Predicting Pedestrian Injury Severity using Machine Learning: A Case Study of the City of Toronto

Ian Christopher Roman (30049116)

April 2024

Abstract

Pedestrian safety in urban areas is a critical concern, prompting the need for advanced predictive methods to mitigate road accidents. This term paper uses machine learning (ML) techniques to predict the severity of pedestrian injuries from road collisions using a detailed dataset provided by the Toronto Police Service. The study rigorously evaluates several ML models, including K Nearest Neighbors (KNN), Random Forest (RF), Extreme Gradient Boosting (XGB), and Artificial Neural Networks (ANN), across both binary and multiclass classification frameworks. The analysis focuses on the model's ability to accurately predict injury outcomes, revealing that the ANN model, particularly in binary classification mode, significantly outperforms others by effectively distinguishing between *fatal* and *non-fatal* injuries. This superior performance is attributed to the model's sophisticated handling of complex data interactions, enhanced by dropout regularization techniques and multiple hidden layers. Furthermore, the research identifies critical variables influencing predictions, such as environmental conditions and pedestrian behaviors, which are crucial for understanding the dynamics of road accidents. These findings underscore the potential of advanced ML models in improving pedestrian safety by enabling more accurate predictions of injury severities, thus contributing value insights to urban safety planning and intervention strategies.

1 Introduction

Pedestrians are vulnerable road users (motorists, cyclists, micro-mobility), meaning they do not have a protective shell to reduce the severity of their injuries in a collision. According to the Canadian Association of Road Safety Professionals (CARSP, 2021), 32 percent of *collision fatalities* were experienced by vulnerable road users, with pedestrians (16%) having the most records. From 2018 to 2020, Canada recorded an average of about 300 pedestrian fatalities, with the primary crash location at an intersection or midblock (Statistics Canada, 2023). These statistical measures prompt better road interventions and policies to mitigate pedestrian fatalities further and achieve *Vision Zero's* goals, especially for the vulnerable population. Crash information data are collected after any collision occurs. Such information records the severity of the vehicle crash and captures the injury severity of any pedestrian involved. Using classical statistical methods (CSM), crash data provide insights into crash frequency and hotspots within a specified area. Although CSMs provide a generalized insight into crash frequency, machine learning (ML)/deep learning models outperform CSMs, especially in interpreting complex data patterns and relationships when CSMs assume *linear* relationships in modeling crash data (Zhang et al., 2019; Wahab & Jiang, 2019). This study aims to utilize machine learning models to predict pedestrian injury severity and determine the most accurate model and its architecture. This paper is divided into three (3) sections: **Section 1** provides background to the study, reviews relevant literature on contributing factors to varying pedestrian injury severity and the use of machine learning methods in pedestrian crash predictions, and presents the gap in the literature that this study aims to bridge. **Section 2** defines machine learning models, the models used in this study, and the performance metrics to assess each model. **Section 3** summarizes the methodology used in this study, from data augmentation to modeling and results. **Section 4** outlines the result of the data analysis and ML modeling. Finally, **Section 5** provides the conclusion to this study, including a summary of the study, limitations, and future works.

1.1 Literature Review

1.1.1 Existing Practice in Pedestrian Crash Modeling and Contributing Factors

Research in pedestrian crash modeling has significantly improved the understanding of factors contributing to injury severity. It employs statistical methodologies to analyze the correlation between urban design, traffic management, and pedestrian and driver behavior. Studies such as Forbes and Habib (2015) and Shrinivas et al. (2023) have highlighted the critical role of localized conditions, including geographic, infrastructure, and socio-economic factors, in pedestrian safety outcomes. Through a combination of spatial and classical statistical models, these studies highlight the importance of urban design and planning interventions to mitigate pedestrian injuries in specific urban contexts. The comprehensive insights from these studies highlight a complex strategy towards improving pedestrian safety, incorporating knowledge of how the environmental and demographic factors, the relationship between urban design and injury severity, and the impact of community characteristics

all play a role in determining pedestrian injury level. Studies such as Tiwari (2020) and Toran Pour et al. (2017) suggest the growing emphasis on comprehensive models that account for the complex interactions between pedestrians and their urban surroundings. Contributing factors that affect pedestrian injury severity are crucial to pedestrian crash modeling. Forbes and Habib (2015) identified *speed limits* and *road visibility* as a critical factor in modeling pedestrian crashes in the Regional Municipality of Halifax, while Grisé et al. (2018) focused on the relationship between *spatial properties* and injuries among *children and elderly* in the City of Toronto. Jang et al. (2013) found that *environmental (e.g., temperature, visibility)* and *temporal (e.g., hour in the day)* risk factors affect pedestrian crash hot spots. Prato et al. (2018) examined the effects of *built environment (e.g., land use)* and its spatial aspects on pedestrian injury severity. Lastly, Toran Pour et al. (2017) explored the effect of incorporating *neighborhood characteristics* on crash severity. These studies show the depth and breadth of various factors, from urban infrastructure to demographics, as indicators of pedestrian injury levels.

1.1.2 Pedestrian Crash Prediction using Machine Learning

In recent years, machine learning and artificial intelligence have increasingly been applied to various transportation and traffic applications, such as forecasting travel demand and traffic flow and crash prediction. By analyzing extensive datasets that capture a wide range of factors, these sophisticated algorithms can reveal underlying insights and risk factors that must be apparent when employing classical statistical methodologies. Existing studies about using ML techniques in predicting pedestrian injury severity are summarized in **Table 1** below.

Table 1: Summary of Prediction of Pedestrian Injury Severity

Study	**Approach Summary**	**Most Accurate Model**
Zhang et al. (2018)	Utilized both statistical and machine learning models in performing a comparative study to predict *crash injury severity on freeway diverge areas in Florida*. Ordered probit (OP) and multinomial logit (MNL) models were the statistical models compared to K Nearest Neighbor (KNN) and Random Forest (RF) models.	Random Forest (RF)
Chakraborty et al. (2019)	Employed an Artificial Neural Network (ANN) to predict *pedestrian crash fatalities* in Kolkata, India, and *applied it to urban intersections*, incorporating attributes such as average daily traffic volume, pedestrian-vehicle volume ratio, and approach speed.	Bayesian Regularization Neural Network with a hyperbolic tangent-sigmoid activation function and 13 nodes in the hidden layer

Continued on next page

Table 1: Summary of Prediction of Pedestrian Injury Severity (Continued)

Study	**Approach Summary**	**Most Accurate Model**
Das et al. (2020)	Utilized Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB) machine learning models to *classify pedestrian crash types based on unstructured text data (crash narrative reports)* in select cities in Texas.	Extreme Gradient Boosting (XGB)
Guo et al. (2021)	Utilized Extreme Gradient Boosting (XGB) to classify the *severity of traffic crashes with older pedestrians in Colorado*. The most essential features are driver and pedestrian characteristics, vehicle movement, and environmental conditions.	Extreme Gradient Boosting (XGB)
Komol et al. (2021)	Focused on the vulnerable road users (VRUs): pedestrians, cyclists, and motorists and utilized KNN, SVM, and RF models to *predict the injury severity of each VRU and combined VRUs* in Queensland, Australia. Key influencing factors are time of day, speed limit, and age group.	Random Forest (RF)
Meocci et al. (2021)	Created an *accident prediction model to enhance pedestrian safety* in Italy using the CatBoost (CB) algorithm. The number of lanes, traffic conditions, and presence of traffic lights are critical in determining a road section's risk level.	CatBoost (CB)
Hossain et al. (2021)	Utilized the chi-square technique for feature selection and employed RF, SVM, and Decision Tree (DT) models to *predict and analyze pedestrian injuries from road accidents in Great Britain from 2016 to 2018*.	Random Forest (RF) and Decision Tree (DT)

Continued on next page

Table 1: Summary of Prediction of Pedestrian Injury Severity (Continued)

Study	**Approach Summary**	**Most Accurate Model**
Tao et al. (2022)	Employed a Bayesian Neural Network (BNN) to *predict pedestrian fatalities from road crashes.* The BNN model was compared to ANN, RF, and standard Bayesian networks, suggesting that BNN is the most accurate model. This study reveals that individual characteristics, time, circumstances, and road and crash attributes are significant factors in predicting pedestrian fatalities.	Bayesian Neural Network (BNN)

1.1.3 Research Gap

Using ML/AI to predict pedestrian injury severity is new and has been a topic of interest in recent years. While previous studies have focused on utilizing machine learning for injury severity prediction of VRUs (Komol et al., 2021), there is a need for a more comprehensive approach to understanding the myriad of factors influencing pedestrian injury severity. Some studies above do not tackle predicting the pedestrian injury level. The works of Chakraborty et al. (2019) and Tao et al. (2022) only focused on predicting whether there is a pedestrian crash and if the road fatality involved is a pedestrian or non-pedestrian, respectively. Other studies also highlighted the purpose of their prediction models, such as whether the crash involved a pedestrian intended or unintendedly caused (Das et al., 2020) and classified whether pedestrians are at high, medium, or low risk depending on location (Meocci et al., 2021). While these studies were helpful, the models developed in these papers **do not** predict pedestrian injury levels, which is the objective of this study. The works of Zhang et al. (2018) and Rao et al. (2022) have predicted pedestrian injury levels. However, these studies are limited to considering ML algorithms (e.g., nearest neighbors, random forest), and explanatory variables are too small and often focused on one (1) type of explanatory variables, such as road and traffic characteristics, and not incorporating crucial variables such as demographics and the built environment. This study will bridge the gap in knowledge by incorporating various explanatory variables and creating not just ML models but also neural network models to predict pedestrian injury levels.

