EHEI University

Thesis Defense

By Fatemeh Radboy

Title: Comparative Analysis of Machine Learning Models for Predicting Workplace Accident Severity

Name: Fatemeh Radboy

Program: MSc in Data Science and Artificial Intelligence

Supervisor: professor Damiano Laviola

Date: June 20, 2025

Draft

01 Introduction

02 Literary Review

03 Research Methods

04 Discussion

05 Conclusion

Background & Motivation

- Over 2.7 million workplace injuries occur annually in the U.S.
- Economic cost exceeds $170 billion per year
- Not all injuries are equal – hospitalization is a key severity indicator
- Goal: Predict hospitalization to support prevention, planning, and response

Research Objectives

- Analyze OSHA severe injury data from 2015 to 2025
- Build a complete ML pipeline: preprocessing to deployment
- Compare four classifiers: LR, DT, RF, XGBoost
- Optimize and evaluate models
- Deploy the best model as a RESTful API for real-time predictions

Significance of Study

- Enables real-time severity prediction of incidents
- Assists safety teams in proactive decision-making
- Converts unstructured injury data into actionable insight
- Contributes a practical ML tool to occupational safety management

Literature Review & Research Gaps

- Previous studies focus on injury detection, not hospitalization
- Limited use of unstructured text and OSHA code data
- Few include model deployment and real-time systems
- This project addresses all these gaps

Data Source & Overview

- Dataset: OSHA Severe Injury Reports (2015–2025)
- Records: ~93,000
- Target: Hospitalized (1) vs. Not Hospitalized (0)
- Features: Date, State, Injury Codes, NAICS, Narrative, Geo-coordinates

Sample OSHA Codes

- Nature_111 → Fracture
- Body_4422 → Fingers
- Event_4330 → Fall from height
- Source_5721 → Co-worker
- Used as categorical variables after encoding

Data Preprocessing Steps

- Extracted temporal features (day, month, weekday, weekend)
- Imputed missing values using median/mode/‘Unknown’
- Text cleaning using NLTK (lowercase, stopwords, regex)
- Geo-clustering via KMeans on lat/long
- Vectorized narrative using TF-IDF (top 100 terms)

Feature Engineering Summary

- Categorical: State, NAICS, Nature, Event, Body, Source, Geo Cluster
- Numerical: Latitude, Longitude, Day, Month, Year
- Textual: 100 TF-IDF narrative tokens
- Combined using ColumnTransformer in Scikit-learn

Models & Pipeline

- Models: Logistic Regression, Decision Tree, Random Forest, XGBoost
- Preprocessing and modeling built into a unified pipeline
- GridSearchCV used for hyperparameter tuning
- Stratified train-test split with evaluation on test data

Evaluation Metrics

- Metrics: Accuracy, Precision, Recall, F1 Score, ROC AUC
- F1 prioritized due to class imbalance (80% hospitalized)
- ROC curve used to validate discrimination ability

Performance Before Optimization

- XGBoost: F1 = 0.9518, AUC = 0.9669
- Logistic Regression: F1 = 0.9516, AUC = 0.9648
- Random Forest: F1 = 0.9493, AUC = 0.9628
- Decision Tree: F1 = 0.9423, AUC = 0.8610

Performance After Optimization

- XGBoost (best): F1 = 0.9525, AUC = 0.9671
- Decision Tree: F1 = 0.9517, AUC = 0.9585 (after tuning)
- Random Forest and LR slightly improved
- XGBoost selected for deployment

Feature Importance (Top Predictors)

- Nature_1311 (Fracture injuries)
- Body_4422 (Fingers)
- Latitude and Longitude (regional risk patterns)
- Event_6412 (Struck by person)
- Text terms (TF-IDF) added contextual power

Key Insights & Patterns

- Certain injuries (fractures, falls) are strong hospitalization indicators
- Geographic clusters reflect regional safety or healthcare access
- Narrative text enhances model understanding of severity context

Deployment Architecture

- Trained model saved using Joblib (XGBoost)
- Flask REST API built with /predict endpoint
- Input: JSON injury record → Output: Probability & prediction
- Modular, scalable, and easy to integrate into existing systems

API Prediction Example

- Input JSON: OSHA code values, narrative, location
- Output: {"hospitalized": true, "probability": 0.9973}
- Enables real-time use in safety reporting systems

Practical Applications

- On-site risk scoring for safety officers
- Prioritizing incidents for rapid response
- Integrating into dashboards, mobile apps, or inspection software
- Can be scaled across industries or departments

Project Limitations

- Imbalanced Classes: 80% hospitalized → F1 used to address bias
- OSHA Codes: Numeric codes are not intuitive for non-technical users
- U.S.-Only Data: Limits generalization to other countries
- XGBoost Interpretability: Acts as a black-box without SHAP/LIME

Future Work

- Add SHAP/LIME for model explainability
- Map OSHA codes to human-readable labels
- Use deep NLP models (BERT) for narrative analysis
- Expand to other regions or industry-specific models

Conclusion

- Built and deployed a full ML pipeline for OSHA severity prediction
- XGBoost achieved F1 = 0.9525, AUC = 0.9671
- Model deployed via REST API for real-time use
- Provides tangible value for occupational safety applications

Question and Answer...

EHEI University

Thank You

By Fatemeh Radboy