Assessing Pedestrian Safety on Roads Through Machine Learning Approaches for State Highways in Washington State

FINAL PROJECT REPORT

by

Yinhai Wang, Wei Sun, Sam Ricord, Cesar Maia de Souza
University of Washington

for

Center for Safety Equity in Transportation (CSET)
USDOT Tier 1 University Transportation Center
University of Alaska Fairbanks
ELIF Suite 240, 1764 Tanana Drive
Fairbanks, AK 99775-5910

In cooperation with U.S. Department of Transportation, Research and Innovative Technology Administration (RITA)

rural ◦ isolated ◦ tribal ◦ indigenous

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Center for Safety Equity in Transportation, the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle Assessing Pedestrian Safety on Roads Through Machine Learning Approaches for State Highways in Washington State	**5. Report Date** July 1, 2024
	6. Performing Organization Code
7. Author(s) and Affiliations Yinhai Wang, Wei Sun, Sam Ricord, Cesar Maia de Souza	**8. Performing Organization Report No.** INE/CSET 24.09
9. Performing Organization Name and Address Center for Safety Equity in Transportation ELIF Building Room 240, 1760 Tanana Drive Fairbanks, AK 99775-5910	**10. Work Unit No. (TRAIS)**
	11. Contract or Grant No. 69A34520501020620
12. Sponsoring Organization Name and Address United States Department of Transportation Research and Innovative Technology Administration 1200 New Jersey Avenue, SE Washington, DC 20590	**13. Type of Report and Period Covered** Final report, Sep 2019 – Sep 2022
	14. Sponsoring Agency Code

15. Supplementary Notes

Report uploaded to:

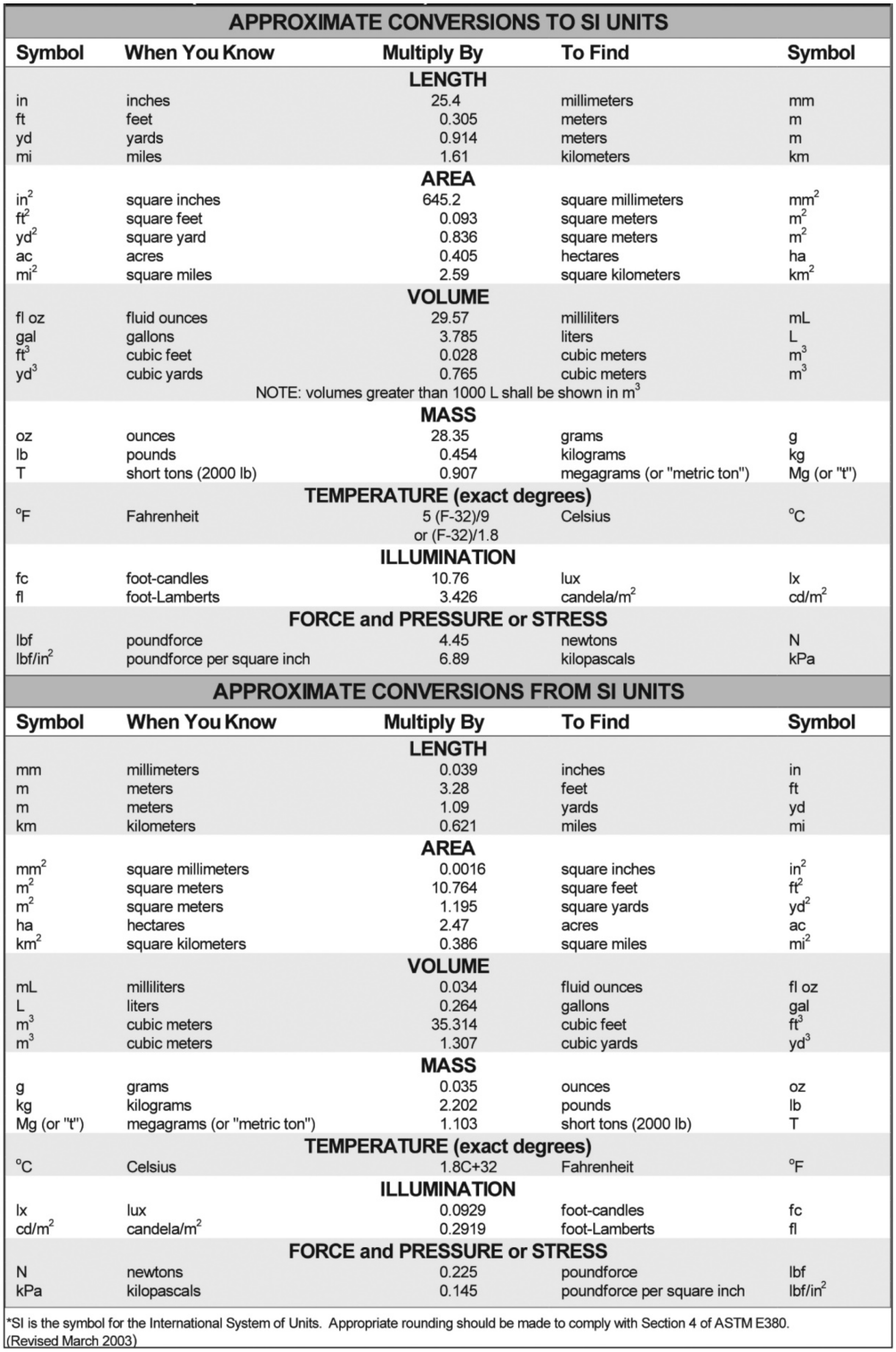
16. Abstract

The report presents a unique contrast for several Machine Learning approaches aiming at understanding pedestrian fatal collisions. Four classification techniques are applied to assess how roadway features mainly correlate to pedestrian fatal crashes: Logistic Regression, Nearest Neighbor Classification, Decision Tree, and Random Forest Classifier. The data used in this project was collected from the Highway Safety Information System (HSIS) database, which provides both collision data for the entire state of Washington and roadway characteristics for all state highways. Each of the four modeling approaches was implemented using K-fold cross-validation, a process that allows choosing the best parameters for the model. Their results were evaluated and then compared in terms of accuracy score and confusion matrices for the testing data set. It was found that the Decision tree had consistent results and the best performance among all models, showing how the distinct predictors relate to each other to predict fatal pedestrian collisions.

17. Key Words Safety Data Tool; Roadway Safety Assessment; Pedestrian safety, collisions, severity, HSIS database, Machine Learning, Statistical and Machine Learning Modeling, Classification methods;	**18. Distribution Statement**

19. Security Classification (of this report)	**20. Security Classification (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified.	Unclassified.	40	N/A

SI* (MODERN METRIC) CONVERSION FACTORS



APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
^{o}F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^{o}C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
^{o}C	Celsius	1.8C+32	Fahrenheit	^{o}F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Pedestrian injuries have emerged as a major traffic safety concern, particularly in rural, isolated, tribal, and indigenous (RITI) communities. These communities face several challenges, including scattered accidents on RITI roads, and the absence of a data-driven safety assessment method. As most RITI communities' road safety analyses remain incomplete, it is crucial to enhance safety assessment methods for data analysis and management.

The project's objective was to employ various methods to evaluate pedestrian safety in RITI communities. Four classification techniques are used to determine the relationship between roadway features and pedestrian fatal crashes: Logistic Regression, Nearest Neighbor Classification, Decision Tree, and Random Forest Classifier. Data for this study were gathered from the Highway Safety Information System (HSIS) database, which includes collision data for the entire state of Washington and roadway features for all state highways. K-fold cross-validation was used to implement each of the four modeling approaches, allowing for the selection of optimal model parameters. The results were assessed and compared in terms of accuracy scores and confusion matrices for the test dataset. The Decision Tree model demonstrated consistent results and superior performance compared to other models, revealing the relationships between unique predictors and fatal pedestrian collisions.

