Predictive Modeling of Road Accident Severity

**Machine Learning Course - Group 12** (Rafick Jungul)

### Introduction and Project Objectives

Road accidents pose a significant public health and safety challenge, leading to economic costs, injuries, and fatalities. The ability to accurately predict the severity of a road accident can provide city officials, urban planners, and emergency services with a powerful tool to make data-driven decisions. This report details the comprehensive development of a predictive model for classifying the severity of over 200,000 road accidents in Montreal. The project's primary goal was to create an accurate and robust machine learning system capable of categorizing accidents into five distinct severity levels:

* Minor material damage (Dommages matériels inférieurs au seuil de rapportage)
* Material damage only (Dommages matériels seulement)
* Minor injury (Léger)
* Serious injury (Grave)
* Fatal (Mortel)

A central and critical challenge was the severe **class imbalance** within the dataset. Fatal and serious injury accidents, while the most important to predict, are statistically rare events. Consequently, a successful model's performance was not judged solely on overall accuracy, but more importantly on its ability to correctly identify these rare, high-consequence events, a task that requires a targeted and nuanced modeling approach. The insights gained from this project have the potential to inform proactive road safety initiatives, optimize emergency response resource allocation, and guide infrastructure improvements.

### Data Acquisition, Cleaning, and Exploratory Data Analysis (EDA)

**Source and Scope**

The Road Collision dataset is sourced from the Government of Quebec website (<https://www.donneesquebec.ca/recherche/dataset/vmtl-collisions-routieres>). The dataset is a subset and only includes accidents that occurred within the city of Montréal and it excludes collisions that happened on the highway network. The project began with a comprehensive dataset containing 218,272 accident records. Each record included a wide range of features (68) related to the accident's context.

**Features**

The dataset's 68 features are categorized into:

* Temporal Data: Accident date, time, day, and month, allowing exploration of seasonal and daily patterns.
* Environmental Conditions: Weather codes indicate conditions like clear, rainy, or snowy, assessing weather's role in accidents.
* Road Characteristics: Includes road type, surface condition, and lighting, offering insight into physical road aspects affecting safety.
* Demographic and Vehicle Data: Information on involved parties and vehicles enhances understanding of affected groups.

**Initial Exploration**

* Data Size and Structure: With 218,272 entries, the dataset supports robust statistical analysis. Features include both numerical and categorical data, necessitating varied preprocessing strategies.
* Target Variable: 'GRAVITE' (severity), which classifies severity from minor to fatal, guides the analysis towards understanding accident impacts.

**2.1. Data Cleaning and Preprocessing**

The initial dataset required extensive cleaning to be usable for machine learning. This process included:

**Handling Missing Data**

* **Imputation Strategies:** Utilized median imputation for numerical values to preserve distribution shape, essential for unsupervised learning tasks.
* **Categorical Handling:** Imputed missing categorical values with 'Not specified', avoiding artificial biases while maintaining dataset integrity.
* **Dropped Columns:** Columns with over 70% missing data, such as 'NO\_CIVIQ\_ACCDN', were excluded to streamline analysis focus.

**Encoding and Transformation**

* **OneHotEncoding:** Essential for converting categorical variables into a machine-friendly format. This transformation involved turning nominal features like weather and road type into binary vectors, facilitating model processing.
* **Dimensionality Reduction with PCA:** Reduced computational burden by compressing the feature space, maintaining 95% variance, and preventing overfitting.

