NOVA IMS Information Management School

2024/2025

MACHINE LEARNING

TO GRANT OR NOT TO GRANT

MASTER IN DATA SCIENCE AND ADVANCED ANALYTICS

GROUP 48

- AFONSO LOPES, 20211540
- ANA ISABEL DUARTE, 20240545
- DIOGO FERNANDES, 20220507
- PEDRO CAMPINO, 20240537

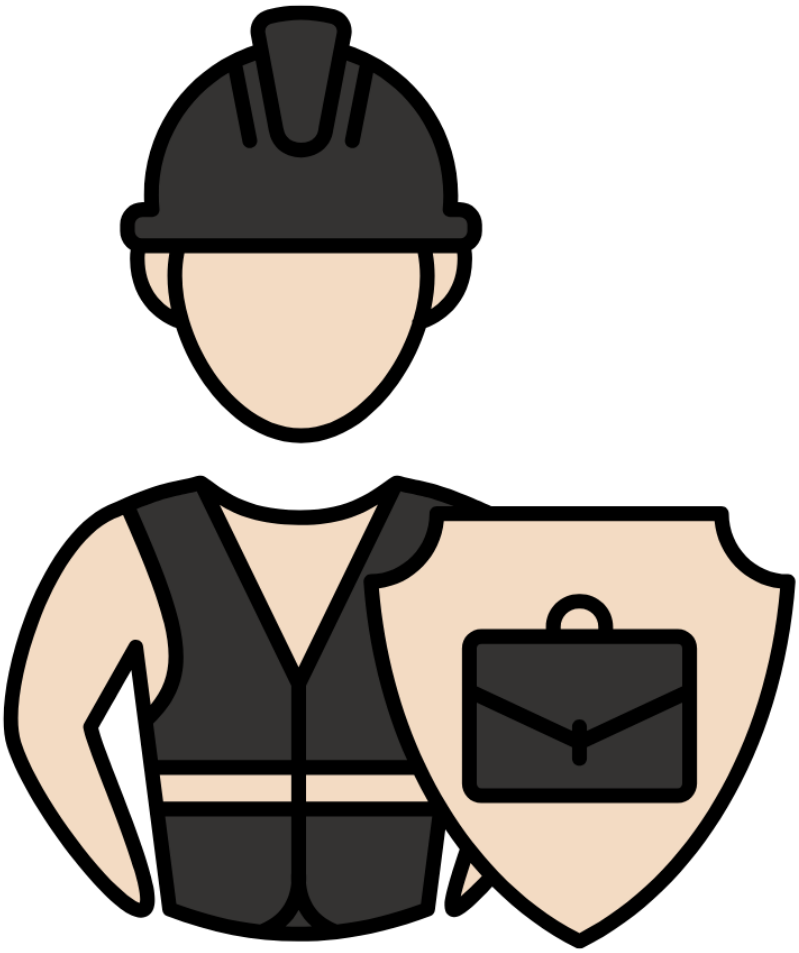

Table of Contents

1. Introduction

This project aims to develop a model to assist the New York Workers' Compensation Board in automating decisions on new workplace injury claims. We will begin with an in-depth Exploratory Data Analysis (EDA) to gain insights that guide the model's development phases. Next, we'll conduct Data Preprocessing to prepare the dataset, followed by Feature Selection to identify the most relevant variables to train our model. Finally, in the Modeling, Assessment, and Deployment phases, we'll evaluate several predictive models to choose the best one for unseen data. Additionally, we will include an Open-Ended Section to apply advanced techniques to further enrich the project.

2. Exploratory Data Analysis

The Data Exploration phase begins with **Univariate Analysis**, examining the main statistics and distributions of numeric features, as well as the frequency of each categorical variable. We then move to **Bivariate Analysis** to explore relationships between variables, including correlations and potential nonlinear associations among numeric features. Next, we analyze the distribution of numeric features across categorical feature groups and examine interaction effects between categorical features. As this is a supervised learning project, we will focus on each feature's discriminative power with respect to the target variable.

3. Data Pre-Processing

Using insights from the EDA, we'll proceed with Data Pre-Processing to clean and transform the dataset, improving data quality for model development. We'll address missing values and outliers, filling some based on logical inference or provided information and imputing others based on data type and percentage of NaNs. High-cardinality categorical features will be grouped, and irrelevant columns removed to simplify the dataset. We'll also apply Feature Engineering to create new features that capture meaningful relationships and scale numeric features where needed. This preprocessing stage is crucial to model performance, and if results are unsatisfactory, we may revisit and refine our approach to ensure optimal data quality.

4. Feature Selection

In the Feature Selection stage, we will select the most relevant features for our model:

- **Filter Methods**

We will begin by identifying **Constant** and **Quasi-Constant features** using variance analysis. For exploring relationships between numeric variables, we will apply

Spearman's correlation. Although direct correlation with the categorical target variable isn't appropriate due to its categorical nature, we will use this approach to assess redundancy and prevent multicollinearity. For categorical features, we will utilize visualizations to examine how the target variable's categories are distributed across the feature categories. Additionally, we will employ the **Chi-Squared Test of Independence** and **Cramér's V** to assess the strength of association between two categorical variables.

- **Wrapper Methods**

We will also apply wrapper methods, including **Recursive Feature Elimination** (RFE) to iteratively remove the least important features and optimize the feature subset. Additionally, **Recursive Feature Elimination with Cross-Validation** (RFECV) will be used to provide a more stable feature ranking by incorporating cross-validation.

- **Embedded Methods**

We will also use embedded methods, such as **Lasso Regression**, which select important features during model training by applying regularization to reduce complexity.

Alternatively, we will use dimensionality reduction techniques such as Factorial analysis of mixed data to reduce the feature space and capture underlying patterns across both categorical and numeric variables.

5. Modelling, Assessment and Deployment

In this stage, we will experiment with different types of models to identify which one best captures the patterns in the dataset, although our main focus is Neural Networks due to yielding the best results. To avoid overfitting and ensure a fair evaluation, we plan to use cross-validation to assess the models across different data subsets. We will also perform some hyperparameter tuning and probably some pre-processing and feature selection adjustments in order to optimize our model. Given the unbalanced nature of our data and the fact that the model's performance will be evaluated using a macro approach, using different metrics, such as accuracy and F1 Score, is very important.

6. Open-Ended Section

For this section we will execute the last step on the Machine Learning pipeline, model deployment. We will use Steamlit to create an application with a user-friendly interface where users can insert their data and get a prediction. This application will be hosted on Streamlit Community Cloud.