CHAPTER 1. INTRODUCTION

1.1. Research Background

Pedestrian injury has become a significant traffic safety issue, especially in the rural, isolated, tribal, and indigenous (RITI) communities. Conventional wisdom has considered pedestrian safety to be a more severe issue for the urban areas, where there are more conflicts between vehicles and pedestrian. Therefore, considerably more attention has been given to urban cities regarding the implementation of countermeasures to improve pedestrian safety conditions. However, studies indicated that pedestrian-involved crashes in RITI communities often lead to severe injuries or fatalities [1][2][3]. The lack of accommodation, such as sidewalks, marked crosswalk, lighting condition, and traffic control, etc., makes pedestrian and bicyclists at huge disadvantage in these RITI communities. Over seventy percent of pedestrian fatalities on tribal lands occurred in the rural areas and approximately seventy-five percent of them happened at night[4]. In addition to limited pedestrian facilities, cultural and human behavior factors, such as speeding, driving under the influence, pedestrian behavior, etc., also contribute to the high pedestrian injury and fatality rates[3]. Iragavarapu conducted a review of tribal transportation safety and found that over fifty percent of pedestrian fatalities occurred in reservations were related to driving under the influence [5]. Despite of the significance of high pedestrian injury and fatality rates, pedestrian safety issue in RITI communities is still a relatively under-recognized issue [6].

1.2. Problem Statement

Through the outreach and research activities in previous years' CSET projects, some of the existing challenges and issues of the RITI communities on traffic safety that require immediate actions are fully understood. One of the outstanding issues is pedestrian safety, which is getting increasing attention from the government agencies to transportation researchers and practitioners to the residents living in the RITI communities. According to the FHWA (Federal Highway Administration), 54% of collisions occurred on rural roads while only 19% of the country's population lives in rural areas[7]. Per capita, it is much more likely that fatalities occur in rural areas rather than urban areas. The disparity is even more stark when considering Native American populations in the U.S. Native Americans are three times more likely to be killed in a traffic incident than Non-Native populations according to the Washington Safety Traffic Commission[8].

The research team has been in close communication with engineers from Yakama Nation Department of Natural Resources (DNR) Engineering department and learned that Yakama Nation has the highest number of pedestrian fatalities in Washington State. The research team conducted a site visit to Yakama

Nation and several teleconferences with engineers and planners of the Yakama Nation DNR Engineering department and recognized the pressing need to improve the pedestrian safety conditions. One of the most critical issues faced by pedestrians in Yakama Nation is the lack of pedestrian facilities. For most of the roadways throughout the reservation there are no existing pedestrian facilities. Most roads don't even have a shoulder, and instead have an embankment or a drainage ditch. This forces pedestrians to walk essentially on the fog line or in the live traffic lane along most of these roads. Additionally, there is a relatively high population that have no access to a car, instead relying on either public transportation or walking. Considering public transportation has limited routes and does not run every day and high levels of amenities along roads without pedestrian infrastructure, high levels of pedestrians use these roadways and are exposed to vehicles. Winter months are particularly problematic due to fog regularly limiting visibility for drivers. Additionally, there are several intersections that are only stop controlled by stop signs that have poor visibility, even without fog, that also increase the risk for pedestrians. In addition to the lack of accommodations for pedestrian, other factors such as roadway geometrics, traffic characteristics, roadway and intersection operation characteristics, weather conditions, and cultural and human behaviors, also contribute greatly to the pedestrian safety issues. However, due to practical challenges, most of the relevant data to measure the pedestrian exposure risk are limited/lacking in Yakama Nation.

Many state and local governments, especially in the rural and tribal areas, are experiencing pedestrian safety issues similar to Yakama Nation. There is an urgent need for systemic and data-driven pedestrian safety assessment methods that provide guidance on the collection and analysis of necessary data, identification and investigation of contributing factors, development of pedestrian safety indices, identification of high-risk roadways and intersections, etc.[9]

1.3. Research Objectives

The research objectives are:

1. Conducting outreach activities to strengthen /establish long-term relationship with RITI communities and to have a thorough understanding of the challenges that the RITI communities are facing and their current practices regarding pedestrian safety.
2. Establishing research collaboration with interested RITI communities, sign data share agreement, and conduct case study of pedestrian assessment.
3. Developing data-driven pedestrian safety assessment methods for RITI communities to support decision making, such as data collection and management guidance, contributing factors

identification and analysis, pedestrian safety risk levels estimation, pedestrian safety indices development, high-risk roadway/intersection identification, countermeasures analysis, etc.

CHAPTER 2. LITERATURE REVIEW

2.1. Traditional Statistical Pedestrian Safety Analysis

There have been many different studies related to pedestrian safety. This has been a subject of study for many decades. M Snyder conducted a study in 1971 to identify the causes and countermeasures of pedestrian collisions in Maryland. For this study over 2000 pedestrian collisions were analyzed, mostly focusing on pedestrian behavior. It was found that over 50% of crashes were caused by some form of pedestrians entering the roadway inappropriately[10]. In 1983, Hall conducted a study to measure rural pedestrian safety in New Mexico. The study described the discrepancy between rural pedestrian fatalities and urban fatalities, with those results for each region being 49% and 34% respectively[11]. These studies show how the topic of pedestrian safety has been a subject of study for a long time. These studies highlight the importance of pedestrian safety studies and some of the initial methodologies aimed at understanding them.

More recently, an NCHRP study was conducted for the National Academies of Sciences, Engineering, and Medicine that investigated the correlation between site-specific characteristics and pedestrian collisions. They found that site-specific characteristics increased the likelihood of pedestrian collisions[12]. A similar study was conducted in Bangladesh. This study also found that specific characteristics precipitate pedestrian collisions in the city of Dhaka[13]. Additionally, several studies were conducted utilizing different modeling techniques to assess pedestrian safety. Zajac created an ordered probit model that evaluated roadway features that are prevalent in pedestrian collisions[14]. This model showed different characteristics that influence pedestrian collisions that are more appropriate for a rural setting than the other above studies. Chen conducted a similar study in 2019, which used the alternative method of a mixed logit model to predict rural pedestrian collisions[15]. Baireddy conducted a study in rural Illinois that identified several factors that increase the pedestrian collision likelihood using multiple correspondence analysis[16]. These previous studies hold several implications for this project. Firstly, they show that using roadway characteristics as a method to predict pedestrian collisions is a valid and well-documented methodology. Secondly, it highlights the research gaps filled by the study of this project where the severity of pedestrian collisions is not considered in any of these previous studies mentioned.

2.2. Machine Learning based Pedestrian Safety Analysis

It can be noted that most of the literature relies on traditional statistical modeling approaches to address pedestrian safety issues. Nevertheless, with the recent advent of Machine Learning, some researchers have started applying these latest approaches to this type of problem. In 2018, Ding developed a study to examine built environmental effects on the frequency of crashes involving automobiles and pedestrians by applying Multiple Additive Poisson Regression Trees (MAPRT), a Machine Learning approach based on decision trees. Using data from Seattle, Washington, the study helped to detect non-linear relationships between the built environment and pedestrian collisions frequency, confronting the linearity assumption frequently used in studies that use statistical models[17]. Das applied in 2020 distinct Machine Learning techniques to classify pedestrian collision types (intended vs. untended, pedestrian at fault vs. motorist at fault) using pedestrian crashing data from two locations in Texas[18]. These reference studies were essential for the development of our methodology applied specifically to fatal collisions, an unprecedented approach so far.

2.3. State of the Art and the Practice

Recognizing the needs to improve pedestrian safety and reduce pedestrian injury and fatality, the Federal Highway Administration (FHWA) published a guide that documented scalable risk assessment methods for pedestrians and bicyclists[9]. The guide provides guidance on the steps to estimate the exposure to risk of pedestrian and bicyclists, including determining uses of risk values, selecting geographic scale, selecting risk definition, selecting exposure measure, selecting analytical method to estimate exposure, using analytic method to estimate selected exposure measure, and calculating risk values.