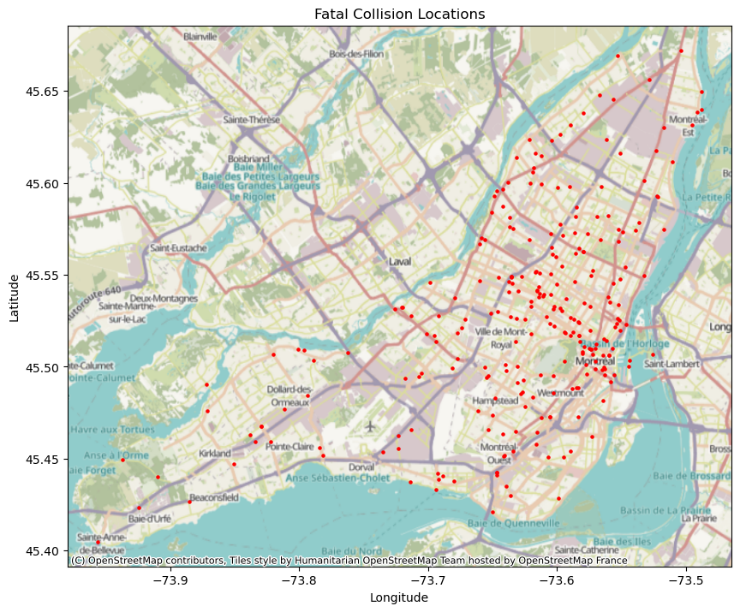
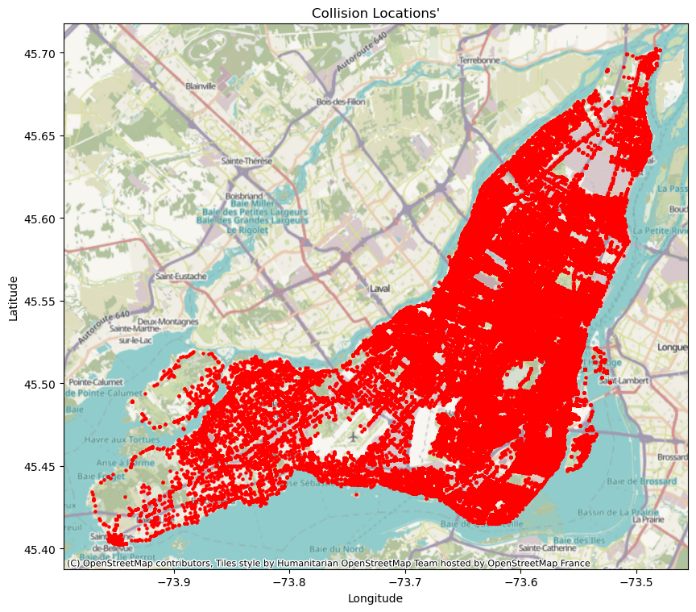
**Scaling**

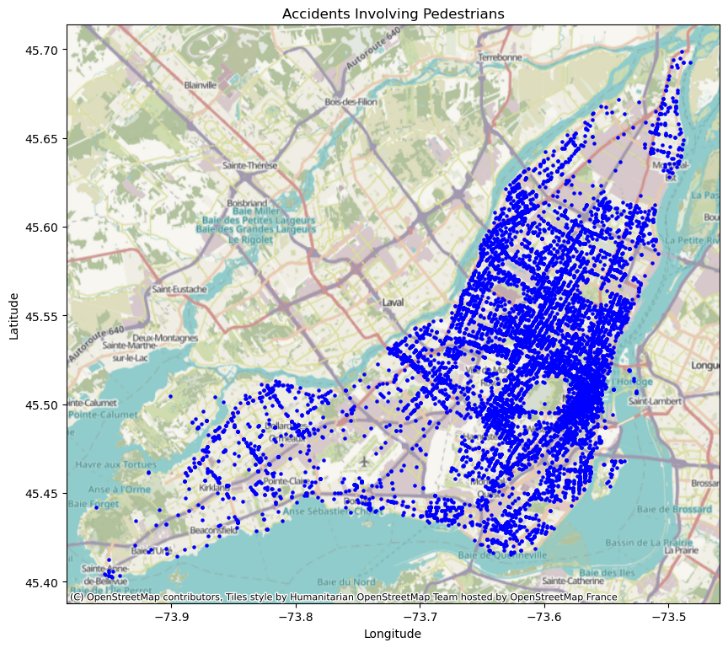
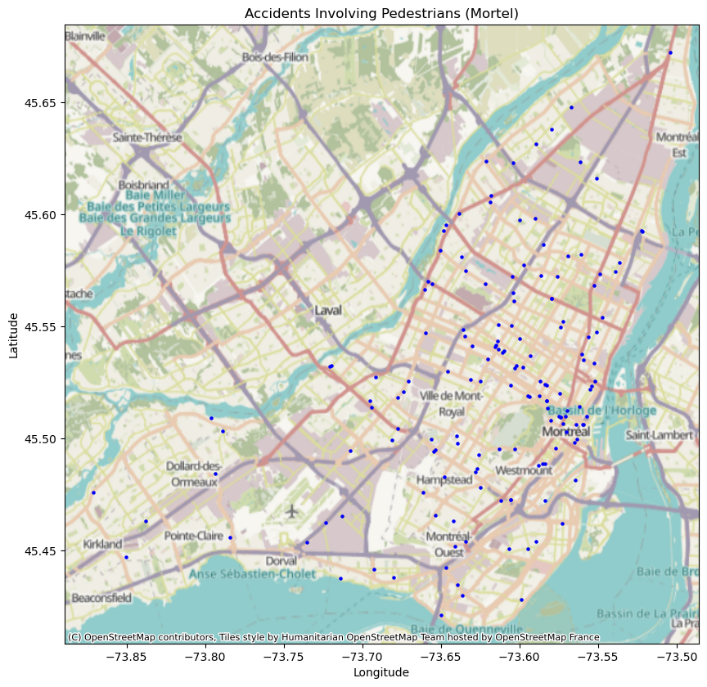
* **Feature Scaling with StandardScaler:** Ensured normalized feature contributions across models, a crucial step particularly for gradient-based optimization methods, enhancing model convergence and stability.

