TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

THESIS NO.: T23/079

Evaluation of Pedestrian Gap Acceptance and Crossing Path Choice at Uncontrolled Midblock Crossings - A Case Study of Kamalpokhari & Mitrapark

by

Sudarshan Tamang

A THESIS

SUBMITTED TO THE DEPARTMENT OF CIVIL ENGINEERING

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE IN TRANSPORTATION ENGINEERING

DEPARTMENT OF CIVIL ENGINEERING

LALITPUR, NEPAL

MAY, 2026

COPYRIGHT

The author has agreed that the library, Department of Civil Engineering, Pulchowk Campus, Institute of Engineering may make this thesis freely available for inspection. Moreover, the author has agreed that permission for extensive copying of this thesis for scholarly purpose may be granted by the professor(s) who supervised the work recorded herein or, in their absence, by the Head of the Department wherein the thesis was done. It is understood that the recognition will be given to the author of this thesis and to the Department of Civil Engineering, Pulchowk Campus, and Institute of Engineering in any use of the material of this thesis. Copying or publication or the other use of this thesis for financial gain without approval of the Department of Civil Engineering, Pulchowk Campus, Institute of Engineering and author's written permission is prohibited.

Request for permission to copy or to make any other use of the material in this thesis in whole or in part should be addressed to:

Head
Department of Civil Engineering
Pulchowk Campus
Pulchowk, Lalitpur
Nepal

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

DEPARTMENT OF CIVIL ENGINEERING

The undersigned certify that they have read, and recommended to the Institute of Engineering for acceptance, a thesis entitled "Evaluation of Pedestrian Gap Acceptance and Crossing Path Choice at Uncontrolled Midblock Crossings - A Case Study of Kamalpokhari & Mitrapark" submitted by Sudarshan Tamang (079MSTRE023) in partial fulfillment of the requirements for the degree of Master of Science in Transportation Engineering.

Supervisor: Dr. Pradeep Kumar Shrestha
Department of Civil Engineering,
Institute of Engineering

External Examiner: Prof. Dr. Thusitha Chandani Shahi
Department of Civil Engineering, Nepal Engineering College

Program Co-ordinator: Anil Marsani
Coordinator: MSc in Transportation Engineering,
Department of Civil Engineering

Date: 11 July 2026

ABSTRACT

Pedestrians in Kathmandu frequently cross at uncontrolled midblock locations under heterogeneous mixed traffic conditions often without formal crossing facilities. This study evaluates pedestrian behavior at two such sites, Kamalpokhari and Mitrapark with a dual focus on accepted vehicular gap and crossing path choice. Video based observations of 950 pedestrian crossing events (460 at Kamalpokhari and 490 at Mitrapark) were used to extract pedestrian, traffic and contextual variables. Accepted gap was modeled using Multiple Linear Regression (MLR) and Generalized Additive Models (GAM) while crossing path was analyzed using Multinomial Logistic Regression (MNL) and CatBoost.

For accepted gap, both MLR and GAM performed strongly. The reduced MLR models achieved R^2 of 0.668 at Kamalpokhari and 0.734 at Mitrapark while the reduced GAM models slightly improved explanatory power with R^2 of 0.684 and 0.772 respectively. Accepted gap behavior was mainly influenced by vehicle speed, safety distance, average rejected gap, pedestrian speed, waiting time and number of crossing attempts.

For crossing path choice, the MNL models showed moderate explanatory fit with satisfactory predictive performance with test accuracies of 0.543 at Kamalpokhari and 0.612 at Mitrapark. In contrast, the reduced CatBoost models achieved slightly stronger results with test accuracies of 0.667 and 0.646 and macro F1-scores of 0.675 and 0.644 respectively. Overall, the findings show that conventional statistical models remain useful for interpretable analysis particularly for accepted gap modeling whereas CatBoost is more effective for classifying complex pedestrian crossing paths under mixed traffic midblock conditions.

Keywords: Pedestrian crossing path, gap acceptance, crossing path choice, Multiple Linear Regression, Multinomial Logistic Regression, Generalized Additive Model, CatBoost.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to our Program Coordinator Mr. Anil Marsani, Assistant Professor Dr. Pradeep Kumar Shrestha and Assistant Professor Dr. Rojee Pradhananga, whose guidance, encouragement, and support have been invaluable throughout the preparation of this thesis. I am equally grateful to all the course instructors in the Master of Science in Transportation Engineering program for their dedicated teaching and insightful feedback, which have greatly contributed to the development of this work. I would also like to express my sincere appreciation to Mr. Bikal Adhikari, whose sustained support throughout the course of this research, especially in Python coding and constructive feedback, played an important role in enhancing the quality of this thesis.

I would like to express my sincere gratitude to my peers for their invaluable support, insightful feedback and continuous encouragement throughout the development of this thesis. Their collaboration and camaraderie have been a source of inspiration, and I am grateful for the collective learning journey we have shared.

Name: Sudarshan Tamang

Roll No.: 079MsTrE023

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AADT	Annual Average Daily Traffic
ANN	Artificial Neural Network
CatBoost	Categorical Boosting
DoTM	Department of Transport Management
F1-score	Harmonic mean of precision and recall
GAM	Generalized Additive Model
IIA	Independence of Irrelevant Alternatives
JICA	Japan International Cooperation Agency
KMC	Kathmandu Metropolitan City
LOS	Level of Service
MLR	Multiple Linear Regression
MNL	Multinomial Logistic Regression
MTPD	Metropolitan Traffic Police Division
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
SPSS	Statistical Package for the Social Sciences
VIF	Variance Inflation Factor
VRUs	Vulnerable Road Users
WHO	World Health Organization

CHAPTER ONE: INTRODUCTION

1.1. Background

A person using modification aids for walking such as wheelchairs, walkers, canes, skateboards, etc. are also considered as pedestrian (Shahi & Gautam, 2020). In Kathmandu valley, about 40% of journeys are made on foot (JICA, 2012). Hence, pedestrians are an integral element of the transport system and are at a greater risk of being involved in a crash compared to other road users.

Historically, the focus on road safety has predominantly centered on vehicular traffic often disregarding pedestrian safety. However, the increase in pedestrian casualties has caused a shift compelling policymakers and researchers to prioritize pedestrians. Road accidents in Nepal have claimed thousands of lives with pedestrians accounting for a significant proportion of these tragedies. In 2016, Vulnerable Road Users (VRUs) accounted for around 72% of all road fatality victims in Nepal with pedestrian accounting for about half (The World Bank, 2020). Similarly, approx. 23% of victims of fatal crash are pedestrians in Kathmandu (MTPD, 2019).

As per the traffic police record, 16388 pedestrians were caught who violated the rules in the first two weeks of June 2019 (Shahi & Gautam, 2020). Very often, pedestrians choose the shortest and most direct path to reach their destination crossing the roads. Studies have shown that pedestrians crossing at uncontrolled midblock crosswalks are dangerous in developing countries due to the lack of traffic lights and low yielding rate of vehicles (Aziz, Ukkusuri, & Hasan, 2013; Kadali & Vedagiri, 2016). The crossing maneuver is largely dependent on the pedestrian assessment and perception of the surrounding condition. They take calculated risks while crossing the road which may be different for different pedestrians. Their movement across the road intersects the vehicular path offering high risk of conflict with vehicles. If the pedestrian underestimates the risk then there is a greater chance of a pedestrian-vehicle interaction (Shaaban, Muley, & Mohammed, 2018).

Pedestrians tend to feel less secure when crossing unmarked roadways compared to marked ones as the presence of a crosswalk provides a greater sense of safety and confidence (Zhuang & Wu, 2011). Midblock locations are roadway segments where pedestrians cross away from designated intersections or marked crosswalks. These locations often lack traffic control devices resulting in unregulated interactions between pedestrians and vehicles. At such locations, pedestrians must rely on their own judgment to identify acceptable gaps in traffic which increases the risk of conflict. Factors such as limited visibility due to parked vehicles or roadside obstructions, high vehicle speeds and unpredictable driver behavior further increase the complexity of crossing decisions. Pedestrian behavior at midblock crossings is influenced by urgency, convenience and familiarity with the area while characteristics such as age, gender and physical ability also affect crossing decisions. These conditions make midblock crossings particularly challenging and unsafe in mixed traffic environments.

Thus, pedestrian safety improvement in urban roads demands the understanding of pedestrian gap acceptance behavior as well as the choice of crossing path pedestrians take to cross the road at midblock. Once both the pedestrian gap acceptance behavior as well as choice of crossing paths is understood, it can be used to enhance the existing crosswalk facilities, to support the design of safer crossing facilities and inform the implementation of appropriate traffic control measures.

This study integrates both statistical models such as Multiple Linear Regression (MLR) and Multinomial Logistic Regression (MNL) and machine learning models such as Generalized Additive Model (GAM) and CatBoost to investigate gap acceptance and crossing path in Kathmandu where midblock crossings remain largely uncontrolled and unpredictable.

1.2. Problem Statement

Pedestrian safety remains a pressing concern globally and the problem is especially acute in Nepal. Uncontrolled midblock crossings which lack traffic signals or pedestrian infrastructure are hotspots for accidents. This study explores whether machine learning can better capture the complexities of pedestrian behavior and deliver practical insights for urban safety improvements. Studies from Indian urban corridors have shown that frequent unsignalised midblock crossings reduce average midblock speeds and create operational disturbances for vehicles, underscoring the system-wide impact of pedestrian gap forcing under mixed traffic (Kadali & Vedagiri, 2016). Kadali, Chiranjeevi, & Rajesh (2015) likewise reported that unsignalized midblock pedestrian crossings can significantly influence vehicular speeds and interrupt normal traffic progression, emphasizing that pedestrian crossing path at such locations has important implications not only for safety but also for roadway operational performance.

According to available traffic reports, over 30% of pedestrian accidents in Kathmandu occur at uncontrolled midblock crossings. Studies have shown that the absence of proper crossing facilities significantly increases the risk of pedestrian injuries and fatalities. Experimental evidence from (Oxley, Ihsen, Fildes, Charlton, & Day, 2005) further indicates that older pedestrians' walk more slowly, require longer decision times and frequently accept smaller safety margins which increase their exposure to risk when crossing in dense traffic. The high rate of pedestrian accidents at these crossings not only affects public health but also hampers urban mobility and safety. Addressing this issue is crucial for creating safer urban environments and promoting sustainable transportation.

While several studies have examined pedestrian behavior at crossing locations, most rely on conventional statistical models to analyze gap acceptance and crossing path choice. Although these models are valuable for their interpretability, they assume linear relationships and may not fully capture the complex, non-linear interactions present in mixed traffic conditions. In this study, both statistical and machine learning models are applied using a range of pedestrian, traffic and environmental variables. Machine learning approaches are explored to better represent these interactions and improve predictive performance for gap acceptance and crossing path choice at uncontrolled midblock locations.

1.3. Objectives Of Study

The main objective of this study is the evaluation of pedestrian gap acceptance and crossing path choice at uncontrolled midblock crossings.

The specific objectives of this study are as follows:

i. To identify the factors influencing pedestrian crossing path at uncontrolled midblock crossings with particular focus on accepted gap and crossing path choice.
ii. To develop gap acceptance model and predict the size of the accepted gap.
iii. To develop crossing path model and predict the path taken by the pedestrians.

1.4. Scope of Study

The scopes of the study are as follows:

i. The study focuses on pedestrian crossing behavior at uncontrolled midblock crossings.
ii. Data collection was conducted through videographic survey at the Kamalpokhari and Mitrapark segments.
iii. The analysis utilizes a dataset of 950 pedestrian crossing events recorded during weekday morning peak hours.
iv. The study examines the influence of pedestrian, traffic and environmental variables on accepted gaps and crossing path choices.
v. Accepted gap was modeled using Multiple Linear Regression (MLR) and Generalized Additive Model (GAM) while crossing path was modeled using Multinomial Logistic Regression (MNL) and CatBoost.

1.5. Limitations of Study

The limitations of the study are as follows:

i. The study is based on two selected midblock locations in Kathmandu and therefore may not capture the full range of variation in traffic conditions,

pedestrian behavior, roadway geometry and roadside environment present at other midblock sites. Also, the findings should be interpreted as representative of similar uncontrolled urban midblock environments rather than all midblock crossings.

ii. Data were collected during daytime and under favorable weather conditions only and pedestrian behavior during nighttime or adverse weather was not analyzed.

iii. Only pedestrian related variable were considered and does not include vehicle related factors such as driver behavior, speed control or responses to pedestrian actions.

1.6. Organization of Report

The report consists of total five chapters which are listed as follows:

Chapter 1: Introduction

It briefly explains the pedestrians' behavior, challenges faced by them along with the problem statement, the research objectives, the study's aims, scope and limitations.

Chapter 2: Literature Review

It discusses the relevant literature of pedestrian behavior along with previous researchers' efforts regarding pedestrian gap acceptance and crossing path analysis.

Chapter 3: Research Methodology

It outlines the methodological approach for this research which includes the site selection, data collection and extraction, variables selection and analysis of the data used in this study.

Chapter 4: Results and Discussion

It shows an overview of the data, results obtained from modeling the extracted data, as well as the validation results of all the developed models which includes MLR model, MNL model, GAM model and CatBoost model.

Chapter 5: Conclusion and Recommendation

It provides a concise summary of the findings from the obtained results and provides recommendations for future research designs.

CHAPTER TWO: LITERATURE REVIEW

2.1. Pedestrian Behavior Studies

Several researchers have attempted to model pedestrian gap acceptance and choice of crossing path using different modeling techniques. These studies were mainly focused on modeling various aspects of pedestrian behavior, vehicular gap size accepted and choice of crossing path and decision for crossing the road. In these studies, the variables are selected depending on the site location, traffic characteristics and scope of their research work. Several such literatures have been studied and reviewed. A broader critical review of pedestrian behavior models covering gap acceptance, route choice and facility characteristics is provided by Papadimitriou, Yannis, & Golias (2009), who argue that linking microscopic behavior with safety outcomes is essential for designing effective pedestrian facilities. Govinda & Ravishankar (2022) also presented a critical review of pedestrian crossing path and pedestrian-vehicle interactions highlighting that crossing speed varies with crossing location and that improved modeling of pedestrian-vehicle interaction is essential for addressing pedestrian safety under mixed traffic conditions.

Evan & Norman (1998) used the Theory of Planned Behavior (TPB) to study pedestrians' road crossing decisions. The TPB suggests behavior is influenced by attitudes, subjective norms and perceived control. Survey data showed these factors significantly predicted pedestrians' intentions to cross at signalized crossings. Additionally, perceived risk and the number of waiting cars impacted crossing path. The study concludes that the TPB is valuable for understanding and predicting pedestrian behavior, recommending interventions to promote safe crossings by influencing attitudes, norms and control.

Cherry, Donlon, Yan, Moore, & Xiong (2012) investigates illegal midblock pedestrian crossings in China focusing on gap acceptance and crossing path choices. It highlights that increased urbanization and motorization in China have led to conflicts between pedestrians and vehicles particularly at midblock crossings on long urban blocks known as superblocks. The study uses a probit discrete outcome model to analyze factors

influencing gap acceptance, such as gap size, vehicle speed, wait time and gap lane position while noting that group presence or crossing from the roadside or median did not significantly affect gap acceptance. The average accepted gap was found to be 8.8 seconds with a rejected gap averaging 5.3 seconds which is higher than previous studies. The study also notes that pedestrians employ various crossing strategies including crossing diagonally or lane by lane and sometimes retreating to wait for a sufficient gap. The findings suggest that reducing vehicle speeds could increase the probability of accepting smaller gaps and reduce crash severity. Complementary experimental work by Oxley, Ihsen, Fildes, Charlton, & Day (2005) showed that older pedestrians tend to walk more slowly, take longer to make crossing decisions and often accept shorter time gaps resulting in smaller safety margins compared with younger adults, which underscores the heightened vulnerability of certain road user groups.

Zhuang & Wu (2011) discusses pedestrian crossing paths focusing on factors influencing gap acceptance and the choice of crossing paths. It highlights that pedestrians' crossing decisions are influenced by characteristics such as age and gender with females and older pedestrians generally waiting longer at crosswalks. The study also emphasizes the importance of gap acceptance theory where pedestrians assess whether the gap between vehicles is sufficient for safe crossing. Additionally, it notes that pedestrians perceive crossing roadways as more challenging than using crosswalks indicating a preference for marked paths when available. These findings contribute to understanding pedestrian behavior in selecting crossing paths and evaluating gaps in traffic.

Yannis, Papadimitriou, & Theofilatos (2013) investigated pedestrian gap acceptance and midblock crossing path in urban Athens focusing on a less pedestrian friendly road environment. Using field observations and regression models, the study identified key factors influencing gap acceptance and crossing decisions. The study found that pedestrians’ gap acceptance was strongly influenced by the distance of incoming vehicles rather than their speed. Other factors such as vehicle size, illegal parking and pedestrian characteristics such as gender and accompaniment also played significant roles. Binary logistic regression revealed that traffic gaps, waiting times, vehicle type and illegal parking significantly affected the likelihood of pedestrians crossing the road. Longer waiting times were associated with more cautious behavior as pedestrians preferred to wait for safer gaps. The findings emphasize the importance of road and

traffic conditions in shaping pedestrian behavior and suggest that addressing factors like illegal parking and visibility could improve pedestrian safety. Similarly, Hamed (2001) analyzed pedestrian behavior at unsignalised midblock locations in Amman and reported that accepted gaps, waiting time and the number of rejected opportunities are closely interrelated with longer delays often prompting pedestrians to accept shorter and riskier gaps.

Shaaban, Muley, & Mohammed (2018) examined pedestrian crossing path on a six lane arterial road in Qatar focusing on illegal crossings and decision making under high speed traffic. The study revealed that demographic factors, group size and traffic conditions influenced gap acceptance and crossing path choices. Pedestrians adjusted walking speed and waiting times based on available gaps particularly during group crossings or when starting from the median. Over 50% of pedestrians opted for the shortest path (perpendicular crossing) influenced by the desire to minimize exposure to traffic. Conflicts with vehicles and low driver yielding rates significantly impacted pedestrian decisions leading to risky crossing paths. The findings highlight the importance of understanding pedestrian decision making to design interventions that improve safety such as creating safer crossings and educating pedestrians on gap selection. Tezcan, Elmorssy, & Aksoy (2019) analyzed pedestrian crossing path at four marked midblock crosswalks on one-way urban streets in Istanbul using detailed video observations and multinomial logit models. Two hours of recording at each site were used to extract information on pedestrian platooning at the curbside, individual crossing decisions, traffic volume, crosswalk occupancy, illegal parking and pedestrian characteristics such as age, gender and distraction status. The results showed that the likelihood of pedestrians forming platoons increases with traffic volume and platoon size and those individuals who started crossing with little or no waiting time while one or more lanes were already occupied tended to lose time during the crossing potentially increasing their exposure to risk. The authors concluded that under non-yielding driver behavior, the safety benefits of marked midblock crosswalks are questionable and that interventions such as enforcement or demand responsive signals may be required to manage pedestrian streams effectively. Zhang, Li, Sze, & Ren (2023) further examined pedestrian crossing route choice at midblock locations without crossing facilities and reported that roadside environmental conditions play an important role in shaping the path selected by pedestrians. Their findings reinforce the importance of contextual

factors such as roadside obstructions and surrounding street environment when analyzing crossing path decisions at uncontrolled midblock locations. Similarly, Muraleetharan & Hagiwara (2007) linked perceived level of service on sidewalks and crosswalks to factors such as obstructions, opposing flows and turning vehicles while Rastogi, Thaniarasu, & Chandra (2011) observed that pedestrian walking speeds, group composition and roadway geometry strongly influence midblock crossing decisions. Kadali & Vedagiri (2016) further examined unsignalised midblock crossings in India and reported that pedestrian gap forcing behavior significantly reduced midblock operating speeds highlighting the combined safety and efficiency implications of uncontrolled crossings in mixed traffic environments. Torres, et al. (2020) evaluated pedestrian behavior at four types of midblock crossing facilities in Fortaleza, Brazil using logistic regression models and found that raised crosswalks and signalized treatments substantially reduced aggressive or risky pedestrian crossings while raised facilities also increased driver yielding behavior. Their findings highlight the importance of crossing facility design in shaping pedestrian safety and behavior at midblock locations. Alver, Onelcin, Cicekli, & Abdel-Aty (2021) investigated pedestrian critical gap and crossing speed at two midblock crossings in Izmir, Turkey using image processing techniques and reported critical gaps ranging from 4.1 s to 6.2 s with 15th percentile crossing speeds between 0.78 m/s and 0.80 m/s. Their results further showed that accepted gap was significantly influenced by roadway context and vehicle type, reinforcing the importance of site specific behavioral analysis at uncontrolled midblock crossings. These studies support the importance of context specific behavioral analysis for densely trafficked urban streets.

2.2. Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) has been a widely used statistical method for modeling continuous outcomes particularly the size of accepted gaps by pedestrians (Shaaban, Muley, & Mohammed, 2018); (Kadali & Vedagiri, 2016). Studies applying MLR to accepted gap size typically reported R^2 values ranging approximately between 60% and 80%. While MLR's simplicity in application and the straightforward interpretation of its coefficients have made it a popular choice, this range of R^2 values indicated that a significant portion of the variability in accepted gap sizes often remains

unexplained indicating that its inherent assumption of linearity may not fully capture the complex, non-linear relationships in pedestrian decision making.

For modeling the binary decision of whether a pedestrian accepts or rejects a gap, Binary Logistic Regression has been frequently employed (Zhang, Zhou, Qiu, & Liu, 2018); (Yannis, Papadimitriou, & Theofilatos, 2013). These models generally achieved high prediction accuracies often ranging from 85% to 100%. However, a key limitation of binary logistic regression is its inability to quantify the actual accepted gap size in seconds or to predict specific, multi class crossing path choices focusing solely on the binary decision. For instance, Kadali & Vedagiri (2013) modeled minimum accepted gap at an uncontrolled midblock location using MLR in combination with a binary logit formulation for gap acceptance probability demonstrating that pedestrian characteristics and traffic conditions can be consistently incorporated within a unified regression and discrete choice framework.

To overcome some of these limitations and enhance predictive accuracy, researchers have explored more advanced machine learning models. For instance, Kadali & Vedagiri (2016) applied Artificial Neural Networks (ANN) to model accepted gap size obtaining a model with an R^2 value of approximately 85% which represents an improvement over typical MLR performance. Their study introduced the concept of Pedestrian Safety Margin at unprotected midblock crosswalks and showed using both MLR and artificial neural networks that variables such as vehicle speed, traffic volume, pedestrian rolling behavior and speed change are critical in explaining unsafe gap acceptance decisions. Similarly, when ANN was utilized for pedestrian behavior modeling, an accuracy of 98% was reported. The complex and often non-linear nature of pedestrian behavior coupled with the prevalence of diverse influencing factors including categorical variables highlights the continuous need for robust and interpretable modeling approaches that can capture these intricacies effectively.

2.3. Multinomial Logistic Regression (MNL)

For modeling the categorical decision of a pedestrian's crossing path (e.g., perpendicular, oblique, irregular), Multinomial Logistic Regression (MNL) has been a frequently employed statistical method. This approach extends binary logistic regression to handle dependent variables with more than two nominal categories allowing

researchers to predict the probability of a pedestrian choosing a specific path among several options based on various independent variables (Agresti, 2007). Studies investigating pedestrian behavior such as those by Shaaban, Muley, & Mohammed (2018) and Zhuang & Wu (2011) have highlighted the influence of factors like demographic characteristics, group size and traffic conditions on crossing path choices. Tezcan, Elmorssy, & Aksoy (2019) estimated separate multinomial logit models for pedestrian platooning decisions and individual crossing path at midblock crosswalks in Istanbul using waiting time, platoon size, traffic conditions and distraction as explanatory variables. Their work demonstrates the suitability of MNL for representing discrete pedestrian responses to complex midblock traffic environments which motivates the use of MNL in the present study to model crossing path choice at uncontrolled midblock locations in Kathmandu.

Multinomial Logistic Regression provides interpretable coefficients that indicate how a change in an independent variable affects the log odds of choosing one specific path relative to a reference path. This interpretability has made it a valuable tool for understanding the drivers behind different pedestrian behaviors. While specific accuracy ranges for multinomial logistic regression models applied to pedestrian path choice are less consistently reported than for binary decisions, these models are generally robust for multi class classification when the relationships are predominantly linear and the assumption of independence of irrelevant alternatives (IIA) holds. However, for highly complex interactions or when dealing with a large number of categorical features, more advanced machine learning algorithms may offer enhanced predictive power and flexibility.

2.4. Generalized Additive Model (GAM)

Generalized Additive Models (GAM) represents a more flexible statistical approach for modeling continuous outcomes particularly when relationships between predictors and the response are expected to be non-linear. Unlike MLR, GAMs have been explored in behavioral modeling due to their flexibility in handling non-linear relationships between variables. Their potential application in modeling pedestrian gap acceptance lies in their ability to capture complex interactions often not explained by linear models (Hastie & Tibshirani, 1986); (Wood, 2017). This capability is particularly valuable in behavioral

studies where human responses such as accepted gap sizes may not change linearly with factors like vehicle speed or waiting time (e.g., a pedestrian's sensitivity to vehicle speed might diminish at very high speeds).

While specific applications of GAM directly to pedestrian accepted gap sizes might be less prevalent in the core literature compared to linear models, their utility has been demonstrated in various fields requiring flexible modeling of continuous, often skewed outcomes. For instance, GAMs have been successfully applied in transportation research to model traffic flow dynamics or driver behavior where non-linear effects are common. GAM can handle different types of data, including skewed values like gap sizes and helps show how each variable affects the outcome through partial effect plots. This makes it useful for identifying patterns that may not be captured by simple linear models. This flexibility allows for a more accurate and robust understanding of how various pedestrian, vehicle and environmental factors collectively influence the decision to accept a specific gap size.

For GAM, early applications by Xie & Zhang (2008) illustrated how the model outperformed generalized linear models in crash frequency analysis by uncovering non-linear relationships between traffic exposure variables and crash counts. Similarly, Zhang, Xie, & Li (2012) extended the use of GAM to crash severity analysis showing that the flexibility of smooth functions allowed for better insights into how road conditions and flow patterns affect crash outcomes. They further demonstrated the usefulness of GAM in roadway safety analysis by showing that crash frequency exposure relationships may vary by roadway segment type and may not always follow simple monotonic forms. Their findings also indicated that GAM generally provides greater flexibility and slightly better modeling performance than generalized linear models. More recently, Laflamme, Villamagna, & Kim (2024) applied GAM to crash severity prediction and confirmed that combining additive smooth functions with linear terms leads to models that are both interpretable and statistically robust making GAM a strong candidate for safety modeling where threshold effects and gradual behavioral shifts exist. These studies collectively support the use of GAM for accepted gap modeling in this research as pedestrian decisions are rarely linear and often shaped by subtle variations in speed, distance and waiting time.

In transportation safety research, GAMs have been employed to analyze non-linear influences on driving speeds, signal compliance and pedestrian delays (e.g., Schepers et

al., 2017; Gao & Lee, 2020). While applications to pedestrian gap acceptance remain limited, studies have confirmed GAM ability to flexibly model skewed, non Gaussian outcomes like accepted gap sizes under varying traffic scenarios.