In addition, the FHWA developed the guidebook summarizing the data-driven approaches for identifying high-risk locations for pedestrians [19]. However, despite the intention of the guidebooks to make general approaches that could suit most agencies with different analysis capabilities and resources, through the research team’s communication with tribal leaders and agencies, it is necessary to recognize that most agencies in the RITI communities do not have the practice to collect and manage pedestrian safety related data and lack the guidance of doing so. Without the practice of collecting necessary data and developing data-driven solutions, it is not easy for RITI communities to follow the approaches of such guidebooks. Besides, the uniqueness of roadway geometrics and operational characteristics, traffic characteristics, environmental conditions, and cultural and human behavior characteristics of RITI communities deserve unique data collection and assessment approaches in order to obtain accurate risk assessment results. These guidebooks are based on the classical Empirical Bayes safety performance function based approaches, while studies have suggested that machine learning based approaches could provide more insights from the multi-source pedestrian safety data[17][20][21][22], as introduced in the previous section.

CHAPTER 3. ROADWAY PEDESTRIAN SAFETY ASESSMENT METHODS

3.1. Data

This study utilizes collision and roadway data from the Highway Safety Information System (HSIS)[23]. The HSIS is a project by the Federal Highway Administration (FHWA) that collects collision and roadway data from participating states to offer larger sample sizes for studies relating to safety, both historic and current. Washington State is one of the participating states of HSIS, which is where we retrieved our data.

This data is split into several different files that can all be connected through various identifiers in each file. One critical file is the 'accident file', which stores all of the information relating to the details and circumstances of a collision. This information is gathered directly from police reports of collisions, so it captures the same information a police officer captures when responding to a collision. There are several subfiles in addition to the 'accident file' that also give additional information about each crash. These are the 'vehicle subfile', the 'occupant subfile', and the 'pedestrian subfile'. These files are all connected to the 'accident file' through the case number and give more detailed information on individual vehicles, occupants, and pedestrians respectively that are involved in a collision.

The other major file included in the HSIS data is the 'roadlog file'. This file gives the characteristics for every state-owned roadway down to a segment level, where each segment is determined to have a homogenous set of roadway characteristics based upon the attributes used in HSIS. These segments are differentiated by the specific state route and the beginning and end milepost of each segment. There are roughly 42,125 segments overall in this dataset of Washington State, of which 20,730 segments are rural and 21,356 are urban (39 are not classified).

The 'roadlog file' has several subfiles that supplement this information as well. Firstly, there is the 'grade file' and 'curve file'. These files give information on the gradient and curvature on the roadway respectively for roadway segments. It is important to note that the segments for each of these files are different, as roadway characteristic changes do not always match up with curve and grade changes on a roadway. Several additional files further supplement the 'roadlog file', including the 'ramp file', the 'special-use lanes file', the 'features file', the 'left/right file', and the 'railroad crossing file'.

For this study, we utilized the 'accident file' and the 'roadlog file'. Our data spans 5 years, from 2013 through 2017. Though the 'roadlog file' remained very similar throughout those five years, the number of collisions recorded in the accident file for each year did vary some. 2013 had the lowest total number of recorded collisions at 43,469. The highest number occurred in 2016 at 57,415. The other three years of 2014, 2015, and 2017 had collision records of 48,292, 53,010, and 55,548 respectively.

Over all five years, however, only 2,387 of these collisions involved pedestrians somehow, where 92% occurred in urban areas and 8% in rural areas. This means that there is on average approximately only one collision for every 17 segments. Also, since some segments have more than one collision, that means that even less than 6 percent of segments have any sort of pedestrian collision on it. This shows how collisions are inherently random, especially when considering a fairly unique subset of collisions such as pedestrian collisions. This study is only possible because we have all of the collisions across an entire state for five years which creates a dataset that is large enough to avoid errors due to randomness of collision occurrence.

Regarding collisions' severity, roughly 84% of the crashes involving pedestrians lead to some type of injury, whilst 7% are fatal for either pedestrians or vehicles' occupants. This number is close to the total of pedestrian collisions generating property damage only (9%), which indicates the relevance of understanding the factors that may contribute to fatal crashes.

3.2. Methodology

The main goal of this study is to analyze distinct Machine Learning (ML) approaches to predict the occurrence of fatal collisions involving pedestrians using mainly roadway features from the HSIS database. Fatal crashes have the most severe impacts in terms of human losses, therefore identifying locations with a high occurrence of deadly crashes is crucial to prioritize spots to implement pedestrian safety countermeasures. We linked the 'accident' and 'roadlog' HSIS files by matching the RD_INV of both files and then locating where the collision's milepost is between the Beginning Milepost (BEGM) and the Ending Milepost (ENDMP) of a segment in this same road. Nevertheless, an issue occurs when the collision's milepost is exactly either the BEGM or the ENDMP for two consecutive segments as the BEGM for one segment is the ENDMP for the other (or vice-versa), there is no way to define exactly one road segment for that collision. Although randomly assigning the crash to one of the two segments might be an option, we would be overlooking the characteristics of the other segment without knowing their relevance. To avoid any loss of information or wrong assumptions, we decided to keep both roads segments assigned to a crash when this situation happened. This implies that a single collision may have two observations in the final database, each one associated with a distinct road segment.

Some data cleaning was required to carry out this analysis. We selected only collisions involving pedestrians from the database. This was possible by using the Number of Pedestrians variable (NO_PEDS) originally from the accident file, which indicates the number of pedestrians involved in the crash (zero or greater). By filtering this variable for values greater than zero, we could obtain a data set showing only collisions involving one or more pedestrians. It is worth mentioning that every collision having 'NO_PEDS' greater than zero was considered, which includes vehicles directly hitting pedestrians and also other crashes involving pedestrians (e.g., a collision between two vehicles that ends up affecting a nearby pedestrian). However, the vast majority of cases (roughly 92% of crashes involving accidents) are due to vehicles directly hitting pedestrians. Our final data set with pedestrian cashes only contained 2,067 observations after data cleaning.

Our response variable was built based on the "REPORT" variable from the 'accident file', which defines the severity of the crash as described in Table 3 - 1 below:

Table 3 - 1: Report Variable Designations

Variable Number	1	2	3
Description	Property Damage Only	Injury Accident	Fatal Accident

Source: HSIS Guidebook – WA, 2014

The "REPORT" variable was then recoded into a dichotomous variable called 'Fatal', showing True (1) for fatal collisions (i.e., for REPORT's values equal to 3) and False (0) for other collisions (i.e., for REPORT's values other than 3). This approach summarizes well the set of collisions presented in the data set and is aligned with the purpose of this study. Nonetheless, it must be explained that a "fatal collision" does not necessarily imply that the pedestrians were the fatality victims. It tells us that an accident culminated in someone's death, but this could correspond either to the pedestrians or the vehicles' occupants. However, pedestrians are the most likely victims of collisions of this type due to their vulnerable conditions on the roads.

The next step was to determine which variables were to be used in the initial models. There are over 500 variables in all of the files of the HSIS data set and we initially chose 21 variables from the accident file and 25 variables from the roadlog file. The variables were then visualized to understand their distribution and to initially select only those that might be of interest to the project. Since most of the techniques used in this study (e.g., logistic regression and random forest) require independent predictors, correlation matrices were built to detect whether the predictors were overall independent of each other or not, eliminating some of the variables with high correlation (generally greater than 0.6). After having selected and re-coded some variables, we reduced the number of variables that were used to 18 initial variables that are described in the following table.