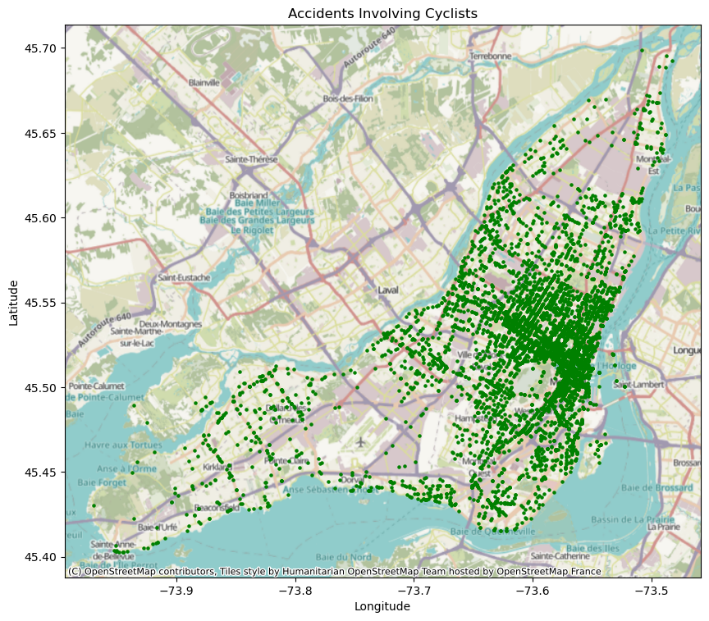
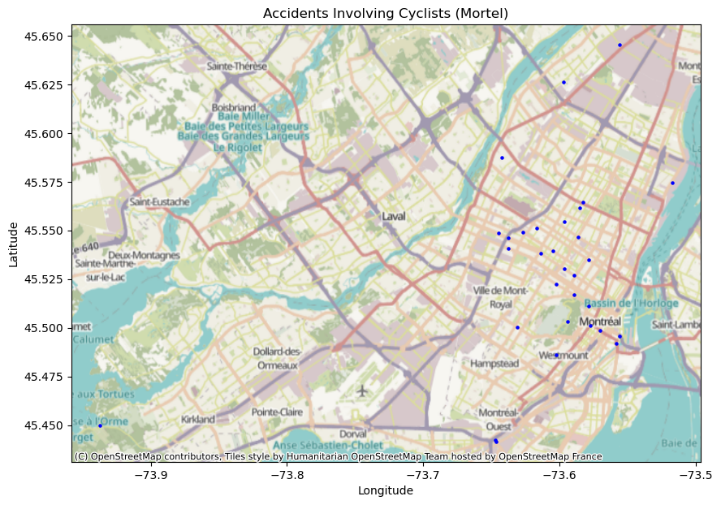
**2.2. Exploratory Data Analysis (EDA)**