2.5. CatBoost

CatBoost is a machine learning algorithm developed by Yandex that has emerged as a powerful tool for analyzing tabular datasets with both categorical and numerical variables (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2018). It has shown promise in behavioral modeling though its application in pedestrian behavior studies is still emerging. In this study, CatBoost is used to predict crossing path choices as it can effectively handle multiple influencing factors such as gender, phone use, vehicle type, etc. without requiring manual encoding. Its strength lies in uncovering non obvious patterns from real world data.

While direct applications of CatBoost specifically to pedestrian crossing path choice are still emerging in academic literature, its effectiveness has been widely demonstrated in various multi class classification problems across diverse domains including risk prediction, fraud detection and other behavioral modeling tasks where complex, non-linear relationships and numerous categorical features are present. For predicting the categorical crossing path chosen by pedestrians (e.g., Perpendicular, Oblique, Irregular, etc.), CatBoost's ability to automatically process and leverage the information from categorical variables like gender, mobile phone use, vehicle type, roadside obstructions, etc. without extensive manual preprocessing makes it highly advantageous. Its robustness to overfitting and superior performance in capturing intricate patterns and interactions among variables positions CatBoost as a strong candidate for developing accurate and reliable predictive models for pedestrian behavior.

Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin (2018) formally introduced the algorithm and demonstrated how ordered boosting and categorical encoding reduce bias and overfitting laying the foundation for its widespread adoption. Building on this, (Li, Wu, Bai, & Zhang (2023) employed a ReMAHA-CatBoost framework to address class imbalance in crash severity data showing that the model could reliably identify severe crashes that traditional methods often misclassified. Samerei, Aghabayk, & Montella (2024) combined CatBoost with SHAP values to study pile up crash severity

highlighting its dual advantage of predictive accuracy and interpretability, a critical feature for policy applications. Zhao, Qi, Yao, Guo, & Su (2023) further demonstrated CatBoost's strength in traffic safety modeling by applying it to analyze the influence of driver and roadway features on crash outcomes confirming its ability to handle mixed categorical and continuous predictors. Taken together, these works reinforce CatBoost's suitability for modeling crossing path in this study where interactions between pedestrian characteristics, vehicle dynamics and contextual factors are complex and non-linear. Recent studies such as Lin et al. (2021) and Abeykoon et al. (2023) have applied CatBoost in transport mode choice, driver behavior prediction and crash severity classification. Its efficient handling of mixed type data and non-linear interactions makes it a suitable candidate for modeling pedestrian path choices where both behavioral and infrastructural variables are influential. Singh, Das, & Ghosh (2024) analyzed pedestrian crossing path at unsignalized intersections using several machine learning algorithms and found that random forest achieved the highest prediction accuracy (81.72%) followed by XGBoost (77.19%) and binary logit (74.95%). Their findings further showed that variables such as arrival order, pedestrian delay, vehicle speed, pedestrian speed, age, gender, traffic hour and vehicle category were influential in predicting crossing path reinforcing the value of advanced machine learning approaches for behavior classification. Zhang, Sprenger, Ni, & Berger (2024) further demonstrated the usefulness of machine learning approaches for pedestrian behavior prediction at unsignalized crossings by showing that neural networks achieved the best performance for both accepted gap prediction and zebra crossing usage. Their findings also indicated that waiting time, walking speed, missed gaps and group behavior are influential factors in pedestrian crossing decisions.

While CatBoost excels in capturing non-linear relationships and handling categorical variables efficiently, it lacks the interpretability of simpler models such as MNL. This tradeoff between performance and explainability should be considered when interpreting the results. Abeykoon et al. (2023) demonstrated CatBoost's effectiveness in crash severity classification, suggesting its strong potential for pedestrian behavior prediction as well.

2.6. Summary and Research Gap

This literature review has examined various modeling approaches for pedestrian behavior. Multiple Linear Regression (MLR) has been widely used to predict accepted gap sizes typically showing R^2 values of 60-80%. While simple and interpretable, its linear assumption may not fully capture the complex, non-linear dynamics of pedestrian decisions. For predicting categorical crossing path choices, Multinomial Logistic Regression has been employed offering interpretability for multi class outcomes. However, like MLR, it assumes linear relationships and may struggle with highly complex interactions or numerous categorical features.

To address these limitations, more flexible and robust models such as Generalized Additive Models (GAMs) offering the ability to capture non-linear relationships for continuous outcomes like accepted gap size providing a more accurate fit than MLR and for multi class classification, CatBoost known for handling categorical data efficiently provides a robust alternative to traditional models especially when predicting pedestrian crossing paths that involve complex behavioral and environmental interactions. Govinda and Ravishankar (2022) similarly emphasized that pedestrian crossing behavior and pedestrian-vehicle interactions are shaped by a combination of pedestrian, traffic, geometric and environmental factors, and highlighted the need for improved analytical approaches to better understand pedestrian safety at crossing locations.

Despite the growing use of advanced modeling techniques, limited research has applied both interpretable non-linear models and machine learning classifiers simultaneously to study pedestrian behavior in Nepalese urban conditions. This study addresses this gap by integrating statistical and machine learning approaches to provide both explanatory and predictive insights into pedestrian gap acceptance and crossing path behavior at uncontrolled midblock locations in Kathmandu.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1. Research Methodology

All the prior studies were meticulously studied to identify the crucial factors required to be considered for the intended analysis. The framework outlines the overall research process, beginning with site selection and data collection followed by data processing and analysis. It then proceeds to model development for both gap acceptance and crossing path analysis along with validation of the developed models and concludes with interpretation and reporting of results. This structured approach ensures a systematic progression from data acquisition to model evaluation. The adopted methodological framework of the study is shown in Figure 3.1.

Research variables were identified considering the study's objectives and relevant literature studies. Subsequently, the study areas were assessed. A target sample size was identified based on practical feasibility and prior studies to ensure sufficient observations for meaningful analysis. Following that, a video-graphic survey was conducted at the selected crosswalk locations and the video recordings were used to gather the data. All relevant characteristics were extracted from the footage and recorded on observational sheets for further analysis and interpretation.

To understand the factors influencing pedestrian crossing behavior, multiple modeling approaches were employed, including traditional statistical methods such as Multiple Linear Regression (MLR) and Multinomial Logistic Regression (MNL) as well as machine learning models such as GAM and CatBoost. This approach enabled a comparative evaluation of the models to assess their ability to capture the complexity of real world pedestrian crossing behavior.

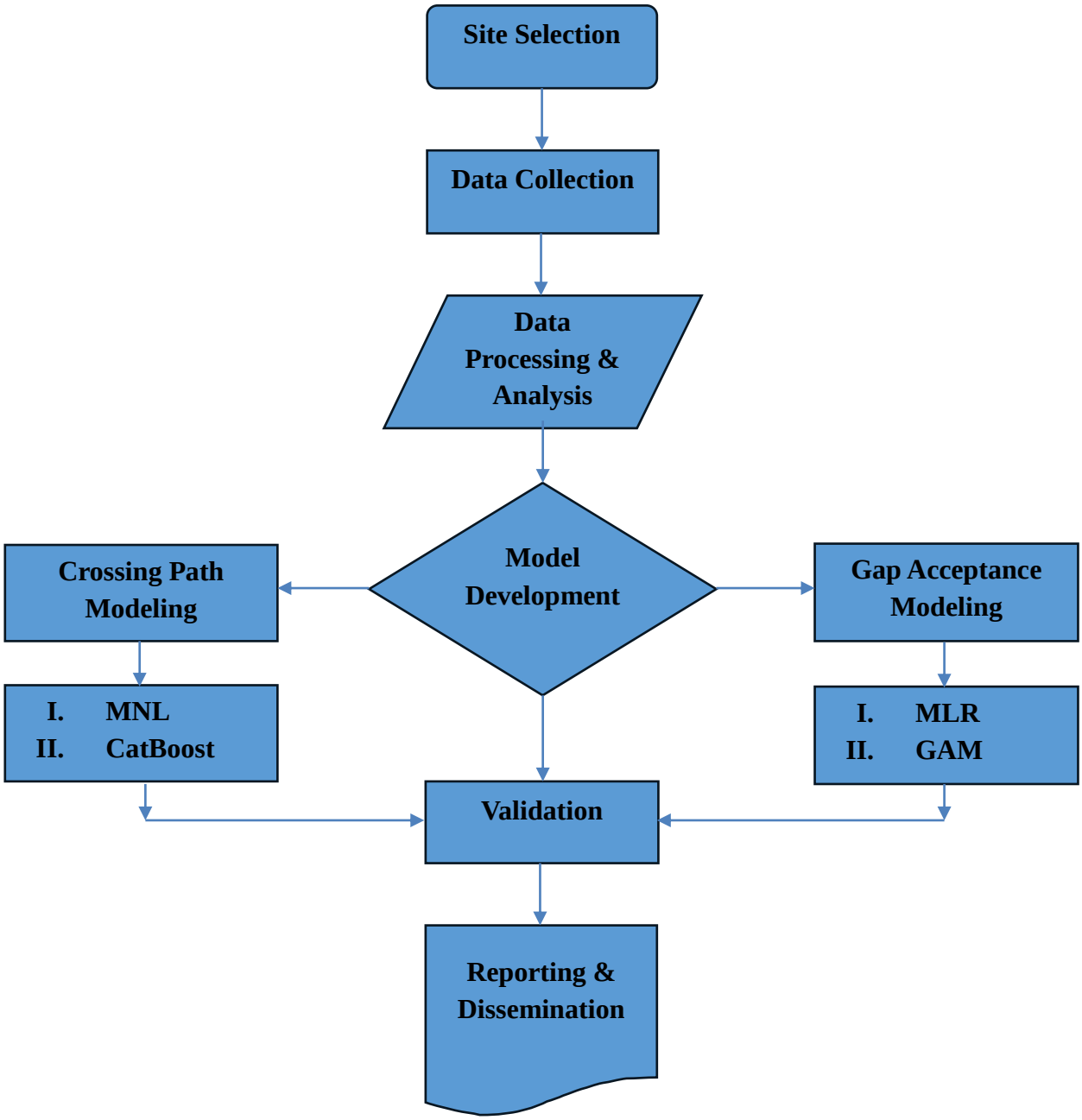


Figure 3.1. Framework of Research Methodology

3.2. Study Area

In order to finalize the study locations to analyze pedestrian gap acceptance behavior and crossing path choice, a short pilot survey was conducted to assess several potential midblock sites within Kathmandu. Each site was observed for approximately 15-30 minutes during selected daytime periods to identify locations with relatively high pedestrian crossing activity. During the pilot survey, key indicators such as pedestrian volume, frequency of midblock crossings, vehicle flow characteristics, pedestrian-vehicle interaction patterns and visibility constraints were recorded. The observations helped confirm whether the locations experienced frequent uncontrolled crossings with minimal traffic control measures.

Based on these criteria, the uncontrolled midblock sections at Kamalpokhari and Mitrapark were selected for further study of gap acceptance and crossing path as they exhibited high pedestrian demand, recurring conflict-prone crossing behavior and mixed traffic flow conditions representative of typical urban midblock environments in Kathmandu. Figure 3.2 shows the locations of the selected midblock sections.

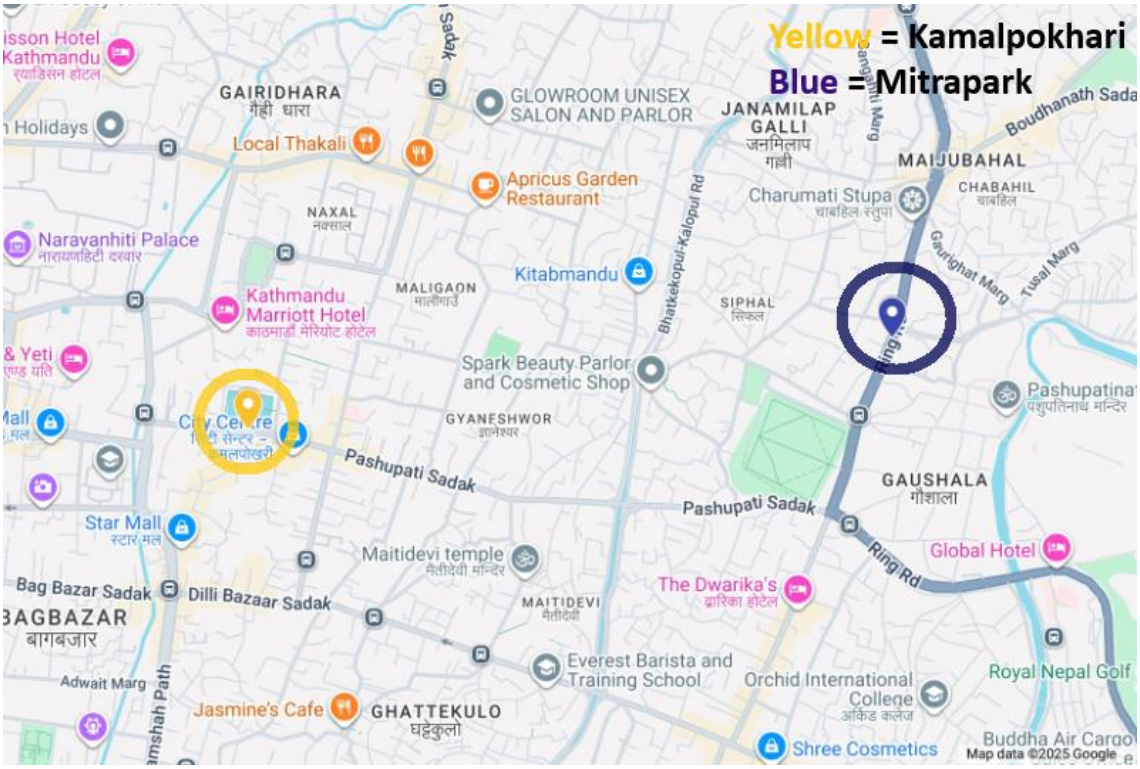


Figure 3.2. Location of the selected midblock sections

The Kamalpokhari midblock segment had a total length of 370m between the Hattisar and Gyaneshwor intersection. The roadway was a four-lane two-way road with a width of 15 m where the travel directions are separated by a solid median line and individual lanes demarcated by dashed white lines. There were three crosswalks present along the entire length of the midblock. For the purpose of this research, the section located between the western crosswalk near the Hattisar intersection and central marked crosswalks which had a distance of 170m between them was selected for observation. While the physical distance between the two crosswalks was 170m, the actual analysis was focused on a specific sub-segment within the camera's optimal field of view to ensure maximum visibility and data accuracy. It lacked physical barriers such as pedestrian railings or a raised median, which allows for unrestricted and informal crossings. Video data was captured from an elevated position on a nearby commercial building near the central crosswalk oriented towards the general direction of the western crosswalk providing a clear view of the roadway. The sidewalk along one side of the road was in satisfactory condition while the opposing side were deteriorated and discontinuous.

Similarly, the Mitrapark midblock segment had a total length of 600m situated between the Gaushala and Guheswori intersections. Similar to the first site, the roadway width was 15 m. While there were five marked crosswalks along the total length of this segment, the observation for this research focused on a specific segment that was bounded by two marked crosswalks separated by a distance of 175m. While the physical distance between the two crosswalks was 175m, the actual analysis was focused on a specific sub-segment within the camera's optimal field of view to ensure maximum visibility and data accuracy. Pedestrian railings were present throughout the midblock. However, these barriers were discontinuous or damaged in several locations, allowing pedestrians to enter the roadway at informal points. Video data was captured from an elevated commercial building with the camera oriented toward the crosswalk in the general direction of the Gaushala intersection to ensure a clear view of the roadway. Unlike the Kamalpokhari site, the pedestrian footpaths on both sides of this specific section were in satisfactory condition.

3.3. Data Collection and Extraction

Videographic surveys were conducted at the midblock sections of Kamalpokhari and Mitrapark during typical weekday and clear weather conditions. Based on the pilot survey, the period from 9:00AM to 11:00AM was selected because it showed relatively high pedestrian crossing activity and frequent pedestrian-vehicle interactions at both sites. As this window represents the site's most challenging traffic conditions, it serves as a reliable baseline, such that modeling behavior during these high-pressure hours ensures the findings are robust enough to cover off-peak periods where crossing is generally less complex. Furthermore, focusing on the morning peak ensured consistent daylight and visibility for consistent measurement of pedestrian and vehicular movements. Video recording was carried out for 2 hours per day over 4 days at Kamalpokhari and 3 days at Mitrapark resulting in a total of 7 survey days and 840 minutes of footage. The difference in the number of survey days by site was intended to obtain an adequate number of usable pedestrian crossing observations under comparable traffic conditions. The final dataset consisted of 950 pedestrian crossing observations, including 460 from Kamalpokhari and 490 from Mitrapark. As the total pedestrian population at the study sites was large and undefined, an infinite population assumption was adopted. Using a standard sample size formula with a 95% confidence level and 5% margin of error, the required sample size was estimated at approximately 384 observations. While the minimum requirement for an infinite population is 384, a larger sample was used to support the complexity of the GAM and CatBoost models. This ensures the models can capture non-linear behavioral patterns without overfitting and provides enough data for statistically significant analysis of less frequent sub-groups, such as 'Irregular' crossing paths.

The data were extracted from the recorded videos using Kinovea v.2023.1.2, a frame-by-frame motion analysis software. It was selected because its frame-by-frame capability allowed precise extraction of temporal and spatial variables from both pedestrian and vehicle movements including pedestrian speed, vehicle speed, waiting time, accepted and rejected gaps and safety distance. This enabled the timings and relative positions of pedestrians and approaching vehicles to be measured more accurately from the video footage. It is an open-source motion analysis software capable of analyzing video recordings with time precision up to 0.01 seconds. Previous research

in biomechanics and transport behavior analysis has reported that Kinovea provides measurement errors typically within ±1-2% when compared with professional motion analysis systems (Puig-Divi, Escalona-Marfil, Padullés-Riu, Busquets, Padullés-Chando, & Marcos-Ruiz, 2019). Therefore, its use in this study provided a reliable basis for extracting the temporal and spatial variables required for the analysis. All variables were extracted using the next frame option in the software and the videos were analyzed at the millisecond level to reduce error in time measurement. The extracted data were then entered manually into MS Excel using a structured data entry form to ensure consistency and minimize errors during the data recording process.

3.4. Variables Selection

The study was conducted based on the data extracted from the video-graphic survey on the selected midblock locations. The choice of variables for the prediction of pedestrian gap acceptance and choice of crossing path depends upon a number of factors such as traffic conditions, road geometry, etc. These variables were selected based on extensive literature review and preliminary field observations at the selected midblock sections. The outcome (dependent) and predictor (independent) variables were extracted from the video recordings.

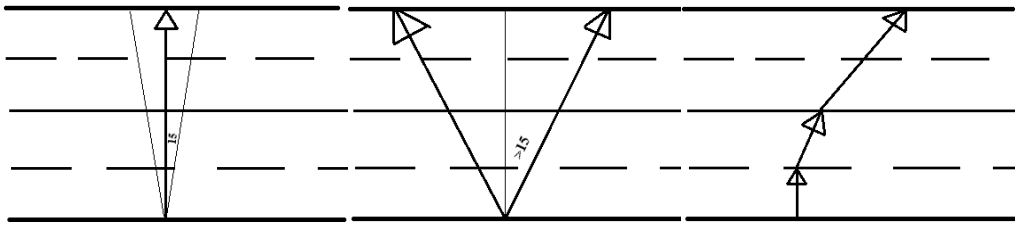
The dependent variables are Accepted Gap and Crossing Path while all other factors served as independent variables. For clarity and analysis, the independent variables are grouped into three categories: Behavioral, Traffic-related and Environmental. All categorical variables were coded using binary or ordinal representations as needed for each model type. For example, Gender was coded as 0 for Female and 1 for Male. This coding scheme ensured consistency across statistical and machine learning models allowed for direct comparison of variable effects. The variables that are considered in this study are described below:

A. Dependent Variables:

    1. Accepted Gap (AG): It refers to the time gap (time headway) measured in seconds between two consecutive vehicles that a pedestrian accepts to initiate crossing. It represents the critical time threshold that pedestrians consider safe for crossing and forms the basis for gap acceptance modeling.

Since pedestrians at multi-lane divided roads interact with multiple approaching vehicles simultaneously, the study considered the most critical conflicting vehicle within each relevant traffic stream during data extraction. The accepted gap was determined based on the available crossing opportunity created by the critical vehicle influencing the pedestrian's immediate crossing decision. This approach reflects the real-world behavior where pedestrians navigate complex multi-lane crossings through sequential lane-by-lane decision-making.

2. Crossing Path (CP): It is a categorical variable that describes the path a pedestrian follows while crossing the road. The observed crossing types were categorized as 1 for Perpendicular, 2 for Oblique and 3 for Irregular. Perpendicular crossing refers to a straight crossing path across the roadway, generally within an approximate ±15° deviation from the roadway-normal direction. Oblique crossing refers to a diagonal crossing path with a more noticeable angular deviation, while irregular crossing refers to a non-uniform path with noticeable changes in direction or movement during crossing. This variable serves as the outcome for the crossing path model. A small tolerance is necessary because pedestrians rarely maintain a perfectly 90° trajectory in real mixed-traffic conditions. The ±15° range provides a practical observational distinction while still preserving clear separation between straight and diagonal crossing behavior.



Perpendicular Crossings Oblique Crossings Irregular Crossings

While Accepted Gap and Crossing Path are modeled as separate dependent variables, they are treated as reciprocal predictors. This accounts for the behavioral trade-off where the available time gap influences a pedestrian's trajectory (path), and the chosen path geometry dictates the minimum gap required. Including them as mutual predictors allows the models to capture the simultaneous relationship between time pressure and crossing distance.

B. Behavioral Variables:

1. Gender (G): It was recorded as a binary variable coded as 0 for Female and 1 for Male. Gender was included as a demographic variable to examine whether crossing decisions varied between male and female pedestrians.
2. Age (A): It was visually estimated from the video footage and categorized into two groups: pedestrians aged 30 years or below and pedestrians above 30 years. This grouping was used to distinguish younger and older adult pedestrians in a simplified manner given that exact age could not be obtained from video-based observation.
3. Carrying Object (CO): It denotes whether the pedestrian was carrying an item such as a bag or parcel while crossing coded as 0 for No and 1 for Yes. Carrying objects may limit reaction time or visibility influencing crossing decisions.
4. Mobile Phone Use (MPU): It represents whether a pedestrian was using a mobile phone during crossing coded as 0 for No and 1 for Yes. Mobile phone use can distract pedestrians and delay their reaction to oncoming traffic.
5. Pedestrian Speed (S): It is a continuous variable representing the average crossing speed of the pedestrian measured in meters per second (m/s). In this study, speeds between 1.2m/s and 1.5m/s were considered typical walking speeds. These reference ranges align with prior studies of pedestrian movement in urban mixed-traffic conditions (Rastogi, Thaniarasu, & Chandra, 2011; Hamed, 2001). It was calculated as the effective crossing distance divided by the total time taken to cross with crossing time extracted from the video using frame-by-frame analysis in Kinovea. Any temporary slowing, stopping or acceleration during crossing is therefore reflected in the average speed value.
6. Pedestrian Speed Change (ΔS): It indicates whether a pedestrian altered their walking speed (e.g., slowing down or speeding up) while crossing coded as 0 for No and 1 for Yes. Changes in speed often occur in response to perceived vehicle threat or hesitation. In this study, pedestrian speed change was noted when the pedestrian's speed varied by ±25% from their average crossing

speed with a decrease in speed indicating hesitation and an increase in speed indicating urgency.

7. Running (R): Running reflects urgency and risk-taking behavior. In this study, a pedestrian was classified as running when pedestrian speed exceeded 2.2m/s for at least two seconds during crossing, while speeds between 0.9m/s and 2.2m/s were treated as walking, following benchmarks from mixed traffic environments (Rastogi, Thaniarasu, & Chandra, 2011). It identifies whether the pedestrian ran at any point during crossing, coded as 0 for No and 1 for Yes.
8. Pedestrian Size (PSi): It reflects the number of pedestrians crossing together as a group recorded as a continuous variable. For practical behavioral interpretation, pedestrian groups consisting of three or more than three individuals were considered relatively large crossing group. Larger groups may encourage riskier or more assertive crossing decisions due to group influence.
9. Number of Crossing Attempts (NCA): It captures how many times a pedestrian attempted to initiate a crossing before successfully finding an acceptable gap. A higher number of attempts can indicate cautious behavior or unfavorable traffic conditions.
10. Waiting Time (WT): It measures the duration, in seconds that a pedestrian waited at the curb or median before accepting a suitable gap to cross. Longer waiting times suggest cautiousness or high traffic flow. In this study, waiting time values greater than 5 seconds were considered high waiting representing cautious behavior.
11. Flow Against (FA): It indicates whether another pedestrian was simultaneously crossing from the opposite direction coded as 0 = No and 1 = Yes. The presence of another pedestrian may psychologically influence decision making by offering a sense of safety.

C. Traffic-Related Variables:

1. Average Rejected Gap (ARG): It represents the average of vehicular gaps that a pedestrian chose not to accept before finally initiating a crossing

expressed in seconds. It provides insight into risk tolerance and perception of safe crossing intervals.

2. Speed 1 and Speed 2 (S1 & S2): It indicates the speed (in m/s) of approaching vehicles in the nearer and farther lanes respectively for which the gap was accepted. These directly affect the perceived level of safety and time available for crossing.
3. Vehicle Yield (VY): It describes whether the driver of the approaching vehicle yielded or changed speed/lane to allow the pedestrian to cross coded as 0 = No and 1 = Yes. Yielding behavior influences pedestrian confidence and timing.
4. Vehicle Type (VT): It categorizes the critical vehicle approaching the crossing as either a two-wheeler coded as 0 or a four-wheeler/heavy vehicle coded as 1. Vehicle type affects pedestrian perception of risk with larger vehicles perceived as more threatening.

D. Environmental Variables:

1. Presence of Roadside Obstructions (PRO): It indicates whether any roadside element such as parked vehicles or vegetation obstructed the pedestrians' view coded as 0 = No and 1 = Yes. Reduced visibility increases perceived risk and influences waiting behavior. It was treated as a binary categorical variable because the observation recorded only whether a roadside obstruction was present or absent at the crossing location.
2. Road Surface Condition (RSC): It reflects the physical condition of the road along the pedestrian's path. It was treated as a binary categorical variable and coded as 0 for no visible potholes and 1 for presence of potholes. Poor road surface conditions can make pedestrians walk more cautiously and sometimes change their path to avoid uncomfortable areas. This may increase the gap they are willing to accept and also affect how they cross the road.
3. Presence of Crosswalk Nearby (PCN): It specifies whether a designated pedestrian crossing was available close to the pedestrians' chosen crossing point. If a crosswalk was present within a distance of 35m from the point of

crossing, it was coded as 1 = Yes, otherwise 0 = No. The presence of a formal crossing facility may influence decision making and path choice.

4. Safety Distance 1 and Safety Distance 2 (SD1 & SD2): It measures the spatial distance in meters between the pedestrian and the critical vehicle when the pedestrian enters the nearer and farther lanes respectively. These variables indicate the level of pedestrian-vehicle clearance at crossing initiation and are useful for assessing perceived risk.