2 Machine Learning Models

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that uses data and algorithms to mimic how humans make inferences and decisions and gradually improve their accuracy. ML is divided into three categories: supervised learning (classification and regression), unsupervised learning (clustering), and reinforcement learning. The focus of this study will be supervised learning, specifically classification. This study will employ four (4) ML models: K Nearest Neighbors (KNN),

Random Forest (RF), Extreme Gradient Boosting (XGB), and Artificial Neural Networks (ANN). These models will aim to predict the pedestrian injury level of a given dataset. This section will introduce each machine learning algorithm's intuition and modeling performance metrics in classification problems.

2.1 K Nearest Neighbors (KNN)

K Nearest Neighbors (KNN), the nearest neighbor classifier, belongs to the family of instance-based or lazy learning algorithms where the function is only locally approximated. The KNN algorithm finds the k closest training examples in the feature space to a given test point and makes predictions based on these k nearest neighbors. In classification, KNN assigns a class to the test point based on the majority class among its k nearest neighbors. If $k = 1$, the test point is simply assigned the class of its nearest neighbor. For larger values of $\boldsymbol{k}$, a majority vote mechanism is typically used (Tan et al., 2014). **Figure 1** illustrates the KNN classifier with a large $\boldsymbol{k}$ value.

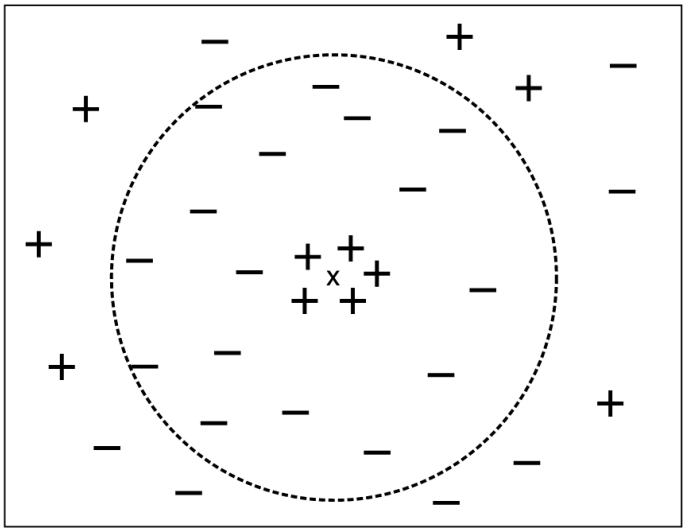

Figure 1: k nearest neighbor classification with large k (Tan et al., 2014)

2.2 Random Forest (RF)

Random Forest (RF) is a class of ensemble methods designed explicitly for decision tree classifiers. It combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors, as shown in **Figure 2** below. The random vectors are generated from a fixed probability distribution. Due to its nature, RF can deal with massive datasets with many features, making the model highly scalable. It can also determine the importance level of each feature for classification tasks, making it helpful in identifying significant variables (Tan et al., 2014).

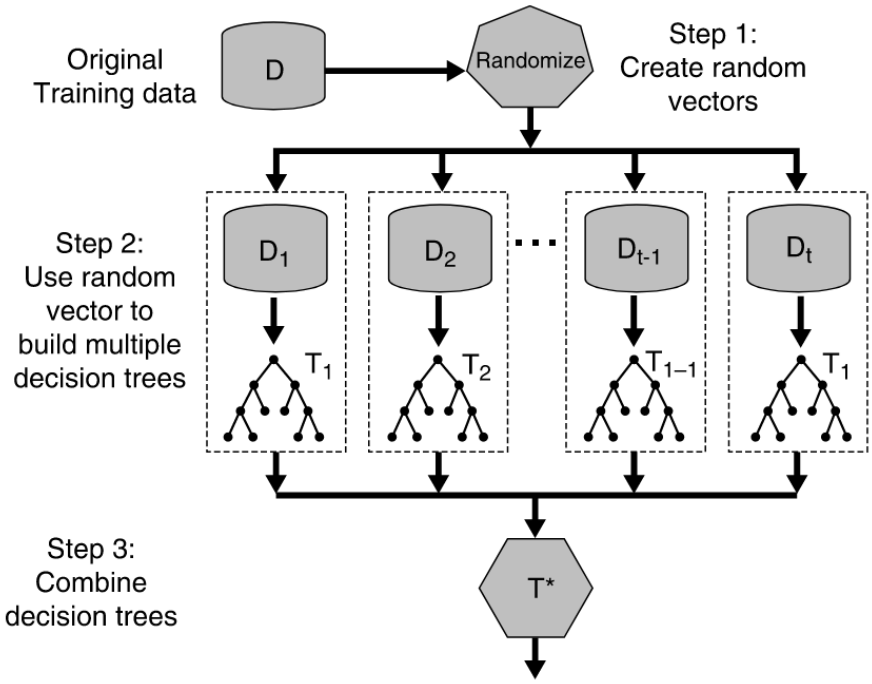

Figure 2: Random Forest Intuition (Tan et al., 2014)

2.3 Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting or XGBoost (XGB) is a decision tree-based ensemble algorithm that leverages the concepts of gradient boosting to create a more robust model. The algorithm works by creating a set of decision trees iteratively, with each tree attempting to correct the mistakes of the previous tree, as illustrated in **Figure 3** below. XGB employs a *gradient descent* algorithm to minimize the sum of errors of each tree in the ensemble. The final output of the model is a weighted combination of all decision trees, with each tree assigned a weight based on its contribution to the cost function. XGB is known for its high accuracy, as the model can capture complex relationships between variables, speed, and efficiency. XGB also incorporates regularization techniques to prevent overfitting.

2.4 Artificial Neural Network (ANN)

Biological neural systems inspired the Artificial Neural Network (ANN) model. An ANN is composed of an interconnected assembly of nodes and directed links. **Figure 4** below illustrates a simple perceptron model. In this figure, the perceptron comprises three (3) input nodes and one (1) output node. The perceptron computes its output value, y, by performing a weighted sum on its inputs, adding a bias factor, t, to the sum, and then examining the calculated output. Training the simple perceptron model requires constantly updating the weight parameters (w) until the perceptron outputs become consistent with the actual outputs of the training sample (Tan et al., 2014). The equation gives the weight update equation (1) below:

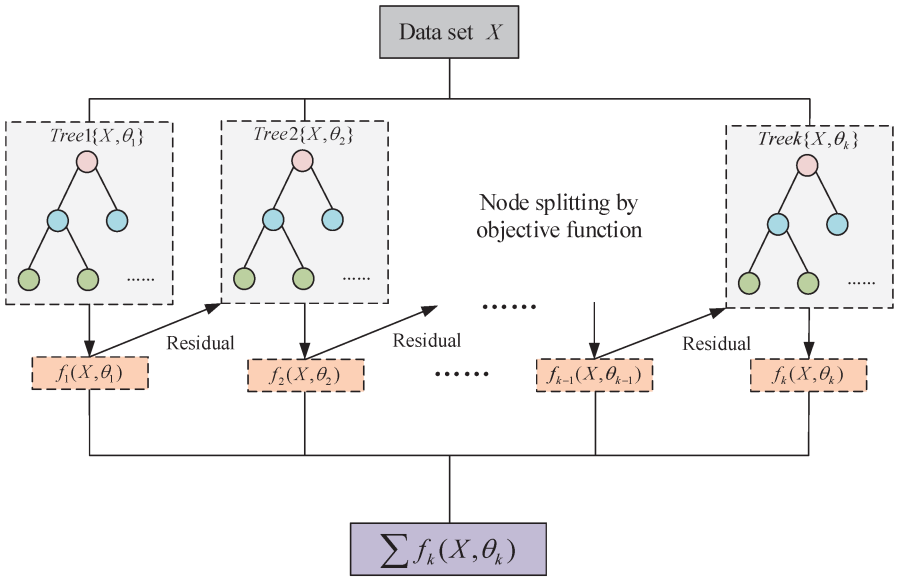

Figure 3: XGBoost Flowchart (Guo et al., 2020)

