

**Msc - Data Science and Advanced Analytics  
2024/25**

**Machine Learning Final Project**

**To Grant or Not to Grant**

**Group 45**

* 20240536 Inês Jacinto [20240536@novaims.unl.pt](mailto:20240536@novaims.unl.pt)
* 20240561 Antônio Ramos [20240561@novaims.unl.pt](mailto:20240561@novaims.unl.pt)
* 20240598 Sofia Jacinto [20240598@novaims.unl.pt](mailto:20240598@novaims.unl.pt)
* 20240750 Marisa Marques [20240750@novaims.unl.pt](mailto:20240750@novaims.unl.pt)

Index

[1. Introduction 3](#_Toc185275579)

[2. Exploration and Data Preprocessing 4](#_Toc185275580)

[Artefacts 4](#_Toc185275581)

[Imbalanced target (“Claim Injury Type”) 4](#_Toc185275582)

[Duplicates & datasets alignment 4](#_Toc185275583)

[Anomalies & Missed Values 4](#_Toc185275584)

[3. Multiclass Classification 7](#_Toc185275585)

[4. Open-Ended Section 7](#_Toc185275586)

[5. Conclusion 8](#_Toc185275587)

[6. References 8](#_Toc185275588)

[7. Annexes (if needed) 9](#_Toc185275589)

# Introduction

Handling workplace injury claims is a critical and complex task for organizations and regulatory bodies. In New York, the Workers’ Compensation Board (WCB) manages these claims, ensuring proper categorization and fair compensation. With over 5 million claims reviewed since 2000, the manual nature of this process demands significant time and resources, creating opportunities for automation.

This project seeks to develop a machine learning solution capable of classifying injury types associated with claims submitted to the WCB. Using a dataset spanning claims from 2020 to 2022, the primary objective is to create and benchmark multiclass classification models that predict WCB decisions with precision. Furthermore, the project will investigate model optimization strategies and assess feature significance, providing deeper insights into the data and its impact on prediction accuracy. The ultimate aim is to design a predictive tool that improves the efficiency and reliability of WCB’s claim evaluation process.

The initiative draws inspiration from successful applications of machine learning in insurance and healthcare, where automated models have demonstrated their ability to streamline operations. By leveraging these technologies, this project aspires to reduce the manual burden on the WCB while improving the consistency and fairness of its decision-making processes.

Besides our work, its results, with trials and our better predictions, will be in a Kaggle competition.

**Similar Works or Applications**

Many insurance companies use machine learning to classify and process claims. For example, models are trained to categorize claims based on severity or type (e.g.), accidents, thefts, or injuries). A notable example is the use of Natural Language Processing (NLP) to analyze claims documentation and predict outcomes.

* + A study by Verma et al. (2020) developed a machine learning framework to automate claims categorization, achieving high accuracy with Random Forest models.
  + Research by Xu et al. (2018) explored injury severity prediction using Support Vector Machines (SVM) and ensemble methods, leveraging structured medical datasets.
  + A project by Kube et al. (2021) used machine learning models to analyze occupational injury data and identify factors influencing claim outcomes. Their models also highlighted feature importance to guide policy-making.
  + Zhang et al. (2019) employed deep learning techniques to classify claims, demonstrating that hybrid approaches combining feature engineering and deep models outperform traditional methods.

# Exploration and Data Preprocessing

## Artefacts

As the basis for this work, we have two datasets: ***train\_df*** & ***test\_df*** (csv files). The 1st, which contains supervised values for the target feature, will serve as the training and testing artifact for the applied models. The 2nd dataset, which lacks the target feature (“Claim Injury Type”), also exhibits some structural differences in its features, which must be evaluated and adapted to ensure the success of the project.

The ***train\_df*** dataset structure may be seen in the References

## Imbalanced target (“*Claim Injury Type*”)

The first thing to notice is the significant imbalance among the possible outcomes for the target feature in the working dataset, as demonstrated below:

A graph of injury type

Description automatically generated

Regarding the work itself, this characteristic implies that the underrepresented classes have fewer data points, and accuracy is not the ideal metric for evaluating the model.

## Duplicates & datasets alignment

* 18350 “*Claim Identifier*” duplicates in ***train\_df*** will be dropped
* The features “Agreement Reached” & “WCB Decision” are not in the ***test\_df*** dataset and will be dropped from ***train\_df***.

## Anomalies & Missed Values

* 3.27% of the training data has NaN on the target “*Claim Injury Type*” that will be dropped.
* Other anomalies & pre-processing issues will be seen below:

|  |  |
| --- | --- |
| *Alternative Dispute Resolution* | * Not isna() * Dropped lines == “U” (annomaly) * MultiEncoder(Binary) – LabelEncoder (bin/bool) |
| *Attorney Representative* | * Not isna() * MultiEncoder(Binary) – LabelEncoder (bin/bool) |
| *Carrier Name* | * Not isna() * MultiEncoder(frequency) |
| *Carrier Type* | * Not isna() * MultiEncoder(Binary) – LabelEncoder (int) |
| *Claim Injury Type* | * Isna() treated in item 1 * Target feature * MultiEncoder(Binary) – LabelEncoder (int) |
| *County of Injury* | * Not isna() * MultiEncoder(frequency) |
| *COVID-19 Indicator* | * Not isna() * MultiEncoder(Binary) – LabelEncoder (bin/bool) |
| *District Name* | * Not isna() * MultiEncoder(frequency) |
| *Gender* | * Not isna() * Replace(“U”, “X”) » domain: (“M”, “F”, ”X”) * MultiEncoder(frequency) |
| *Industry Code Description* | * Strongly correlated with “*Industry Code*” * Description feature * Dropped |
| *Medical Fee Region* | * Not isna() * MultiEncoder(frequency) |
| *OIICS Nature of Injury Description* | * Only NaN * Dropped |
| *WCIO Cause of Injury Description'* | * Strongly correlated with “*WCIO Cause of Injury Code*” * Description feature * Dropped |
| *WCIO Nature of Injury Description'* | * Strongly correlated with “WCIO Nature of Injury Code” * Description feature * Dropped |
| *WCIO Part Of Body Description'* | * Strongly correlated with *“WCIO Part Of Body Code”* * Description feature * Dropped |

**Categorical and also Numeric Features**

|  |  |
| --- | --- |
| *Industry Code* | * 1,73% NaN |
| *WCIO Cause of Injury Code* | * Not isna() |
| *WCIO Part Of Body Code* | * Not isna() |
| *WCIO Nature of Injury Description* | * Not isna() |
| *Zip Code* | * not.isnumeric() to NaN * Not isna() (fillna(0)) |

Obs.: Missed Values on *Industry Code* filled after split (train & val) for train data integrity maintenance (Rui).

**EDA – Numerical Features**

|  |  |
| --- | --- |
| *Age at Injury* | * Dropped values <16 * Filled NAs with the average (because the distribution is simetric) * Handled outliers with zscore |
| *Average Weekly Wage* | * No missing values * Handled outliers with zscore * KNNImputer |
| *Birth Year* | * Dropped values < 1903 * Filled NAs with Accident Date – Age at Injury * Handled outliers with zscore |
| *Claim Identifier* | * No missing values * Handled outliers with zscore |
| *Industry Code* | * No missing values * Handled outliers with zscore * KNNImputer |
| *IME-4 Count* | * Filled NAs with zero * Handled outliers replacing all values > 15 with 15 |
| *WCIO Cause of Injury Code* | * Filled NAs with zero (resolver) * Handled outliers with zscore |
| *WCIO Nature of Injury Code* | * Filled NAs with zero (resolver) * Handled outliers with zscore |
| *WCIO Part of Body Code* | * Filled NAs with zero (resolver) * Handled outliers with zscore |
| *Number of Dependents* | * Filled NAs with zero (resolver) * Handled outliers with zscore |

**Novas:**

|  |  |
| --- | --- |
| *Accident Date* | * Dropped NAs |
| *Years Past Accident* | * Accident Date – reference\_date |

**Missing values before:**

* Accident Date: 0.64%
* Age at Injury: 0.94%
* Average Weekly Wage: 4.99%
* Birth Year: 0.38%
* Industry Code: 1.73%
* Years Past Accident: 0.64%

**Missing values after:**

* Average Weekly Wage: 4.99%
* Industry Code: 1.46%
  + Data cleaning and preprocessing steps.

# Multiclass Classification

* + Overview of classification objectives and metrics.
  + Additional preprocessing steps.
  + Feature selection strategy.
  + Model assessment strategy (e.g., cross-validation, holdout).
  + Comparison of candidate algorithms.
  + Optimization efforts: methods and results.

# Open-Ended Section

* + Defined objectives and additional analyses.
  + Explanation of approaches (e.g., feature importance analysis, alternate prediction tasks).
  + Presentation and discussion of findings.

**Results and Discussion**

* + Comparison of performance across models.
  + Insights derived from optimization and open-ended analyses.
  + Alignment with initial objectives and expectations.

# Conclusion

This project set out to address the challenge of automating injury classification for workers' compensation claims using machine learning techniques. By leveraging historical data from the New York Workers’ Compensation Board (WCB), we developed and optimized multiclass classification models to predict claim injury types, offering a solution that can potentially streamline and enhance the decision-making process.

The results demonstrate the feasibility of employing machine learning to achieve accurate predictions, with optimized models significantly improving over initial baselines. The exploration of feature importance provided valuable insights into the key factors influencing injury classifications, while additional analyses, such as alternative prediction tasks, further highlighted the flexibility and robustness of the proposed methodologies. However, the project also revealed limitations, such as the dependency on the quality of input data and the challenge of model generalization to unseen claims.

Looking forward, several avenues for future work can be pursued. Incorporating more advanced modeling techniques, such as ensemble learning or neural networks, could further improve prediction accuracy. Additionally, integrating external data sources, such as real-time industry trends or macroeconomic indicators, may enhance the model's contextual understanding. Finally, developing a user-friendly analytics interface could make these models more accessible to the WCB, enabling seamless integration into their operational workflows.

In conclusion, this project not only validates the potential of machine learning in automating claims classification but also sets the foundation for broader applications in regulatory and insurance contexts. By continuing to refine and expand upon these efforts, the WCB could achieve greater efficiency, transparency, and accuracy in its operations.

* + Summary of findings relative to initial goals.
  + Limitations of the approach.
  + Recommendations for future work.

# References

* + List of all referenced materials (peer-reviewed articles, datasets, etc.).

# Annexes (if needed)

* + Additional figures, tables, or analyses that support the report but are not critical to the main narrative.