Table 3.1. Description of Variables

S.No.	Variable	Variable Type	Unit/Code
1	Accepted Gap	Continuous	Seconds
2	Crossing Path	Categorical	1-Perpendicular
			2-Oblique
			3-Irregular
3	Gender	Categorical	0-Female
			1-Male
4	Carrying Object	Categorical	0-No
			1-Yes
5	Average Rejected Gap	Continuous	Seconds (s)
6	Speed 1	Continuous	Meter/seconds (m/s)
7	Speed 2	Continuous	Meter/seconds (m/s)
8	Pedestrian Size	Continuous	Number
9	Age	Categorical	0-Age>30
			1-Age<=30
10	Mobile Phone Use	Categorical	0-No
			1-Yes
11	Waiting Time	Continuous	Seconds
12	Flow Against	Categorical	0-No
			1-Yes
13	No. of crossing attempts	Continuous	Number
14	Running	Categorical	0-No
			1-Yes
15	Pedestrian Speed	Continuous	Meter/seconds (m/s)
16	Pedestrian Speed Change	Categorical	0-No
			1-Yes
17	Vehicle Yield	Categorical	0-No
			1-Yes
18	Vehicle Type	Categorical	0-2W
			1-4W+HV
19	Presence of Roadside Obstructions	Categorical	0-No
			1-Yes
20	Road Surface Conditions	Categorical	0-No visible potholes
			1-Presence of potholes
21	Presence of Crosswalk Nearby	Categorical	0-No
			1-Yes
22	Safety Distance 1	Continuous	Meters (m)
23	Safety Distance 2	Continuous	Meters (m)

It should be noted that the study locations did not have raised medians or pedestrian refuge islands and the opposing traffic streams were separated only by road markings. Therefore, median presence was not considered as a separate variable in this study. Any temporary slowing, pausing or hesitation near the center of the roadway was treated as part of the overall crossing movement rather than as formal median waiting.

The threshold values used in this study were determined by combining published parameters and field evidence from the Kathmandu midblock context. The 2.2m/s running threshold distinguishes purposeful acceleration from normal walking speeds. The safety distance categories reflect reaction distances at common midblock speeds and the ±25% criterion for pedestrian speed change captures meaningful behavioral adaptation without oversensitivity to normal walking variation. The operational definition of safety distance in this study aligns with the findings of Oxley, Ihsen, Fildes, Charlton, & Day (2005) who observed that pedestrians judge approaching vehicle speed and distance to maintain a minimal safety margin while initiating crossings. These operational definitions were adopted to ensure consistency in coding and reproducibility in behavioral interpretation across both study sites.

To evaluate model performance objectively, the dataset was divided into training and testing subsets using 70:30 ratio. The training set was used to develop the models and estimate their parameters while the testing set was used to assess predictive performance on unseen data. This distinction is important because a model may perform well on the data used for calibration but poorly on new observations if it is overfitted. Therefore, reporting both training and testing results provides a more reliable basis for comparing the statistical and machine learning models used in this study.

To improve model interpretability and reduce potential multicollinearity, reduced models were developed using only statistically significant variables. For statistical models, variable selection was based on p-values obtained from SPSS outputs, retaining variables significant at the 95% confidence level. For machine learning models, feature importance rankings were used to identify the most influential variables. These reduced models were then compared with full models to evaluate tradeoffs between model simplicity and predictive performance.

3.5. Gap Acceptance Modeling

This section presents the development of models for pedestrian gap acceptance at uncontrolled midblock crossings. Since accepted gap is a continuous variable measured in seconds, two modeling approaches were adopted: Multiple Linear Regression (MLR) as a conventional statistical model and Generalized Additive Model (GAM) as a more flexible semi-parametric model. The purpose of using both approaches was to examine how pedestrian, traffic-related and environmental variables influence accepted gap size and to compare the ability of linear and non-linear methods to represent pedestrian decision making under mixed-traffic conditions.

3.5.1. Multiple Linear Regression (MLR) Model

Multiple Linear Regression (MLR) is employed as a baseline model to predict the continuous accepted gap (Y) in seconds by pedestrians at midblock crossings. In this study, the accepted gap size was modeled as a linear function of several predictor variables including pedestrian age, group size, traffic speed, road width, waiting time, etc. The general form of the multiple linear regression equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad \ldots(3.1)$$

where,

Y is the dependent variable,

β_0 is the intercept,

$\beta_0,\beta_1,\beta_2,\cdots,\beta_n$ are the variable coefficients,

$X_1,X_2,\cdots,X_n$ are the independent variables.

The procedure for MLR Model development includes:

1. Data Splitting: The collected dataset was randomly split into training and testing sets typically using a ratio of either 70:30 or 80:20 for training and testing the model. A ratio of 70:30 was used for training/testing the model in this study. The training set was used to fit the model while the testing set was reserved for unbiased evaluation of its performance on unseen data.

2. Model Training: The MLR model was trained on the training dataset to estimate the optimal regression coefficients ($\beta_0,\beta_1,\ldots,\beta_n$) that minimize the sum of squared residuals.
3. Model Evaluation: The performance of the fitted MLR model was evaluated on the testing set using standard regression metrics:
    - Root Mean Squared Error (RMSE): Measures the average magnitude of the errors.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(Y_i - \hat{Y}_i)^2} \qquad \ldots(3.2)$$

    - R-squared (R^2): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y})^2}{\sum(Y_i - \bar{Y})^2} \qquad \ldots(3.3)$$

Where Y_i is the actual accepted gap size, $Y\square_i$ is the predicted accepted gap size, and $Y\square$ is the mean of the actual accepted gap sizes.

4. Interpretation: The significance and direction of the estimated coefficients ($\beta\square_j$) was examined to understand the linear influence of each independent variable on accepted gap size.

3.5.2. Generalized Additive Model (GAM)

To capture potential non-linear relationships between independent variables and the continuous accepted gap size, a Generalized Additive Model (GAM) was employed. GAM is a flexible extension of the generalized linear model that allows the linear predictor to depend on smooth functions of the independent variables. This makes it especially suitable for behavioral studies where variables like traffic speed or waiting time may influence pedestrian decisions in a non-linear way.

In this study, the dependent variable is the continuous accepted gap size (in seconds) and the independent variables include pedestrian demographics (e.g., age, group size,

etc.), environmental factors (e.g., presence of roadside obstructions, road surface condition, etc.) and traffic related characteristics (e.g., speed, vehicle type, waiting time, etc.).

The general form of the GAM is:

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_n(X_n) + \varepsilon \quad \ldots(3.4)$$

where,

Y is the accepted gap size,

β_0 is the intercept,

$f_1, f_2, \ldots, f_n$ are smooth (non-parametric) functions of the predictor variables $X_1, X_2, \ldots, X_n$,

ε is a random error term.

This above equation is for a GAM with an identity link and Gaussian errors. However for Accepted Gap Size which are positively skewed, a Gamma distribution with a log link is more appropriate. Given that Accepted Gap Size is a continuous, strictly positive variable often exhibiting a skewed distribution, a Gamma distribution will be chosen for the error term. This choice ensures that predicted gap sizes remain positive. Correspondingly, a log link function will be used to relate the expected value of the response to the additive predictor thus making the form of GAM to be:

$$\ln(E[Y_i]) = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_n(X_{in}) + \varepsilon_i \quad \ldots(3.5)$$

where,

$\ln(E[Y_i])$ is the expected accepted gap size in seconds,

$E[Y_i]$ is the expected accepted gap size for the i-th observation,

The procedure for development of Generalized Additive Model includes:

1. Define the goal: Predict accepted gap size Y (in seconds) using pedestrian and traffic-related variables.

2. Select independent variables: Use inputs such as:

- Traffic speed (X_1)
- Road width (X_2)

- Pedestrian age (X_3)
- Group size (X_4)
- Waiting time (X_5), etc.

3. Choose the GAM structure: Use the formula:

$$\ln(E[Y_i]) = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_n(X_{in}) + \varepsilon_i$$

where each $f_i(X_{in})$ is a smooth function capturing the nonlinear effect of the variable.

4. Prepare the dataset: Clean the data, handle missing values and ensure variables are properly formatted (e.g., continuous, categorical).

4. Fit the GAM model: Use software like R (mgcv) or Python (pygam) to train the model on the dataset by estimating the best fitting smooth functions $f_i(X_{in})$.

5. Interpret the smooth terms: Visualize each $f_i(X_{in})$ to understand how the variable impacts gap size (e.g., diminishing returns or thresholds).

6. Predict accepted gap size: Use the fitted model to calculate:

$$\widehat{Y_\iota} = \exp(\hat{\eta}) = \exp(\beta_0 \sum \mathrm{f_i(X_{in})}) \quad \text{...(3.6)}$$

7. Evaluate model performance: Compare predicted and actual gap sizes using metrics such as:

- Root Mean Squared Error (RMSE): Measures the average magnitude of the errors.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(\mathrm{Y_i} - \widehat{\mathrm{Y}}_\mathrm{i})^2} \quad \text{...(3.7)}$$

- R-squared (R^2): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$\mathrm{R^2} = 1 - \frac{\sum\left(\mathrm{Y_i} - \widehat{\mathrm{Y}}\right)^2}{\sum(\mathrm{Y_i} - \overline{\mathrm{Y}})^2} \quad \text{...(3.8)}$$

3.6. Crossing Path Modeling

This section presents the development of models for pedestrian crossing path choice at uncontrolled midblock crossings. Since crossing path is a categorical variable with three possible outcomes; Perpendicular, Oblique and Irregular, two modeling approaches were adopted: Multinomial Logistic Regression (MNL) as a conventional statistical model and CatBoost as a more flexible machine learning model. The purpose of using both approaches was to examine how pedestrian, traffic-related and environmental variables influence crossing path choice and to compare the ability of conventional and advanced models to represent complex pedestrian behavior under mixed-traffic conditions.

3.6.1. Multinomial Logistic Regression (MNL) Model

Multinomial Logistic Regression (MNL) was applied to model pedestrian crossing path at midblock locations where the dependent variable is nominal with three unordered categories: Perpendicular, Oblique and Irregular. In this study, the dependent variable represents the crossing path category chosen by the pedestrian while the independent variables include accepted gap, pedestrian age, group size, traffic speed, road width, waiting time, etc. The general form of the multinomial logistic regression model is:

$$P(Y = {}^{j}/_{X}) = \frac{e^{(\beta_{0j}+\beta_{1j}x_1+\beta_{2j}x_2+\cdots+\beta_{kj}x_k)}}{\sum_{m=1}^{M} e^{(\beta_{0m}+\beta_{1m}x_1+\beta_{2m}x_2+\cdots+\beta_{km}x_k)}} \quad \ldots(3.9)$$

where,

P ($Y={}^{j}/_{X}$) is the probability of a pedestrian taking crossing path j given the predictor X,

j is the type of crossing paths such as perpendicular, oblique and irregular,

M is the total no. of types of crossing path,

β_{0j} is the intercept of crossing path j,

β_{1j}, β_{2j}…β_{kj} are the coefficients of the independent variables for crossing path j,

($\beta_{kj}>0$ indicates as the independent variable increases; the probability of selecting the crossing path j increases and vice-versa),

X_1, X_2…X_k are the independent variables.

The procedure for Multinomial Logistic Regression Model development includes:

1. Data Splitting: The collected dataset was randomly split into training and testing sets, typically using a ratio of either 70:30 or 80:20 for training and testing the model. A ratio of 70:30 was used for training/testing the model in this study. The training set was used to fit the model while the testing set was reserved for unbiased evaluation of its performance on unseen data.
2. Model Training: The Multinomial Logistic Regression model was trained on the training dataset to estimate the coefficients (β_{kj}) for each non-reference category.
3. Model Evaluation: The performance of the fitted model was assessed on the testing set using standard classification metrics:
    - Accuracy:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad \text{...(3.10)}$$

    - Confusion Matrix
4. Interpretation: The estimated coefficients (β_{kj}) was interpreted to understand the influence of each independent variable on the likelihood of choosing a particular crossing path relative to the reference path.

3.6.2. CatBoost Model

For the multi-class classification of the categorical Crossing Path chosen by pedestrians, the CatBoost algorithm was employed. This choice is driven by CatBoost's robust handling of categorical features and its ability to capture complex, non-linear relationships which are expected in pedestrian behavior data. It is a gradient boosting algorithm developed by Yandex known for its strong performance with heterogeneous datasets that include both categorical and numerical variables. It handles categorical data natively without the need for one-hot encoding and is particularly effective with relatively small datasets making it suitable for this study.

The dependent variable in this model is the type of crossing path taken by the pedestrian; perpendicular, oblique and irregular while the independent variables include pedestrian age, group size, road width, vehicle speed, waiting time, etc.

CatBoost builds an ensemble of decision trees in a sequential manner where each new tree corrects the residual errors of the previous one. It uses ordered boosting and target statistics to prevent overfitting especially on small datasets with categorical variables.

The general prediction formula of a boosted decision tree ensemble can be represented as:

$$\hat{y} = \sum_{t=1}^{T} \eta . h_t\,(x) \qquad \ldots(3.4)$$

where,

y is the predicted output,

T is the total number of trees,

$h_t(x)$ is the prediction from the t-th tree,

η is the learning rate,

x represents the feature vector.

The procedures for modeling crossing paths using CatBoost are as follows:

1. Define the objective: Predict the crossing path category (perpendicular, oblique and irregular) chosen by pedestrians.
2. Identify the dependent variable: $Y \in \{1,2,3\}$ (for each crossing path class).
3. Select independent variables: Use inputs like:
    - Traffic speed (X_1)
    - Road width (X_2)
    - Group size (X_3),
    - Waiting time (X_4)
    - Pedestrian demographics (X_5), etc.
4. Format data: Encode categorical features and split data into training and testing sets.

5. Initialize CatBoostClassifier: Specify categorical features to CatBoost using cat_features parameter; CatBoost handles them internally via ordered target statistics.
6. Hyperparameter tuning: Tune hyperparameters (e.g., learning rate, depth, iterations, l2_leaf_reg) using cross-validation or grid/random search to optimize model performance.
7. Train the model: Fit using the training data:

$$\hat{Y} = CatBoost(X) \quad \text{...(3.5)}$$

Internally, CatBoost uses gradient boosting over decision trees to minimize a multiclass loss function:

$$L = -\sum_{i-1}^{n}\sum_{k=1}^{K} y_{ik}\log(p_{ik}) \quad \text{...(3.6)}$$

where y_{ik} is the true label and p_{ik} is the predicted probability for class k.

8. Predict crossing path class: For new data X, output is:

$$\hat{Y} = \arg\max p(Y = k \mid \mathrm{X}) \quad \text{...(3.7)}$$

9. Evaluate performance: Use classification metrics:
    - Accuracy:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad \text{...(3.8)}$$

    - Confusion Matrix
    - Precision, Recall and F1-Score
10. Interpret results: Use built-in tools to understand which variables most influenced the predictions.

CHAPTER FOUR: RESULT AND DISCUSSION

4.1. Overview

This chapter presents the results of the analysis of pedestrian behavior at uncontrolled midblock crossings in Kamalpokhari and Mitrapark. The analysis focuses on two key aspects: the size of the vehicular gaps that pedestrians accept for crossing and the crossing paths they choose under mixed-traffic conditions. To provide a clear structure, the chapter begins with descriptive statistics followed by the results of accepted gap modeling and crossing path modeling.

The findings from the statistical and machine learning models are then compared and interpreted in relation to pedestrian decision making at uncontrolled midblock locations. Tables and figures are used to support the analysis and to highlight the relative performance of the different models, the important contributing variables and the broader behavioral patterns observed in the study.

4.2. Descriptive Statistics for Variables

Tables 4.1 and 4.2 summarize the continuous and categorical characteristics of 460 pedestrian crossings observed at Kamalpokhari. The mean accepted gap was 11.01 seconds with an average rejected gap of 1.35 seconds indicating moderate variation in gap acceptance behavior. Mean vehicle speeds in both the nearer and farther lanes were similar at 6.86m/s while the average waiting time was 10.91 seconds. Pedestrians crossed at an average speed of 1.37m/s with mean safety distances of 36.44m and 35.19m in the nearer and farther lanes respectively. The sample was dominated by male pedestrians and oblique crossings, with most individuals under 30 years of age, not using mobile phones, and not crossing against pedestrian flow.

Table 4.1. Descriptive Statistics of Continuous Variables: Kamalpokhari

	N	Unit	Minimum	Maximum	Mean	Std. Deviation
Accepted_Gap	460	seconds	2.848	34.824	11.012	4.434
Average_Rejected_Gap	460	seconds	0.000	3.794	1.345	0.994
Speed_1	460	m/s	2.793	12.997	6.860	2.195
Speed_2	460	m/s	2.618	11.965	6.863	2.313
Pedestrian_Size	460	No.	1.000	4.000	2.141	1.137
Waiting_Time	460	seconds	0.500	37.571	10.913	6.998
No._of_Crossing_Attempts	460	No.	1.000	3.000	2.002	0.839
Pedestrian_Speed	460	m/s	0.280	4.336	1.372	1.012
Safety_Distance_1	460	meters	4.404	104.073	35.786	23.011
Safety_Distance_2	460	meters	4.684	108.587	34.604	22.046

Table 4.2. Frequency Distribution of Categorical Variables: Kamalpokhari

S. No.	Variables	Category	Number	%
1	Gender	Female	173	37.6
		Male	287	62.4
2	Carrying Object	No	252	54.8
		Yes	208	45.2
3	Crossing Path	Perpendicular	125	27.2
		Oblique	194	42.2
		Irregular	141	30.7
4	Age	Age>30	152	33.0
		Age<=30	308	67.0
5	Mobile Phone Use	No	392	85.2
		Yes	68	14.8
6	Flow Against	No	409	88.9
		Yes	51	11.1
7	Running	No	381	82.8
		Yes	79	17.2
8	Pedestrian Speed Change	No	116	25.2
		Yes	344	74.8
9	Vehicle Yield	No	265	57.6
		Yes	195	42.4
10	Vehicle Yield	2W	294	63.9
		4W+HV	166	36.1
11	Presence of Roadside Obstructions	No	390	84.8
		Yes	70	15.2
12	Road Surface Conditions	Good	449	97.6
		Bad	11	2.4
13	Presence of Crosswalk Nearby	No	368	80.0
		Yes	92	20.0

Overall, the Kamalpokhari dataset reflects a midblock environment in which pedestrians frequently adopt indirect crossing paths while interacting with moderate vehicle speeds and variable waiting times.

Tables 4.3 and 4.4 present the summary statistics at Mitrapark. The mean accepted gap was 10.27 seconds slightly lower than at Kamalpokhari while the mean rejected gap was 1.34 seconds. Vehicle speeds were marginally higher, averaging 7.15m/s in the nearer lane and 7.02m/s in the farther lane. The average waiting time was longer at 12.95 seconds and pedestrians crossed at a slightly lower speed of 1.26m/s. Mean safety distances were 33.69m and 32.38m in the nearer and farther lanes, respectively.

Table 4.3. Descriptive Statistics of Continuous Variables: Mitrapark

	N	Unit	Minimum	Maximum	Mean	Std. Deviation
Accepted_Gap	490	seconds	2.321	50.370	10.274	4.624
Average_Rejected_Gap	490	seconds	0.000	3.794	1.343	1.060
Speed_1	490	m/s	2.394	12.997	7.146	2.261
Speed_2	490	m/s	2.214	13.109	7.018	2.416
Pedestrian_Size	490	No.	1.000	4.000	2.161	1.146
Waiting_Time	490	seconds	0.500	47.166	12.952	8.196
No._of_Crossing_Attempts	490	No.	1.000	3.000	1.992	0.836
Pedestrian_Speed	490	m/s	0.301	4.447	1.259	0.910
Safety_Distance_1	490	meters	4.813	102.697	32.726	21.538
Safety_Distance_2	490	meters	3.777	100.293	31.746	21.492

Table 4.4. Frequency Distribution of Categorical Variables: Mitrapark

S. No.	Variables	Category	Number	%
1	Gender	Female	192	39.2
		Male	298	60.8
2	Carrying Object	No	281	57.3
		Yes	209	42.7
3	Crossing Path	Perpendicular	148	30.2
		Oblique	208	42.4
		Irregular	134	27.3
4	Age	Age>30	159	32.4
		Age<=30	331	67.6
5	Mobile Phone Use	No	423	86.3
		Yes	67	13.7
6	Flow Against	No	445	90.8
		Yes	45	9.2
7	Running	No	417	85.1
		Yes	73	14.9
8	Pedestrian Speed Change	No	155	31.6
		Yes	335	68.4
9	Vehicle Yield	No	229	46.7
		Yes	261	53.3
10	Vehicle Yield	2W	291	59.4
		4W+HV	199	40.6
11	Presence of Roadside Obstructions	No	422	86.1
		Yes	68	13.9
12	Road Surface Conditions	Good	466	95.1
		Bad	24	4.9
13	Presence of Crosswalk Nearby	No	381	77.8
		Yes	109	22.2

Similar to Kamalpokhari, oblique crossing was the most common behavior with most pedestrians being male, under 30 years of age and not using mobile phones while crossing. Overall, the descriptive statistics indicate that the two sites share broadly similar pedestrian and traffic characteristics while Mitrapark exhibits slightly longer waiting times, higher vehicle speeds and a slightly lower average accepted gap.

4.3. Gap Acceptance Modeling

This section presents the results of the models developed to analyze pedestrian gap acceptance behavior at the selected midblock locations. Both Multiple Linear Regression (MLR) and Generalized Additive Models (GAM) were applied to examine the influence of pedestrian, traffic and environmental variables on accepted gap size. The performance of full and reduced models is evaluated and compared to assess their ability to capture the underlying behavioral patterns.

4.3.1. Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) was used to model accepted gap size at the two study sites using a 70:30 train-test split. The purpose of this analysis was to identify the variables that significantly influence accepted gap size and to compare the explanatory performance of full and reduced model specifications. Detailed coefficient tables and additional diagnostic plots are provided in the appendix while the key results are summarized here.

A. Model I: Kamalpokhari (Considering all variables)

Full models were developed by including all available independent variables to capture the complete set of potential influences on the dependent variable. This approach provides a comprehensive baseline for assessing overall model performance and identifying statistically significant predictors before developing reduced specifications.

i. **Training the model**

For Kamalpokhari, the training MLR model also showed strong explanatory performance with $R = 0.824$, $R^2 = 0.679$, Adjusted $R^2 = 0.661$, a standard error of 2.58 and a Durbin-Watson statistic of 1.961. The results indicate that the model fits the

training data well providing a reliable basis for evaluating its predictive performance on the testing dataset. The model summary is presented in Table 4.5.

Table 4.5. Summary of MLR Model: Kamalpokhari (Full Model - Training)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.824	0.679	0.661	2.58	1.961

ii. **Testing the model**

For Kamalpokhari, the testing MLR model showed strong explanatory performance with R = 0.838, R^2 = 0.702, Adjusted R^2 = 0.643, a standard error of 2.737 and a Durbin-Watson statistic of 1.956. Significant predictors included Speed 1, Speed 2, Crossing Path, Pedestrian Speed Change, Safety Distance 1 and Safety Distance 2. The model summary is presented in Table 4.6.

Table 4.6. Summary of MLR Model: Kamalpokhari (Full Model - Testing)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
2	.838	0.702	0.643	2.737	1.956

B. Model II: Kamalpokari (Considering significant variables only)

Reduced models were developed using only the statistically significant variables identified from the full model. This approach focuses on the most important factors while removing less relevant ones resulting in a simpler and more efficient model without losing key information.

i. **Training the model**

The reduced training MLR model also showed strong explanatory performance with R = 0.817, R^2 = 0.668, Adjusted R^2 = 0.662, a standard error of 2.576 and a Durbin-Watson statistic of 1.967. The results indicate that the reduced model fits the training data well while maintaining consistency with the testing performance, suggesting good generalization with minimal overfitting. The model summary is presented in Table 4.7.

Table 4.7. Summary of MLR Model: Kamalpokhari (Reduced Model - Training)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.817	0.668	0.662	2.576	1.967

ii. **Testing the model**

The reduced testing model with Speed 1, Crossing Path, Pedestrian Speed Change, Safety Distance 1 and Safety Distance 2 continued to show strong performance, with R = 0.817, R^2 = 0.668, Adjusted R^2 = 0.662, a standard error of 2.576 and a Durbin-Watson statistic of 1.967. Speed 1, Crossing Path, Safety Distance 1, and Safety Distance 2 remained significant while Pedestrian Speed Change showed reduced stability. The slight decrease in explanatory power compared to the full model indicates that the reduced model retains most of the relevant information while providing a more compact representation of accepted gap behavior. The model summary is presented in Table 4.8.

Table 4.8. Summary of MLR Model: Kamalpokhari (Reduced Model - Testing)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
2	.817	0.668	0.662	2.576	1.967

Overall, the reduced Kamalpokhari model performs nearly as well as the full specification while remaining easier to interpret. The final regression equation for Kamalpokhari is:

AG=9.311 - 0.305(S1) - 0.237(S2) – 0.766(CB) + 0.099 (SD1) + 0.086(SD2)...(4.1)

C. Model III: Mitrapark (Considering all variables)

Full models were developed by including all available independent variables to capture the complete set of potential influences on the dependent variable. This approach provides a comprehensive baseline for assessing overall model performance and identifying statistically significant predictors before developing reduced specifications.

i. **Training the Model**

The training MLR model also showed strong explanatory performance, with R = 0.861, R^2 = 0.742, Adjusted R^2 = 0.729, a standard error of 2.228 and a Durbin-Watson statistic of 1.899. The results indicate that the model fits the training data well and remains consistent with the testing performance suggesting good generalization with minimal overfitting. The model summary is presented in Table 4.9.

Table 4.9. Summary of MLR Model: Mitrapark (Full Model - Training)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.861	0.742	0.729	2.228015	1.899

ii. **Testing the Model**

The testing MLR model showed strong performance with R = 0.867, R^2 = 0.752, Adjusted R^2 = 0.705, a standard error of 2.13 and a Durbin-Watson statistic of 2.089. Significant predictors included Speed 1, Speed 2, Safety Distance 1, and Safety Distance 2, while Running and Waiting Time were only marginally significant. The model summary is presented in Table 4.10.

Table 4.10. Summary of MLR Model: Mitrapark (Full Model - Testing)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
2	.867	0.752	0.705	2.13	2.089

D. Model IV: Mitrapark (Considering significant variables only)

Reduced models were developed using only the statistically significant variables identified from the full model. This approach focuses on the most important factors while removing less relevant ones resulting in a simpler and more efficient model without losing key information.

i. **Training the Model**

The reduced training MLR model also showed strong explanatory performance with R = 0.857, R^2 = 0.734, Adjusted R^2 = 0.730, a standard error of 2.223 and a Durbin-Watson statistic of 1.921. The results indicate that the reduced model fits the training data well and remains consistent with the testing performance suggesting good generalization with minimal overfitting. The model summary is presented in Table 4.11.