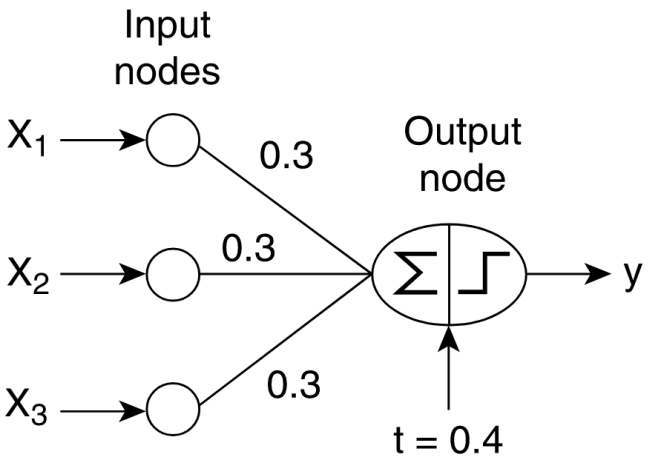

Figure 4: Modeling with a Simple Perceptron (Tan et al., 2014)

$$w_j^{(k+1)} = w_j^k + \lambda(y_i - y_i^k)x_{ij} \tag{1}$$

The ANN model can undoubtedly be more complex than the simple perceptron model. The network may contain several intermediate layers between input and output, called hidden layers. The resulting structure is known as a **multilayer neural network** or, most times, a **deep neural network** model. In a feed-forward propagation network, the nodes in one layer are connected only to the nodes in the next layer, as shown in **Figure 5** below. The network also utilizes **activation functions** that calculate the node's output based on its inputs and their weights. Examples of the most commonly used activation

functions in classification exercises are **sigmoid** and **rectified linear units (ReLU)**.

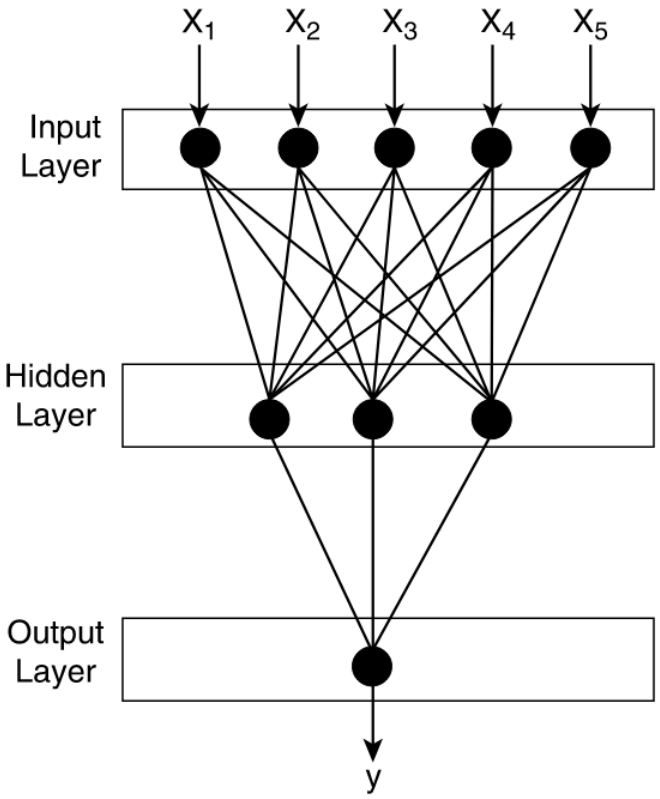

Figure 5: Example of a Multilayer Feed-Forward Artificial Neural Network (Tan et al., 2014)

2.5 Model Performance Metrics

Defining metrics to determine how well the model performs is common in any modeling exercise. In machine learning, specifically in classification problems, the most common indicator is *accuracy*, defined as the proportion of cases classified correctly (**Equation 2**). However, the model's accuracy is one of many indicators of success in training and testing ML models. A helpful visualization for binary classification problems is a confusion matrix showing the predicted and actual classification status. **Figure 6** illustrates the matrix structure where

- **True Positive** (TP) and **True Negative** (TN) are correctly classified as positive and negative, respectively.
- **False Positive** (Type I Error, FP) occurs when the predicted value is positive (1); however, the actual value is negative (0).
- **False Negative** (Type II Error, FN) is the opposite of False Positive and occurs when the predicted value is negative (0); however, the actual value is positive (1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

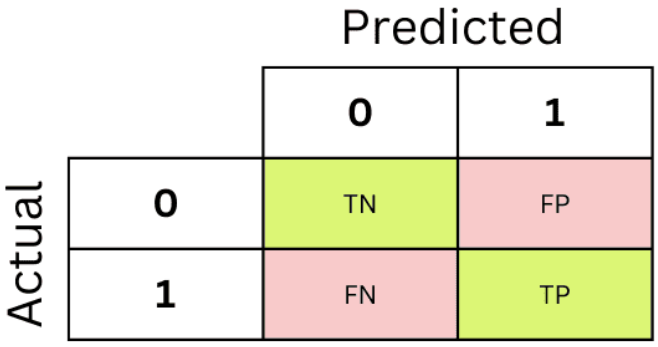

Figure 6: Confusion Matrix

The confusion matrix above can derive other performance metrics. **Precision** measures the proportion of **predicted** positive (1) that is **actually** positive (1) and can be calculated using **Equation 3** below.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Another metric derived from the confusion matrix was **Recall**, which measures the proportion of all positive (1) measures **correctly** classified as positive (1). *Recall* can be calculated using Equation 4 below.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

Lastly, another metric is the **F1 Score**, defined as the harmonic mean of the *Precision* and *Recall* scores and given by **Equation 5** below.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{5}$$

This study will perform binary and multiclass classification and use the above metrics to evaluate each ML model's performance. The accuracy score will be assessed for both the training and testing sets, while the rest of the metrics will only focus on the test set to determine how each model performs with unseen data.

3 Methodology

3.1 Dataset and Processing

This study utilized data from the *Toronto Police Service: Public Safety Data Portal*, where a wide array of crash data gathered by the Toronto Police are available on the website. The pedestrian dataset was used, a subset of the Killed and Seriously Injured (KSI) data containing all road collision records for different users. The raw dataset contains about 75 attributes, some of which are either not useful to the analysis, repeated attributes, or too many missing values, which can impact the effectiveness of any model created. Most of the raw dataset's attributes are categorical and need numerical variables, such as temperature, population, and traffic volume, to improve the generalization of each model in predicting. Supplementary temperature, traffic flow, road characteristics, and demographics data were extracted and joined to the pedestrian dataset. **Figure 7** illustrates the entire process, from data sources to processing to ML modeling.

