**Maritime Hackathon 2025**

**Team MA1522**

**Adhikrish Rethinakumar**

**Cheong Kang Jie**

**Chua Chloe**

**Justin Goh Guan Eng**

```
Validation Accuracy: 0.9891636565712698
Classification Report:
                   precision    recall  f1-score   support

             High       0.99      0.98      0.99       910
              Low       0.99      0.99      0.99      1451
           Medium       0.99      0.99      0.99      1232
Not a deficiency        1.00      1.00      1.00         6

         accuracy                           0.99      3599
        macro avg       0.99      0.99      0.99      3599
     weighted avg       0.99      0.99      0.99      3599
```

## Abstract

This study develops a severity prediction model for maritime deficiencies using a combination of structured data and unstructured text analysis. Severity levels (Low, Medium, High) are encoded numerically as 0, 1, and 2, and predictor variables include time since the last inspection, vessel type weights, port-specific risk scores, annotator reliability, regional authority performance, and textual descriptions. Text features are vectorized using TF-IDF, while structured features are engineered for relevance and comparability.

A linear regression model determines the contribution of each factor, producing coefficients that reflect their relative importance. These coefficients are integrated into a unified framework, which, combined with a Random Forest Classifier, predicts severity levels for new deficiencies. The approach balances interpretability and accuracy, enabling actionable insights for maritime inspections. The model lays the foundation for prioritizing high-risk deficiencies, enhancing safety and compliance in maritime operations.

## 1. Introduction

This paper presents a machine learning-based approach to predict the severity of maritime deficiencies identified during inspections. By integrating structured data and unstructured textual descriptions, the model predicts severity levels ("Low," "Medium," or "High") to assist inspectors in prioritizing high-risk issues.

## 2. Methodology

### 2.1 Data Preparation

Training and test datasets are preprocessed for downstream analysis.

### 2.2 Feature Engineering

Key features were engineered to represent domain-specific risks:

1. **Time Difference Normalization**: Calculated time since the last inspection (time_diff_normalized).
2. **Vessel Type Weights**: Assigned risk-based weights to vessel categories (e.g., Chemical, Dry Bulk).
3. **Port-Based Risk Scores**: Derived port risk scores based on the proportion of high-severity incidents.
4. **Authority-Based Risk Scores**: Evaluated inspection authorities for historical association with high-severity deficiencies.
5. **Textual Features**: Vectorized deficiency descriptions (def_text) using TF-IDF to extract meaningful patterns.

### 2.3 Weight Optimization

A **linear regression model** was applied to quantify the impact of each predictor on severity. The optimized coefficients were normalized into weights, enabling a combined feature (combined_weight) for severity prediction.

### 2.4 Text and Feature Combination

Textual features (TF-IDF matrix) were combined with structured numeric features, ensuring both data types contributed to predictions.

### 2.5 Model Training

A Random Forest Classifier was trained to capture nonlinear interactions between features. The model was evaluated using accuracy, precision, recall, and F1-scores to ensure robustness.

### 2.6 Prediction

The trained model predicted severity levels for the test dataset, with predictions mapped back to "Low," "Medium," or "High."

```
Validation Accuracy: 0.9891636565712698
Classification Report:
                   precision    recall  f1-score   support

             High       0.99      0.98      0.99       910
              Low       0.99      0.99      0.99      1451
           Medium       0.99      0.99      0.99      1232
Not a deficiency        1.00      1.00      1.00         6

         accuracy                           0.99      3599
        macro avg       0.99      0.99      0.99      3599
     weighted avg       0.99      0.99      0.99      3599
```

## 3. Results

**Validation Accuracy**: 98.9%

## 4. How Unstructured Text is Handled

The model integrates NLP, machine learning, and graph database technologies to process textual descriptions:

1. **Text Severity Analysis**: Text is tokenized using SpaCy to remove stop words and punctuation. Tokens are classified using Hugging Face's distilbert-base-uncased model, mapping probabilities to severity scores. A weighted average of token scores determines the overall severity.
2. **Neo4j Knowledge Graph**: Metadata and severity scores are stored in a Neo4j graph. Nodes represent entities (e.g., vessels, incidents), while edges capture relationships (e.g., "IMPACTS"). The graph allows contextual queries, such as retrieving high-risk deficiencies by vessel type.