Table 4.11. Summary of MLR Model: Mitrapark (Reduced Model - Training)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.857	0.734	0.730	2.223	1.921

ii. **Testing the Model**

The reduced testing model with Speed 1, Speed 2, Safety Distance 1, Safety Distance 2, Waiting Time and Running remained highly effective with R = 0.857, R^2 = 0.734, Adjusted R^2 = 0.730, a standard error of 2.223 and a Durbin-Watson statistic of 1.921. All retained predictors remained significant with Safety Distance 1 showing the strongest influence. The minimal reduction in explanatory power compared to the full model indicates that accepted gap behavior can be effectively represented using a reduced set of key variables. The model summary is presented in Table 4.12.

Table 4.12. Summary of MLR Model: Mitrapark (Reduced Model - Testing)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
2	.857	0.734	0.730	2.223	1.921

The reduced Mitrapark model provides a compact and statistically reliable representation of accepted gap behavior. The final regression equation for Mitrapark is:

$$AG=8933 - 0.412(S1) - 0.243 (S2) + 0.117(SD1) + 0.081(SD2) \quad ...(3.6)$$

Overall, the MLR results indicate that accepted gap behavior at both study sites is strongly influenced by variables that directly affect perceived safety and available crossing time. At Kamalpokhari, both behavioral and traffic-related variables contributed to the model, particularly vehicle speed, crossing path and safety distance. At Mitrapark, the dominant predictors were more concentrated around vehicle speed and safety spacing. In both cases, the reduced models performed nearly as well as the full models showing that a smaller set of key predictors is sufficient to explain most of the variation in accepted gap size.

4.3.2. Generalized Additive Model (GAM)

Generalized Additive Model (GAM) was used to predict accepted gap size while allowing for non-linear relationships between the dependent variable and the predictor set. The objective of this analysis was to evaluate whether a flexible non-linear framework could better capture pedestrian gap acceptance behavior than the corresponding linear MLR specification. Unlike the other models, the GAM was evaluated using cross-validation rather than a fixed training-testing split. The cross-validated R^2 provides a robust measure of model performance by assessing how well the model generalizes across different subsets of the data. This approach is particularly suitable for GAM as it enables more reliable estimation of predictive performance in the presence of non-linear relationships. Detailed coefficient tables and additional diagnostic plots are provided in the appendix, while the key results are summarized here.

A. Model I: Kamalpokhari (Considering all variables)

The full GAM model showed satisfactory predictive performance with a Mean Squared Error of 7.520, an R^2 of 0.672 and a cross-validated R^2 of 0.618. These results indicate that the model explains a substantial portion of the variation in accepted gap and generalizes well across different data folds. The performance statistics are presented in Table 4.13.

Table 4.13. Model Performance Summary: Kamalpokhari (Full Model)

Metric	Value
Mean Squared Error	7.520
R-squared	0.672
CV R-squared (Mean)	0.618

B. Model II: Kamalpokhari (Considering significant variables only)

Examination of the partial dependence functions showed that only a subset of variables had meaningful influence, particularly those related to pedestrian movement, waiting behavior, traffic speeds and safety distance. Accordingly, a reduced GAM was estimated using the key predictors: Average Rejected Gap, Speed 1, Speed 2, Pedestrian Speed, Waiting Time, Number of Crossing Attempts, Safety Distance 1 and Safety Distance 2. The reduced model improved performance yielding an MSE of 7.256, an R^2 of 0.684 and a cross-validated R^2 of 0.653. The close agreement between these values indicates strong generalizability and a more focused representation of gap acceptance behavior at Kamalpokhari. The performance statistics are presented in Table 4.14.

Table 4.14. Model Performance Summary: Kamalpokhari (Reduced Model)

Metric	Value
Mean Squared Error	7.256
R-squared	0.684
CV R-squared (Mean)	0.653

Overall, the Kamalpokhari GAM results indicate that the reduced model offers a clearer and more efficient representation of accepted gap behavior than the full specification. By excluding weakly contributing predictors, the reduced model improved explanatory power and preserved strong predictive stability highlighting the importance of traffic speeds, safety distances, pedestrian movement and waiting related variables in shaping accepted gap decisions at this site.

C. Model III: Mitrapark (Considering all variables)

The full GAM model showed strong predictive performance with an MSE of 4.443, an R^2 of 0.760 and a cross-validated R^2 of 0.739. These results indicate that the model explains a substantial portion of the variation in accepted gap and generalizes well without significant overfitting. The performance statistics are presented in Table 4.15.

Table 4.15. Model Performance Summary: Mitrapark (Full Model)

Metric	Value
Mean Squared Error	4.443
R-squared	0.760
CV R-squared (Mean)	0.739

D. Model IV: Mitrapark (Considering significant variables only)

The full GAM model indicated that only a subset of variables contributed meaningfully to accepted gap behavior. Accordingly, the reduced model retained the key predictors: Average Rejected Gap, Speed 1, Speed 2, Pedestrian Speed, Waiting Time, Number of Crossing Attempts, Safety Distance 1 and Safety Distance 2. The refined model improved performance with an MSE of 4.216, an R^2 of 0.772 and a cross-validated R^2 of 0.743 indicating strong explanatory power and generalizability. The performance summary is presented in Table 4.16.

Table 4.16. Model Performance Summary: Mitrapark (Reduced Model)

Metric	Value
Mean Squared Error	4.216
R-squared	0.772
CV R-squared (Mean)	0.743

Overall, the reduced Mitrapark GAM provides a more focused and effective representation of accepted gap behavior than the full model. The results indicate that traffic speeds, safety distance, pedestrian speed, waiting time and rejected gap measures are the primary drivers of gap acceptance while demographic and contextual variables add limited value once these core predictors are included.

Taken together, the GAM results across both sites demonstrate strong explanatory performance with reduced models outperforming the full specifications in both accuracy and interpretability. At Kamalpokhari, R^2 increased from 0.672 to 0.684 while at Mitrapark it increased from 0.760 to 0.772. In both cases, the reduced models retained only the most behaviorally relevant variables and showed closely aligned cross-

validated R^2 values indicating good generalization. These findings support the use of reduced GAM models for comparison with the MLR results.

4.4. Crossing Path Modeling

This section presents the results of the models developed to analyze pedestrian crossing path behavior at the selected midblock locations. Both Multinomial Logistic Regression (MNL) and CatBoost were applied to examine the influence of behavioral, traffic-related and environmental variables on crossing path choice. The performance of full and reduced models is evaluated and compared to assess their ability to capture and classify crossing behavior.

4.4.1. Multinomial Logistic Regression (MNL)

Multinomial Logistic Regression (MNL) was applied to model pedestrian crossing path choice where the dependent variable Crossing Path was categorized as Perpendicular, Oblique or Irregular. The dataset was divided using a 70:30 train-test split. The purpose of this analysis was to examine the explanatory power of the MNL framework, identify statistically significant predictors of crossing path and compare full and reduced model specifications for the two study sites. Detailed coefficient tables and additional diagnostic plots are provided in the appendix while the key results are summarized here.

A. Model I: Kamalpokhari (Considering all variables)

Full models were developed by including all available independent variables to capture the complete set of potential influences on the dependent variable. This approach provides a comprehensive baseline for assessing overall model performance and identifying statistically significant predictors before developing reduced specifications.

i. Training the Model

The full MNL model showed a significant improvement over the intercept-only model reducing the -2 Log Likelihood from 695.667 to 403.529. The Likelihood Ratio Chi-Square was 292.138 with 44 degrees of freedom and was highly significant ($p<0.001$). The McFadden's Pseudo R^2 of 0.4199 indicates moderate explanatory power for behavioral data. The model-fit statistics are presented in Table 4.17.

Table 4.17. Model Fit Summary: Kamalpokhari (Full Model - Training)

Metric	Value
-2 Log Likelihood (Intercept Only)	695.667
-2 Log Likelihood (Final Model)	403.529
Likelihood Ratio Chi-Square	292.138
Degrees of Freedom	44
Sig. (p-value)	0
McFadden's Pseudo R^2	0.4199

The training confusion matrix in Figure 4.1 shows that the model identifies perpendicular crossings fairly well, correctly classifying 76 of 87 cases. Oblique crossings are moderately captured with 88 of 136 correctly predicted, though 36 oblique observations are mislabeled as irregular and 12 as perpendicular. For irregular crossings, the model correctly assigns 59 of 99 instances but 39 are confused with oblique behaviour. Overall accuracy reaches 69% indicating that while perpendicular movements are the most distinguishable, the boundary between oblique and irregular crossings remains blurred, leading to more frequent misclassification between these two classes.

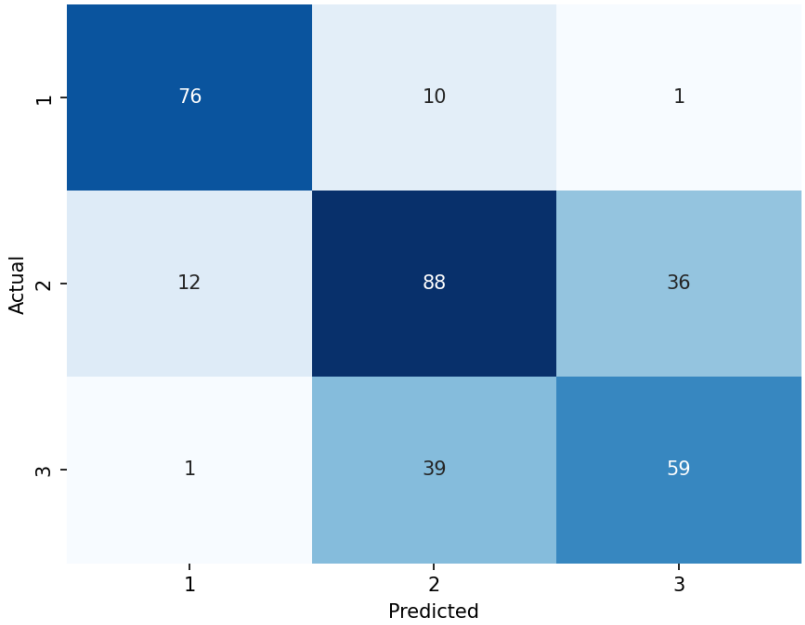


Figure 4.1. Confusion Matrix: Kamalpokhari (Full Model - Training)

iii. **Testing the Model**

When applied to unseen test data, perpendicular accuracy drops to 68.4% (26 of 38), oblique drops slightly to 60.3% (35 of 58) and irregular falls to 54.8% (23 of 42).

Although the full model retains some ability to discriminate, the decline in perpendicular performance highlights limited generalizability, while irregular crossings remain the most challenging class.

Table 4.18. Per Class Accuracy: Kamalpokhari (Full Model - Testing)

Class	Correct	Total	Accuracy
1 (Perpendicular)	26	38	68.40%
2 (Oblique)	35	58	60.30%
3 (Irregular)	23	42	54.80%

The test confusion matrix in Figure 4.2 confirms the model's ability to generalize to unseen data. Perpendicular crossings are again well identified with 26 of 38 cases correctly classified and no instances mislabeled as irregular. Oblique crossings are picked up moderately, with 35 of 58 correctly predicted, though 12 are mistaken for perpendicular and 11 for irregular. Irregular crossings remain the most challenging where 23 of 42 are correctly identified but 19 are confused with oblique movements. The overall test accuracy is 61% which is lower than the training performance but still indicates a reasonable separation between perpendicular and the other two crossing types, while the overlap between oblique and irregular behavior persists.

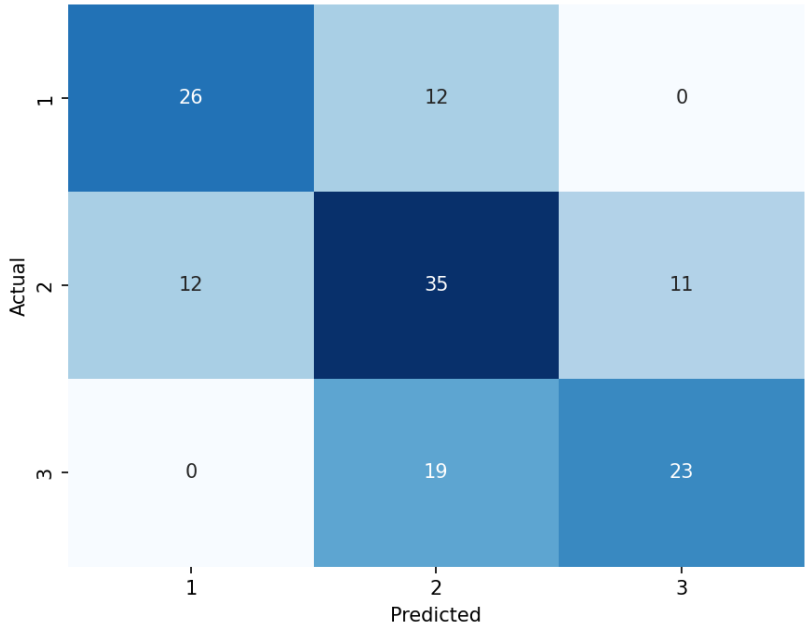


Figure 4.2. Confusion Matrix: Kamalpokhari (Full Model - Testing)

B. Model II: Kamalpokari (Considering significant variables only)

The reduced models were developed using only the statistically significant variables identified from the full model. This approach focuses on the most important factors while removing less relevant ones, resulting in a simpler and more efficient model without losing key information.

i. Training the Model

The reduced MNL model was estimated using only the significant predictors. The model maintained a strong fit reducing the -2 Log Likelihood from 695.667 to 447.398. The Likelihood Ratio Chi-Square was 248.269 with 20 degrees of freedom and remained highly significant ($p<0.001$), with a McFadden's Pseudo R^2 of 0.3569. Although slightly lower than the full model, this indicates that the reduced specification preserves most of the explanatory power while offering a more compact representation of crossing path choice. The fit statistics are presented in Table 4.19.

Table 4.19. Model Fit Summary: Kamalpokhari (Reduced Model - Training)

Metric	Value
-2 Log Likelihood (Intercept Only)	695.667
-2 Log Likelihood (Final Model)	447.398
Likelihood Ratio Chi-Square	248.269
Degrees of Freedom	20
Sig. (p-value)	0
McFadden's Pseudo R^2	0.3569

The training confusion matrix for the simplified model in Figure 4.3 reflects a similar pattern to the full specification. Perpendicular crossings are well identified with 74 of 87 cases correctly classified and only one instance mislabeled as irregular. Oblique crossings are captured moderately with 85 out of 136 correct, though 33 are confused with irregular and 18 with perpendicular. Irregular crossings remain the most difficult category where 48 of 99 are correctly predicted but an equal number are mistaken for oblique. The overall accuracy reaches 64% indicating that while the significant predictors capture most of the separation between crossing types, the overlap between oblique and irregular behavior persists even in the more parsimonious model.

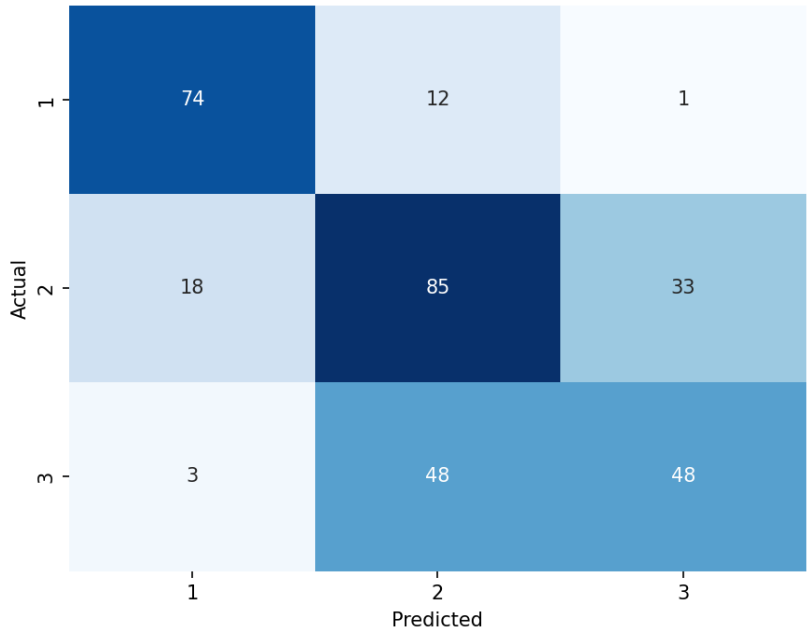


Figure 4.3. Confusion Matrix: Kamalpokhari (Reduced Model - Training)

ii. Testing the Model:

On the test set, the reduced model correctly classifies 26 of 38 perpendicular crossings (68.4%), 31 of 58 oblique crossings (53.4%) and 18 of 42 irregular crossings (42.9%). Perpendicular accuracy holds steady with the full model but oblique and irregular performance decline noticeably. The model still finds irregular crossings the hardest to predict, capturing fewer than half of them which confirm that the omitted predictors, though not individually significant, contributed meaningful information for distinguishing irregular from oblique behavior.

Table 4.20. Per Class Accuracy: Kamalpokhari (Reduced Model - Testing)

Class	Correct	Total	Accuracy
1 (Perpendicular)	26	38	68.40%
2 (Oblique)	31	58	53.40%
3 (Irregular)	18	42	42.90%

The test confusion matrix for the reduced model in Figure 4.4 shows a noticeable drop in performance compared to the full model. Perpendicular crossings remain reasonably well identified with 26 of 38 cases correctly classified and the remaining 12 mislabeled as oblique. Oblique crossings are only moderately captured with 31 of 58 are correctly predicted while 15 are mistaken for irregular and 12 for perpendicular. Irregular

crossings were correctly predicted 18 out of 42 cases while the remaining 24 were mistaken for oblique. The overall test accuracy drops to 54.3% indicating that the seven significant predictors alone lack the information needed to reliably separate oblique and irregular crossing behaviors on unseen data and that the omitted variables contribute meaningfully to generalization.

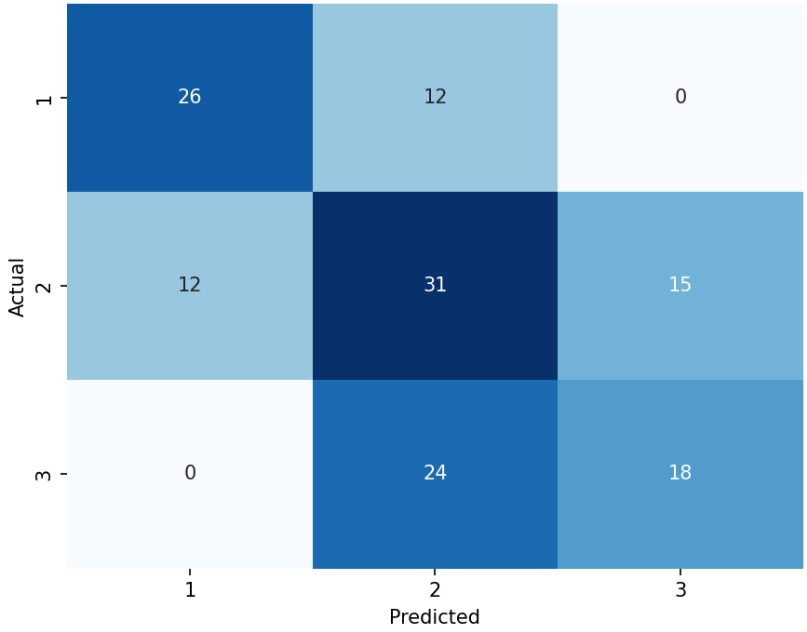


Figure 4.4. Confusion Matrix: Kamalpokhari (Reduced Model - Testing)

Overall, the Kamalpokhari MNL results indicate that the model serves well as an explanatory tool than as a predictive classifier. The full model achieved a reasonable training accuracy of 69.3% and a test accuracy of 60.9% with perpendicular crossings consistently well identified. However, irregular crossings remained the most difficult class, and in the reduced model test performance dropped substantially where irregular accuracy fell to 42.9% and overall test accuracy declined to 54.3%. The model's failure to reliably separate oblique and irregular behaviors, particularly when only significant predictors are used, suggests that pedestrian crossing path choice at this site is highly nuanced and that variables beyond those retained in the reduced specification contribute meaningfully to distinguishing the more complex crossing types.

C. Model III: Mitrapark (Considering all variables)

The full model was developed by including all available independent variables to capture the complete set of potential influences on the dependent variable. This approach provides a comprehensive baseline for assessing overall model performance

and identifying statistically significant predictors before developing reduced specifications.

i. Training the Model

The full training MNL model showed a significant improvement over the intercept-only model reducing the -2 Log Likelihood from 741.258 to 446.533. The Likelihood Ratio Chi-Square was 294.724 with 44 degrees of freedom and was highly significant ($p<0.001$). The McFadden's Pseudo R^2 of 0.3976 indicates moderate to strong explanatory power for behavioral data. The model-fit statistics are presented in Table 4.21.

Table 4.21. Model Fit Summary: Mitrapark (Full Model - Training)

Metric	Value
-2 Log Likelihood (Intercept Only)	741.258
-2 Log Likelihood (Final Model)	446.533
Likelihood Ratio Chi-Square	294.724
Degrees of Freedom	44
Sig. (p-value)	0
McFadden's Pseudo R^2	0.3976

The training confusion matrix for the Mitrapark full model in Figure 4.4 shows that perpendicular crossings are well identified, with 86 of 104 cases correctly classified and only one instance mislabeled as irregular. Oblique crossings are captured reasonably well with 94 out of 145 are correctly predicted, though 28 are mistaken for irregular and 23 for perpendicular. Irregular crossings prove more challenging where 58 of 94 are correct but 35 are confused with oblique behavior. The model achieves a training accuracy of 69.4% indicating that while perpendicular movements are the most distinguishable, the overlap between oblique and irregular patterns persists even within the training data.

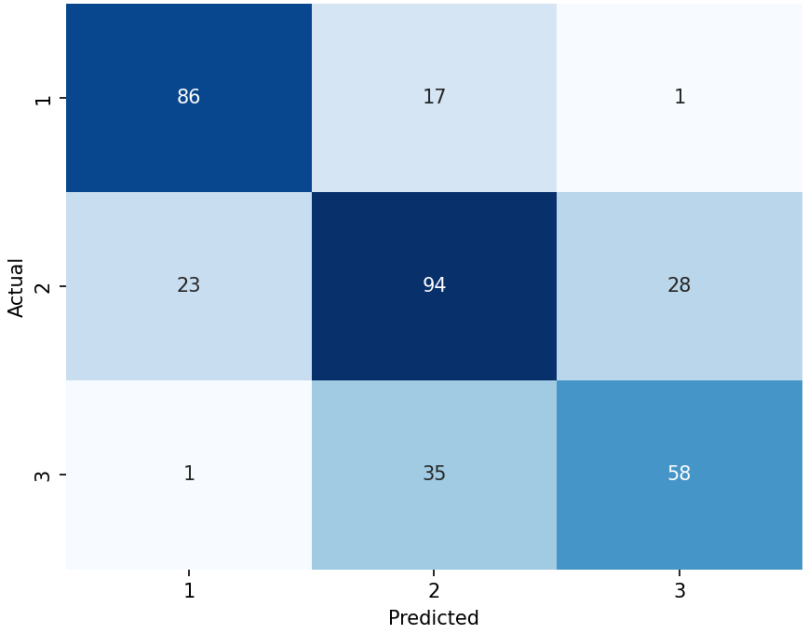


Figure 4.5. Confusion Matrix: Mitrapark (Full Model - Training)

ii. Testing the Model

The full model maintains strong performance for perpendicular crossings, correctly predicting 37 of 44 cases (84.1%). Oblique accuracy falls to 54.0% (34 of 63) and irregular accuracy remains at 62.5% (25 of 40). The gap between training and test accuracy for oblique crossings highlights some overfitting to the training patterns while perpendicular and irregular predictions stay fairly stable.

Table 4.22. Per Class Accuracy: Mitrapark (Full Model - Testing)

Class	Correct	Total	Accuracy
1 (Perpendicular)	37	44	84.10%
2 (Oblique)	34	63	54.00%
3 (Irregular)	25	40	62.50%

The test confusion matrix for the Mitrapark site in Figure 4.5 confirms that perpendicular crossings remain the most distinguishable class with 37 of 44 cases correctly identified and only one mislabeled as irregular. Oblique crossings are moderately captured with 34 out of 63 correct, though 20 are confused with irregular and 9 with perpendicular. Irregular crossings continue to pose the greatest difficulty with 25 of 40 are correctly predicted but 15 are mistaken for oblique behavior. The test

accuracy of 65.3% indicates that the model generalizes reasonably well from training, yet the persistent overlap between oblique and irregular patterns highlights the inherent variability in these two crossing types.

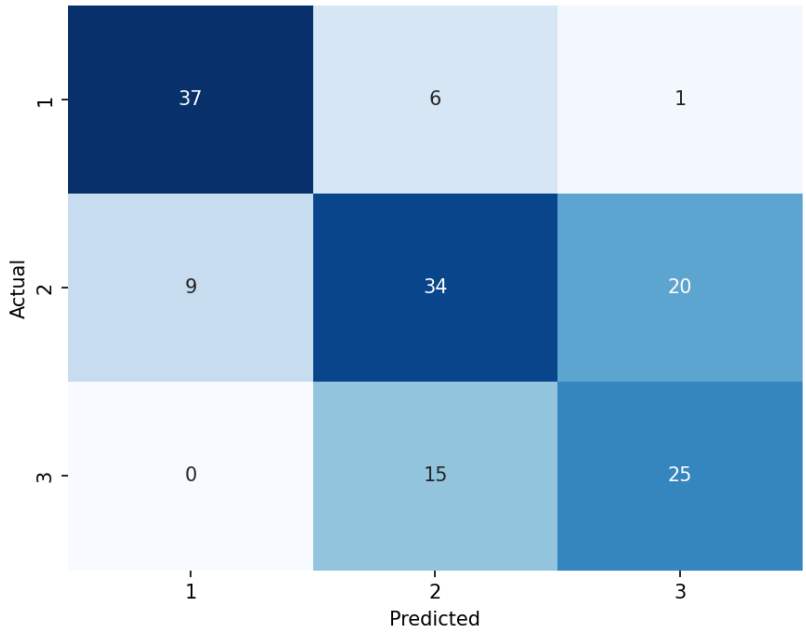


Figure 4.6. Confusion Matrix: Mitrapark (Full Model - Testing)

D. Model IV: Mitrapark (Considering significant variables only)

The reduced models were developed using only the statistically significant variables identified from the full model. This approach focuses on the most important factors while removing less relevant ones resulting in a simpler and more efficient model without losing key information.

i. Training the Model

The reduced MNL model for Mitrapark was estimated using only the significant predictors and remained statistically strong reducing the -2 Log Likelihood from 741.258 to 480.655. The Likelihood Ratio Chi-Square was 260.603 with 10 degrees of freedom and was highly significant ($p<0.001$), with a McFadden's Pseudo R^2 of 0.3516. This indicates that the reduced model still captures a meaningful proportion of variation in crossing path choice. The fit statistics are presented in Table 4.23.