Table 3 - 2: List of All Variables Used in Initial Models

Variable	**Description of Variable (Unit)**
AADT	Annual Average Daily Traffic (vehicles)
SPD_LIMT	Speed Limit (mph)
LANEWID	Lane Width (ft)
RDWY_WID	Roadway Width (ft)
RSHLDWID	Right Shoulder Width (ft)
TRKPCTS	Truck percentage for the roadway segment (%)
RURAL	Rural road (categorical: Yes-1; No-0)
LIGHT_DAYLIGHT	Collision at daylight (categorical: Yes-1; No-0)
LIGHT_DARKLIGHTSON	Collision at dark with street lights on (categorical: Yes-1; No-0)
LIGHT_DARKNOLIGHT	Collision at dark with street lights off or no street lights (categorical: Yes-1; No-0)
WEATHER_CLEAR[1]	Weather conditions when the crash occurred were clear or partly cloudy (categorical: Yes-1; No-0)
TERRAIN_LEVEL	Terrain type: Level (categorical: Yes-1; No-0)
TERRAIN_MOUNTAINOUS	Terrain type: Mountainous (categorical: Yes-1; No-0)
TERRAIN_ROLLING	Terrain type: Rolling (categorical: Yes-1; No-0)
FREEWAY	Roadway classification: Freeway (categorical: Yes-1; No-0)
2-LANEROAD	Roadway classification: 2-Lane Roads (categorical: Yes-1; No-0)
MULTILANE_NON- FREWAY	Roadway classification: Multilane Non-Freeways (categorical: Yes-1; No-0)
FATAL	Fatal collision (categorical: Yes-1; No-0)

We developed this study by applying several Machine Learning approaches to our data set. ML techniques have been recently used for transportation data mining and analysis, particularly due to the massive data growth as a result of enhancements in data generation and collection technologies. Machine Learning can have many definitions, such as "Non-trivial extraction of implicit, previously unknown and potentially useful information from data" and "Exploration & analysis, by automatic or semi-automatic means, of data to discover meaningful patterns"[24]. Among the predictive ML techniques, classification modeling approaches are used to fit a model for a class attribute as a function of the values of other attributes. Since our response variable is dichotomous (Fatal =1/0), we can define it as a class labeled "Fatal", so we can use miscellaneous classification methods that apply to our database and evaluate which one performs the best.

Machine Learning approaches should be able to predict future or unobserved values and also be capable of evaluating the accuracy of these predictions. This is possible by randomly splitting the initial dataset into two: a training/validation data set and a testing data set. The model is built and validated using the training data set, and then used to predict the "unobserved" values of the test data set. Accuracy scores and confusion matrices are typical measures of precision for classification techniques. In our case, the training subset corresponded to 80% of the total data set, while the remaining 20 % were used as the testing data. An issue that may occur when dealing with classification models is related to imbalanced data, which happens when the distribution of examples across the known classes is skewed. Indeed, our response variable ("Fatal") has approximately 93% of the accidents' observations labeled as 0 (Non-Fatal) and only 7% labeled as 1 ("Fatal), which is our target class. This imbalance may make the model biased to predict the majority group. To deal with this issue, we used one of the most common practice techniques to handle imbalanced data called the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE acts over-sampling the minority class (Fatal = 1) so that new data instances of the minority group are created by copying some existing minority instances with small changes.

For this study, we used 4 different Machine Learning classification techniques to assess how roadway features and some other factors correlate to pedestrian fatal collisions. An overview of each of these modeling techniques is presented below.

3.2.1. Logistic Regression

Logistic regression is a classical method for binary classification whose parameters are estimated by maximizing the likelihood estimation through the following equation:

$$log\ \frac{Pr\ (y=1)}{1-P\ r(y=1)} = \ \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

Where: Pr represents the probability of a sample to belong to class 1, β_0 = intercept and β_i = coefficients. The expression $\frac{Pr\ (y=1)}{1-P\ r(y=1)}$ corresponds to the odds, where $log\ \frac{Pr\ (y=1)}{1-P\ r(y=1)}$ represents the log odds. The goal is to predict the log-odds, which is converted to probability through the logistic function.

Logistic regressions can also have a penalty term related to the model complexity. It is represented with the hyperparameter C that controls the inverse of model complexity (smaller values imply stronger regularization). The hyperparameter selection was made using the k-fold cross-validation method, where the training data is split into k groups to select the best value of C. The value of k usually varies from 5 to 10, and due to the data size used in this study, k = 5 was selected. Cross-validation is also an appropriate technique to avoid overfitting issues, a Machine Learning sign of poor performance occurring when the model fits perfectly the training data set, including its noise or outliers.

We built an initial logit regression with all the variables presented in Table 3 - 2, having Fatal as the response variable. Results were then analyzed to verify the significance of each of the variables' coefficients. As a common practice, a level of significance of 0.05 was established, so that each variable with a statistical p-value less than 0.05 was considered significant. A second logistic regression model was built using only the significant variables identified in the initial model.

3.2.2. Nearest Neighbor Classification

Nearest Neighbor Classification is a method that uses class labels of the K nearest neighbors to determine the class label of an unknown record using proximity metrics to calculate distance/similarity between records. The number of nearest neighbors (K) is a hyperparameter that must be provided, along with the distance metric (Minkowski distances are usually used). Choosing the values of K can be difficult, since a too-small K may lead to neighborhoods that are sensitive to noise points, whereas a too-large K may make a neighborhood include points from other classes. K-fold cross-validation is an effective method to handle this issue, where distinct values of K can be analyzed using the training data set, as well as distinct values for parameters such as the exponent factor "p" for the Minkowski distance and the weights associated with the distances. A final model with the best parameters is then fitted and can be applied to the testing data set.

Therefore, a third model using Nearest Neighbor Classification with k-fold cross-validation was built using the same set of significant variables applied to the second logistic regression.

3.2.3. Decision Tree

Decision trees are another effective tool to handle classification problems. The goal is to classify data (leaf nodes of the tree) from the characteristics of the predictor variables (decision nodes). Following Hunt's algorithm (12), if Dt is the set of training data points reaching a node t, two options exist: if Dt contains data points that belong to the same class yt, then t is a leaf node labeled as yt ; if Dt contains records that belong to more than one class, we should use an attribute test to split the data into smaller subsets and recursively apply the procedure to each subset. However, early terminations are often applied to stop the splitting procedure to avoid overfitting issues.

To define the best split, we follow the Greedy approach which establishes that nodes with "purer" class distribution are preferred, i.e., nodes with samples mainly distributed towards one of the classes. We applied the Gini Index as a measure of node impurity, computed as follows:

$$GINI(t) = 1 - \sum_{j} [p(j|t)]^2$$

Where $p(j|t)$ is the relative frequency of the class j at node t. GINI is maximized when the points are equally distributed among all classes, showing the least interesting information. The minimum (0) occurs when all records belong to one class, indicating the most interesting information for that node.

To avoid overfitting and to select an appropriate number of parameters such as minimum samples per leaf nodes and the maximum depth, we applied k-fold cross-validation as was done for the other models to choose the optimal hyper-parameters (always with k=5). A new model was thus fitted using the same set of final variables that were applied for the previous two models.

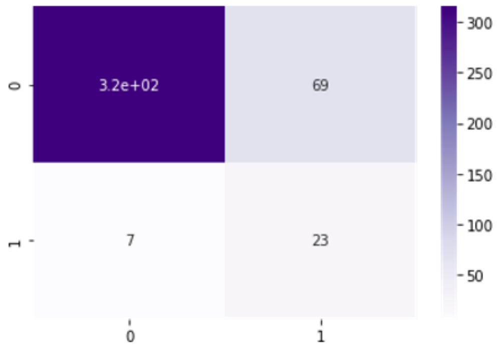
3.2.4. Random Forest Classifier

Random Forest classifiers are part of the so-called Ensemble Methods (12), ML classification techniques aiming at building a set of base classifiers from the training data set and predicting the class label of test records by combining the predictions made by all base classifiers (through majority vote). Ensemble methods also aim to reduce the variance of complex models by aggregating responses of multiple base classifiers.

Ensemble methods generally need independent base classifiers, and Random Forest techniques are well aligned with this. They fit a full decision tree by randomizing which predictors would be available for a given node, which alleviates the split on similar predictors for bagged trees.

We developed a final model by applying a Random Forest Classifier for the same final set of significant variables used in the previous models. Likewise, we used k-fold cross-validations to select parameters needed for this method, such as maximum depth for the trees and the number of trees in the forest ("number of estimators").