EDA was crucial for understanding the data's characteristics and challenges. Key findings include:

* **Geospatial Analysis**: Visualizing accident locations on a map revealed a distinct pattern of concentration. These maps also highlighted specific accident hotspots for vulnerable road users like cyclists and pedestrians, which were often found in areas with limited bike infrastructure or high foot traffic. This geospatial insight suggests that urban planning and traffic management can be targeted to these specific zones to reduce risk.



**Temporal Patterns**

* **Day-of-Week Trends:**Thursday and Friday showed heightened accident volumes, corresponding to increased commuter activity. Saturday and Sunday tend to have less accident as majority of people does not work during week end.

A graph of a number of people

AI-generated content may be incorrect.

* **Seasonal Trends:**Evaluated monthly variations, revealing winter peaks likely due to challenging driving conditions such as snow and ice.
* **Hourly Patterns:**Accidents clustered during morning and evening rush hours, underlining the link between traffic density and accident frequency.

**Environmental and Road Conditions**

* **Weather Conditions:**Clear weather had the highest accident rate, suggesting road usage volume rather than poor conditions contributed.

**A graph with blue bars

AI-generated content may be incorrect.**

* **Lighting:**With a significant number of accidents in daylight and well-lit roads at night, the data suggested overconfidence among drivers during these times, potentially leading to faster driving and riskier behavior.

A graph with blue rectangles

AI-generated content may be incorrect.

* **Surface Conditions:**Dry roads reported most accidents, but wet and icy conditions significantly impacted severity levels, hinting at differential risks based on surface types.

**A graph with blue squares

AI-generated content may be incorrect.**

**Accident Types and Locations**

* **Collision Analysis:** Vehicle-to-vehicle collisions were predominant, suggesting areas for vehicle safety technology improvements.

A graph with numbers and a bar

AI-generated content may be incorrect.

* **High-Risk Areas:** Intersections and busy arterials were identified as critical areas for intervention, highlighting the need for enhanced traffic management strategies to decrease conflict points.

A graph with blue bars

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

### Methodology

**3.1. Dimensionality Reduction with PCA**

Given the high number of features generated by one-hot encoding, **Principal Component Analysis (PCA)** was employed to reduce the dimensionality of the dataset. PCA transforms a set of correlated variables into a smaller set of uncorrelated variables, called principal components. This step was critical for:

* **Combating Multicollinearity**: PCA effectively decorrelates the features, which can negatively impact the performance of linear models and some tree-based methods.
* **Improving Model Efficiency**: Decreasing the computational time for training and inference by reducing the number of features the model needs to process.
* **Preventing Overfitting**: By reducing the noise and redundant information, PCA helps models focus on the most important underlying patterns, leading to better generalization on unseen data.

The PCA was configured to retain **95% of the total variance**, striking a crucial balance between data compression and information preservation.

**3.2. Train-Test Split**

The final preprocessed dataset was partitioned into an 80% training set (X\_train, y\_train) and a 20% test set (X\_test, y\_test). This split ensures that the model's performance is evaluated on unseen data, providing an accurate measure of its generalization ability and preventing the model from simply memorizing the training data.

### Model Training and Evaluation

A rigorous comparison of various models was conducted to identify the most effective solution. Each model's performance was evaluated using a **Classification Report**, which provides detailed metrics for precision, recall, and F1-score for each class. The **macro average** of these metrics was used to assess overall performance, as it gives equal weight to each class, preventing the results from being dominated by the majority classes

**4.1. Baseline Models and Hyperparameter Tuning**

Initial modeling focused on establishing a baseline for comparison.