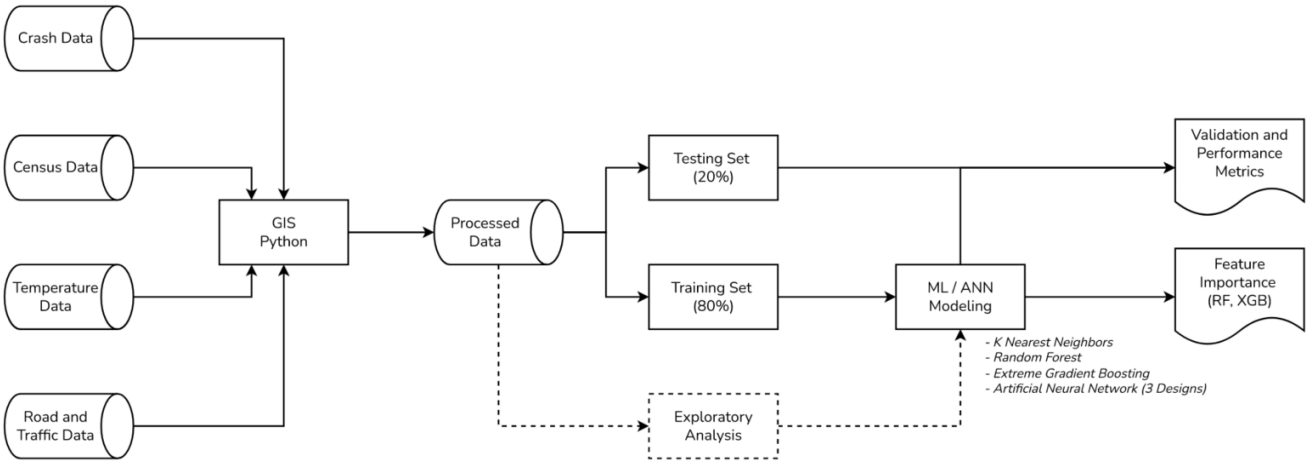

Figure 7: Predicting Pedestrian Injury Level Process

The complete dataset is processed to clean and organize the data in preparation for ML modeling. Techniques such as missing value imputation, unique value grouping, and dropping features are employed to complete the data processing phase. These techniques are summarized below:

- Preliminary processing was conducted in GIS through the spatial join feature to join information to the pedestrian dataset, such as dissemination area unique ID (DAUID), land use, and vehicular and pedestrian flow.
- The raw dataset did not have a datetime attribute; however, it had date and time columns to yield a datetime column where information such as year, month, and day can be extracted. This process is also crucial because temporal information is necessary to join the weather (temperature, relative humidity, dew point temperature) dataset.

- The vehicle's speed **at the time** of the collision is an unknown variable; however, the speed limit is available based on the road classification. The speed limit will depend on which year an incident record occurs since amendments to the *City's Vision Zero Safety Plan* in 2019 lowered the speed limit to most of the City's road class.
- Most missing data are categorical variables, and "Unknown" is imputed for any missing data.
- Some values were grouped into common topics to reduce the unique values of each categorical attribute (e.g., for the variable LIGHT, Dark, and Dark, artificial are combined to just Dark).

3.2 Machine Learning Modeling Process

This study used the Python programming language to perform the machine learning modeling tasks. Packages such as *Scikit Learn* and *TensorFlow* were used for preprocessing and model training.

3.2.1 Preprocessing

Data preprocessing is crucial before training any machine learning model because this process improves model performance. Some techniques proven to improve model performance are converting the dataset into vector format (vectors/matrices) instead of the data frame format, one hot encoding categorical variables, and scaling numerical features. The dataset comprises 3113 records, and before any preprocessing operation, the dataset was vectorized. The vectorized dataset is then divided into train and test sets for training and validating the models. This study used an 80-20 split to create the train and test sets (80% training, 20% testing). Following the train-test split, the next preprocessing operation is scaling the numerical features, and this study utilized the StandardScaler function, which uses the z-score equation to scale the features. Finally, the categorical variables need to be converted to a format the computer can understand, such as numerical features. This process is called one-hot encoding. **Figure 8** below illustrates a simple example of one-hot encoding such that the 'Houston' record is encoded as 1 0 0 0 (vertical) where the placement of "1" indicates which value it is from the original dataset.

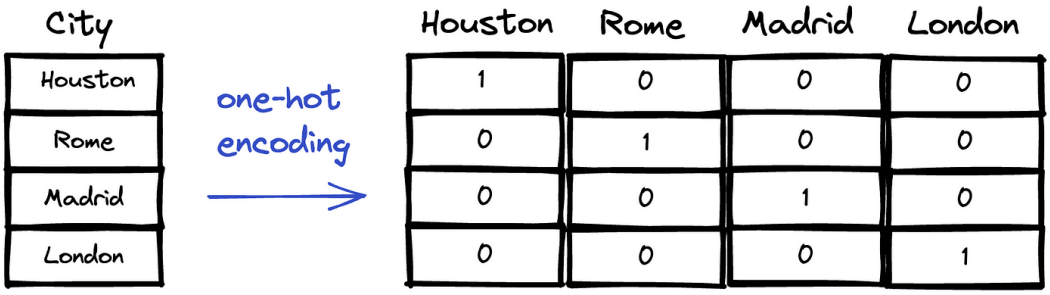

Figure 8: One Hot Encoding Example

3.2.2 Hyperparameter Tuning

Hyperparameter Tuning is a process that allows an ML model to manipulate specific parameters to experiment with and find the optimal solution. For the machine learning models (KNN, RF, XGB), the RandomizedSearchCV function was used to perform hyperparameter tuning, where parameter grids containing parameters of each model are summarized and randomly fitted until the "best parameters" are determined. The optimal model that results from this function yielded the highest accuracy. The hyperparameters for each model are summarized in **Tables 2** to **4** below, where default values are also included. For neural networks, the hyperparameters to be tuned are the *number of hidden layer(s)*, *the number of node(s) for each hidden layer(s)*, *the regularization factor*, and *the number of epochs*.

Table 2: KNN Hyperparameters

Parameter	**Default Value**
weights	'uniform'
n_neighbors	5
metric	'minkowski'

Table 3: RF Hyperparameters

Parameter	**Default Value**
n_estimators	100
min_samples_split	2
min_samples_leaf	1
max_features	'sqrt'
max_depth	None
criterion	'gini'
bootstrap	True

Table 4: XGB Hyperparameters

Parameter	Default Value
subsample	1
min_child_weight	1
max_depth	3
learning_rate	0.1
colsample_bytree	1

3.2.3 Modeling Scenarios

According to the dataset, there are five (5) levels of pedestrian injury: None, Minimal, Minor, Major, and Fatal. From Forbes and Habib (2015), the description of injury levels are the follows:

- **None:** No injuries sustained.
- **Minimal:** Injuries include minor abrasions, bruises, and complaints of pain; did not require medical assistance.
- **Minor:** The injuries required a trip to the hospital and treatment in the emergency room. The patient was not admitted to the hospital.
- **Major:** Injuries required that the person involved be admitted to hospital; includes persons admitted for observation.
- **Fatal:** Death occurred as a result of injuries from the collisions.

The dataset has an imbalanced proportion of the target variable (injury level), as shown in **Figure 9** below. About 77 percent of the injury level belongs to the **Major** category, meaning the injuries sustained by the pedestrian involved required emergency care and hospital admission. With this information, it is necessary to test different scenarios that can improve model performance, given the imbalanced dataset. There will be **five (5) scenarios**, including the base scenario, a multinomial logit model, and one of the many tools frequently used in crash modeling. Each scenario will be compared to determine the well-performing scenario and model, respectively. Scenario descriptions are summarized below.

- **Scenario B0:** Base Scenario. This scenario will utilize a *multinomial logistic regression model (multinomial logit)* to predict pedestrian injury levels.
- **Scenario M1A:** *Multiclass* Classification using *all* features. This scenario will utilize the conventional machine learning models (KNN, RF, XGB) and artificial neural networks (ANN) to predict pedestrian injury levels (None to Fatal) with *all* of the features from the dataset included.

- **Scenario M1B:** *Binary* Classification using *all* features. This scenario is similar to **M1A**, except the target variables are labeled such that we have a binary choice: *Fatal* (1) or *Non-Fatal* (None, Minimal, Minor, Major) (0).
- **Scenario M2A:** *Multiclass* Classification using *select* features. This scenario will utilize the same models as **M1A** and **M1B** but with chosen features according to Random Forest's Feature Importance. Instead of all the features, the **Top 10** features from RF's feature importance property will be used in modeling.
- **Scenario M2B:** *Binary* Classification using *select* features. This scenario is similar to **M2A** and **M1B**.