All the models in this study were evaluated and compared in terms of the accuracy score for the training and testing data sets. Accuracy scores vary from 0 to 1, and values close to one denote effective predictions. However, models with accuracy scores approximately equal to 1 may indicate overfitting issues and are not suitable. Additionally, we applied confusion matrices to the testing data of each of the models to evaluate their level of predictions. A confusion matrix is a representation of the classification rate for a classifier method and has 4 quadrants indicating the number of right and wrong predictions for the class, as seen below:



Legend:

- Top left quadrant: True Negative
- Bottom Right quadrant: True Positive
- Top Right: False Positive
- Bottom Left: False Negative

Figure 3 - 1: Example of a confusion matrix

We compared each model to verify which one performed better when predicting the occurrence of fatal collisions involving pedestrians. Importantly, all the modeling techniques in this study were developed using Python programming language, which has powerful and widely used libraries related to Machine Learning approaches, such as "sklearn".

3.3. Results

After exploring the four methods mentioned above, we were able to find what variables were significant using the logit model. Since the first logit model contains all of the variables and was used to find insignificant variables, its results are not included here as they do not show useful information for the outcome of this study. A summary of findings for the second logit model as well as for the other models is shown below.

3.3.1. Logistic regression

The following table presents the coefficients for the final logistic model using k-fold cross-validation, where all the variables are significant to a level of 0.05, except for the “rural” variable (p-value = 0.1272). However, since we are equally interested in understanding whether rural areas may have a distinct impact on pedestrian fatal collisions compared to urban areas, we decided to keep this variable in the model. Furthermore, the variables shown in Table 3 - 3 are the final set of variables used in all the other models that will be presented below.

Table 3 - 3: Results for the Logistic Regression

Variable	Coefficient
Intercept	-0.0026
SPD_LIMT	-0.0124
LANEWID	-0.0273
RSHLDWID	0.1999
AADT	0.0000
TRKPCTS	0.0706
RURAL	-0.0313
LIGHT_DAYLIGHT	-0.3674
LIGHT_DARKLIGHTSON	-0.2014
LIGHT_DARKNOLIGHT	0.0002
WEATHER_CLEAR	-0.2106
FREEWAY	0.0142
2-LANEROAD	-0.0580
MULTILANE_NON-FREWAY	-0.1868
Penalty value: 5.0	
Accuracy on training data set: **0.759**	
Accuracy on testing data set: **0.826**	

The results show noteworthy insights about the relationship between each variable and their impact on the occurrence of fatal crashes involving pedestrians. SPD_LIMIT and LANEWID have surprisingly negative impacts on the outcome, suggesting that roadways with higher speed limits and lane width tend to be related to fewer fatal collisions. This outcome should be interpreted with caution though, particularly because this analysis does not include features that may be linked to these variables' effects. For example, demographic predictors are not present in this model. However, more populated areas may have roads with lower speed limits and lane width but with greater pedestrian circulation that can be associated with higher fatal crashes frequency.

On the other hand, RSHLDWID and TRKPCTS all have positive coefficients, which may indicate that roads with wider right shoulders and a higher percentage of trucks are more likely to have pedestrian fatal collisions. Rural roads seem to be less likely to have fatal crashes compared to urban roads, whereas daylight periods have the most negative impact on the response variable. However, dark periods when street lights are off or absent are more likely to generate fatal crashes, which intuitively makes sense. As expected, days with clear or partly cloudy weather are less likely to have fatal collisions, and freeways are more likely to be associated with this type of crash, which may be explained by their general higher volume compared to other roads (although AADT has a positive coefficient, its value is practically equal to zero).

Regarding model accuracy, we observe that the testing data set has a higher accuracy when compared to the training data set (0.826 against 0.759, respectively). Since the model is fitted using the training data and then used to predict testing samples, having a higher accuracy on the testing data set does not seem to be conceivable. This may indicate that logistic regression is not a suitable model type for this study.

Additionally, the below figure represents the confusion matrix related to the testing data set. Although the number of correct predictions (higher than 340) surpasses the number of wrong predictions (72), we note the model is biased predicting 0 values (non-fatal collisions), with only 21 correct predictions for fatal crashes.

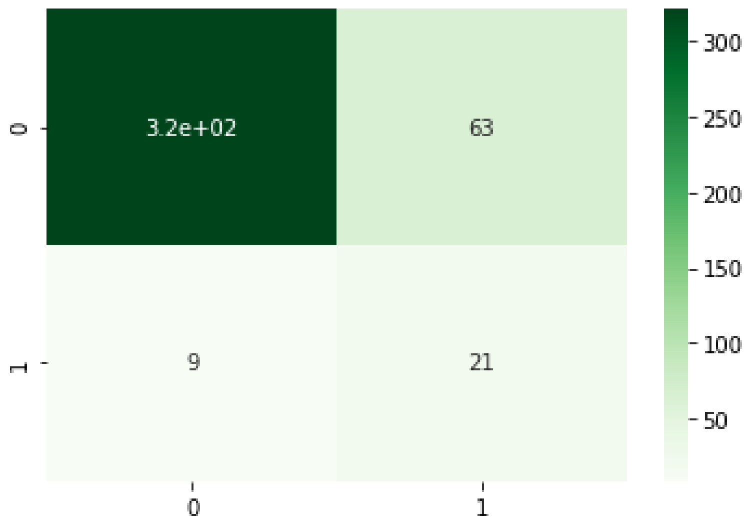
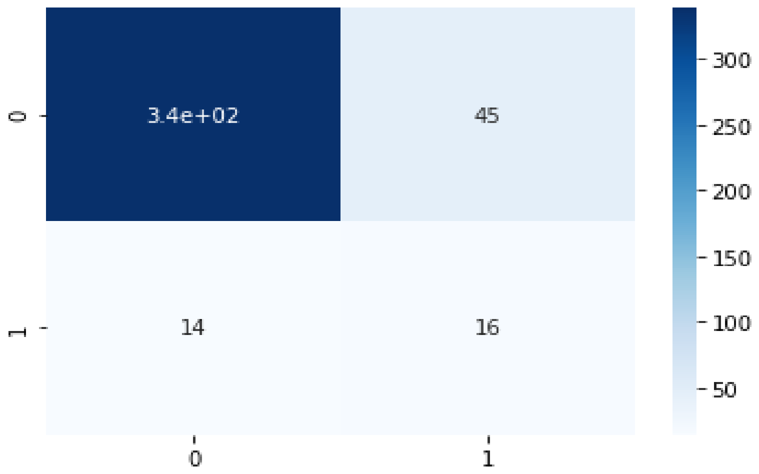


Figure 3 - 2: Confusion matrix for the Logistic regression model

3.3.2. Nearest Neighbor Classification

The following figure is a dashboard with the results for the Nearest Neighbor Classification. It shows the best value of parameters selected after cross-validation using Minkowski as the distance metric: number of neighbors, the exponent factor "p" for the Minkowski distance, and the criteria of weights (two possible options: uniform and distance). The table also presents the accuracy scores for the training and testing data sets, in addition to the confusion matrix (right column) showing the right and wrong predictions for the testing data set.



Best parameters

N. neighbors: 3

p: 1

Metric: Minkowski

Weights: Distance

Accuracy on training data set: **0.997**

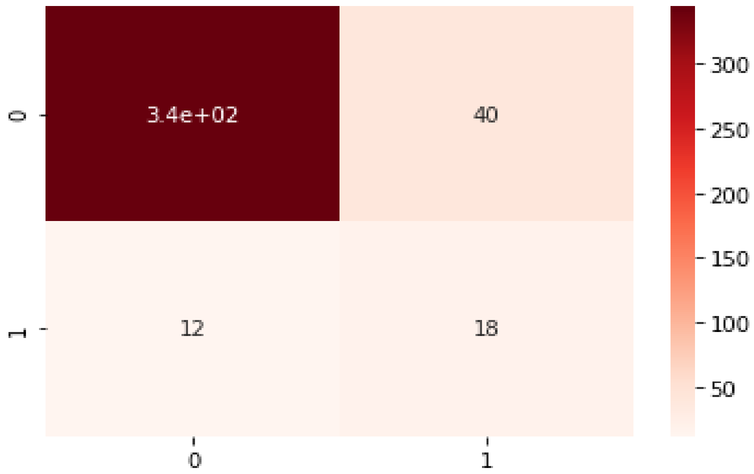
Accuracy on testing data set: **0.857**

Figure 3 - 3: Dashboard with the results for the Nearest Neighbor Classification

We observe from Figure 3 - 3 that the accuracy of the training data set is higher than the testing data set, which is reasonable. However, the accuracy for the training data is practically equal to 1, almost a "perfect" fit to the training data. As previously mentioned, this is a sign of overfitting, which suggests that this model, even after cross-validation, may not represent a good fit for the studied data set. Like the logistic regression, the confusion matrix for the testing predicted values shows that the model is biased predicting 0 values (non-fatal collisions).

