

# 

# MARITIME HACKATHON 2025

Team: Quadrifoglio

`

# 

# Abstract

The objective of this report is to present the methodology for deriving the consensus severity for the deficiencies identified during Port State Control (PSC) inspection. (Clustering issue)

To ensure that all annotations related to a single deficiency were accounted for, the dataset was grouped by two key identifiers , PscInspectionId and deficiency\_code. Each group corresponds to a specific deficiency and includes severity ratings assigned by different SMEs. Using majority voting, all severity ratings for a given deficiency code were tailed and the most frequently occurring severity rating were assigned as the consensus severity.

Once the consensus severity for each deficiency was established, the processed dataset was analysed to assess patterns and trends in severity. Factors such as Age, Vessel Type and def\_text were taken into account. Natural Language Processing (NLP) techniques were applied to analyse the def\_text for deficiencies.

This approach aims to streamline the decision-making process in vessel inspections, enhancing the consistency and efficiency of severity assessments while reducing the subjectivity inherent in manual evaluations. Ultimately, the methodology contributes to improving safety in maritime operations by supporting more accurate and reliable deficiency assessments.

# 

# 

# Part 1: Study severity assigned by SME for identified deficiencies

To ensure the integrity of the dataset, we began by addressing any missing or invalid values. Such as, rows with missing or invalid severity labels (such as "Not a deficiency") were removed from the dataset. Only deficiencies with valid severity ratings (i.e., "High", "Medium", or "Low") were considered for further analysis.

To ensure that all annotations related to a single deficiency are considered, the dataset is first grouped by two key identifiers - PscInspectionId and Deficiency\_Code. Each group corresponds to a single deficiency and may contain multiple severity ratings provided by different annotators. Both the def\_text and the multiple SME votes are used to determine a single consensus severity for each def\_code.

The next step was to aggregate the severity ratings provided by the SMEs. The method chosen for this was majority voting. For each group (Deficiency code of each inspection), the severity ratings from all annotators are collected (High, Medium, Low). The most frequent severity label was assigned as the consensus severity for that deficiency. In cases where there were equal numbers of votes for different severities, the highest severity would be chosen. Majority voting was applied to establish a baseline consensus severity for each deficiency code, ensuring that the most agreed-upon severity was captured.

To further analyse the consensus severity derived from majority voting, the def\_text were analysed using Natural Language Processing (NLP) techniques.The def\_text field was converted into numerical data using **TF-IDF (Term Frequency-Inverse Document Frequency)**. This process captures important keywords and phrases related to severity. This method highlights terms frequently used in individual deficiency descriptions but less common across the entire dataset, helping to detect critical severity indicators (e.g., "leak," "fracture," "urgent"). A **Logistic Regression** model was chosen for its efficiency and interpretability in text classification tasks. The model was trained using the TF-IDF-transformed text data as input features and the **majority-voted severity** as the target labels. The training process aimed to help the model learn patterns and severity-related terms within the text descriptions.

**Prediction Refinement:** The trained model was used to predict severity ratings for all deficiency entries. These predictions were then aggregated for each **deficiency\_code** to refine the initial majority-voted severity and ensure consistency across the dataset.

#### **Integration with Original Data** The original annotation\_severity column was replaced with the derived consensus severity for each deficiency code. The rest of the dataset remained unchanged to maintain the integrity of other relevant information.

**Assumptions Made:**

1. **Equal SME Weighting:** All SME ratings were considered equally valid for majority voting.
2. **Priority in Ties:** Severity levels were prioritized in the order of **High > Medium > Low** to prioritize safety.
3. **Contextual Analysis:** Severity ratings can vary for the same deficiency code depending on the text description.
4. **TF-IDF Assumption:** Important severity cues are present in the **def\_text** and can be captured effectively by TF-IDF.

# Part 2: Using dataset along with the labels derived from Part 1 and building a model that predicts the severity when provided with a new deficiency.

The second phase of this study aims to analyze the relationships between consensus severity levels and other factors included in the dataset, such as vessel age and vessel type. By examining these features alongside the consensus severity, the goal is to uncover trends and patterns that could lead to actionable insights.

The dataset final\_consensus\_full\_data.csv was loaded for training, while the psc\_severity\_test.csv dataset was used for testing. Data cleaning ensured that only valid severity labels (High, Medium, Low) were included. Rows with other labels were removed, and the dataset was reset to ensure a clean index.

The text column (def\_text) in the training dataset was vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) technique. The TfidfVectorizer was applied to transform the textual data into numerical features, which can be used by machine learning models. The vectorizer was set with a maximum of 5000 features and a stop-word filter to remove common, non-informative words.

The target variable annotation\_severity, representing the severity of deficiencies, was encoded into numerical labels using LabelEncoder. This allowed the models to process the categorical severity labels as numerical values.

The dataset was split into training and validation sets, with 80% used for training and 20% for validation. This split ensured that the model could be evaluated on unseen data during training, helping to assess its performance and prevent overfitting.

The following three models were trained to predict deficiency severity based on text features, with each model chosen to bring different strengths to the prediction process. First, Logistic Regression was used, which is a classic and well-established classification algorithm. This model was trained using the vectorized text features obtained from the TF-IDF transformation of the "def\_text" column. Logistic Regression was selected because of its simplicity, efficiency, and interpretability, making it a good baseline model for text classification tasks. It works by estimating the probability of a data point belonging to a specific class, which is useful for understanding how the severity is related to text patterns in the data.

Secondly, a Gradient Boosting Classifier was chosen, which is an ensemble learning method. It constructs a series of decision trees, each trained to correct the errors of the previous one. The predictions of these individual trees are combined to create a stronger overall model. Gradient Boosting was selected for its ability to handle complex data patterns and its robustness in terms of predictive accuracy. By iteratively improving on the weaknesses of the earlier trees, it has the potential to deliver better performance compared to single-model approaches, particularly in capturing non-linear relationships between the features and the target variable.

Lastly, a Hybrid Model, implemented as a Voting Classifier, was developed to combine the predictions of both Logistic Regression and Gradient Boosting. The ensemble approach of the Voting Classifier was chosen to leverage the strengths of both models. In this case, hard voting was used, meaning the final prediction was determined by the majority class predicted by the individual models. This method was employed to improve the overall robustness and accuracy of the predictions by reducing the likelihood of errors that could result from relying on a single model. The use of the ensemble model allows for a more balanced decision-making process by considering the outputs from both a simpler, interpretable model (Logistic Regression) and a more complex, high-performance model (Gradient Boosting).

After training and validation, the hybrid model was applied to the test dataset (psc\_severity\_test.csv) for final predictions. The predicted severity levels were stored in the test dataset.

To facilitate future use, the trained hybrid model and the vectorizer were saved using joblib. This allows the models to be loaded and used for predictions on new data without retraining. The hybrid model was saved as hybrid\_severity\_model.pkl. The vectorizer was saved as tfidf\_vectorizer.pkl.

The hybrid model was used to predict the severity of deficiencies in the test dataset. The predictions were saved to a new file (predicted\_severity\_test.csv).