot to compare the true incident labels with the predicted labels generated by the model. Each bar in the plot represents an incident sample, with the blue bars representing the true labels and the orange bars representing the predicted labels. The side-by-side comparison allows for a visual assessment of how well the model's predictions align with the actual ground truth. This plot can be useful for evaluating the overall performance of the incident classification model and identifying areas of success or improvement.

****

1. **Risk Evaluation Visualizations:**

* Risk Score Distribution: If you have a risk score or probability associated with each prediction, visualize the distribution of risk scores for each class. This can give you an idea of the model's confidence in its predictions and help set risk thresholds.



**Actionable Recommendations:**

1. **Incident Data Utilization:**

* The implementation of the GBNN model in incident classification offers valuable insights for enhancing health and safety practices in mining operations.
* By leveraging incident data, the model identifies high-risk patterns and potential hazards, enabling prompt and informed responses. Integrating the model into AMC's health and safety management system automates incident classification, leading to timely risk mitigation strategies.
* The model's predictions and confidence scores prioritize incidents based on severity, allowing for optimal resource allocation. Furthermore, the model continuously updates and refines itself with new incident data, adapting to changing conditions.
* Emphasizing a data-driven approach, stakeholders can make informed decisions to proactively manage risks.
* Establishing benchmarks and promoting data quality through training ensures the model's accuracy and effectiveness. Collaboration with experts and data sharing within the industry can collectively improve health and safety practices, making the GBNN model a transformative tool for continuous improvement.

**Conclusion:**

The Analytic Enigmas, comprising Rakulan Srinivasan, Preethi KCS, and Bharathi Sriram, successfully developed a robust incident classification model and risk evaluation framework for AMC. Our solution empowers the organization to make informed decisions, prioritize risk mitigation efforts, and embrace a data-driven approach to health and safety. By implementing our recommendations, AMC can significantly enhance its health and safety management system, ultimately leading to improved safety standards and overall operational efficiency.