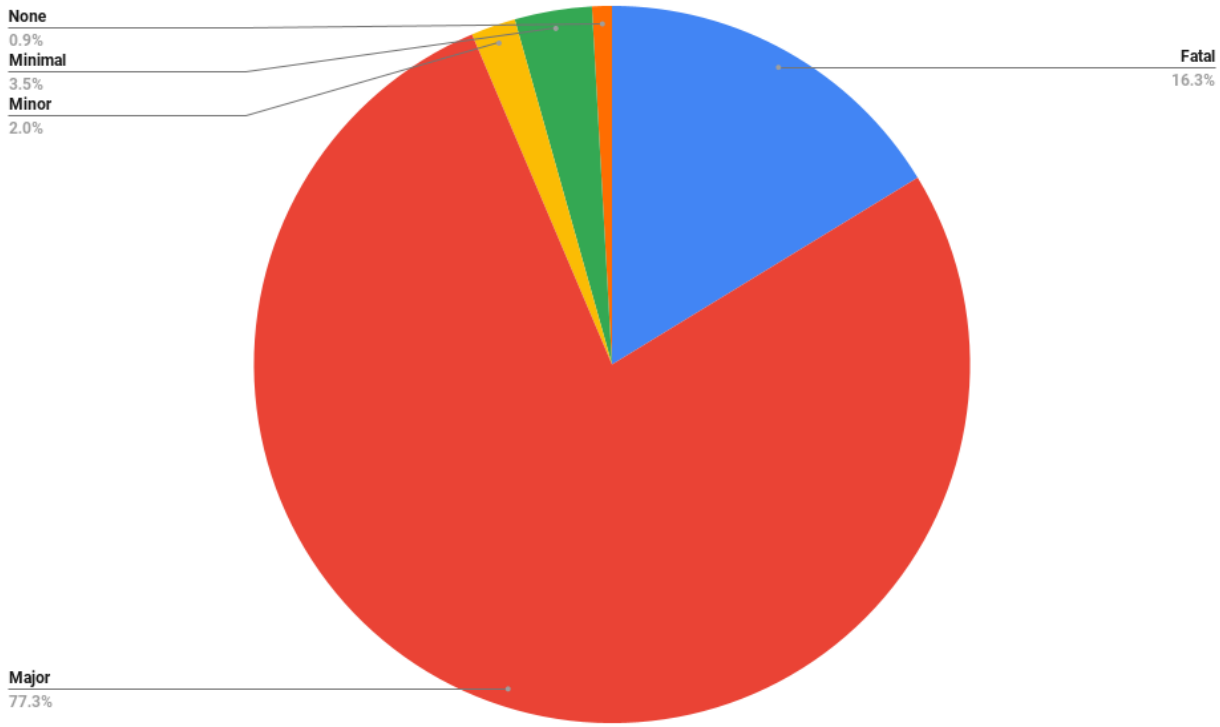

Figure 9: Target Variable (Injury Level) Proportions

4 Results and Discussions

4.1 Exploratory Analysis

Exploratory Data Analysis (EDA) is the process of analyzing, investigating, and summarizing datasets commonly through various data visualizations. EDA helps determine the best way to treat the data sources for analysis and modeling. As shown in the data mining process for this term paper, the EDA process is an intermediate step before proceeding to the next phase (ML Modeling) because it allows for another way to look at the behavior of each feature and its relationship to each other. This study's features were grouped into six (6) distinct groups: Temporal (T), Environment (E), Traffic and Road

Characteristics (TRC), Built Environment (BE), Demographics (D), Pedestrian Characteristics (PC), Crash Information (CI) and the Target Variable. The features are summarized in **Table 1**, where the description, unique values, and mean/mode are included.

Table 5: Summary of Crash Data Attributes

Variable	**Description**	**Values**	**Mean\Mode**
TEMPORAL (C)			
HOUR	Hour of crash	0 - 23	18
DAYOFWEEK	Specifies the day of the week the crash occurred	weekday, weekend	weekday
ENVIRONMENTAL (E)			
VISIBILITY	Atmospheric conditions allowing for road users to look ahead as observed on the crash	Clear, Drifting snow, (Fog, Mist, Smoke, Dust), Freezing rain, Rain, Snow, Strong wind, Other, Unknown	Clear
LIGHT	Lighting condition	Daylight, Dark, Dusk, Dawn, Other	Daylight
RDSFCOND	Road surface conditions at the time of the crash	Dry, Ice/Slush, Loose Sand or Gravel, Snow, Wet, Other, Unknown	Dry
TEMP	Temperature (in Celsius) at the time (hour) of the crash	-22.8 - 37.1	10.297
REL_HUMID	Relative humidity (in %) at the time (hour) of the crash	0.17 - 1.00	0.689
TRAFFIC AND ROAD CHARACTERISTICS (TRC)			
LOCCOORD	Location of the crash	Intersection, Midblock	Intersection
TRAFFCTL	Traffic control (if any) at the crash location	Traffic Signal including Transit, Pedestrian Crossover, No Control, Other Traffic Control	No Control

Continued on next page

Table 5: Summary of Crash Data Attributes (Continued)

Variable	**Description**	**Values**	**Mean\Mode**
ROADCLASS	City of Toronto Road Classification of the crash location	Major Arterial, Minor Arterial, Collector, Local, Expressway, Other	Major Arterial
SPEEDLMT	Speed limit based on the road classification at the crash location	40 - 110	49.232
VEH_ADT	Average vehicular flow near or at the crash location	0 - 1464	496.521
PED_ADT	Average pedestrian trafffic flow near or at the crash location	0 - 1032	74.953
BUILT ENVIRONMENT (BE)			
LAND_USE	Land use designation at the crash location	Commercial, Institutional, Mixed Use, Natural Areas, Other Open Space, Parks, Regeneration, Residential, Unknown	Residential
DEMOGRAPHICS (D)			
POP_2021	Population (Census, 2021) within the dissemination area	0 - 9625	1095.327
PRIV_DWELL	Number of private dwelling within the dissemination area	0 - 6612	569.261
LAND_AREA	Land area of the DA in square kilometers (sq. km.)	0 - 5892	510.780
PEDESTRIAN CHARACTERISTICS (PC)			
INVAGE	Pedestrian age group involved in the crash	Under 15, 15 to 29, 30 to 44, 45 to 64, Over 65, Unknown	45 to 64

Continued on next page

Table 5: Summary of Crash Data Attributes (Continued)

Variable	Description	Values	Mean\Mode
PEDCOND	Pedestrian condition before the crash	Distracted, Fatigue, Impaired - Alcohol (BAC=Normal), Impaired - Alcohol (BAC>0.08), Impaired - Drugs, Medical or Physical Disability, Normal, Other, Unknown	Normal
CRASH INFORMATION (CI)			
PEDACT	Pedestrian action(s) that caused the crash	Crossing with ROW, Crossing with ROW but no control, Crossing without ROW, Other, Unknown	Crossing with ROW
VEHINV	Type of vehicle involved in the crash	automobile, cyclist, emergency vehicle, motorcycle, transit vehicle, truck	automobile
VIOL	Violation type (if any) that lead to the crash	speeding, aggressive driving, alcohol-related, disability-related	speeding
TARGET VARIABLE			
INJURY	Pedestrian injury level	Fatal, Major, Minor, Minimal, None	Major

Since we have temporal data, it is also worthwhile to visualize the number of pedestrian collisions through the years. The dataset contains pedestrian crash records from 2006 to 2022, and the fluctuation of crash frequency is shown in **Figure 10**. Years 2006, 2013, and 2018 are the top three (3) years where pedestrian crashes are at their peak for the City of Toronto. According to the City's Vision Zero Portal, pedestrians are more frequently involved in collisions among road users. In 2016, the City's Vision Zero Plan was approved by the council and amended in 2019, where particular road classes and corridors have undergone speed limit reduction. More funding has been allocated to make roadways safe to prevent pedestrian crash fatalities and casualties[1]. As seen in **Figure 10**, there was a significant decrease in pedestrian crashes after 2018 and

[1] https://www.toronto.ca/services-payments/streets-parking-transportation/road-safety/vision-zero/vision-zero-plan-overview/

continuously going down to 2022.

Pedestrian crashes with major injuries account for about 77% of the crash data, as shown in **Figure 9**. Major injuries refer to injuries requiring that a person be admitted to the hospital, including those that need to be admitted for observations (Forbes & Habib, 2015). It is also worthwhile to visualize the distribution of crashes with respect to pedestrian age groups, which is illustrated in **Figure 11** below. It can be observed that major injuries are most common across age groups and that the age group with the most major and fatal injuries sustained is age 45 and above. Lastly, visualizing specific locations or communities where these crashes are more prevalent is also helpful. **Figure 12** shows the distribution of incidents per community. The Central Business District (e.g., *South Riverdale, Moss Park, Downtown Yonge East, Yonge-Bay Corridor, Kensington-Chinatown*), Some east communities (e.g., *Wexford/Maryvale, Clairlea-Birchmount, Milliken*), and West Humber-Clairville (proximity to YYZ) are the communities where more than 45 pedestrian crashes are recorded.

