Summary Paper: Predictive Modeling of Road Accident Severity

**Machine Learning Course - Group 12**

Rafick Jungul

1. **Introduction and Objectives**

This project aimed to develop a predictive machine learning model to classify the severity of road accidents in Montreal using a dataset of over 200,000 records. The primary goal was to build a robust system that could categorize accidents into five distinct severity levels, ranging from minor material damage to fatal. A central challenge was the significant class imbalance in the dataset, where severe and fatal accidents were rare. The project's success was therefore measured not just by overall accuracy, but by the model's ability to effectively identify these high-impact, low-frequency events. The ultimate purpose of this model is to provide a tool for proactive urban planning and emergency response.

1. **Methodology and Analysis**

The project followed a standard machine learning pipeline, starting with a dataset from the Government of Quebec website.

**Data Preprocessing**

The raw data underwent extensive cleaning, including handling missing values, standardizing categorical features, and converting them using One-Hot Encoding. To manage the high number of features, Principal Component Analysis (PCA) was applied, retaining 95% of the data's variance to improve model efficiency and prevent overfitting. The dataset was then split into an 80% training set and a 20% test set.

**Exploratory Data Analysis (EDA)**

Key observations from the EDA included:

* **Geospatial patterns** revealed that accidents were concentrated in the urban core and along major roads, with a disproportionately high number of pedestrian and cyclist victims in these areas.
* **Temporal trends** showed higher accident volumes during weekday rush hours and a seasonal peak during winter months.
* A **Chi-square test** confirmed that environmental factors like weather, lighting, and road conditions were statistically significant predictors of accident severity.

1. **Model Training and Evaluation**

A variety of supervised learning models were trained and evaluated to find the best approach for the imbalanced dataset. The models were judged on their overall accuracy and, more importantly, their ability to correctly predict the minority classes.

* Random Forest Classifier: A baseline Random Forest model achieved 71% accuracy, showing strong performance on severe accidents but struggling with the majority classes. A hyperparameter search with GridSearchCV confirmed the baseline's performance but did not significantly improve it.
* Logistic Regression: This model performed slightly better, with an accuracy of 73%, suggesting a potential linear relationship between the principal components and the target variable.
* Addressing Imbalance: To combat the class imbalance, a Random Forest model was retrained with class\_weight='balanced'. This drastically improved the recall for the Mortel class to 1.00, correctly identifying every fatal accident, though it came at a small cost to overall accuracy (68%).
* Ensemble and Neural Networks: The project culminated in training more complex models:
  + A Soft-Voting Ensemble of Random Forest, Gradient Boosting, and Logistic Regression achieved 71% accuracy.
  + A scikit-learn MLP Classifier also achieved 71% accuracy.
  + The best performance was achieved by a TensorFlow-based Neural Network. This model yielded a test accuracy of 74%. Crucially, it maintained perfect recall for the Mortel class while also improving performance on the majority classes, proving to be the most robust and well-rounded solution.

1. **Conclusion and Future Work**

The project successfully developed a powerful and accurate predictive model for road accident severity. The final TensorFlow model demonstrates that it is possible to create a reliable system that can identify the factors leading to severe accidents, which can be invaluable for urban planning and public safety.

Future recommendations include exploring advanced techniques like Synthetic Minority Oversampling Technique (SMOTE) for data augmentation, further hyperparameter tuning on the neural network, and developing a production-ready API for real-time risk assessment.