* **Random Forest Classifier**: The initial Random Forest model, with default parameters, achieved an overall accuracy of **71%**. While a good starting point, its performance on the minority classes was moderate, struggling to correctly identify all instances of fatal accidents.
* **Logistic Regression**: This linear model, trained on the PCA-transformed data, slightly outperformed the Random Forest with a test accuracy of **73%**, suggesting that the linear relationships captured by PCA were well-suited to this model's assumptions.
* **Handling Imbalance**: To explicitly handle the class imbalance, a Random Forest model was retrained with the class\_weight parameter set to 'balanced'. This adjustment penalizes the model less for misclassifying minority classes. The results were dramatic:
  + The recall for the Mortel class improved to **1.00**, meaning every fatal accident was correctly identified.
  + The overall accuracy slightly decreased to **68%**, demonstrating the classic trade-off of prioritizing recall for critical events over general accuracy.

**4.2. Advanced Models: Ensemble and Neural Networks**

To further push the performance envelope, more complex models were implemented.

* **Soft-Voting Ensemble**: A VotingClassifier was created using **soft voting** to combine the predictive power of a Random Forest, a Gradient Boosting machine, and a Logistic Regression model. The soft voting approach averages the predicted probabilities from each model, providing a more robust final prediction. This ensemble achieved a balanced performance with an overall accuracy of **71%** and high scores for minority classes.
* **Neural Network (scikit-learn)**: A basic MLPClassifier with two hidden layers (100, 50) and a relu activation function achieved a commendable **71%** accuracy, demonstrating its ability to capture complex, non-linear patterns.
* **Neural Network (TensorFlow)**: The most successful model was a deep neural network built using the TensorFlow framework. This model's architecture consisted of multiple dense layers with a relu activation and Dropout regularization to prevent overfitting. It was trained on one-hot encoded labels (to\_categorical) to handle the multi-class problem. The final model achieved a test accuracy of **74%**, the highest of all models tested. Crucially, it maintained near-perfect recall for the most severe classes while also improving performance on the majority classes, offering the best overall solution.

### Summary of Results and Conclusion

The comparative analysis clearly indicates that the **TensorFlow-based neural network** is the most effective model for this classification task. It successfully navigates the challenge of class imbalance, providing a superior balance of overall accuracy and high-impact predictions for severe accidents. The table below summarizes the key performance metrics for the top-performing models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Overall Accuracy** | **Mortel Recall** | **F1-Score (Macro Avg)** | **Key Insight** |
| **Logistic Regression** | 73% | 1.00 | 0.86 | Strong on linear data. |
| **Random Forest (Balanced)** | 68% | 1.00 | 0.83 | Excellent for fatal accidents, but lower overall accuracy. |
| **Soft-Voting Ensemble** | 71% | 1.00 | 0.84 | Solid, balanced performance. |
| **TensorFlow Neural Network** | **74%** | **1.00** | **0.86** | Highest overall accuracy with robust minority class prediction. |

In conclusion, this project has successfully developed a powerful and accurate predictive model for road accident severity. The final neural network model can serve as a vital tool for data-driven decision-making in urban planning and road safety initiatives, allowing authorities to better allocate resources and target high-risk areas to prevent future tragedies.

### Future Work and Recommendations

To further enhance this project, the following areas are recommended for future exploration:

* **Data Augmentation Techniques**: Investigate using the Synthetic Minority Oversampling Technique (SMOTE) to synthetically balance the dataset. SMOTE works by creating new synthetic examples of the minority class, which could potentially improve model performance without the negative impact on overall accuracy seen with class\_weight.
* **Advanced Hyperparameter Tuning**: Utilize more advanced optimization techniques, such as Bayesian optimization, to conduct a more extensive and efficient search for the optimal neural network architecture, including the number of layers, neurons per layer, and dropout rates.
* **Model Deployment**: Develop a production-ready API for the final model, allowing it to be integrated into real-time systems for predicting accident severity on the fly. This would enable city planners to monitor conditions and identify patterns that contribute to severe accidents as they occur.