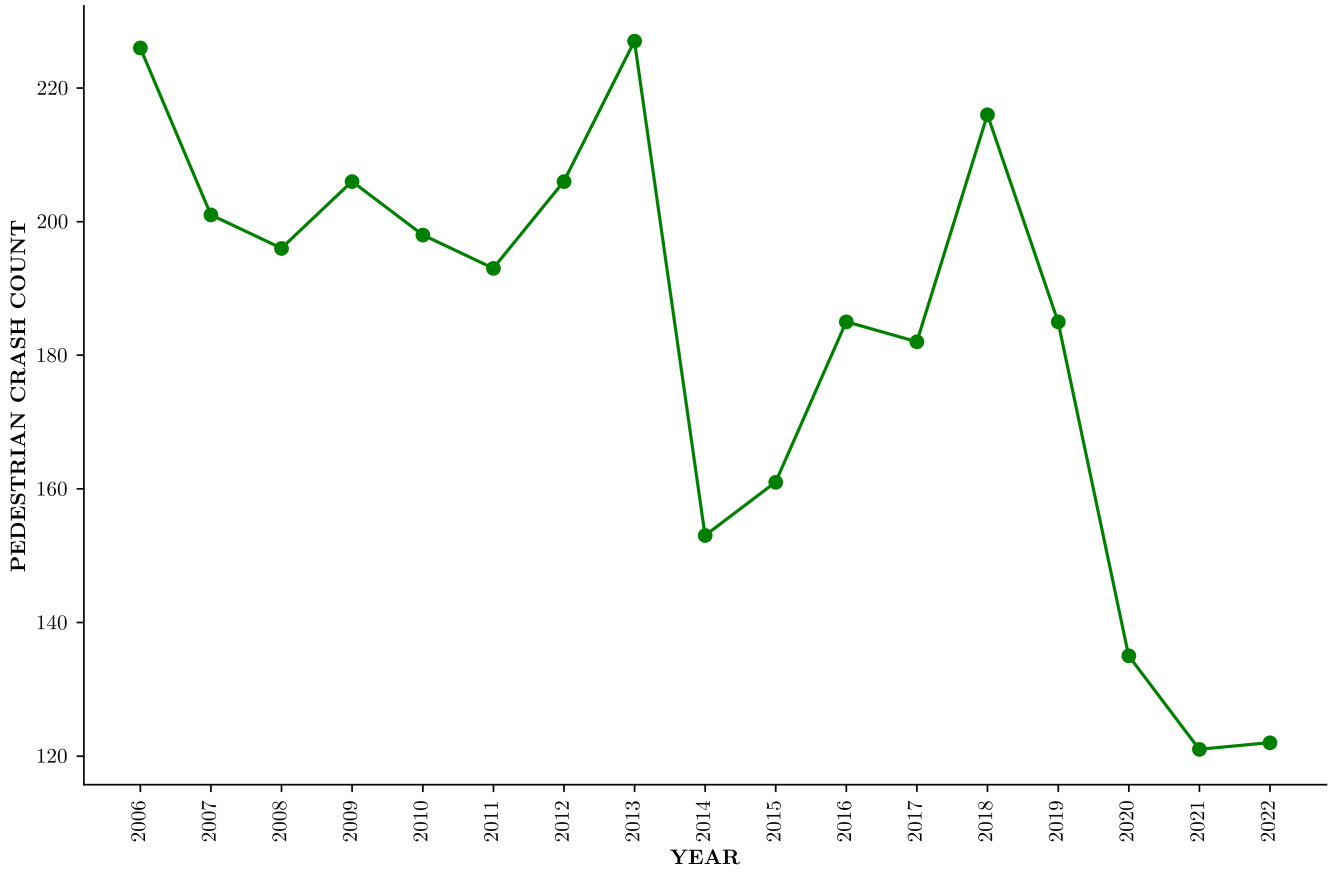

Figure 10: Annual Pedestrian Crash Numbers, City of Toronto

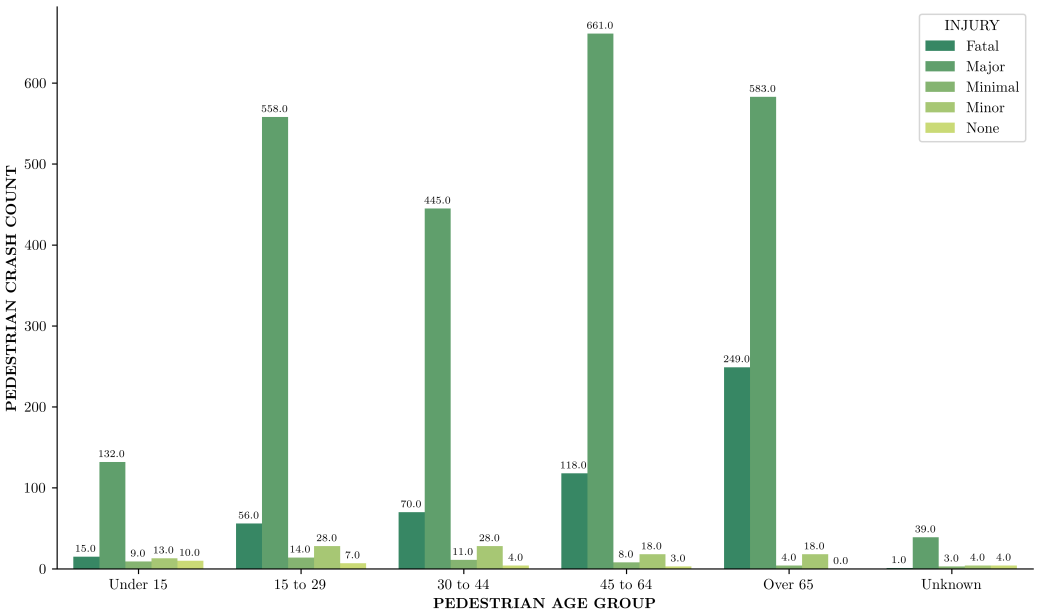

Figure 11: Pedestrian Age and Injury Distribution

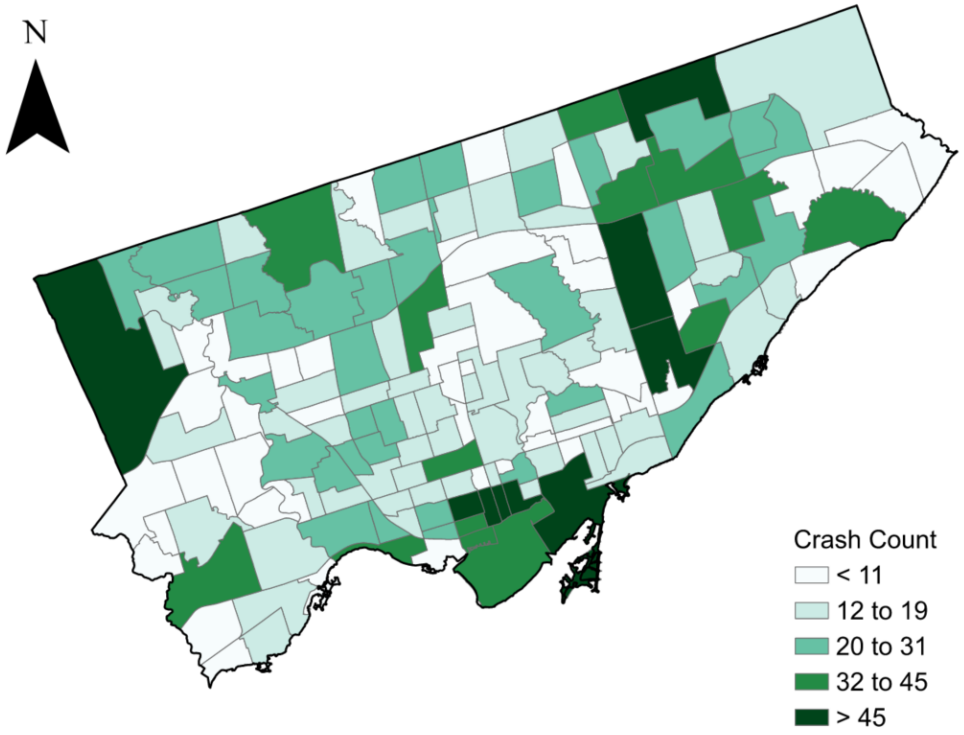

Figure 12: Crash Count per Neighborhood

4.2 Modeling Results

For each model (except neural networks), hyperparameter tuning was performed to determine the optimized solution to the modeling problem depending on the metrics assigned. For this study, the metric used for each model is accuracy. This study also utilized `RandomizedSearchCV` in the *Sci-Kit Learn* package of *Python*, where each model and model's parameter grid are iterated such that we "search" for the ideal solution based on the metric. Table 6 below shows a sample parameter grid (K Nearest Neighbors) used to find the optimal solution to the KNN model. The code snippet (**Figure 14**) shows how the `RandomizedSearchCV` was implemented where `cv` is the cross-validation strategy (i.e., five (5) folds), `n_iter` is the number of parameter settings sampled, `scoring` is the strategy to evaluate the performance of the cross-validated model and `random_state` is the assigned random seed for reproducibility. The best parameters for each model can also be extracted once fitted with the training dataset. The best parameters indicate the hyperparameter values that maximize the metrics provided in hyperparameter tuning.