Table 4.23. Model Fit Summary: Mitrapark (Reduced Model - Training)

Metric	Value
-2 Log Likelihood (Intercept Only)	741.258
-2 Log Likelihood (Final Model)	480.655
Likelihood Ratio Chi-Square	260.603
Degrees of Freedom	10
Sig. (p-value)	0
McFadden's Pseudo R^2	0.3516

The training confusion matrix in Figure 4.6 shows that perpendicular crossings are still the easiest to identify with 82 of 104 cases correctly classified and only three mislabeled as irregular. Oblique crossings are captured reasonably well with 92 out of 145 correctly predicted, though 28 are mistaken for irregular and 25 for perpendicular. Irregular crossings remain the most difficult class with 54 of 94 are correctly identified but 39 are confused with oblique behavior. The overall training accuracy is 66.5% slightly lower than the full model which reflects the trade-off between simplicity and predictive detail.

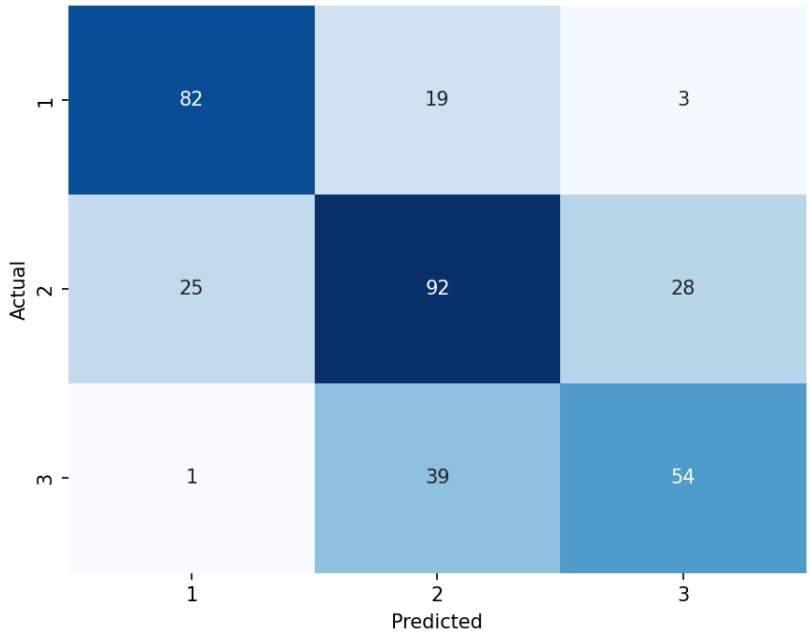


Figure 4.7. Confusion Matrix: Mitrapark (Reduced Model - Training)

ii. Testing the Model

The reduced model correctly classifies 36 of 44 perpendicular crossings (81.8%), 33 of 63 oblique crossings (52.4%) and 21 of 40 irregular crossings (52.5%). While perpendicular accuracy remains high, the decline in oblique and irregular performance compared to the full model confirms that the omitted predictors, though not individually

significant, collectively aid in distinguishing between these two more complex behaviors on unseen data.

Table 4.24. Per Class Accuracy: Mitrapark (Reduced Model - Testing)

Class	Correct	Total	Accuracy
1 (Perpendicular)	36	44	81.80%
2 (Oblique)	33	63	52.40%
3 (Irregular)	21	40	52.50%

The test confusion matrix in Figure 4.7 shows that perpendicular crossings remain the most reliably identified class with 36 of 44 cases correctly predicted and only one instance mislabeled as irregular. Oblique crossings are moderately captured with 33 of 63 are correct, though 21 are mistaken for irregular and 9 for perpendicular. Irregular crossings continue to present a challenge with 21 of 40 are correctly predicted but 18 are confused with oblique behavior. The test accuracy drops to 61.2% confirming that while the significant predictors retain much of the explanatory power, the reduced model loses some ability to discriminate between oblique and irregular crossings on unseen data.

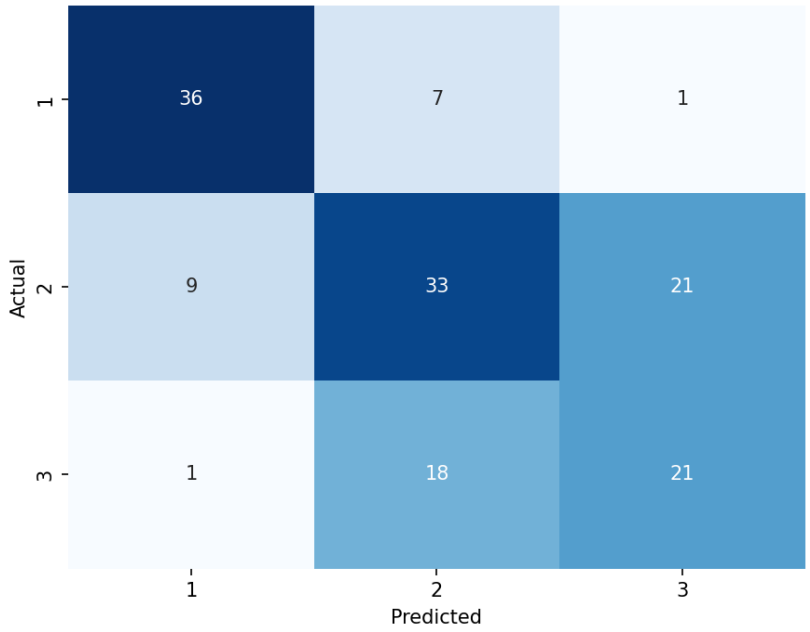


Figure 4.8. Confusion Matrix: Mitrapark (Reduced Model - Testing)

4.4.2. CatBoost

CatBoost was used to classify pedestrian crossing paths into three categories: Perpendicular, Oblique and Irregular. The model was selected for its ability to handle mixed categorical and continuous predictors and capture non-linear interactions in behavioral data. A 70:30 train-test split was applied and both full and reduced models were evaluated for both sites. The key results are summarized below.

A. Model I: Kamalpokhari (Considering all variables)

For Kamalpokhari, the full CatBoost model was developed using 460 observations with 322 for training and 138 for testing. The model classified three crossing path categories and was trained with 1000 iterations, a learning rate of 0.05 and a tree depth of 6. It achieved a McFadden's Pseudo R^2 of 0.2935 indicating moderate explanatory power for complex behavioral data. The model summary statistics are presented in Table 4.25.

Table 4.25. Model Summary: Kamalpokhari (Full Model)

Metric	Value
Total Samples	460
Training Samples	322
Testing Samples	138
Number of Classes	3
Model Parameters	{'iterations': 1000, 'learning_rate': 0.05, 'depth': 6, 'random_seed': 42, 'verbose': 100, 'eval_metric': 'Accuracy'}
Pseudo R^2 (McFadden)	0.2935

The overall predictive performance of the full Kamalpokhari CatBoost model was strong. The model achieved an accuracy of 0.659, a macro F1-score of 0.669 and a weighted F1-score of 0.659 indicating balanced classification performance across the three behavioral categories. These results show that the full model correctly classified approximately two-thirds of the observed crossing paths while maintaining similar performance across classes despite class size differences. The overall metrics are summarized in Table 4.26.

Table 4.26. Overall Metrics: Kamalpokhari (Full Model)

Metric	Value
Accuracy	0.659
Macro F1-score	0.669
Weighted F1-score	0.659

The CatBoost full model achieved strong classification performance on the training data at Kamalpokhari, with an overall accuracy of 86.4%. Perpendicular crossings were almost perfectly identified (94.2% correct) and oblique crossings also showed high accuracy (86.6%). Irregular crossings were the most challenging category with 59.5% correctly classified indicating that even the non-linear model struggles with the most complex pedestrian paths. Nonetheless, the performance across all three classes is markedly more balanced than that of the MNL models.

Table 4.27. Per Class Accuracy: Kamalpokhari (Full Model)

Class	Correct	Total	Accuracy
1 (Perpendicular)	130	138	94.20%
2 (Oblique)	136	157	86.60%
3 (Irregular)	25	42	59.50%

The classification report indicates strong predictive performance with an overall accuracy of 0.659. The model performs well across all classes with perpendicular crossings achieving the highest precision and recall while oblique and irregular crossings are also classified with reasonable accuracy. The balanced macro and weighted F1-scores further confirm consistent performance across categories. The classification summary is presented in Table 4.28.

Table 4.28. Classification Report: Kamalpokhari (Full Model)

	precision	recall	f1-score	support
1	0.732	0.789	0.759	38
2	0.590	0.621	0.605	58
3	0.694	0.595	0.641	42
accuracy	0.659	0.659	0.659	138
macro avg	0.672	0.668	0.669	138
weighted avg	0.661	0.659	0.659	138

The confusion matrix shows strong classification performance across all crossing types. Perpendicular crossings are identified accurately, while oblique crossings show moderate misclassification with some overlap into other categories. Irregular crossings are also well captured, though some instances are misclassified as oblique. Overall, the model demonstrates good class separation with relatively balanced performance across categories. The confusion matrix is illustrated in Figure 4.9.

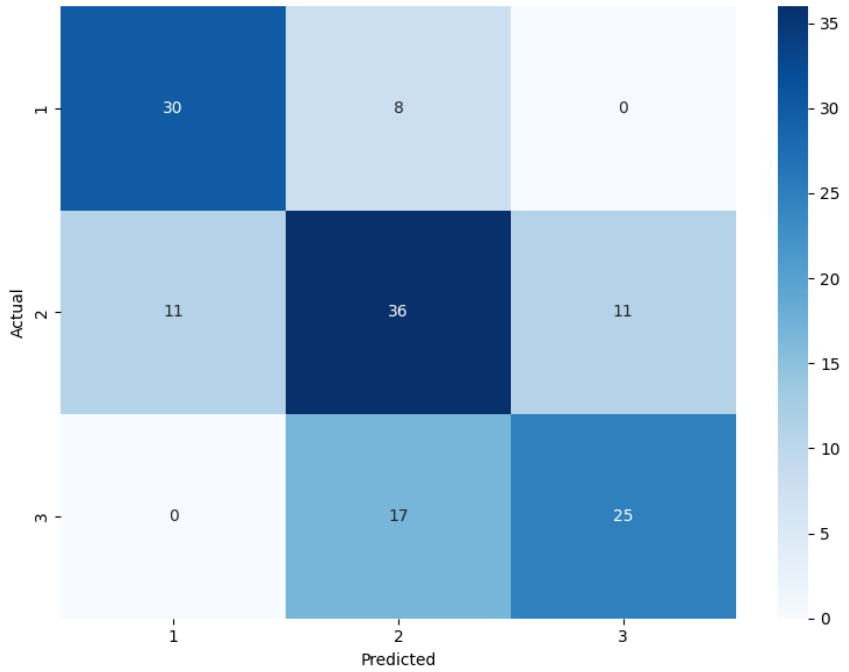


Figure 4.9. Confusion Matrix: Kamalpokhari (Full Model)

B. Model II: Kamalpokhari (Considering significant variables only)

A reduced CatBoost model for Kamalpokhari was estimated using the most influential predictors identified from the full model. The specification retained the same training-testing structure and parameter settings. Although the full model achieved a slightly higher pseudo R^2, the reduced model preserved the core predictive structure while improving interpretability and reducing complexity. The model summary is presented in Table 4.29.

Table 4.29. Model Summary: Kamalpokhari (Reduced Model)

Metric	Value
Total Samples	460
Training Samples	322
Testing Samples	138
Number of Classes	3
Model Parameters	{'iterations': 1000, 'learning_rate': 0.05, 'depth': 6, 'random_seed': 42, 'verbose': 100, 'eval_metric': 'Accuracy'}
Pseudo R^2 (McFadden-like)	0.2857

The reduced Kamalpokhari CatBoost model maintained similar overall accuracy of 0.667 as the full model with macro and weighted F1-scores remaining comparably stable. This indicates that the selected core predictors were sufficient to explain most of

the variation in crossing path and that removing weaker predictors did not compromise empirical performance. The reduced model overall metrics are presented in Table 4.30.

Table 4.30. Overall Metrics: Kamalpokhari (Reduced Model)

Metric	Value
Accuracy	0.667
Macro F1-score	0.675
Weighted F1-score	0.666

The reduced CatBoost model maintained virtually identical accuracy achieving 86.6% overall. Perpendicular and irregular accuracies remained unchanged while oblique accuracy improved marginally to 87.3%. This suggests that the significant predictors alone contain nearly all the information needed for CatBoost to separate the crossing types, and the excluded variables contributed little additional predictive power in the tree-based framework.

Table 4.31. Per Class Accuracy: Kamalpokhari (Reduced Model)

Class	Correct	Total	Accuracy
1 (Perpendicular)	130	138	94.20%
2 (Oblique)	137	157	87.30%
3 (Irregular)	25	42	59.50%

The reduced Kamalpokhari CatBoost model performs best for perpendicular crossings, moderately for oblique crossings and shows the greatest difficulty for irregular trajectories. Despite this, the overall accuracy and F1-scores indicate balanced classification performance across all categories. The classification results are presented in Table 4.32.

Table 4.32. Classification Report: Kamalpokhari (Reduced Model)

	precision	recall	f1-score	support
1	0.75	0.789	0.769	38
2	0.607	0.638	0.622	58
3	0.676	0.595	0.633	42
accuracy	0.667	0.667	0.667	138
macro avg	0.677	0.674	0.675	138
weighted avg	0.667	0.667	0.666	138

The performance of the reduced Kamalpokhari model is further illustrated in Figure 4.10 where the confusion matrix shows that Perpendicular crossings are classified most

reliably while some overlap persists between Oblique and Irregular trajectories. This pattern is consistent with the inherently more ambiguous and variable nature of non-perpendicular crossing path.

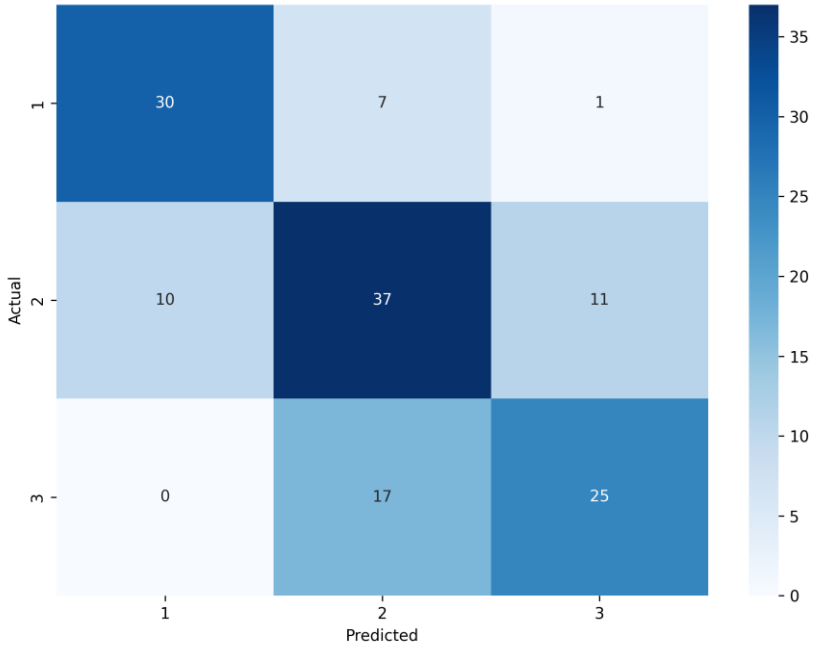


Figure 4.10. Confusion Matrix: Kamalpokhari (Reduced Model)

Overall, the Kamalpokhari CatBoost results show that reducing the predictor set did not compromise performance. The reduced model maintained similar accuracy and balanced classification across all crossing categories indicating that a smaller set of key predictors is sufficient to capture pedestrian crossing behavior at this site.

C. Model III: Mitrapark (Considering all variables)

The full CatBoost model was developed using 490 observations with 343 for training and 147 for testing, while retaining the same parameter settings as Kamalpokhari. The model achieved a McFadden's Pseudo R^2 of 0.1456 indicating moderate explanatory power for complex behavioral data. The model summary is presented in Table 4.33.

Table 4.33. Model Summary: Mitrapark (Full Model)

Metric	Value
Total Samples	490
Training Samples	343
Testing Samples	147
Number of Classes	3
Model Parameters	{'iterations': 1000, 'learning_rate': 0.05, 'depth': 6, 'random_seed': 42, 'verbose': 100, 'eval_metric': 'Accuracy'}
Pseudo R^2 (McFadden)	0.1456

The full-variable Mitrapark CatBoost model achieved an overall accuracy of 0.673 with a macro F1-score of 0.674 and a weighted F1-score of 0.673. These results indicate that the model correctly classified roughly two-thirds of the observed pedestrian trajectories while maintaining balanced performance across classes. The overall performance metrics are summarized in Table 4.34.

Table 4.34. Overall Metrics: Mitrapark (Full Model)

Metric	Value
Accuracy	0.673
Macro F1-score	0.674
Weighted F1-score	0.673

The CatBoost full model at Mitrapark achieved an overall accuracy of 66.4%. Perpendicular crossings were classified with high reliability (81.8%) but oblique accuracy fell to 65.1% and irregular crossings remained the weakest category at 52.4%. While the model outperforms the MNL specification, the progressive decline in accuracy toward the more complex crossing types shows that even a non-linear approach finds it difficult to fully separate oblique and irregular paths at this site.

Table 4.35. Per Class Accuracy: Mitrapark (Full Model)

Class	Correct	Total	Accuracy
1 (Perpendicular)	36	44	81.80%
2 (Oblique)	41	63	65.10%
3 (Irregular)	22	42	52.40%

The classification report indicates strong predictive performance with an overall accuracy of 0.673. The model performs best for perpendicular crossings while oblique and irregular crossings are also classified with good accuracy. The closely aligned

macro and weighted F1-scores indicate balanced performance across all classes. The classification summary is presented in Table 4.36.

Table 4.36. Classification Report: Mitrapark (Full Model)

	precision	recall	f1-score	support
1	0.818	0.818	0.818	44
2	0.631	0.651	0.641	63
3	0.579	0.550	0.564	40
accuracy	0.673	0.673	0.673	147
macro avg	0.676	0.673	0.674	147
weighted avg	0.673	0.673	0.673	147

The confusion matrix shows strong classification performance across all crossing types. Perpendicular crossings are identified with high accuracy while oblique crossings exhibit some misclassification particularly into the irregular category. Irregular crossings are also well captured, though a portion is misclassified as oblique. Overall, the model demonstrates good class separation with balanced performance across categories. The confusion matrix is illustrated in Figure 4.11.

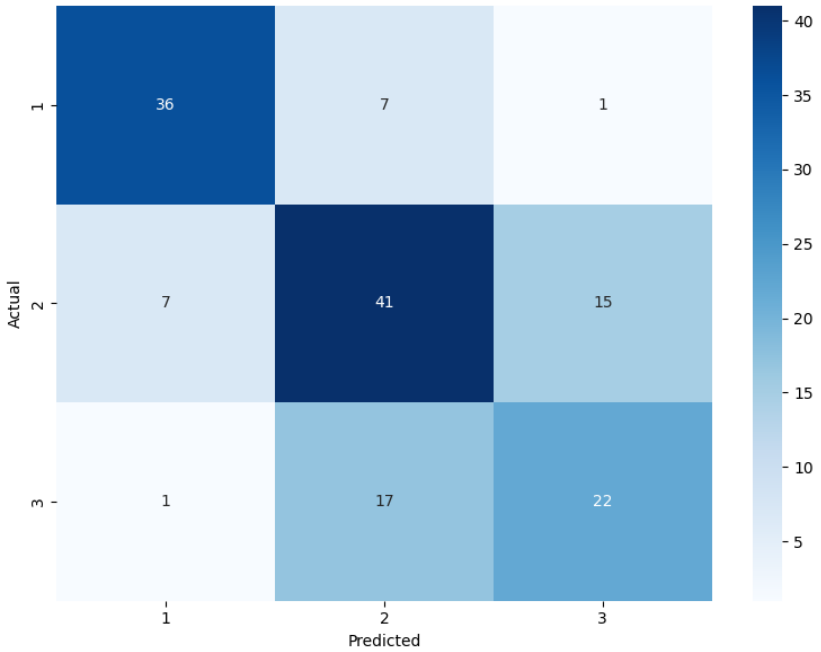


Figure 4.11. Confusion Matrix: Mitrapark (Full Model)

D. Model IV: Mitrapark (Considering significant variables only)

A reduced CatBoost model for Mitrapark was estimated using the most influential predictors from the full model. With the same 70:30 train-test split and parameter settings, it achieved a slightly higher McFadden-like Pseudo R^2 of 0.262 indicating a

more efficient representation of the underlying behavior. The model summary is presented in Table 4.37.

Table 4.37. Model Summary: Mitrapark (Reduced Model)

Metric	Value
Total Samples	490
Training Samples	343
Testing Samples	147
Number of Classes	3
Model Parameters	{'iterations': 1000, 'learning_rate': 0.05, 'depth': 6, 'random_seed': 42, 'verbose': 100, 'eval_metric': 'Accuracy'}
Pseudo R^2 (McFadden-like)	0.262

The reduced Mitrapark CatBoost model achieved an overall accuracy of 0.646, only slightly lower than the full model. The macro F1-score of 0.644 and weighted F1-score of 0.642 indicate balanced performance across all crossing categories while improving interpretability and simplicity. The performance metrics are presented in Table 4.38.

Table 4.38. Overall Metrics: Mitrapark (Reduced Model)

Metric	Value
Accuracy	0.646
Macro F1-score	0.644
Weighted F1-score	0.642

The CatBoost reduced model saw a modest decline in overall accuracy to 62.5%. Perpendicular performance remained strong (84.1%) while oblique accuracy dropped to 58.7% and irregular accuracy to 46.7%. The greater loss of performance at Mitrapark compared to Kamalpokhari suggests that the excluded variables contributed more to the predictive separation of the harder classes at this location and that the reduced set alone is less sufficient for capturing the variability in crossing behavior here.

Table 4.39. Per Class Accuracy: Mitrapark (Reduced Model)

Class	Correct	Total	Accuracy
1 (Perpendicular)	37	44	84.10%
2 (Oblique)	37	63	58.70%
3 (Irregular)	21	45	46.70%

The reduced Mitrapark CatBoost model performs best for perpendicular crossings, moderately for oblique crossings and least effectively for irregular crossings. Despite this, it maintains stable accuracy and balanced macro and weighted F1-scores indicating

that the selected predictors retain sufficient power to classify crossing paths effectively. The classification results are presented in Table 4.40.

Table 4.40. Classification Report: Mitrapark (Reduced Model)

	precision	recall	f1-score	support
1	0.725	0.841	0.779	44
2	0.627	0.587	0.607	63
3	0.568	0.525	0.545	40
accuracy	0.646	0.646	0.646	147
macro avg	0.640	0.651	0.644	147
weighted avg	0.640	0.646	0.642	147

The behavior of the reduced Mitrapark model is further illustrated in Figure 4.12 where the confusion matrix shows strong identification of Perpendicular crossings and moderate overlap between Oblique and Irregular movements. This again reflects the greater behavioral variability associated with less structured crossing trajectories.

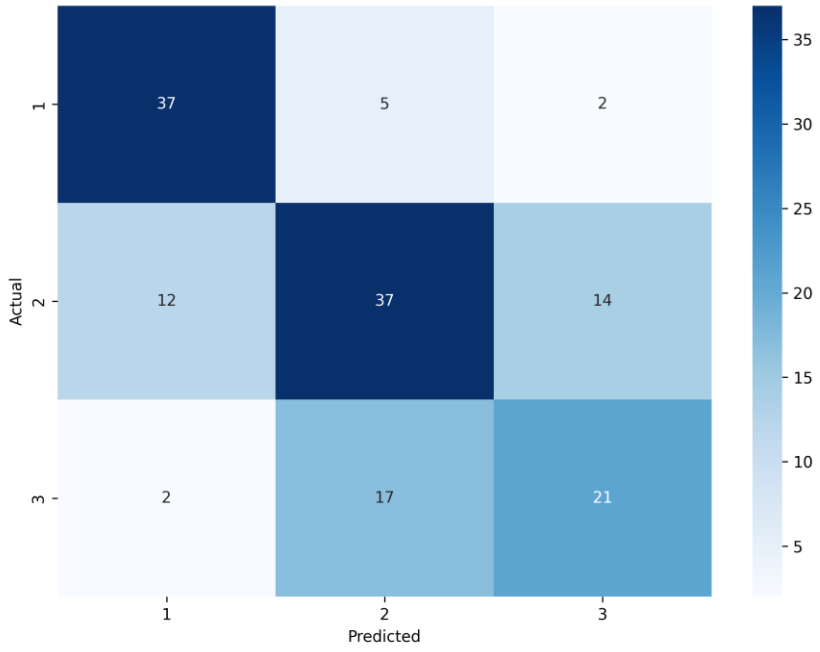


Figure 4.12. Confusion Matrix: Mitrapark (Reduced Model)

Overall, the Mitrapark CatBoost results show that the reduced model remains highly effective despite using fewer predictors. Although its accuracy is slightly lower than the full model, it maintains balanced performance across all classes and achieves a higher pseudo R^2 indicating that key determinants of crossing path are captured by a smaller, more interpretable set of variables.

Taken together, the CatBoost results across both sites demonstrate substantially stronger predictive performance than the MNL models. At Kamalpokhari, the full and reduced models achieved accuracies of 0.659 and 0.667 while at Mitrapark they achieved 0.673 and 0.646 respectively. In both cases, the reduced models preserved most predictive capability while offering simpler and more interpretable representations. These findings confirm that a limited set of key predictors is sufficient for robust classification of pedestrian crossing paths under mixed traffic midblock conditions.

4.5. Model Comparison: MLR vs. GAM for Accepted Gap

The comparison between the Multiple Linear Regression (MLR) model developed in SPSS and the Generalized Additive Model (GAM) implemented in Python for accepted gap size reveals small but meaningful differences in performance and interpretability across both study sites. At Kamalpokhari and Mitrapark, the MLR models provided strong explanatory performance and lower prediction error while the GAM models offered comparable explanatory power with greater flexibility in capturing non-linear relationships. MLR models are more interpretable through direct coefficients and retain a compact linear structure whereas GAMs better represent smooth, non-linear effects of variables such as pedestrian speed, waiting time and safety distance.

In terms of model behavior, GAM produced more balanced residual patterns indicating improved representation of non-linear variation in accepted gap decisions while MLR showed slightly stronger numerical fit and lower RMSE values. Overall, MLR is preferable when a simple and interpretable model is desired whereas GAM is more suitable for capturing behavioral complexity and gradual non-linear responses. A site wise comparison is presented in Table 4.41 for Kamalpokhari and Table 4.42 for Mitrapark.