3.3.3. Decision Tree

The results of fitting a decision tree using cross-validation are summarized in Figure 3 - 4. We can observe that two parameters were tested during the cross-validation process in order to get the best modeling performance: the maximum depth of the tree (i.e., the number of horizontal levels of a top-down tree from its root) and the minimum samples in leaf nodes. Additionally, there are the accuracy scores for the training and testing data sets, as well as the confusion matrix with predictions for the testing data.



Best parameters

Maximum depth: 5

Minimum samples in leaf node: 5

Accuracy on training data set: **0.885**

Accuracy on testing data set: **0.874**

Figure 3 - 4: Dashboard with the results for the Decision Tree

The results of Figure 3 - 4 show that the best tree that fits the training data has a maximum depth of 5 levels, with a minimum of 5 samples for the leaf nodes. These parameters help to avoid overfitting issues, which can be seen for the accuracy score of the training data set: 0.885 (a high value not too close to 1). The accuracy score for the testing data (0.874) is also high and slightly less than the training's score, which is a good performance indicator for this model. However, the confusion matrix once again shows that the model is biased to predicting 0 values (non-fatal collisions).

Figure 3 - 5 provides a visualization of the fitted decision tree. Each decision node presents the variable's criteria of splitting, the Gini Index, the total number of samples in that node, the number of samples per class (0 = "Non-fatal", 1 = "Fatal"), and the class with the majority of the points. Leaf nodes show the final classification of samples (class = 0 or 1), also derived from the Gini Index and the samples' division between the two classes for that node. From the analysis of the tree, we note that RSHLDWID is the root, i.e., the first feature that the tree split on, indicating the importance of right shoulder width when

assessing this type of collision (actually, any shoulder width is relevant since the variable related to left shoulder width was eliminated from the model due to its correlation with right shoulder width). For the dichotomous variables (0/1), a value less than 0.5 is equivalent to 0, while the opposite is equivalent to 1.

The analysis of how leaf nodes are formed is crucial to understand the relevance of each variable for the final classification (Fatal or Non-Fatal collisions). For example, a fatal pedestrian collision may be the result of the combination of RSHLDWID > 0.002 ft, not during the daylight, not during the dark with lights on but during the dark with street lights off/no lights, on a road that is not a Multilane Non – Freeway. At the same time, the same conditions on a road that is a Multilane Non – Freeway do not seem to be related to fatal crashes.

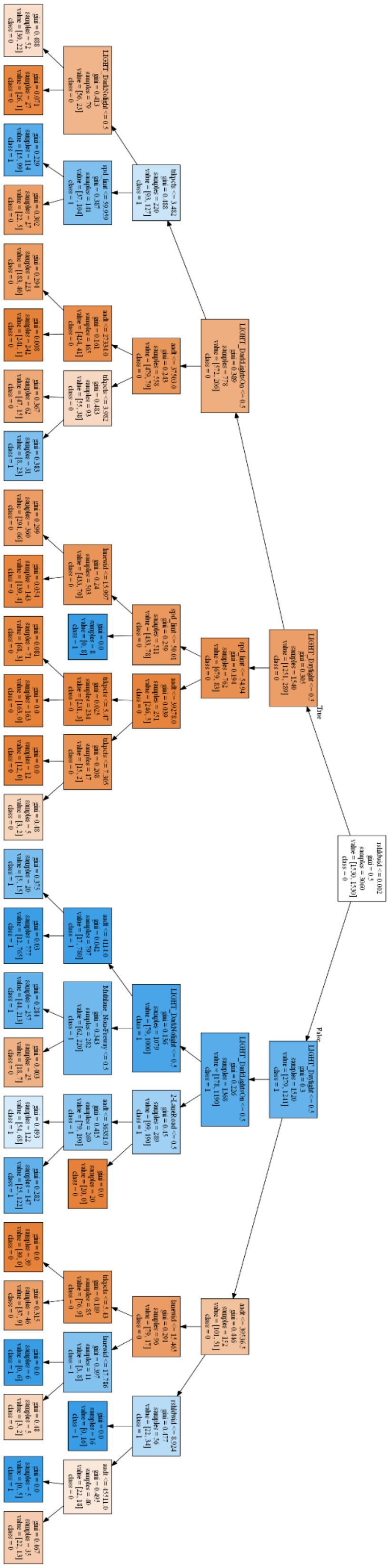
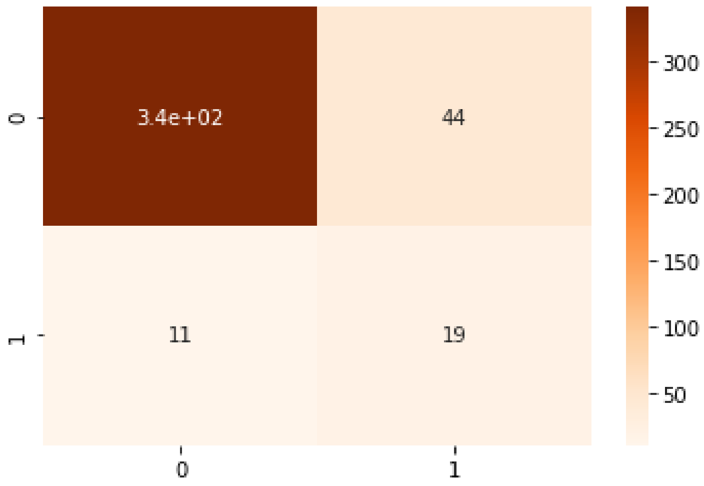


Figure 3 - 5: Decision Tree visualization

3.3.4. *Random Forest Classifier*

The following table summarizes the best parameters related to the Random Forest Classifier after cross-validation (maximum depth and number of estimators), as well as the accuracy score for the training and testing data sets and the confusion matrix for the predicted tested values.



Best parameters

Maximum depth: 5

Number of estimators: 100

Accuracy on training data set: **0.883**

Accuracy on testing data set: **0.867**

Figure 3 - 6: Dashboard with the results for the Random Forest Classifier

We can observe from Figure 3 - 6 that the model that best fits the training data has a maximum depth of 5 levels and the number of estimators equal to 100. The accuracy score of the training data set is 0.883 (a high value not too close to 1) and the accuracy for the testing data is slightly less (0.867), which may represent a model that performs well with apparently no overfitting issues. As observed in the other models, this model is also biased to predicting 0 values (non-fatal collisions), although the total of correct predictions is much more significant than the wrong ones.

3.4. Conclusions

Overall, this study presents an unprecedented contrast between several powerful Machine Learning classification approaches for predicting fatal collisions involving pedestrians using the HSIS database for Washington State. The results show that some of the modeling techniques do not seem to appropriately fit the studied data set for this research, leading to inconsistent outcomes such as higher accuracy scores for the training than for the testing data and strong indications of overfitting. For the models that reasonably fit the data, the accuracy scores for the testing data set as well as their confusion matrices are metrics used to define which one performs better on predicting unobserved values.

The two classification techniques that best fit the data and have consistent and high accuracy scores for the training and testing data sets are the Decision Tree and the Random Forest Classifier. Since both models are based on building classifier trees and have similar parameters, we would expect them to have related performances. Nevertheless, the Decision Tree has a slightly higher accuracy score on the testing data (0.874) when compared to the same metric for the Random Forest Classifier (0.867), thus this is the model with the best performance among all. Furthermore, we noted from the results that the confusion matrices for all the models were alike: even though they have substantially more correct than incorrect predictions, they are all biased to predict 0 values (non-fatal collisions). Since this is happening with all the models, we believe that it is derived from the dataset itself and its errors rather than specificities related to any of the modeling approaches used in this study.