Table 6: KNN Parameter Grid

Parameter	Set of Values
n_neighbors	1 to 36
weights	['uniform', 'distance']
metric	['euclidean', 'manhattan']

```
random_search = RandomizedSearchCV(estimator=model,
                                   param_distributions=grid,
                                   n_iter=100,
                                   cv=5,
                                   scoring='accuracy',
                                   random_state=100)
random_search.fit(X_train, y_train)
```

Figure 13: Hyperparameter Tuning using RandomizedSearchCV: Specifications

4.2.1 Model Performance

This section will summarize the model performance for each modeling scenario experimented with. **Table 7** summarizes the results of the base model (logit model), where the binary classification version is more accurate in predicting fatal or non-fatal pedestrian injury. However, another metric is included: the false negative rate (FNR), which calculates the rate at which fatal injuries are misclassified as non-fatal or other label categories given by **Equation 6**. Apart from accuracy, it is also essential to minimize the FNR value. The base model exhibits an 86% FNR, which is relatively high and means that 86% of the FN and TP values are misclassified.

$$\text{False Negative Rate (FNR)} = \frac{\sum FN}{\sum FN + TP} \tag{6}$$

Table 7: Base Model Performance

Model	TRA	TEA	PRE	REC	F1	FNR
Multiclass	78%	77%	67%	77%	71%	86%
Binary	84%	83%	78%	83%	80%	86%

TRA – Training Accuracy, **TEA** – Testing Accuracy, **PRE** – Precision, **REC** – Recall, **F1** – F1 Score, **FNR** – False Negative Rate

Now, comparing the Base Model to the rest of the scenario, **Tables 8** to **11** summarize the model performance of each model per scenario.

Table 8: **Scenario M1A** Model Performance

Model	TRA	TEA	PRE	REC	F1	FNR
KNN	78%	78%	68%	78%	70%	94%
RF	79%	79%	72%	79%	71%	92%
XGB	78%	78%	69%	78%	72%	85%
ANN D1	84%	78%	70%	78%	72%	81%
ANN D2	**88%**	**76%**	**70%**	**76%**	**73%**	**69%**
ANN D3	87%	78%	70%	78%	73%	77%

Table 9: **Scenario M2A** Model Performance

Model	TRA	TEA	PRE	REC	F1	FNR
KNN	78%	79%	70%	79%	71%	94%
RF	78%	78%	69%	78%	71%	94%
XGB	78%	78%	67%	78%	70%	95%

Continued on next page

Table 9: **Scenario M2A** Model Performance (Continued)

Model	TRA	TEA	PRE	REC	F1	FNR
ANN D1	**78%**	**78%**	**68%**	**78%**	**71%**	**88%**
ANN D2	78%	78%	68%	78%	71%	93%
ANN D3	78%	78%	64%	78%	69%	99%

Table 10: **Scenario M1B** Model Performance

Model	TRA	TEA	PRE	REC	F1	FNR
KNN	84%	85%	79%	85%	79%	15%
RF	84%	85%	83%	85%	80%	14%
XGB	84%	83%	78%	83%	79%	14%
ANN D1	87%	84%	79%	82%	80%	14%
ANN D2	**94%**	**82%**	**79%**	**82%**	**80%**	**13%**
ANN D3	92%	84%	80%	84%	80%	14%

Table 11: **Scenario M2B** Model Performance

Model	TRA	TEA	PRE	REC	F1	FNR
KNN	84%	85%	80%	85%	79%	15%
RF	84%	84%	78%	84%	79%	15%
XGB	**84%**	**84%**	**78%**	**84%**	**79%**	**14%**
ANN D1	84%	84%	75%	84%	78%	15%
ANN D2	83%	84%	72%	84%	78%	15%
ANN D3	84%	85%	77%	85%	78%	15%

As observed from the summary tables, **Scenarios M1A** and **M2A** (multiclassification) yielded some high-accuracy models; however, FNR ranges from 69% to 99%, which is not ideal for correctly identifying *fatal* pedestrian injury. **Scenario M1A**'s most accurate model is **ANN Design 2** (two (2) hidden layers and 50 nodes, each with a dropout rate of 40%), yielding a train and test accuracy of 88% and 76%, respectively, and an FNR value of 69% which is the lowest of all the models in the

scenario. On the other hand, **Scenario M2A**'s most accurate model is **ANN Design 1** (one (1) hidden layer and 50 nodes, with a dropout rate of 40%), yielding a train and test accuracy of 78% and an FNR value of 88% which is relatively high. In contrast with the first two (2) scenarios, the binary classification scenarios (**Scenarios M1B** and **M2B**) have significantly reduced the FNR values when only predicting *fatal* and *non-fatal* (grouped none, minor, minimal, and major) labels while maintaining similar accuracy scores. In **Scenario M1B**, the most accurate model is ANN Design 2 (two (2) hidden layers and eight (8) nodes, with a dropout rate of 10%), yielding a train and test accuracy of 94% and 82%, respectively and an FNR value of 13% which is the lowest among scenarios. In **Scenario M2B**, the most accurate model is **XGB**, with a train and test accuracy of 84% and yielded an FNR value of 14%.

Lastly, feature importance was also determined using the built-in function for *Random Forest* and *XGBoost* to unveil which variable greatly influences predicting pedestrian injury severity. **Figures 15** and **16** below showcase the feature importance score of each attribute in the dataset. For the Random Forest (RF) model (**Figure 15**), it can be perceived that `TEMP` (temperature in Celsius) influences the most in the prediction process, followed by a combination of pedestrian characteristics and demographics attributes such as `PEDCOND` (pedestrian condition before crossing) and `PED_ADT` (pedestrian flow). For XGBoost (**Figure 16**), `PEDCOND` (pedestrian condition before crossing) is the main factor influencing the predictions. Other features that may affect the prediction under the XGBoost model are `VEHINV` (type of vehicle involved in crashes like cars, transit bus, emergency vehicles, etc.) and `INVAGE` (age group of a pedestrian involved in the crash).