Table 4.41. Model Comparison (MLR vs. GAM) for Kamalpokhari

Metric	MLR	GAM
R^2	0.668	0.684
Adjusted R^2	0.662	-
Cross-Validated R^2	-	0.653
RMSE	2.576 sec	7.256 sec
Significant Variables	Speed 1, Crossing Behaviour, Safety Distance 1.	Average Rejected Gap, Safety Distance 1, Speed 1.
Residual Shape	Slight skew	More balanced
Complexity	Simple, linear	Flexible, spline-based
Interpretability	High	Moderate (via plots)

Table 4.42. Model Comparison (MLR vs. GAM) for Mitrapark

Metric	MLR	GAM
R^2	0.734	0.772
Adjusted R^2	0.730	-
Cross-Validated R^2	-	0.743
RMSE	2.223 sec	4.216 sec
Significant Variables	Speed 1, Speed 2, Waiting Time, Safety Distance 1.	Average Rejected Gap, Safety Distance 1, Pedestrian Speed.
Residual Shape	Slight skew	More balanced
Complexity	Simple, linear	Flexible, spline-based
Interpretability	High	Moderate (via plots)

4.6. Model Comparison: MNL vs. CatBoost for Crossing Path

The comparison between the Multinomial Logistic Regression (MNL) model developed in SPSS and the CatBoost model implemented in Python for predicting pedestrian crossing path shows a clear advantage for CatBoost at both study sites. At Kamalpokhari and Mitrapark, the MNL models achieved moderate explanatory fit as indicated by McFadden's Pseudo R^2 and demonstrated reasonable classification performance, though some overlap among crossing categories persisted, particularly for oblique and irregular movements. This suggests that while MNL remains effective for identifying significant predictors and interpreting their effects, its predictive capability is still somewhat limited for highly variable pedestrian behaviour.

In contrast, CatBoost provided substantially stronger and more balanced predictive performance across all crossing categories. Both full and reduced models achieved higher accuracy and F1-scores than MNL, reflecting a greater ability to capture non-linear relationships and interactions among behavioural, traffic-related, and environmental variables. While both approaches identified key determinants such as pedestrian speed change, vehicle yield, waiting time, safety-related variables, and

vehicle speed, CatBoost translated these into more reliable classification performance. Overall, MNL remains valuable for interpretation, whereas CatBoost is more effective for predictive classification under mixed-traffic midblock conditions. A site-wise comparison is presented in Table 4.43 for Kamalpokhari and Table 4.44 for Mitrapark.

Table 4.43. Model Comparison (MNL vs. CatBoost) for Kamalpokhari

Metric	MNL	CatBoost
Accuracy (Test)	0.543	0.667
Macro F1-Score	-	0.675
Pseudo R^2 (McFadden)	0.4199	0.2857
Confusion Matrix	Imbalanced (High misclassification)	More balanced
Top Significant Variables	Vehicle Yield, Pedestrian Speed Change, Pedestrian Speed, Running.	Vehicle Yield, Pedestrian Speed, Waiting Time.
Interpretability	High (coefficients, p-values)	Moderate (feature importance, SHAP)
Complexity	Low (linear, parametric)	Medium (boosted decision trees)
Overfitting Risk	Higher (poor generalization)	Lower (validated accuracy)

Table 4.44. Model Comparison (MNL vs. CatBoost) for Mitrapark

Metric	MNL	CatBoost
Accuracy (Test)	0.612	0.646
Macro F1-Score	-	0.644
Pseudo R^2 (McFadden)	0.3976	0.262
Confusion Matrix	Imbalanced (High misclassification)	More balanced
Top Significant Variables	Vehicle Yield, Pedestrian Speed Change, Waiting Time.	Vehicle Yield, Pedestrian Speed, Average Rejected Gap.
Interpretability	High (coefficients, p-values)	Moderate (feature importance, SHAP)
Complexity	Low (linear, parametric)	Medium (boosted decision trees)
Overfitting Risk	Higher (poor generalization)	Lower (validated accuracy)

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1. Conclusions

This study evaluated pedestrian behavior at uncontrolled midblock crossings in Kathmandu using data collected from the Kamalpokhari and Mitrapark midblock sections. The analysis focused on two behavioral outcomes: accepted vehicular gap and crossing path choice. To examine these outcomes, accepted gap was modeled using Multiple Linear Regression (MLR) and Generalized Additive Model (GAM) while crossing path was modeled using Multinomial Logistic Regression (MNL) and CatBoost. The comparison was intended to assess how conventional statistical models and machine learning models perform under mixed-traffic midblock conditions.

For accepted gap, both MLR and GAM showed strong explanatory performance, although their strengths differed across the two sites. At Kamalpokhari, the reduced MLR model achieved an R^2 of 0.668 and an RMSE of 2.576 seconds while the reduced GAM achieved an R^2 of 0.684, a cross-validated R^2 of 0.653 and an RMSE of 7.256 seconds. At Mitrapark, the reduced MLR model achieved an R^2 of 0.734 and an RMSE of 2.223 seconds while the reduced GAM achieved an R^2 of 0.772, a cross-validated R^2 of 0.743 and an RMSE of 4.216 seconds. These results indicate that MLR provided lower prediction error and clearer interpretability whereas GAM provided slightly higher explanatory power and a more balanced residual structure through its ability to represent non-linear relationships.

For crossing path, the difference between the statistical and machine learning approaches was much more pronounced. The MNL models produced moderate explanatory fit with McFadden's pseudo R^2 values of 0.4199 at Kamalpokhari and 0.3976 at Mitrapark and their test accuracies were satisfactory at 0.543 and 0.612 respectively. In contrast, the CatBoost models performed slightly better with test accuracies of 0.667 at Kamalpokhari and 0.646 at Mitrapark and macro F1-scores of 0.675 and 0.644 respectively. This shows that CatBoost was more effective than MNL for predictive classification of pedestrian crossing path in complex mixed-traffic environments.

Several variables remained important across sites and models. For accepted gap, major contributors included Speed 1, Safety Distance 1, Average Rejected Gap, Waiting Time, Speed 2 and Pedestrian Speed. For crossing path, Vehicle Yield appeared consistently across both sites while Pedestrian Speed Change, Pedestrian Speed, Waiting Time, Running and Average Rejected Gap also emerged as important predictors depending on the site and model.

The key findings of the study are as follows:

i. For accepted gap, both MLR and GAM performed strongly but MLR produced lower RMSE values at both sites whereas GAM produced slightly higher R^2 values and more balanced residual behavior. At Kamalpokhari, MLR achieved R^2 of 0.668 and RMSE of 2.576s while GAM achieved R^2 of 0.684 and RMSE of 7.256s. At Mitrapark, MLR achieved R^2 of 0.734 and RMSE of 2.223s while GAM achieved R^2 of 0.772 and RMSE of 4.216s.

ii. For crossing path, CatBoost clearly outperformed MNL in predictive performance. The MNL models achieved test accuracies of 0.543 at Kamalpokhari and 0.612 at Mitrapark whereas CatBoost achieved 0.667 and 0.646 respectively with macro F1-scores of 0.675 and 0.644.

iii. MNL remained useful as an explanatory model because it provided interpretable coefficients and pseudo R^2 values but its satisfactory classification accuracy indicated that it was not able to fully capture the complexity of actual crossing path choices.

iv. CatBoost was better able to capture non-linear relationships and interactions among behavioral, traffic-related and environmental variables which explain its stronger classification performance across the three crossing categories.

v. Key predictors of accepted gap included Speed 1, Safety Distance 1, Average Rejected Gap, Speed 2, Waiting Time and Pedestrian Speed while key predictors of crossing path included Vehicle Yield, Pedestrian Speed Change, Pedestrian Speed, Waiting Time, Running and Average Rejected Gap.

The findings of this study directly address the problem identified at the outset of the research. The prevalence of oblique and irregular crossings indicates that pedestrians frequently adapt their crossing path in response to inadequate crossing facilities, traffic exposure and limited visibility at uncontrolled midblock locations. The comparison of

modeling approaches shows that while conventional statistical models remain valuable for interpretation, more flexible models such as GAM and CatBoost are better at capturing the complexity of pedestrian behavior under mixed traffic conditions. By using real world data collected from Kathmandu, this study provides context specific evidence that can support safer midblock crossing design, pedestrian facility improvement and future behavior based transport planning in Nepal. Overall, the study confirms that combining interpretable statistical models with more flexible machine learning models provides a stronger basis for understanding and predicting pedestrian behavior at uncontrolled midblock crossings.

5.2. Limitations

Pedestrian behavior at uncontrolled midblock crossings was examined using observations from only two locations, Kamalpokhari and Mitrapark, under daytime and normal weather conditions. Therefore, the findings may not be directly generalizable to night time conditions, adverse weather or other urban and rural settings. Accordingly, the findings should be interpreted as representative of similar uncontrolled urban midblock environments rather than all midblock crossings in Kathmandu.

The behavioral variables were manually extracted from video footage which may introduce observer-related inconsistencies particularly for comparatively subjective variables such as crossing path classification and pedestrian speed change. Although a clear coding procedure was followed during data extraction, some degree of classification error cannot be completely ruled out.

Although the dataset of 950 pedestrian observations was adequate for the present analysis, some crossing path categories were less frequently represented than others. This class imbalance may have affected model training and contributed to weaker predictive performance for minority classes particularly in the multinomial and machine learning based classification models.

In addition, the models developed in this study focused primarily on pedestrian-side variables and did not explicitly incorporate driver-side behavioral factors such as braking response, horn use, attentiveness or evasive action. As a result, pedestrian risk was evaluated mainly from the pedestrian perspective whereas a more comprehensive

assessment of midblock safety would require integrated analysis of both pedestrian and driver behavior.

5.3. Recommendations

Based on the empirical findings from Kamalpokhari and Mitrapark, the following recommendations are proposed:

i. Improve the alignment and visibility of pedestrian crossing facilities with observed pedestrian movement patterns: The high prevalence of oblique and irregular crossings suggests that pedestrians often preferred direct crossing paths under mixed-traffic conditions. Therefore, pedestrian safety measures should focus on improving the accessibility, visibility and usability of existing crossing facilities while minimizing unnecessary detours and supporting more predictable crossing behavior.

ii. Introduce context-sensitive speed management measures near frequent pedestrian crossing locations: Higher vehicle approach speeds were associated with changes in pedestrian crossing behavior and reduced safety margins. Appropriate speed management measures such as rumble strips, advance warning markings, speed display signs, targeted enforcement or localized traffic calming may help improve pedestrian safety under mixed-traffic conditions..

iii. Improve sidewalk continuity and reduce roadside obstructions near pedestrian crossing areas: The presence of roadside obstructions was associated with more oblique and irregular crossing paths, as pedestrians were often required to maneuver around parked vehicles or encroachments. Maintaining continuous pedestrian pathways and minimizing roadside obstructions near likely crossing locations may help support safer and more predictable pedestrian movement.

iv. Address pedestrian behavioral factors through targeted site-level safety measures: Variables such as running, pedestrian speed change and waiting time were found to influence crossing path behavior. Therefore, pedestrian safety interventions should combine improved crossing conditions, local guidance measures and public awareness efforts to encourage safer crossing behavior under mixed-traffic conditions.

v. Promote improved vehicle yielding behavior near pedestrian crossing locations: Vehicle yield was one of the most influential factors associated with safer and more stable crossing choices. Measures such as clearer road markings, advance yield lines, driver awareness initiatives and targeted enforcement may help improve yielding behavior and support safer pedestrian crossing conditions.

vi. Incorporate behavior-based modeling approaches in pedestrian safety assessment: The application of GAM and CatBoost alongside conventional statistical models demonstrated the usefulness of behavior-based approaches in capturing non-linear relationships and complex interactions associated with pedestrian crossing behavior. Such approaches may support future pedestrian safety evaluation and midblock crossing studies under mixed-traffic conditions. Transport agencies and urban planners in Kathmandu may also consider incorporating these modeling approaches when assessing pedestrian safety and evaluating potential midblock crossing improvements.

5.4. Future Work

Building on the insights obtained from Kamalpokhari and Mitrapark and considering the limitations of the present study, the following directions for future research are proposed:

i. Broaden spatial and temporal coverage: Future studies should include additional locations within Kathmandu and other urban corridors as well as nighttime conditions, rainy weather and festival periods in order to capture variations in pedestrian urgency, visibility and risk taking behavior across a wider range of contexts. In exceptionally high volume or high speed corridors, future studies may also examine the feasibility of grade separated pedestrian facilities such as subways provided that issues of accessibility, drainage, security, maintenance and impacts on nearby residential and commercial access are carefully evaluated.

ii. Adopt automated and richer data collection methods: Automated video analysis using computer vision and sensor-based tracking may reduce observer-related

bias, improve the accuracy of trajectory and speed measurements and support larger scale and longer duration data collection.

iii. Incorporate vehicle-side behavior and two-way interactions: Future research should integrate vehicle-side variables such as detailed speed profiles, braking response, horn use and video-based driver behavior observations so that pedestrian-vehicle interactions can be modeled more comprehensively rather than from the pedestrian side alone.

iv. Explore advanced and hybrid modeling approaches: Building on the use of GAM and CatBoost in this study, future work may examine additional machine learning and hybrid modeling frameworks along with external validation across different sites and conditions to improve prediction accuracy, generalizability and model robustness.

REFERENCES

Agresti, A. (2007). *An Introduction to Categorical Data Analysis (2nd ed.).*

Alver, Y., Onelcin, P., Cicekli, A., & Abdel-Aty, M. (2021). Evaluation of pedestrian critical gap and crossing speed at midblock crossing using image processing. *Accident Analysis & Prevention* .

Aziz, H. A., Ukkusuri, V. S., & Hasan, S. (2013). Exploring the determinants of pedestrian–vehicle crash severity in New York City. *Accident Analysis & Prevention* , 1298-1309.

Cherry, C., Donlon, B., Yan, X., Moore, E. S., & Xiong, J. (2012). Illegal mid-block pedestrian crossings in China: gap acceptance, conflict and crossing path analysis. *International Journal of Injury Control and Safety Promotion* , 320-330.

Evan, D., & Norman, P. (1998). Understanding pedestrians' road crossing decisions: an application of the theory of planned behaviour. *Health Education Research, Volume 13, Issue 4* , 481-489.

Govinda, L., & Ravishankar, K. V. (2022). A critical review on pedestrian crossing behaviour and pedestrian-vehicle interactions. *Innovative Infrastructure Solutions* , 313.

Hamed, M. M. (2001). Analysis of pedestrians' behavior at pedestrian crossings. *Safety Science* , 63-82.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statist. Sci.* , 297-310.

Jain, S., Advani, M., & Yadav, L. K. (2022). Pedestrians Safety Analysis at Uncontrolled Midblock Crosswalks. *Recent Advances in Transportation Systems Engineering and Management* (pp. 763-774). Lecture Notes in Civil Engineering, vol 261. Springer, Singapore.

JICA. (2012). *Data Collection Survey on Improvement in Kathmandu Valley.* Kathmandu: Department of Roads.

Kadali, B. R., & Vedagiri, P. (2016). Proactive pedestrian safety evaluation at unprotected mid-block crosswalk locations under mixed traffic conditions. *Safety Science* , 94-105.

Kadali, B. R., Chiranjeevi, T., & Rajesh, R. (2015). EFFECT OF PEDESTRIANS UN-SIGNALIZED MID-BLOCK CROSSING ON VEHICULAR SPEED. *International Journal for Traffic and Transport Engineering* , 170-183.

Kadali, B., & Vedagiri, P. (2013). Modelling pedestrian road crossing behaviour under mixed traffic condition. *European Transport* .

Laflamme, E. M., Villamagna, A., & Kim, H. J. (2024). Predicting severe wildlife vehicle crashes (WVCs) on New Hampshire roads using a hybrid generalized additive model. *Archives of Transport* .

Li, G., Wu, Y., Bai, Y., & Zhang, W. (2023). ReMAHA–CatBoost for imbalanced crash severity prediction. *Applied Sciences* .

MTPD. (2019). *Safe & Sustainable Travel Nepal Annual Report.* Nepal Engineers' Association.

Muraleetharan, T., & Hagiwara, T. (2007). Overall Level of Service of Urban Walking Environment and Its Influence on Pedestrian Route Choice Behavior: Analysis of Pedestrian Travel in Sapporo, Japan. *Transportation Research Record: Journal of the Transportation Research Board* , 7-17.

Oxley, J. A., Ihsen, E., Fildes, B. N., Charlton, J. L., & Day, R. H. (2005). Crossing roads safely: An experimental study of age differences. *Accident Analysis and Prevention 37* , 962-971.

Papadimitriou, E., Yannis, G., & Golias, J. (2009). A critical assessment of pedestrian behaviour models. *Transportation Research Part F* , 242-255.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* , 6639-6649.

Puig-Divi, A., Escalona-Marfil, C., Padullés-Riu, M. J., Busquets, A., Padullés-Chando, X., & Marcos-Ruiz, D. (2019). Validity and reliability of the Kinovea program in obtaining angles and distances using coordinates in 4 perspectives.

Rastogi, R., Thaniarasu, I., & Chandra, S. (2011). Design Implications of Walking Speed. *Journal of Transportation Engineering* .

Samerei, A. S., Aghabayk, K., & Montella, A. (2024). Analyzing Pile-Up Crash Severity: Insights from Real-Time Traffic and Environmental Factors Using Ensemble Machine Learning and Shapley Additive Explanations Method. *Safety* .

Shaaban, K., Muley, D., & Mohammed, A. (2018). Analysis of illegal pedestrian crossing behavior on a major divided arterial road. *Transportation Research Part F: Traffic Psychology and Behaviour* , 124-137.

Shahi, T. C., & Gautam, S. (2020). An Assessment of Pedestrian Non-compliance Behavior along Urban Roads: Case Study of Kathmandu. *IOSR Journal of Mechanical and Civil Engineering* , 23-31.

Singh, D., Das, P., & Ghosh, I. (2024). Prediction of pedestrian crossing behaviour at unsignalized intersections using machine learning algorithms: analysis and comparison. *Journal on Multimodal User Interfaces* , 239-256.

Tezcan, H. O., Elmorssy, M., & Aksoy, G. (2019). Pedestrian crossing behavior at midblock crosswalks. *Journal of Safety Research* , 49-57.

The World Bank. (2020). *Delivering Road Safety in Nepal.* The World Bank.

Theobald, N., Joisten, P., & Abendroth, B. (2022). Measuring Pedestrians' Gap Acceptance When Interacting with Vehicles - A Human Gait Oriented Approach. *HCI International 2022 Posters* (pp. 251-258). Communications in Computer and Information Science, vol 1583. Springer, Cham.

Torres, C., Sobreira, L., Castro-Neto, M., Cunto, F., Vecino-Ortiz, A., Allen, K., et al. (2020). Evaluation of Pedestrian Behavior on Mid-block Crosswalks: A Case Study in Fortaleza—Brazil. *Advances in Road Safety Planning* .

Wood, S. N. (2017). *Generalized Additive Models.* New York: Chapman and Hall/CRC.

Xie, Y., & Zhang, Y. (2008). Crash Frequency Analysis with Generalized Additive Models. *Transportation Research Record: Journal of the Transportation Research Board* , 39-45.

Yannis, G., Papadimitriou, E., & Theofilatos, A. (2013). Pedestrian gap acceptance for mid-block street crossing. *Transportation Planning and Technology* , 450-462.

Zhang, C., Sprenger, J., Ni, Z., & Berger, C. (2024). Predicting and Analyzing Pedestrian Crossing Behavior at Unsignalized Crossings. *IEEE Intelligent Vehicles Symposium (IV)* .

Zhang, C., Zhou, B., Qiu, T. Z., & Liu, S. (2018). Pedestrian crossing behaviors at uncontrolled multi-lane mid-block crosswalks in developing world. *Journal of Safety Research* , 145-154.

Zhang, Y., Xie, Y., & Li, L. (2012). Crash frequency analysis of different types of urban roadway segments using generalized additive model. *J Safety Res.* , 107-114.

Zhang, Z., Li, H., Sze, N. N., & Ren, G. (2023). Investigating pedestrian crossing route choice at mid-blocks without crossing facilities: The role of roadside environment. *Travel Behaviour and Society* .

Zhao, X., Qi, H., Yao, Y., Guo, M., & Su, Y. (2023). Traffic Order Analysis of Intersection Entrance Based on Aggressive Driving Behavior Data Using CatBoost and SHAP. *Journal of Transportation Engineering, Part A: Systems* .

Zhuang, X., & Wu, C. (2011). Pedestrians' crossing behaviors and safety at unmarked roadway in China. *Accident Analysis & Prevention* , 1927-1936.

APPENDIX A: Data Entry Form and Variable Coding

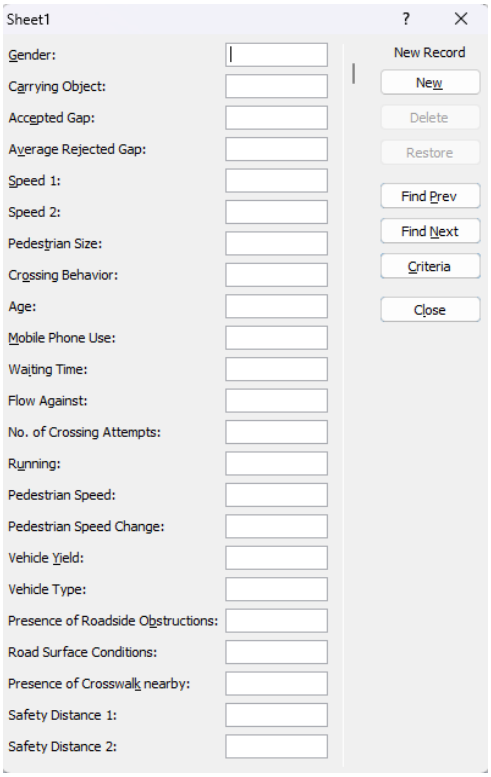
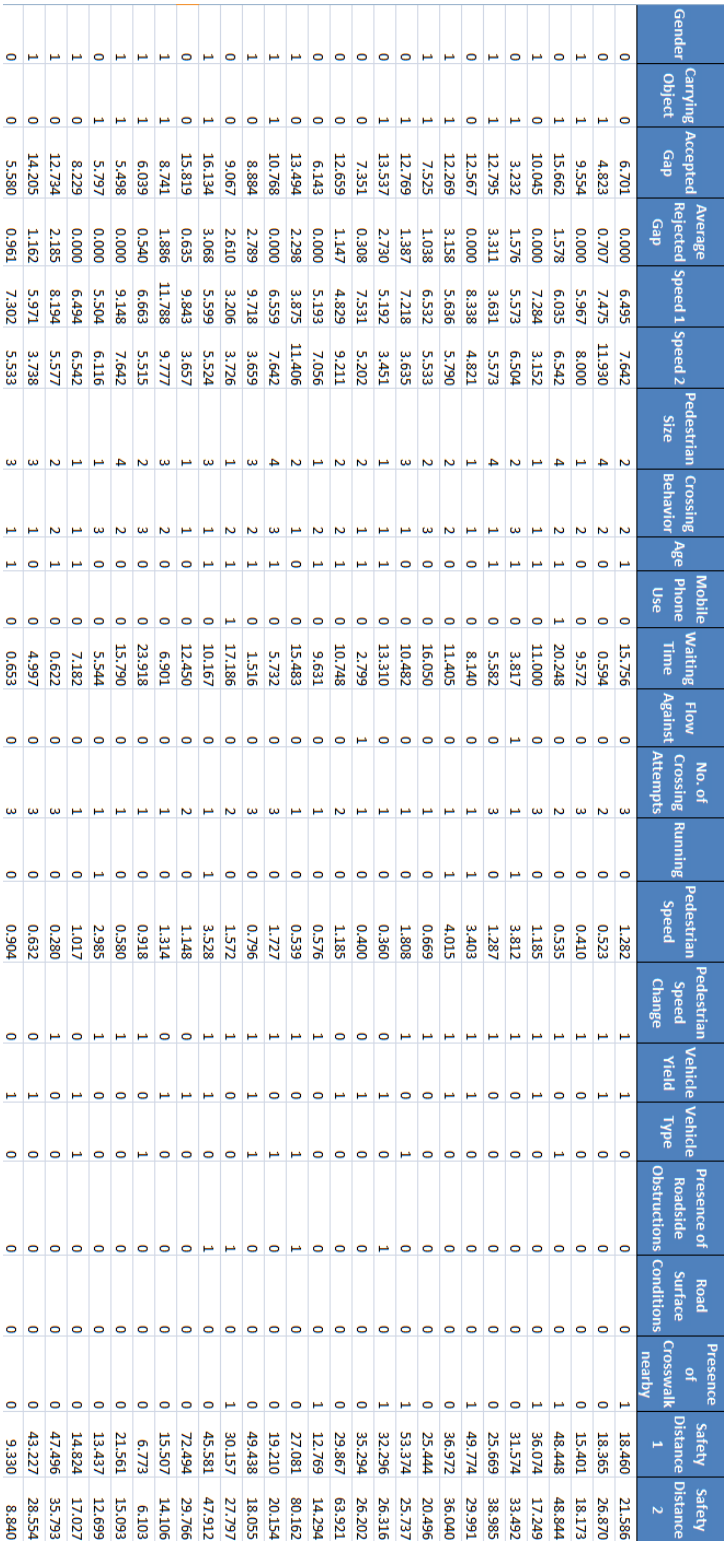


Figure A.1. Data Entry Form

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	1	0	Gender	{0, Female}...	None	12	Right	Nominal	Input
2	CarryingObj...	Numeric	1	0	Carrying_Object	{0, No}...	None	12	Right	Nominal	Input
3	AcceptedGap	Numeric	6	3	Accepted_Gap	None	None	16	Right	Scale	Input
4	AverageRej...	Numeric	6	3	Average_Reject...	None	None	16	Right	Scale	Input
5	Speed1	Numeric	6	3	Speed_1	None	None	16	Right	Scale	Input
6	Speed2	Numeric	6	3	Speed_2	None	None	12	Right	Scale	Input
7	PedestrianS...	Numeric	1	0	Pedestrian_Size	None	None	12	Right	Scale	Input
8	CrossingBe...	Numeric	1	0	Crossing_Beha...	{1, Perpendi...	None	12	Right	Nominal	Input
9	Age	Numeric	1	0	Age	{0, Age>30}...	None	12	Right	Nominal	Input
10	MobilePhon...	Numeric	1	0	Mobile_Phone_...	{0, No}...	None	12	Right	Nominal	Input
11	WaitingTime	Numeric	6	3	Waiting_Time	None	None	16	Right	Scale	Input
12	FlowAgainst	Numeric	1	0	Flow_Against	{0, No}...	None	12	Right	Nominal	Input
13	No.ofCrossi...	Numeric	1	0	No._of_Crossin...	None	None	12	Right	Scale	Input
14	Running	Numeric	1	0	Running	{0, No}...	None	12	Right	Nominal	Input
15	PedestrianS...	Numeric	6	3	Pedestrian_Sp...	None	None	16	Right	Scale	Input
16	PedestrianS...	Numeric	1	0	Pedestrian_Sp...	{0, No}...	None	12	Right	Nominal	Input
17	VehicleYield	Numeric	1	0	Vehicle_Yield	{0, No}...	None	12	Right	Nominal	Input
18	VehicleType	Numeric	1	0	Vehicle_Type	{0, 2W}...	None	12	Right	Nominal	Input
19	Presenceof...	Numeric	1	0	Presence_of_R...	{0, No}...	None	12	Right	Nominal	Input
20	RoadSurfac...	Numeric	1	0	Road_Surface_...	{0, Good}...	None	12	Right	Nominal	Input
21	Presenceof...	Numeric	1	0	Presence_of_C...	{0, No}...	None	12	Right	Nominal	Input
22	SafetyDista...	Numeric	6	3	Safety_Distanc...	None	None	16	Right	Scale	Input
23	SafetyDista...	Numeric	6	3	Safety_Distanc...	None	None	16	Right	Scale	Input