With the best performance among all models, the Decision Tree represents how the combination of predictor variables leads to fatal or non-fatal accidents involving pedestrians. However, not all the predictor variables are used as decision nodes for the final tree, and this is due to the selected parameters after the cross-validation process, particularly for the maximum depth. Indeed, since we have 12 final predictor variables, a tree having almost all of them used as decision nodes would require a higher depth, but large-sized trees are generally not easy to interpret and may lead to overfitting. In this case, the Machine Learning algorithm searches for the set of variables that impacts the outcome for the selected tree depth the most.

Some errors occurred in this study’s methodology, particularly regarding how some of the variables were designated by HSIS and the police reports the accident file is based on. In the Roadlog file, each roadway segment is designated either rural or urban by HSIS, but it is unclear how they make this designation. This could introduce some error into the study as some rural roads may be classified as “non-rural” and urban roads as “rural”. This error however will be negligible because this is only likely to happen at the interface between rural and urban areas, which constitutes a very small portion of the roadway. Moreover, the “rural” variable derived from the classification of urban/rural is not playing a significant role in the Decision Tree. Another similar error lies in how the police report each collision. Although these reports are mostly accurate, there could be errors in reporting the collision location or collision severity. All of these discrepancies could affect this analysis, although this effect is minimal. The vast majority of the time, the police reports are close enough to ground truth that they will not have a noticeable effect on this study. Both the errors from HSIS coding and police reporting will have a negligible effect on this study and were not addressed directly because these errors occurred before we received the data. Another error can be associated with the exclusion of POP_GRP, a variable from the Roadlog file indicating the population group related to a roadway segment. Although our initial intention was to include this variable due to the relevance of evaluating the impact of population on collisions, most of the sections were blank for the rural areas and would not be representative predictors.

3.5. Comparing Other Machine Learning Approaches for Analyzing Pedestrian Safety

In the previous sections, we have emphasized the methodologies and results of the Decision Tree and Random Forest Classifier models to predict fatal collisions involving pedestrians in Washington State. The present section aims to compare these two approaches with other prominent machine learning techniques in analyzing pedestrian safety. The objective is to provide a broader understanding of the advantages, disadvantages, and potential applications of each method in this domain. The techniques under discussion in this section include Support Vector Machines (SVM), Logistic Regression, k-Nearest Neighbors (k-NN), and Neural Networks.

3.5.1. Support Vector Machines (SVM)

Support Vector Machines (SVM) is a powerful and widely-used classification technique. In pedestrian safety studies, SVM can be employed to separate the data points into two categories, representing fatal and non-fatal collisions. SVM uses hyperplanes to define the decision boundaries, maximizing the margin between the two classes. One of the advantages of SVM is its ability to work well with high-dimensional data and complex decision boundaries.

SVM is sensitive to the scale of the input features, making it crucial to normalize or standardize the data before training the model. This step can be achieved using methods like Min-Max scaling or Z-score standardization. In comparison, Decision Trees do not require scaling, as they are less sensitive to the range of input features.

SVM can handle high-dimensional data well, and it automatically selects the most relevant features during the model training process by maximizing the margin between the classes. However, using domain knowledge or other feature selection techniques like Recursive Feature Elimination (RFE) can help improve the model's performance further. Decision Trees, on the other hand, provide built-in feature selection through the splitting process and can be visualized for better interpretability.

SVM offers flexibility in choosing the kernel function, such as linear, polynomial, or radial basis function (RBF) kernels, which enables capturing complex, non-linear decision boundaries. This flexibility allows SVM to adapt to various types of data and classification tasks. In contrast, Decision Trees rely on a single hierarchical structure, making them prone to overfitting, especially with deep trees. However, the Random Forest model, an ensemble of Decision Trees, can mitigate this issue while maintaining high performance.

The performance of SVM in pedestrian safety analysis may vary depending on the choice of kernel, parameters, and the nature of the data. In some cases, SVM can outperform Decision Trees by capturing more complex relationships between the features and the target variable. However, in our study, the Decision Tree model achieved slightly higher accuracy in predicting fatal pedestrian collisions. The performance difference may be attributed to the dataset's characteristics, such as class imbalance, noise, or non-linear relationships that the Decision Tree model could capture more effectively.

One of the main drawbacks of SVM, especially with non-linear kernels, is the lack of interpretability. It can be challenging to understand the decision-making process within the SVM model, making it less suitable for applications where transparency and interpretability are essential, such as understanding the factors contributing to pedestrian safety. In contrast, Decision Trees provide a clear and easily interpretable structure that helps reveal the relationships between the input features and the target variable.

3.5.2. k-Nearest Neighbors (k-NN)

k-Nearest Neighbors (k-NN) is a non-parametric, instance-based learning method that can be used for both classification and regression tasks. In the context of pedestrian safety analysis, the k-NN algorithm can be employed to predict the likelihood of fatal and non-fatal collisions involving pedestrians. In this section, we provide a detailed comparison of k-NN based pedestrian safety analysis with other machine learning approaches, particularly the Decision Tree and Random Forest Classifier models, which were the focus of our study.

The performance of the k-NN algorithm depends on the choice of k (the number of nearest neighbors) and the distance metric used for calculating the similarity between instances. In the pedestrian safety dataset, k-NN may not perform as well as Decision Trees or Random Forests due to the following reasons:

1. High dimensionality: The presence of many predictor variables can negatively impact the performance of k-NN, as the algorithm relies on distance-based similarity measures, which are less meaningful in high-dimensional spaces.
2. Noisy data: The k-NN algorithm is sensitive to noise in the data, as it considers the neighbors' labels when making predictions. In pedestrian safety datasets, the presence of noise, outliers, or mislabeled instances may adversely affect the algorithm's performance.
3. Imbalanced classes: Pedestrian safety datasets may have imbalanced classes, where the number of non-fatal collisions significantly exceeds the number of fatal collisions. k-NN can be biased towards the majority class in such cases, resulting in poor classification performance for the minority class (i.e., fatal collisions).

k-NN models are generally less interpretable than Decision Trees, as they do not provide an explicit decision-making process. It may be challenging to understand the factors contributing to pedestrian safety based on the k-NN model. The main reasons for this limitation are:

1. Lack of explicit decision rules: k-NN does not generate a set of decision rules like Decision Trees, which can be visually represented and easily understood.
2. No feature importance: Unlike Decision Trees or Random Forests, k-NN does not provide a measure of feature importance, making it difficult to determine which predictor variables have the most significant impact on pedestrian safety.
3. Instance-based learning: k-NN is an instance-based learning method, meaning that it uses the entire training dataset to make predictions for new instances. This approach can make it challenging to generalize the insights gained from the model, as it does not generate a compact representation of the decision-making process.

k-NN can be computationally expensive, particularly when dealing with large datasets. The algorithm requires calculating the distance between a new instance and all instances in the training dataset, which can be time-consuming and resource-intensive. In contrast, Decision Trees and Random Forests are generally more computationally efficient, as they create a hierarchical structure during the training process that can be quickly traversed to make predictions for new instances.

As a result, while the k-Nearest Neighbors algorithm has been successfully employed in various classification and regression tasks, its application in pedestrian safety analysis may be limited due to its performance, interpretability, and scalability constraints. In comparison, the Decision Tree and Random Forest Classifier models offer a more transparent and easily interpretable structure for understanding the factors contributing to pedestrian safety, as well as better performance on high-dimensional and noisy data.

3.5.3. Neural Networks

Neural Networks, particularly deep learning models, have demonstrated outstanding performance in various complex classification and prediction tasks. In the pedestrian safety study, Neural Networks can be employed to identify patterns and relationships among variables that contribute to fatal and non-fatal collisions. In this section, we will provide a detailed comparison of Neural Networks-based pedestrian safety analysis with the Decision Tree approach.

