Data Science Project Report: Week 2 Summary

Technical Analysis Team

December 2025

1 Project Objective

Following the data cleaning phase completed in Week 1, the primary objective of Week 2 was to develop a predictive system capable of identifying the **Nature of Injury** and the **Cause of Injury** based on textual descriptions found in the *Final Narrative* column.

2 Feature Engineering and Vectorization

To transform the textual data into a format suitable for machine learning, the **TF-IDF (Term Frequency-Inverse Document Frequency)** technique was employed.

- **Initial Feature Extraction:** The raw vectorization of the *Final Narrative* column produced an initial set of **16,050** features.
- **Frequency Filtering:** To reduce noise and computational complexity, words appearing in fewer than 5 rows or more than 80% of the narratives were ignored.
- **Final Feature Set:** This optimization resulted in a refined feature space of **6,785** features, significantly improving model efficiency.
- **Target Encoding:** Label Encoding was applied to the *Nature Title* and *Event Title* (Cause) to prepare the categorical targets for classification.

3 Model Training

A **Support Vector Machine (SVM)** classifier was trained using an 80/20 train-test split. To ensure model convergence across the high-dimensional feature space, the model was configured with **5,000 iterations**.

4 Evaluation and Results

The model served as a dual-classifier to meet the two primary goals of the week.

4.1 Goal 1: Predicting Nature of Injury

The model demonstrated high predictive power for the nature of injuries, likely due to specific medical terminology in the narratives.

Metric	Value
Accuracy	0.8793 (88%)
Precision (Weighted)	0.8610
Recall (Weighted)	0.8793
F1-Score (Weighted)	0.8658

Table 1: Evaluation Metrics for Nature of Injury

4.2 Goal 2: Predicting Cause of Injury

Predicting the cause of the injury proved more challenging, resulting in lower performance metrics compared to the nature of injury.

Metric	Value
Accuracy	0.5657 (56%)
Precision (Weighted)	0.5234
Recall (Weighted)	0.5657
F1-Score (Weighted)	0.5350

Table 2: Evaluation Metrics for Cause of Injury

5 Conclusion

The objectives for Week 2 were successfully initiated. While the SVM model performs exceptionally well for classifying the **Nature of Injury** with 88% accuracy, there is a significant performance gap in predicting the **Cause of Injury** (56%).