Figure A.2. Variable Coding SPSS

APPENDIX B: Sample Data & Correlation Matrices



Gender	Carrying Object	Accepted Gap	Average Rejected Gap	Speed 1	Speed 2	Pedestrian Size	Crossing Behavior	Age	Mobile Phone Use	Waiting Time	Flow Against	No. of Crossing Attempts	Running	Pedestrian Speed	Pedestrian Speed Change	Vehicle Yield	Vehicle Type	Presence of Roadside Obstructions	Road Surface Conditions	Presence of Crosswalk nearby	Safety Distance 1	Safety Distance 2
0	0	6.701	0.000	6.495	7.642	2	2	1	0	15.756	0	3	0	1.282	1	1	0	0	0	1	18.460	21.586
0	1	4.823	0.707	7.475	11.930	4	2	0	0	0.594	0	2	0	0.523	1	1	0	0	0	0	18.365	26.870
1	1	9.554	0.000	5.967	8.000	1	2	0	0	9.572	0	3	0	0.410	1	0	0	0	0	0	15.401	18.173
0	1	15.662	1.578	6.035	6.542	4	2	1	1	20.248	0	2	0	0.535	1	0	1	0	0	1	48.448	48.844
1	0	10.045	0.000	7.284	3.152	1	1	1	0	11.000	0	3	0	1.185	1	1	0	0	0	1	36.074	17.249
0	1	3.232	1.576	5.573	6.504	2	3	1	0	3.817	1	1	1	3.812	1	0	0	0	0	0	31.574	33.492
1	1	12.795	3.311	3.631	5.573	4	1	1	0	5.582	0	3	0	1.287	1	0	0	0	0	0	25.669	38.985
0	0	12.567	0.000	8.338	4.821	1	1	0	0	8.140	0	1	1	3.403	1	1	0	0	0	1	49.774	29.991
1	1	12.269	3.158	5.636	5.790	2	2	0	0	11.405	0	1	1	4.015	1	1	0	0	0	0	36.972	36.040
1	1	7.525	1.038	6.532	5.533	2	3	0	0	16.050	0	1	0	0.669	1	0	0	0	0	0	25.444	20.496
0	1	12.769	1.387	7.218	3.635	3	1	0	0	10.482	0	1	0	1.808	1	0	1	0	0	1	53.374	25.737
0	1	13.537	2.730	5.192	3.451	1	1	1	0	13.310	0	1	0	0.360	0	1	0	1	0	1	32.296	26.316
0	0	7.351	0.308	7.531	5.202	2	1	1	0	2.799	1	1	0	0.400	0	1	0	0	0	0	35.294	26.202
0	0	12.659	1.147	4.829	9.211	2	2	1	0	10.748	0	2	0	1.185	0	1	0	0	0	0	29.867	63.921
0	0	6.143	0.000	5.193	7.056	1	2	1	0	9.631	0	1	0	0.576	1	0	0	0	0	1	12.769	14.294
1	0	13.494	2.298	3.875	11.406	2	1	0	0	15.483	0	1	0	0.539	1	0	1	1	0	0	27.081	80.162
1	1	10.768	0.000	6.559	7.642	4	3	1	0	5.732	0	3	0	1.727	1	0	1	0	0	0	19.210	20.154
1	0	8.884	2.789	9.718	3.659	3	2	1	0	1.516	0	3	0	0.796	1	1	1	0	0	0	49.438	18.055
0	0	9.067	2.610	3.206	3.726	1	2	1	1	17.186	0	2	0	1.572	1	0	0	1	0	1	30.157	27.797
1	1	16.134	3.068	5.599	5.524	3	1	1	0	10.167	0	1	1	3.528	1	1	0	1	0	0	45.581	47.912
0	0	15.819	0.635	9.843	3.657	1	1	0	0	12.450	0	2	0	1.148	0	1	0	0	0	0	72.494	29.766
1	1	8.741	1.886	11.788	9.777	3	2	0	0	6.901	0	1	0	1.314	0	1	0	0	0	0	15.507	14.106
1	1	6.039	0.540	6.663	5.515	2	3	0	0	23.918	0	1	0	0.918	1	0	1	0	0	0	6.773	6.103
1	1	5.498	0.000	9.148	7.642	4	2	0	0	15.790	0	1	0	0.580	1	0	0	0	0	0	21.561	15.093
0	1	5.797	0.000	5.504	6.116	1	3	0	0	5.544	0	1	1	2.985	1	0	0	0	0	0	13.437	12.699
1	0	8.229	0.000	6.494	6.542	1	1	1	0	7.182	0	1	0	1.017	0	1	1	0	0	0	14.824	17.027
1	0	12.734	2.185	8.194	5.577	2	2	1	0	0.622	0	3	0	0.280	1	0	0	0	0	0	47.496	35.793
1	0	14.205	1.162	5.971	3.738	3	1	0	0	4.997	0	3	0	0.632	0	1	0	0	0	0	43.227	28.554
0	0	5.580	0.961	7.302	5.533	3	1	1	0	0.653	0	3	0	0.904	0	1	0	0	0	0	9.330	8.840

Figure B.1. Sample Data

	Crossing Behavior	Gender	Carrying Object	Age	Mobile Phone Use	Flow Against	Running	Pedestrian Speed Change	Vehicle Yield	Vehicle Type	Presence of Roadside Obstructions	Road Surface Conditions	Presence of Crosswalk nearby	Accepted Gap	Average Rejected Gap	Speed 1	Speed 2	Pedestrian Size	Waiting Time	No. of Crossing Attempts	Pedestrian Speed	Safety Distance 1	Safety Distance 2
Crossing Behavior	1	-0.052	0.072	-0.072	0.061	-0.032	0.041	0.590	-0.634	0.043	-0.179	-0.008	0.056	-0.124	0.016	0.015	0.033	0.108	0.203	-0.012	-0.021	0.017	-0.024
Gender	-0.052	1	0.011	-0.011	-0.018	-0.155	-0.075	-0.110	-0.024	-0.043	0.042	0.063	-0.038	-0.019	-0.032	-0.093	-0.021	0.040	0.079	0.061	0.004	-0.020	0.018
Carrying Object	0.072	0.011	1	-0.012	-0.095	0.041	0.073	0.095	-0.081	-0.037	-0.117	-0.028	-0.105	0.096	0.054	0.045	0.006	0.116	-0.066	-0.049	0.061	-0.004	-0.023
Age	-0.072	-0.011	-0.012	1	-0.033	0.130	0.063	-0.025	0.079	0.037	0.053	-0.011	0.016	0.067	-0.009	0.102	0.010	0.005	-0.004	0.074	0.067	0.103	0.063
Mobile Phone Use	0.061	-0.018	-0.095	-0.033	1	-0.089	-0.043	0.002	-0.060	-0.007	0.045	0.015	0.067	-0.008	-0.027	0.004	0.061	-0.043	0.041	0.006	-0.036	0.014	0.010
Flow Against	-0.032	-0.155	0.041	0.130	-0.089	1	0.060	-0.018	0.075	0.081	-0.130	0.035	-0.038	-0.018	-0.051	0.005	-0.161	-0.034	-0.089	-0.042	0.060	0.003	-0.067
Running	0.041	-0.075	0.073	0.063	-0.043	0.060	1	0.264	-0.029	0.024	-0.032	-0.071	-0.084	-0.083	-0.085	0.022	0.002	-0.013	0.010	-0.070	0.603	-0.024	-0.037
Pedestrian Speed Change	0.590	-0.110	0.095	-0.025	0.002	-0.018	0.264	1	-0.616	0.040	-0.130	-0.007	0.028	-0.084	0.031	-0.033	0.008	0.090	0.168	0.007	0.172	-0.049	-0.084
Vehicle Yield	-0.634	-0.024	-0.081	0.079	-0.060	0.075	-0.029	-0.616	1	0.024	0.175	-0.048	-0.077	0.080	0.012	0.121	-0.067	-0.023	-0.209	0.024	-0.030	0.041	0.006
Vehicle Type	0.043	-0.043	-0.037	0.037	-0.007	0.081	-0.006	0.040	0.024	1	-0.066	0.001	0.043	-0.119	-0.075	0.051	-0.052	0.072	-0.084	-0.045	0.074	-0.064	-0.099
Presence of Roadside Obstructions	-0.179	0.042	-0.117	0.053	0.045	-0.130	-0.032	-0.130	0.175	-0.066	1	-0.066	0.000	0.024	-0.015	-0.043	0.049	-0.057	-0.013	-0.066	-0.034	0.009	0.058
Road Surface Conditions	-0.008	0.063	-0.028	-0.011	0.015	0.035	-0.071	-0.007	-0.048	0.001	-0.066	1	-0.043	-0.048	-0.007	-0.036	0.082	0.003	0.018	0.034	0.013	-0.057	0.010
Presence of Crosswalk nearby	0.056	-0.038	-0.105	0.016	0.067	-0.038	-0.084	0.028	-0.077	0.043	0.000	-0.043	1	0.049	0.021	-0.016	-0.014	0.032	0.108	-0.066	-0.067	0.040	0.033
Accepted Gap	-0.124	-0.019	0.006	0.067	-0.008	-0.018	-0.083	-0.084	0.060	-0.119	0.024	-0.048	0.049	1	0.292	0.069	0.026	0.098	0.060	-0.020	-0.067	0.729	0.722
Average Rejected Gap	0.016	-0.032	0.054	-0.009	-0.027	-0.051	-0.085	0.031	0.012	-0.075	-0.015	-0.007	0.021	0.292	1	-0.035	-0.007	0.298	-0.019	0.053	-0.009	0.208	0.217
Speed 1	0.015	-0.093	0.045	0.102	0.004	0.005	0.022	-0.033	0.121	0.051	-0.043	-0.036	-0.016	0.069	-0.035	1	0.083	0.086	-0.063	0.086	-0.030	0.372	-0.027
Speed 2	0.033	-0.021	0.006	0.010	0.061	0.161	0.002	0.008	-0.067	-0.052	0.049	0.082	-0.014	0.026	-0.007	0.083	1	-0.028	0.040	0.054	-0.115	-0.105	0.371
Pedestrian Size	0.108	0.040	0.116	0.005	-0.043	-0.034	-0.013	0.090	-0.023	0.072	-0.057	0.003	0.032	0.098	0.298	0.086	-0.028	1	0.029	0.022	-0.023	0.128	0.064
Waiting Time	0.203	0.079	-0.066	-0.004	0.041	-0.089	0.010	0.168	-0.209	-0.084	-0.013	0.018	0.108	0.080	-0.019	-0.063	0.040	0.029	1	-0.030	-0.006	0.057	0.097
No. of Crossing Attempts	-0.012	0.061	-0.049	0.074	0.006	-0.042	-0.070	0.007	0.024	-0.045	-0.066	0.034	-0.066	-0.020	0.053	0.086	0.054	0.022	-0.030	1	-0.026	-0.000	-0.043
Pedestrian Speed	-0.021	0.004	0.061	0.067	-0.036	0.060	0.603	0.172	-0.030	0.074	-0.034	0.013	-0.067	-0.067	-0.009	-0.030	-0.115	-0.023	-0.006	-0.026	1	-0.013	-0.050
Safety Distance 1	0.017	-0.020	-0.004	0.103	0.014	0.003	-0.024	-0.049	0.041	-0.064	0.009	-0.057	0.040	0.729	0.208	0.372	-0.105	0.128	0.057	-0.000	-0.013	1	0.741
Safety Distance 2	-0.024	0.018	-0.023	0.063	0.010	-0.067	-0.037	-0.084	0.006	-0.099	0.058	0.010	0.033	0.722	0.217	-0.027	0.371	0.064	0.097	-0.043	-0.050	0.741	1

Figure B.2. Correlation Matrix: Kamalpokhari

	Crossing Behavior	Gender	Carrying Object	Age	Mobile Phone Use	Flow Against	Running	Pedestrian Speed Change	Vehicle Yield	Vehicle Type	Presence of Illegal Obstructions	Road Surface Conditions	Presence of Crosswalk nearby	Accepted Gap	Average Rejected Gap	Speed 1	Speed 2	Pedestrian Size	Waiting Time	No. of Crossing Attempts	Pedestrian Speed	Safety Distance 1	Safety Distance 2
Crossing Behavior	1	-0.058	-0.001	0.009	0.016	0.002	0.024	0.632	-0.607	0.074	-0.124	0.134	-0.038	-0.107	-0.070	-0.123	-0.022	0.035	0.168	0.084	0.000	-0.124	-0.127
Gender	-0.058	1	-0.052	-0.056	-0.033	-0.063	-0.016	0.029	0.002	0.017	0.008	-0.012	-0.063	0.034	-0.021	-0.088	-0.052	-0.019	-0.014	-0.033	0.138	-0.018	-0.006
Carrying Object	-0.001	-0.052	1	-0.010	-0.019	-0.031	-0.036	-0.026	-0.069	-0.083	-0.084	-0.043	-0.064	0.035	0.163	0.080	0.063	0.128	-0.127	0.043	-0.021	0.013	-0.009
Age	0.009	-0.056	-0.010	1	-0.079	0.069	-0.065	0.016	0.041	0.067	0.051	-0.004	0.056	0.088	0.041	0.064	0.014	0.065	0.021	0.014	-0.075	0.103	0.101
Mobile Phone Use	0.016	-0.033	-0.019	-0.079	1	0.017	-0.066	0.041	-0.092	-0.027	-0.091	0.020	-0.041	-0.064	-0.128	-0.034	0.028	-0.055	0.020	-0.032	-0.086	-0.049	-0.039
Flow Against	0.002	-0.063	-0.031	0.069	0.017	1	0.026	-0.027	-0.014	-0.018	0.056	-0.039	0.068	-0.068	0.009	-0.060	-0.112	-0.008	-0.030	-0.023	-0.087	-0.076	-0.078
Running	0.024	-0.016	-0.036	-0.065	-0.066	0.026	1	0.260	-0.033	0.016	-0.002	0.011	0.010	-0.079	0.016	0.027	0.016	-0.020	-0.098	0.039	0.519	-0.042	-0.038
Pedestrian Speed Change	0.632	0.029	-0.026	0.016	0.041	-0.027	0.260	1	-0.540	-0.000	-0.044	0.134	0.037	-0.021	-0.012	-0.125	0.013	0.127	0.141	0.114	0.135	-0.030	-0.011
Vehicle Yield	-0.607	0.002	-0.069	0.041	-0.092	-0.014	-0.033	-0.540	1	0.042	0.068	-0.148	0.049	0.093	0.061	0.170	-0.045	0.002	-0.154	-0.093	-0.079	0.129	0.070
Vehicle Type	0.074	0.017	-0.083	0.067	-0.027	-0.018	0.016	-0.000	0.042	1	0.041	0.043	0.097	-0.097	-0.021	-0.046	0.027	0.025	0.030	-0.091	0.011	-0.094	-0.075
Presence of Illegal Obstructions	-0.124	0.008	-0.084	0.051	-0.091	0.056	-0.002	-0.044	0.068	0.041	1	0.018	0.098	0.024	0.001	-0.089	0.092	0.024	0.037	-0.088	0.007	0.045	0.129
Road Surface Conditions	0.134	-0.012	-0.043	-0.004	0.020	-0.039	0.011	0.134	-0.148	0.043	0.018	1	0.015	-0.057	0.007	-0.073	-0.029	-0.011	0.071	0.048	-0.001	-0.057	-0.048
Presence of Crosswalk nearby	-0.038	-0.063	-0.064	0.056	-0.041	0.068	0.010	0.037	0.049	0.097	0.098	0.015	1	0.029	-0.037	-0.006	-0.016	0.006	0.063	-0.124	-0.018	0.017	0.013
Accepted Gap	-0.107	0.034	0.035	0.088	-0.064	-0.068	-0.079	-0.021	0.093	-0.097	0.024	-0.057	0.029	1	0.155	0.101	0.055	0.095	0.201	-0.059	-0.032	0.745	0.735
Average Rejected Gap	-0.070	-0.021	0.163	0.041	-0.128	0.009	0.016	-0.012	0.061	-0.021	0.001	0.007	-0.037	0.155	1	-0.016	0.045	0.296	0.111	0.103	0.032	0.091	0.128
Speed 1	-0.123	-0.088	0.080	0.064	-0.034	-0.060	0.027	-0.125	0.170	-0.046	-0.089	-0.073	-0.006	0.101	-0.016	1	0.169	0.072	-0.080	0.021	-0.100	0.441	0.086
Speed 2	-0.022	-0.052	0.063	0.014	0.028	-0.112	0.016	0.013	-0.045	0.027	0.092	-0.029	-0.016	0.055	0.045	0.169	1	0.065	-0.005	0.074	-0.091	0.024	0.419
Pedestrian Size	0.035	-0.019	0.128	0.065	-0.055	-0.008	-0.020	0.127	0.002	0.025	0.024	-0.011	0.006	0.095	0.296	0.072	0.065	1	0.137	0.191	0.019	0.096	0.088
Waiting Time	0.168	-0.014	-0.127	0.021	0.020	-0.030	-0.098	0.141	-0.154	0.030	0.037	0.071	0.063	0.201	0.111	-0.080	-0.005	0.137	1	0.035	0.036	0.103	0.121
No. of Crossing Attempts	0.084	-0.033	0.043	0.014	-0.032	-0.023	0.039	0.114	-0.093	-0.091	-0.088	0.048	-0.124	-0.059	0.103	0.021	0.074	0.191	0.035	1	0.074	-0.050	-0.041
Pedestrian Speed	0.000	0.138	-0.021	-0.075	-0.086	-0.087	0.519	0.135	-0.079	0.011	0.007	-0.001	-0.018	-0.032	0.032	-0.100	-0.091	0.019	0.036	0.074	1	-0.030	-0.027
Safety Distance 1	-0.124	-0.018	0.013	0.103	-0.049	-0.076	-0.042	-0.030	0.129	-0.094	0.045	-0.057	0.017	0.745	0.091	0.441	0.024	0.096	0.103	-0.050	-0.030	1	0.791
Safety Distance 2	-0.127	-0.006	-0.009	0.101	-0.039	-0.078	-0.038	-0.011	0.070	-0.075	0.129	-0.046	0.013	0.735	0.128	0.086	0.419	0.088	0.121	-0.041	-0.027	0.791	1

Figure B.3. Correlation Matrix: Mitrapark

APPENDIX C: Detailed Model Outputs

This appendix presents the supplementary outputs removed from Chapter Four for compactness. It includes detailed SPSS outputs for the MLR and MNL models, additional GAM diagnostic plots, and supporting CatBoost classification outputs for both Kamalpokhari and Mitrapark. These materials are provided to preserve transparency and support the interpretation of the main results presented in the thesis.

Appendix C.1: MLR Detailed Outputs

This subsection contains the coefficient tables and diagnostic plots for the full and reduced MLR models at both study sites.

Table C.1. Coefficients of MLR Model: Kamalpokhari (Full Model - Training)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	9.705	1.151		8.435	0.000		
Gender	-0.228	0.260	-0.025	-0.877	0.381	0.914	1.094
Carrying Object	0.355	0.251	0.040	1.418	0.157	0.930	1.075
Average Rejected Gap	0.128	0.133	0.029	0.962	0.337	0.826	1.210
Speed 1	-0.270	0.101	-0.134	-2.683	0.008	0.296	3.376
Speed 2	-0.276	0.100	-0.144	-2.770	0.006	0.272	3.672
Pedestrian Size	-0.075	0.115	-0.019	-0.656	0.512	0.849	1.178
Crossing Path	-0.871	0.228	-0.149	-3.824	0.000	0.484	2.068
Age	0.014	0.264	0.001	0.053	0.958	0.937	1.067
Mobile Phone Use	0.236	0.348	0.019	0.677	0.499	0.950	1.053
Waiting Time	0.005	0.018	0.008	0.290	0.772	0.882	1.133
Flow Against	0.159	0.406	0.011	0.393	0.695	0.890	1.124
No. of Crossing Attempts	0.066	0.149	0.012	0.442	0.659	0.933	1.071
Running	-0.020	0.710	-0.002	-0.029	0.977	0.202	4.959
Pedestrian Speed	-0.328	0.264	-0.075	-1.242	0.215	0.202	4.938
Pedestrian Speed Change	0.866	0.406	0.085	2.131	0.034	0.465	2.150
Vehicle Yield	0.127	0.360	0.014	0.354	0.724	0.457	2.190
Vehicle Type	0.007	0.262	0.001	0.026	0.979	0.915	1.093
Presence of Roadside Obstructions	-0.534	0.354	-0.043	-1.510	0.132	0.895	1.117
Road Surface Conditions	-0.538	0.806	-0.019	-0.667	0.505	0.954	1.048
Presence of Crosswalk nearby	0.093	0.310	0.008	0.300	0.764	0.939	1.065
Safety Distance 1	0.092	0.016	0.475	5.834	0.000	0.111	8.993
Safety Distance 2	0.093	0.016	0.462	5.694	0.000	0.112	8.916

Table C.2. Coefficients of MLR Model: Kamalpokhari (Full Model - Testing)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	12.728	3.251		3.915	0.000		
Gender	-0.460	0.541	-0.049	-0.850	0.397	0.822	1.216
Carrying Object	0.442	0.518	0.048	0.854	0.395	0.858	1.166
Average Rejected Gap	0.012	0.265	0.003	0.044	0.965	0.776	1.289
Speed 1	-0.409	0.224	-0.190	-1.827	0.070	0.249	4.015
Speed 2	-0.256	0.237	-0.123	-1.080	0.282	0.206	4.860
Pedestrian Size	-0.249	0.250	-0.060	-0.997	0.321	0.738	1.355
Crossing Path	-0.558	0.494	-0.094	-1.131	0.260	0.391	2.555
Age	-0.202	0.525	-0.021	-0.384	0.702	0.874	1.144
Mobile Phone Use	-0.645	0.664	-0.054	-0.971	0.334	0.861	1.161
Waiting Time	-0.027	0.039	-0.043	-0.693	0.490	0.697	1.434
Flow Against	-0.293	0.862	-0.020	-0.340	0.735	0.804	1.244
No. of Crossing Attempts	0.236	0.320	0.042	0.736	0.463	0.814	1.229
Running	1.696	1.797	0.138	0.944	0.347	0.126	7.928
Pedestrian Speed	-0.848	0.666	-0.186	-1.273	0.206	0.126	7.925
Pedestrian Speed Change	0.116	0.921	0.011	0.126	0.900	0.348	2.876
Vehicle Yield	0.011	0.805	0.001	0.014	0.989	0.356	2.808
Vehicle Type	0.642	0.551	0.066	1.166	0.246	0.845	1.184
Presence of Roadside Obstructions	-1.034	0.686	-0.090	-1.508	0.134	0.760	1.316
Presence of Crosswalk nearby	-0.814	0.588	-0.078	-1.384	0.169	0.836	1.195
Safety Distance 1	0.106	0.035	0.525	3.009	0.003	0.088	11.357
Safety Distance 2	0.087	0.036	0.447	2.404	0.018	0.077	12.905
rand_split	-1.189	2.776	-0.024	-0.428	0.669	0.876	1.142

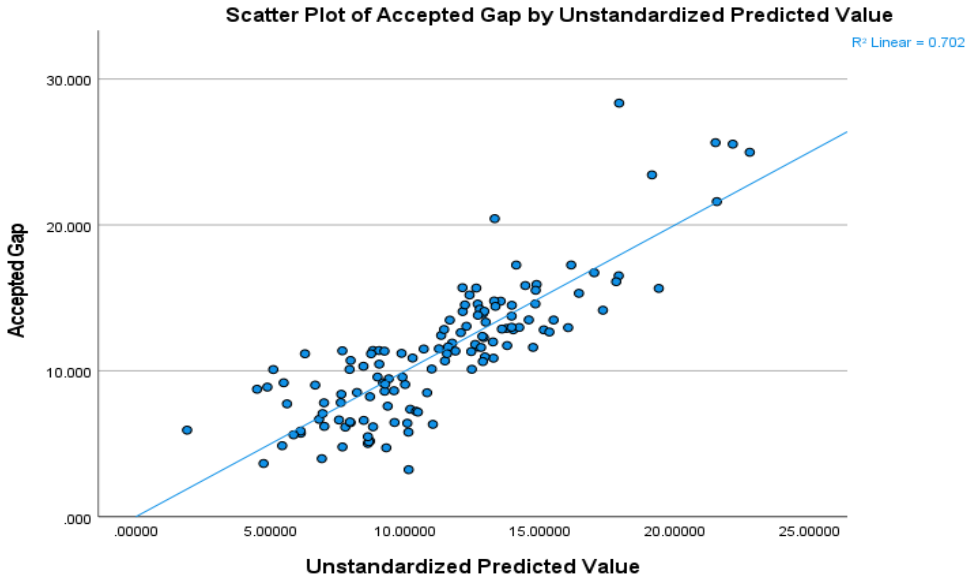


Figure C.1. Scatter Plot: Kamalpokhari (Full Model - Testing)

Table C.3. Coefficients of MLR Model: Kamalpokhari (Reduced Model - Training)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	9.311	0.849		10.967	0.000		
Speed 1	-0.305	0.097	-0.151	-3.152	0.002	0.322	3.110
Speed 2	-0.237	0.095	-0.124	-2.493	0.013	0.299	3.347
Crossing Path	-0.766	0.199	-0.131	-3.850	0.000	0.633	1.581
Pedestrian Speed Change	0.593	0.347	0.058	1.711	0.088	0.637	1.569
Safety Distance 1	0.099	0.015	0.514	6.538	0.000	0.119	8.405
Safety Distance 2	0.086	0.016	0.429	5.507	0.000	0.121	8.255

Table C.4. Coefficients of MLR Model: Kamalpokhari (Reduced Model - Testing)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	9.198	0.642		14.318	0.000		
Speed 1	-0.305	0.097	-0.151	-3.152	0.002	0.322	3.110
Speed 2	-0.237	0.095	-0.124	-2.493	0.013	0.299	3.347
Crossing Path	-0.766	0.199	-0.131	-3.850	0.000	0.633	1.581
Pedestrian Speed Change	0.593	0.347	0.058	1.711	0.088	0.637	1.569
Safety Distance 1	0.099	0.015	0.514	6.538	0.000	0.119	8.405
Safety Distance 2	0.086	0.016	0.429	5.507	0.000	0.121	8.255