Neural Networks, especially deep learning models, generally require large amounts of data to achieve optimal performance. The larger the dataset, the more effectively the model can learn complex patterns and generalize to unseen data. In contrast, Decision Trees can work well with smaller datasets, as they do not require the same level of data complexity to make accurate predictions. In the pedestrian safety analysis, obtaining large amounts of data may be challenging due to limitations in data collection and reporting. Therefore, Neural Networks may not be as effective as Decision Trees in situations where data availability is constrained.

Neural Networks consist of interconnected layers of artificial neurons that learn to make predictions or classify data points. The complexity of these models can range from simple single-layer networks to deep architectures with multiple hidden layers. The increased complexity of deep learning models allows them to capture more nuanced patterns in the data. In comparison, Decision Trees have a more straightforward structure, with a series of binary decisions based on feature values. While this simplicity can be advantageous for interpretability, it may limit the ability to capture complex relationships in the data.

Neural Networks can provide exceptional performance in various classification tasks, particularly when sufficient data and computational resources are available. However, their performance in pedestrian safety analysis may not surpass that of Decision Trees due to the limited size and complexity of the data. Furthermore, Neural Networks can be prone to overfitting, especially in cases where the data is noisy or contains irrelevant features. Decision Trees can also be prone to overfitting but can be easily pruned or combined with ensemble techniques like Random Forests to address this issue.

One significant drawback of Neural Networks is their lack of interpretability. Often referred to as "black boxes," understanding the decision-making process within these models can be challenging. This lack of transparency can be problematic in pedestrian safety analysis, where understanding the factors contributing to accidents is crucial for developing effective interventions.

In contrast, Decision Trees provide a more transparent and easily interpretable structure for understanding the factors contributing to pedestrian safety. Their hierarchical structure allows for a clear visualization of the decision-making process, making it easier to communicate results to stakeholders and policymakers.

Neural Networks have been successfully applied in various real-world applications, including image recognition, natural language processing, and self-driving cars. In the context of pedestrian safety, Neural Networks could be used for tasks such as predicting pedestrian behavior, detecting pedestrians in video feeds, or classifying accident severity based on sensor data.

However, the lack of interpretability of Neural Networks may limit their applicability in situations where understanding the underlying factors contributing to pedestrian safety is crucial. In such cases, Decision Trees may be a more appropriate choice due to their transparency and ease of interpretation.

While Neural Networks offer the potential for exceptional performance in complex classification tasks, their applicability in pedestrian safety analysis may be limited due to factors such as data requirements, model complexity, and interpretability. Decision Trees, on the other hand, provide a more transparent and interpretable structure that can effectively capture relationships among variables contributing to pedestrian safety.

In this section, we have compared the Decision Tree and Random Forest Classifier models used in our study with other machine learning approaches, including Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Neural Networks. Each method has its strengths and weaknesses in terms of performance and interpretability.

The Decision Tree model, which demonstrated the best performance among all models in our study, offers a transparent and easily interpretable structure for understanding the factors contributing to pedestrian safety. While other techniques, such as SVM and Neural Networks, can provide strong performance in various classification tasks, their applicability and interpretability may be limited in the context of pedestrian safety, especially when compared to the Decision Tree approach.

CHAPTER 4. REFERENCES

[1] Marshall, W.E. and Ferenchak, N.N. (2017). Assessing equity and urban/rural road safety disparities in the US. Journal of Urbanism: International Research on Placemaking and Urban Sustainability, 10(4), pp.422-441.

[2] Baireddy, R., Zhou, H. and Jalayer, M. (2018). Multiple correspondence analysis of pedestrian crashes in rural Illinois. Transportation research record, 2672(38), pp.116-127.

[3] Chen, Z. and Fan, W. (2019). Modeling pedestrian injury severity in pedestrian-vehicle crashes in rural and urban areas: mixed logit model approach. Transportation research record, 2673(4), pp.1023-1034.

[4] Awwad-Rafferty, R., Chang, K. and Brown, H. (2019). Reaching Out to Tribal Communities: Lessons Learned and Approaches to Consider.

[5] Iragavarapu, V., Carlson, P. and Schertz, G. (2015). Review of Tribal Transportation Safety. Transportation Research Record, 2531(1), pp.153-160.

[6] Quick, K. and Narváez, G. (2018). Understanding roadway safety in American Indian reservations: Perceptions and management of risk by community, tribal governments, and other safety leaders.

[7] Federal Highway Administration. (2012). Traffic Safety Facts "Rural and Urban Comparison". U.S. Department of Transportation.

[8] Washington Traffic Safety Commission. (2013). Washington State Strategic Highway Safety Plan 2013. Washington State Department of Transportation.

[9] Turner, S.M., Sener, I.N., Martin, M.E., White, L.D., Das, S., Hampshire, R.C., Colety, M., Fitzpatrick, K. and Wijesundera, R.K. (2018). Guide for scalable risk assessment methods for pedestrians and bicyclists (No. FHWA-SA-18-032). United States. Federal Highway Administration. Office of Safety.

[10] Snyder, M., & Knoblauch, R. (1971). Pedestrian Safety The Identification of Precipitating 6 Factors and Possible Countermeasures. US Department of Transportation.

[11] Hall, J. (1983). Pedestrian Accidents on Rural Highways. Transportation Research Record, 8 904, 46–50.

[12] Pedestrian Safety Prediction Methodology. (2008). National Academy of Sciences, 10 Engineering and Medicine.

[13] Bhuiyan, N. (2019). Enhancing Pedestrian Safety in Bangladesh. CONE 2019 Discipline Specific Proposal (Transportation Engineering). https://www.researchgate.net/publication/332057524

[14] Zajac, S., & Ivan, J. (2002). Factors Influencing Injury Severity of Motor-Vehicle Crossing Pedestrian Crashes in Rural Connecticut. Accident Analysis and Prevention, 35, 369–379. https://doi.org/10.1016/S0001-4575(02)00013-1

[15] Chen, Z., & Fan, W. (2019). Modeling Pedestrian Injury Severity in Pedestrian-Vehicle Crashes in Rural and Urban Areas: Mixed Logit Model Approach. Transportation Research Record, 2673(4), 1023–1034. https://doi.org/DOI: 10.1177/0361198119842825

[16] Baireddy, R., Zhou, H., & Jalayer, M. (2018). Multiple Correspondence Analysis of Pedestrian Crashes in Rural Illinois. Transportation Research Record, 2672(38), 116–127. https://doi.org/DOI: 10.1177/0361198118777088

[17] Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. Accident Analysis and Prevention, Vol. 112, 116-126. https://doi.org/10.1016/j.aap.2017.12.026

[18] Das, S., Le, M, & Dai, B. (2020). Application of machine learning tools in classifying pedestrian crash types: A case study. Transportation Safety and Environment, Vol. 2, No. 2, 106–119. https://doi.org/10.1093/tse/tdaa010

[19] Fitzpatrick, K., Avelar, R. and Turner, S.M., 2018. Guidebook on Identification of High Pedestrian Crash Locations (No. FHWA-HRT-17-106). United States. Federal Highway Administration. Office of Safety Research and Development.

[20] Jamali, A. and Wang, Y., 2018. Pedestrian Crash Hotspot Identification Using Two-Step Floating Catchment Area Method and Machine Learning Tools (No. 18-05575).

[21] Shrivastava, K. and Jathar, S., 2019. Implication of machine learning in automobile for enhancing passenger and pedestrian safety. International Journal of Management, IT and Engineering, 9(6), pp.106-117.

[22] Rahman, M.S., Abdel-Aty, M., Hasan, S. and Cai, Q., 2019. Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones. Journal of safety research, 70, pp.275-288.

[23] Highway Safety Information System (HSIS) Guidebook for State Data Files – Washington. (2014). Federal Highway Administration, U.S. Department of Transportation.

[24] Martell, M. (2021). Lecture material: CEE 415 – Machine Learning for Civil Engineers, University of Washington.