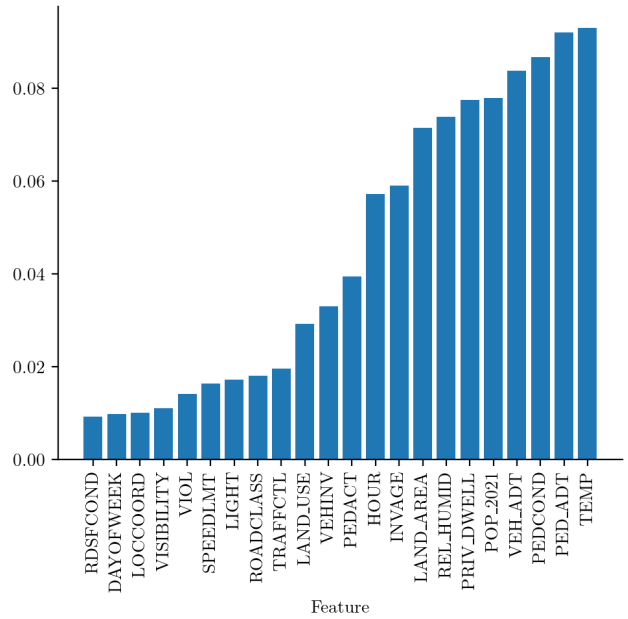

Figure 14: Random Forest Feature Importance

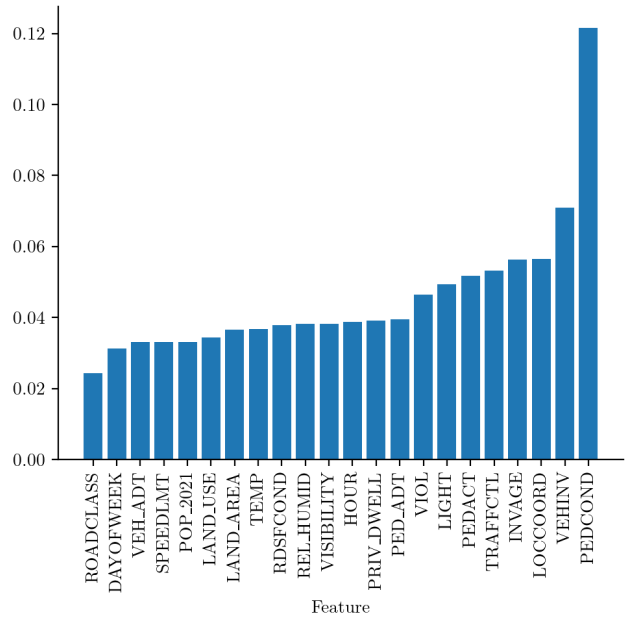

Figure 15: XGBoost Feature Importance

5 Conclusions

This term paper utilized a comparative approach to evaluate different Machine Learning (ML) techniques and their efficacy in predicting pedestrian injury levels. The crash dataset obtained from the Toronto Police Service Open Data portal was merged with the publicly accessible dataset (temperature, demographics, built environment) to form the overall dataset used in ML modeling. It was found that the target variable labels are imbalanced; therefore, multiple scenarios were created to test out which configurations yielded the highest accuracy and the lowest false negative rate (FNR). The ML models used in this study were K Nearest Neighbors (KNN), Random Forest (RF), Extreme Gradient Boosting (XGB), and Artificial Neural Networks (ANN). A logistic regression (logit) model was also used as the base model for all scenario models. Comparing all of the scenarios, it was found that the model with the highest accuracy score and low FNR value is the ANN Design 2 model. ANN Design 2 is a binary classification model (predicting whether pedestrian injury is fatal or non-fatal) and is composed of two (2) hidden layers with 50 nodes in each layer and a dropout rate of 40% for each hidden layer. This model also contains L1 regularization at the output node to reduce model overfitting. The model achieved 94% training accuracy and 82% test accuracy with an FNR value of 13%. Important features were also determined using the built-in function from the RF and XGB modeling packages. Using RF, it was found that the most important feature influencing the prediction process was `TEMP`, which is the temperature at the time of the incident. Finally, using XGB, it was found that the most important feature influencing the prediction process was `PEDCOND`, which is the pedestrian condition right before the incident.

This study has limitations that are worth incorporating for future studies. The dataset's imbalanced proportion of target labels was the main issue during the modeling and prediction phases. This study attempted some remedies, such as oversampling the minority class and undersampling the majority class, which yielded very low accuracy scores. Another remedy used in this study was employing ensemble learning methods such as the RF and XGB models, which consider the imbalanced nature of the target labels. Exploring further treatment for this issue is suggested using performance metrics such as Area Under the Curve (AUC) Receiver Operating Characteristics (ROC), and Matthews Correlation Coefficient and implementing a more advanced algorithm that considers the imbalanced target label proportions. Another limitation of this study was that only conventional ANN architecture was used. A more sophisticated ANN architecture should be considered in future studies, such as changing the activation function to capture the complexity of the dataset in predicting the target label. Activation functions such as hyperbolic tangent should be considered to increase the complexity of the conventional ANN. Bayesian Neural Networks are another type of neural network that incorporates Bayesian inference techniques in neural networks, and this type is also known for its overfitting reduction capabilities. The dataset used in this study should be modeled using this type of neural network algorithm to observe any improvement in predicting with an imbalanced dataset.

References

[1] Chakraborty, A., Mukherjee, & D., Mitra, S. (2019). *Development of pedestrian crash prediction model for a developing country using artificial neural network*. International Journal of Injury Control and Safety Promotion, 26(3), 283–293.

[2] Das, S., Le, M., & Dai, B. (2020). *Application of Machine Learning Tools in Classifying Pedestrian Crash Types: A Case Study*. Transportation Safety and Environment, 2(2), 106–119.

[3] Eduful, J. (2023, October 30). Pedestrian Fatalities in Canada, 2018 to 2020. https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2023059-eng.htm

[4] Forbes, J. J.,& Habib, M. A. (2015). *Pedestrian Injury Severity Levels in the Halifax Regional Municipality, Nova Scotia, Canada: Hierarchical Ordered Probit Modeling Approach*. Transportation Research Record, 2519(1), 172-178.

[5] Guo, R., Zhao, Z., Wang, T., Liu, G., Zhao, J., & Gao, D. (2020). *Degradation State Recognition of Piston Pump Based on ICEEMDAN and XGBoost*. Applied Sciences, 10(18), 6593.

[6] Guo, M., Yuan Z., Janson, B., Peng, Y., Yang, Y., & Wang, W. (2021). *Older Pedestrian Traffic Crashes Severity Analysis Based on an Emerging Machine Learning XGBoost*. Sustainability, 13(2), 926.

[7] Hossain, A., Sun, X., Shahrier, M., Islam, S., & Alam, S. (2023). *Exploring nighttime pedestrian crash patterns at intersection and segments: Findings from the machine learning algorithm*. Journal of Safety Research, 87, 382-394.

[8] Jang, K., Park, S. H., Kang, S., Song, K. H., Kang, S., & Chung, S. (2013). *Evaluation of Pedestrian Safety: Pedestrian Crash Hot Spots and Risk Factors for Injury Severity*. Transportation Research Record, 2393(1), 104-116.

[9] Komol, M., Hasan, M., Elhenawy, M., Yasmin, S., & Masoud M. (2021). *Crash severity analysis of vulnerable road users using machine learning*. PLOS ONE 16(8): e0255828.

[10] Meocci, M., Branzi, V., Martini, G., Arrighi, R., & Petrizzo I. (2021) *A Predictive Pedestrian Crash Model Based on Artificial Intelligence Techniques*. Applied Sciences, 11(23), 11364

[11] Mohamed, M., Saunier, N., Miranda-Moreno, L., & Ukkusiri, S. (2013). *A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada*. Safety Science. 54, 27-37.

[12] Olszewski, P., Osi nska, B., & Zieli nska, A. (2016). *Pedestrian Safety at Traffic Signals in Warsaw*. Transportation Research Procedia. 14, 1174-1182.

[13] Prato, C. G., Kaplan, S., Patrier, A., & Rasmussen, T. K. (2018). *Considering built environment and spatial correlation in modeling pedestrian injury severity*. Traffic injury prevention, 19(1), 88–93.

[14] Rao, A., Sarkar, S., Pramanik, A., & Maiti, J. (2022). *Predicting and Analysing Pedestrian Injury Severity: A Machine Learning-Based Approach. In: Giri, D., Raymond Choo, KK., Ponnusamy, S., Meng, W., Akleylek, S., Prasad Maity, S. (eds) Proceedings of the Seventh International Conference on Mathematics and Computing.* Advances in Intelligent Systems and Computing, vol 1412. Springer, Singapore.

[15] Shrinivas, V., Bastien, C., Davies, H., Daneshkhah, A., & Hardwicke, J. (2023). *Parameters influencing pedestrian injury and severity – A systematic review and meta-analysis.* Transportation Engineering, 11, 100158.

[16] Stipancic, J., Miranda-Moreno, L., Strauss, J., Labbe, & A. (2020). *Pedestrian safety at signalized intersections: Modelling spatial effects of exposure, geometry and signalization on a large urban network.* Accident Analysis & Prevention, 134, 105265.

[17] Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining.* Pearson.

[18] Tao, W., Aghaabbasi, M., Ali, M., Almaliki, AH., Zainol, R., Almaliki, AA., & Hussein, EE. (2022) A*n Advanced Machine Learning Approach to Predicting Pedestrian Fatality Caused by Road Crashes: A Step toward Sustainable Pedestrian Safety.* Sustainability.

[19] Tiwari, G. (2020). *Progress in pedestrian safety research.* International Journal of Injury Control and Safety Promotion, 27(1), 35–43.

[20] Toran Pour, A., Moridpour, S., & Tay, R. (2017). *Neighborhood Influences on Vehicle-Pedestrian Crash Severity.* J Urban Health 94, 855–868.

[21] *Vulnerable road users.* The Canadian Association of Road Safety Professionals CARSP. (2023, May 16). https://carsp.ca/en/news-and-resources/road-safety-information/vulnerable-road-users/

[22] Wahab, L., & Jiang, H. (2019). *A comparative study on machine learning based algorithms for prediction of motorcycle crash severity.* Plos One, 14(4), e0214966.

[23] Zhang, J., Li, Z., Pu, Z., & Xu, C. (2018). *Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods.* IEEE Access, vol. 6, pp. 60079-60087.

[24] Zhang, X., Waller, S. T., & Jiang, P. (2019). *An ensemble machine learning-based modeling framework for analysis of traffic crash frequency.* Computer-Aided Civil and Infrastructure Engineering, 35(3), 258-276.