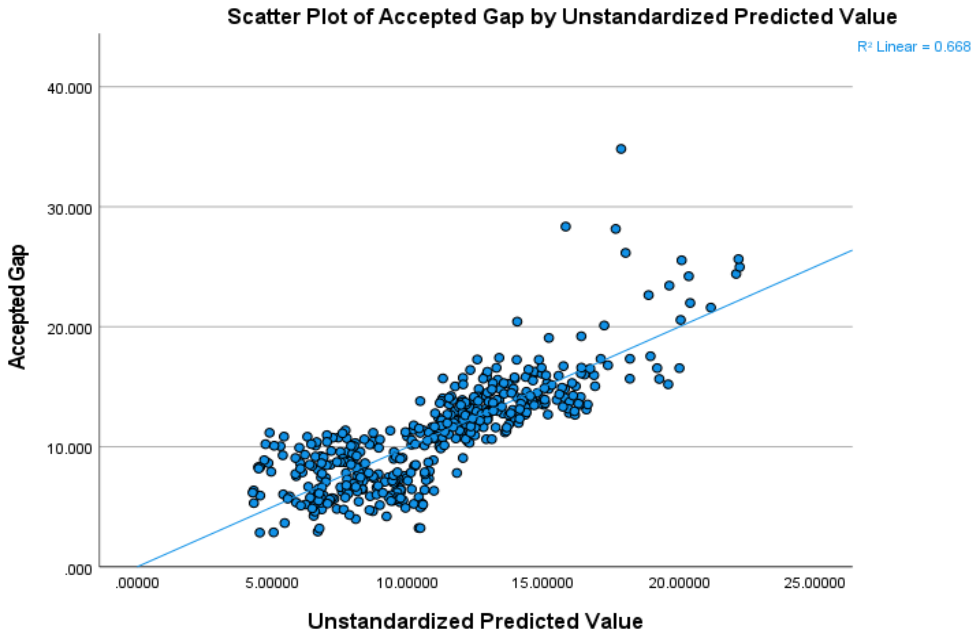


Figure C.2. Scatter Plot: Kamalpokhari (Reduced Model - Testing)

Table C.5. Coefficients of MLR Model: Mitrapark (Full Model - Training)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	8.323	0.916		9.083	0.000		
Gender	-0.008	0.213	-0.001	-0.040	0.968	0.937	1.067
Carrying_Object	0.156	0.216	0.018	0.723	0.470	0.888	1.126
Average_Rejected_Gap	0.035	0.103	0.009	0.338	0.735	0.854	1.171
Speed_1	-0.437	0.081	-0.231	-5.394	0.000	0.302	3.312
Speed_2	-0.197	0.080	-0.111	-2.445	0.015	0.269	3.715
Pedestrian_Size	-0.071	0.097	-0.019	-0.735	0.463	0.829	1.206
Crossing_Path	0.151	0.199	0.027	0.760	0.448	0.444	2.254
Age	-0.020	0.222	-0.002	-0.090	0.928	0.937	1.068
Mobile_Phone_Use	-0.416	0.305	-0.033	-1.364	0.173	0.923	1.083
Waiting_Time	0.028	0.013	0.053	2.055	0.040	0.839	1.192
Flow_Against	0.187	0.362	0.013	0.516	0.606	0.926	1.080
No._of_Crossing_Attempts	-0.080	0.127	-0.016	-0.627	0.531	0.901	1.110
Running	-1.113	0.515	-0.093	-2.160	0.031	0.301	3.325
Pedestrian_Speed	0.311	0.199	0.066	1.566	0.118	0.311	3.217
Pedestrian_Speed_Change	-0.198	0.317	-0.022	-0.626	0.532	0.467	2.142
Vehicle_Yield	0.329	0.274	0.038	1.200	0.231	0.542	1.843
Vehicle_Type	-0.184	0.213	-0.021	-0.866	0.387	0.927	1.078
Presence_of_Roadside_Obstructions	-0.524	0.307	-0.042	-1.705	0.089	0.898	1.114
Road_Surface_Conditions	-0.191	0.478	-0.010	-0.401	0.689	0.954	1.048
Presence_of_Crosswalk_nearby	0.274	0.251	0.027	1.090	0.276	0.929	1.076
Safety_Distance_1	0.120	0.015	0.602	7.990	0.000	0.098	10.232
Safety_Distance_2	0.078	0.015	0.394	5.198	0.000	0.097	10.332

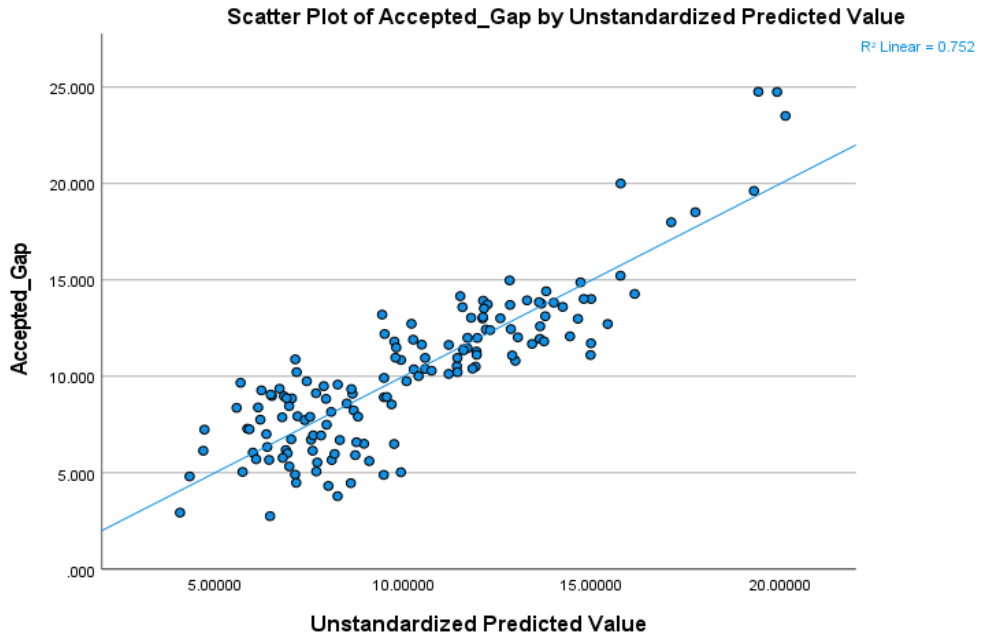


Figure C.3. Scatter Plot: Mitrapark (Full Model - Testing)

Table C.6. Coefficients of MLR Model: Mitrapark (Full Model - Testing)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	6.928	2.113		3.278	0.001		
Gender	-0.201	0.395	-0.025	-0.508	0.612	0.844	1.185
Carrying_Object	0.221	0.408	0.028	0.540	0.590	0.740	1.351
Average_Rejected_Gap	0.214	0.204	0.058	1.050	0.296	0.672	1.488
Speed_1	-0.005	0.136	-0.003	-0.039	0.969	0.303	3.302
Speed_2	-0.408	0.140	-0.241	-2.915	0.004	0.294	3.397
Pedestrian_Size	0.053	0.178	0.016	0.301	0.764	0.754	1.326
Crossing_Path	0.770	0.360	0.148	2.136	0.035	0.419	2.385
Age	0.352	0.397	0.042	0.886	0.378	0.879	1.137
Mobile_Phone_Use	0.528	0.605	0.044	0.873	0.384	0.784	1.275
Waiting_Time	0.021	0.023	0.048	0.884	0.379	0.687	1.455
Flow_Against	-0.736	0.642	-0.055	-1.147	0.253	0.870	1.150
No._of_Crossing_Attempts	-0.325	0.226	-0.071	-1.440	0.152	0.842	1.187
Running	-0.527	0.932	-0.049	-0.565	0.573	0.269	3.717
Pedestrian_Speed	0.299	0.386	0.068	0.774	0.440	0.263	3.805
Pedestrian_Speed_Change	-0.928	0.567	-0.113	-1.637	0.104	0.424	2.358
Vehicle_Yield	0.008	0.517	0.001	0.015	0.988	0.462	2.165
Vehicle_Type	0.059	0.418	0.007	0.141	0.888	0.733	1.365
Presence_of_Roadside_Obstructions	-1.767	0.649	-0.133	-2.723	0.007	0.851	1.176
Road_Surface_Conditions	0.283	1.133	0.012	0.249	0.803	0.908	1.101
Presence_of_Crosswalk_nearby	0.894	0.470	0.091	1.902	0.059	0.883	1.133
Safety_Distance_1	0.043	0.026	0.239	1.670	0.097	0.099	10.148
Safety_Distance_2	0.134	0.027	0.699	4.923	0.000	0.100	9.997

Table C.7. Coefficients of MLR Model: Mitrapark (Reduced Model - Training)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	8.933	0.659		13.564	0.000		
Speed_1	-0.412	0.079	-0.218	-5.220	0.000	0.317	3.154
Speed_2	-0.243	0.077	-0.137	-3.161	0.002	0.292	3.421
Waiting_Time	0.023	0.013	0.045	1.848	0.065	0.942	1.061
Running	-0.531	0.284	-0.044	-1.869	0.062	0.988	1.013
Safety_Distance_1	0.117	0.015	0.591	8.036	0.000	0.102	9.784
Safety_Distance_2	0.081	0.015	0.408	5.580	0.000	0.103	9.668

Table C.8. Coefficients of MLR Model: Mitrapark (Reduced Model - Testing)

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	8.479	0.490		17.296	0.000		
Speed_1	-0.412	0.079	-0.218	-5.220	0.000	0.317	3.154
Speed_2	-0.243	0.077	-0.137	-3.161	0.002	0.292	3.421
Waiting_Time	0.023	0.013	0.045	1.848	0.065	0.942	1.061
Running	-0.531	0.284	-0.044	-1.869	0.062	0.988	1.013
Safety_Distance_1	0.117	0.015	0.591	8.036	0.000	0.102	9.784
Safety_Distance_2	0.081	0.015	0.408	5.580	0.000	0.103	9.668

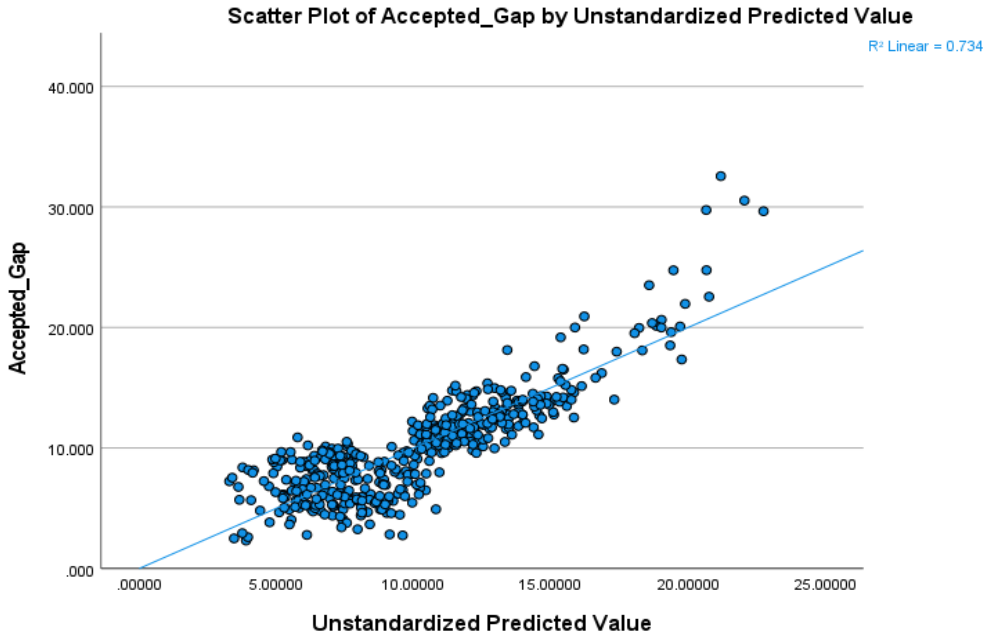


Figure C.4. Scatter Plot: Mitrapark (Reduced Model - Testing)

Appendix C.2: MNL Detailed Outputs

This subsection contains the additional classification reports and confusion matrices for the full and reduced MNL models at both study sites.

Table C.9. Significant Variables: Kamalpokhari (Full Model - Training)

Predictor	Crossing Path Category	Coefficient	Odds Ratio	p-value
const	0	2.857	17.410	0.123
const	1	1.793	6.005	0.429
Gender	0	-1.092	0.335	0.029
Gender	1	-0.854	0.426	0.136
Carrying Object	0	0.773	2.166	0.122
Carrying Object	1	0.469	1.599	0.405
Accepted Gap	0	-0.170	0.844	0.124
Accepted Gap	1	-0.273	0.761	0.024
Average Rejected Gap	0	0.211	1.235	0.396
Average Rejected Gap	1	0.054	1.056	0.847
Speed 1	0	-0.231	0.793	0.176
Speed 1	1	-0.182	0.834	0.372
Speed 2	0	0.443	1.558	0.028
Speed 2	1	0.338	1.402	0.136
Pedestrian Size	0	0.192	1.211	0.446
Pedestrian Size	1	0.288	1.334	0.297
Age	0	0.481	1.618	0.371
Age	1	0.263	1.300	0.662
Mobile Phone Use	0	0.191	1.210	0.799
Mobile Phone Use	1	0.538	1.712	0.510
Waiting Time	0	0.074	1.077	0.075
Waiting Time	1	0.108	1.114	0.017
Flow Against	0	-1.946	0.143	0.022
Flow Against	1	-0.554	0.575	0.539
No. of Crossing Attempts	0	-0.175	0.839	0.543
No. of Crossing Attempts	1	-0.275	0.760	0.396
Running	0	3.995	54.337	0.041
Running	1	5.031	153.112	0.015
Pedestrian Speed	0	-2.394	0.091	0.002
Pedestrian Speed	1	-2.828	0.059	0.000
Pedestrian Speed Change	0	3.278	26.533	0.000
Pedestrian Speed Change	1	4.989	146.724	0.000
Vehicle Yield	0	-2.437	0.087	0.000
Vehicle Yield	1	-4.040	0.018	0.000
Vehicle Type	0	-0.026	0.975	0.959
Vehicle Type	1	0.419	1.521	0.459
Presence of Roadside Obstructions	0	-0.431	0.650	0.481
Presence of Roadside Obstructions	1	-1.313	0.269	0.096
Road Surface Conditions	0	11.947	154326.544	0.982
Road Surface Conditions	1	11.198	72979.301	0.983
Presence of Crosswalk nearby	0	-0.692	0.501	0.314
Presence of Crosswalk nearby	1	-0.551	0.576	0.455
Safety Distance 1	0	0.062	1.064	0.033
Safety Distance 1	1	0.077	1.080	0.022
Safety Distance 2	0	-0.063	0.939	0.060
Safety Distance 2	1	-0.050	0.951	0.183

Table C.10. Significant Variables: Mitrapark (Full Model - Training)

Predictor	Crossing Path Category	Coefficient	Odds Ratio	p-value
const	0	-0.057	0.944	0.966
const	1	-0.729	0.482	0.695
Gender	0	-0.844	0.430	0.032
Gender	1	-0.912	0.402	0.066
Carrying Object	0	-0.065	0.937	0.867
Carrying Object	1	-0.259	0.772	0.602
Accepted Gap	0	-0.022	0.978	0.819
Accepted Gap	1	0.052	1.054	0.640
Average Rejected Gap	0	-0.125	0.882	0.517
Average Rejected Gap	1	-0.390	0.677	0.110
Speed 1	0	-0.013	0.987	0.934
Speed 1	1	0.014	1.014	0.943
Speed 2	0	0.087	1.091	0.573
Speed 2	1	0.059	1.061	0.758
Pedestrian Size	0	0.097	1.102	0.637
Pedestrian Size	1	-0.044	0.957	0.858
Age	0	0.212	1.236	0.589
Age	1	-0.023	0.977	0.962
Mobile Phone Use	0	0.306	1.358	0.626
Mobile Phone Use	1	0.087	1.091	0.909
Waiting Time	0	0.100	1.105	0.002
Waiting Time	1	0.118	1.126	0.001
Flow Against	0	0.157	1.170	0.841
Flow Against	1	-0.321	0.726	0.744
No. of Crossing Attempts	0	0.040	1.041	0.861
No. of Crossing Attempts	1	0.023	1.023	0.938
Running	0	-0.436	0.646	0.654
Running	1	-0.853	0.426	0.488
Pedestrian Speed	0	-0.303	0.739	0.383
Pedestrian Speed	1	-0.268	0.765	0.558
Pedestrian Speed Change	0	3.520	33.793	0.000
Pedestrian Speed Change	1	4.901	134.485	0.000
Vehicle Yield	0	-1.381	0.251	0.009
Vehicle Yield	1	-3.624	0.027	0.000
Vehicle Type	0	-0.234	0.791	0.539
Vehicle Type	1	0.966	2.626	0.050
Presence of Roadside Obstructions	0	-1.451	0.234	0.016
Presence of Roadside Obstructions	1	-2.741	0.065	0.001
Road Surface Conditions	0	1.551	4.717	0.257
Road Surface Conditions	1	1.321	3.748	0.367
Presence of Crosswalk nearby	0	-0.928	0.395	0.061
Presence of Crosswalk nearby	1	-0.647	0.524	0.277
Safety Distance 1	0	-0.010	0.990	0.739
Safety Distance 1	1	-0.009	0.991	0.812
Safety Distance 2	0	-0.005	0.995	0.857
Safety Distance 2	1	-0.029	0.971	0.426

Table C.11. Significant Variables: Kamalpokhari (Reduced Model - Training)

Predictor	Crossing Path Category	Coefficient	Odds Ratio	p-value
const	0	3.561	35.196	0.010
const	1	2.262	9.603	0.208
Vehicle Yield	0	-2.198	0.111	0.000
Vehicle Yield	1	-3.877	0.021	0.000
Pedestrian Speed Change	0	3.276	26.471	0.000
Pedestrian Speed Change	1	5.094	163.097	0.000
Pedestrian Speed	0	-1.912	0.148	0.002
Pedestrian Speed	1	-2.372	0.093	0.000
Running	0	2.698	14.855	0.102
Running	1	3.951	52.001	0.025
Flow Against	0	-1.838	0.159	0.014
Flow Against	1	-0.456	0.634	0.568
Speed 2	0	0.154	1.167	0.086
Speed 2	1	0.068	1.070	0.522
Accepted Gap	0	-0.229	0.795	0.003
Accepted Gap	1	-0.319	0.727	0.000
Waiting Time	0	0.047	1.048	0.188
Waiting Time	1	0.081	1.085	0.040
Safety Distance 1	0	0.020	1.020	0.144
Safety Distance 1	1	0.044	1.045	0.004
Gender	0	-0.997	0.369	0.031
Gender	1	-0.824	0.439	0.120

Table C.12. Significant Variables: Mitrapark (Reduced Model - Training)

Predictor	Crossing Path Category	Coefficient	Odds Ratio	p-value
const	0	-0.163	0.850	0.797
const	1	-1.738	0.176	0.055
Pedestrian Speed Change	0	2.681	14.600	0.000
Pedestrian Speed Change	1	4.152	63.535	0.000
Vehicle Yield	0	-1.637	0.195	0.001
Vehicle Yield	1	-3.831	0.022	0.000
Waiting Time	0	0.085	1.089	0.001
Waiting Time	1	0.099	1.104	0.001
Presence of Roadside Obstructions	0	-1.272	0.280	0.012
Presence of Roadside Obstructions	1	-2.823	0.059	0.000
Gender	0	-0.602	0.548	0.094
Gender	1	-0.756	0.469	0.094
Vehicle Type	0	-0.123	0.884	0.725
Vehicle Type	1	1.066	2.903	0.019

Appendix C.3: GAM Detailed Outputs

This subsection contains the additional predicted-versus-observed plots, QQ plots, and residual plots for the full and reduced GAM models at both study sites.

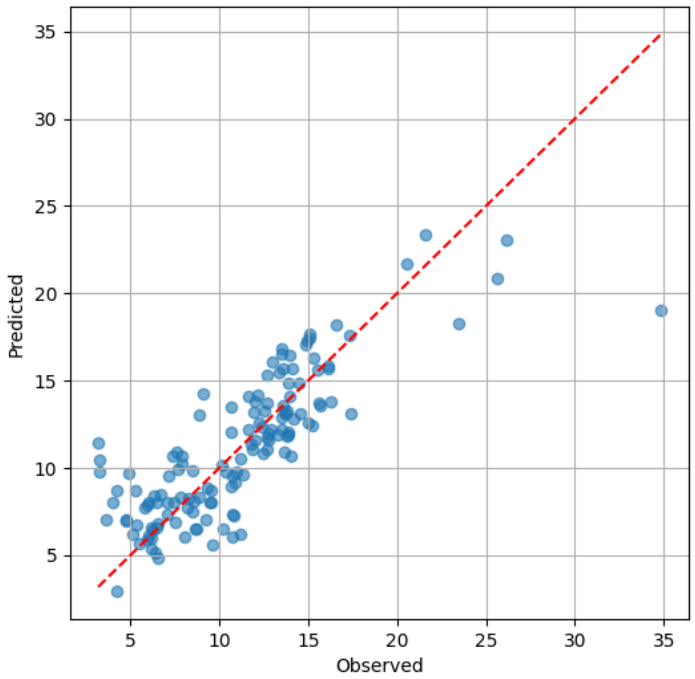


Figure C.5. Predicted vs. Observed: Kamalpokhari (Full Model)

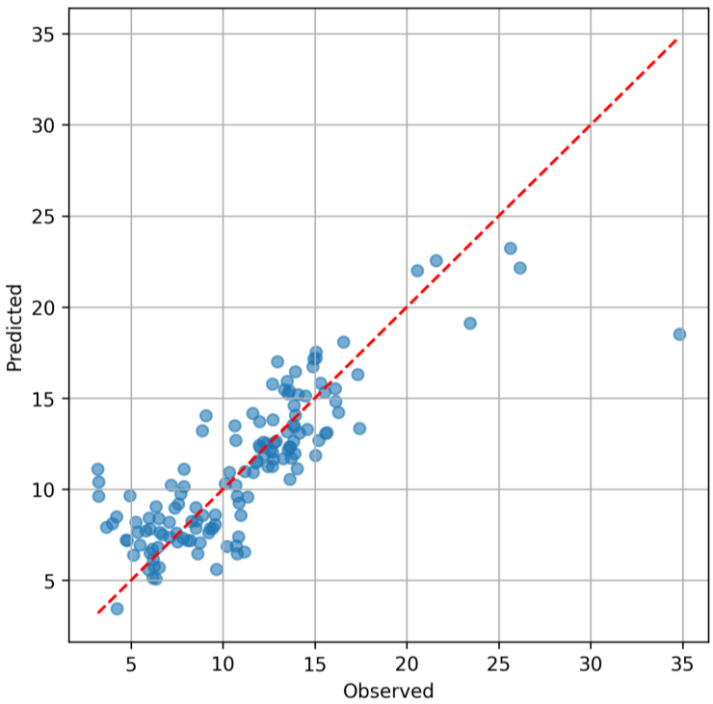


Figure C.6. Predicted vs. Observed: Kamalpokhari (Reduced Model)

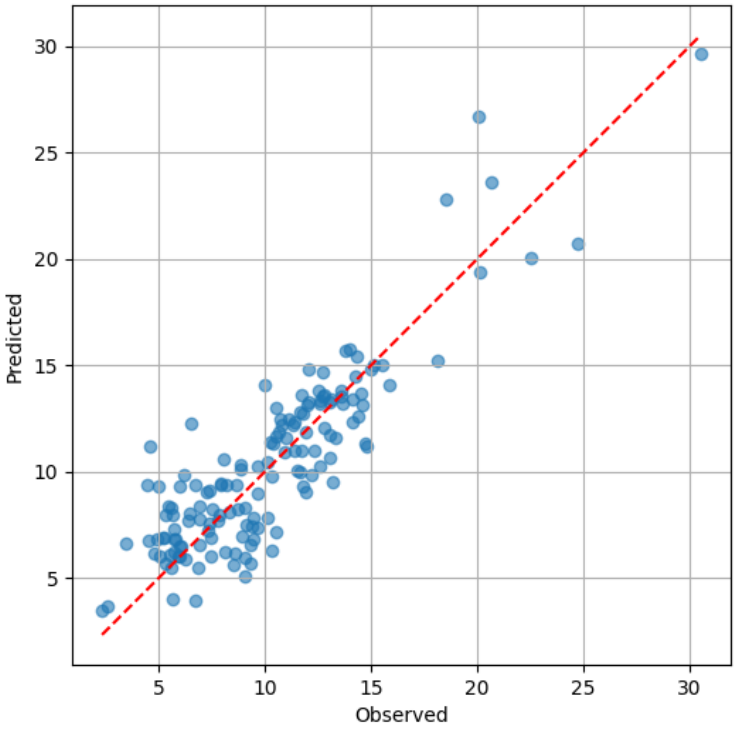


Figure C.7. Predicted vs. Observed: Mitrapark (Full Model)

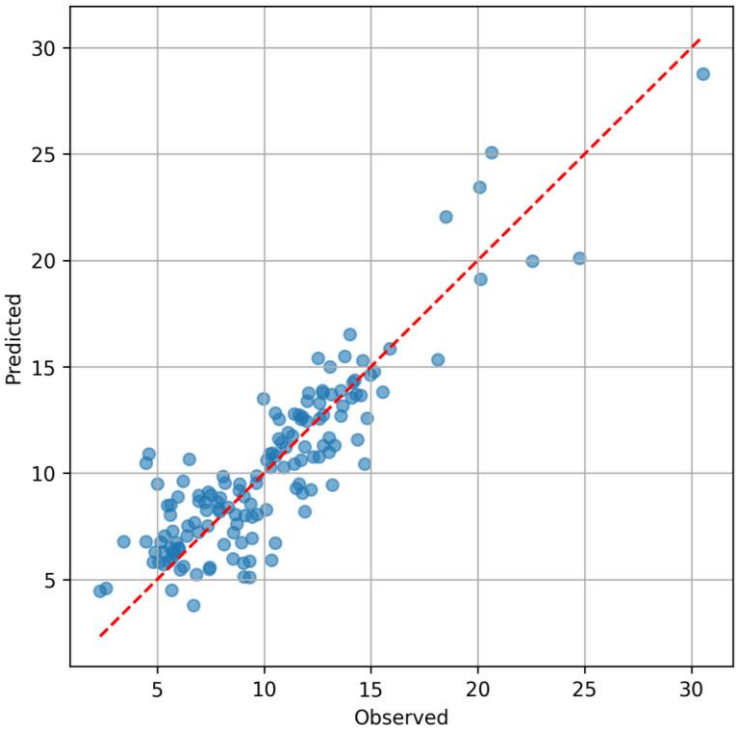


Figure C.8. Predicted vs. Observed: Mitrapark (Reduced Model)

APPENDIX D: Python Scripts

Appendix D.1: GAM Python Scripts

```
from __future__ import annotations


from pathlib import Path


import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from pygam import LinearGAM, s
from scipy import stats
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import KFold, train_test_split


SITE = "Kamalpokhari"   # "Kamalpokhari" or "Mitrapark"
DATA_PATH = "your_dataset.xlsx"
TARGET = "Accepted Gap"
TEST_SIZE = 0.30
RANDOM_STATE = 42
N_SPLITS_CV = 5


SITE_FEATURES = {
    "Kamalpokhari": [
        "Safety Distance 1", "Safety Distance 2", "Speed 1", "Speed 2",
        "Pedestrian Speed", "Waiting Time", "Average Rejected Gap",
        "No. of Crossing Attempts",
    ],
```

```
    "Mitrapark": [
        "Average Rejected Gap", "Speed 1", "Speed 2", "Pedestrian Speed",
        "Waiting Time", "No. of Crossing Attempts", "Safety Distance 1",
        "Safety Distance 2",
    ],
}



def build_terms(n: int):
    terms = s(0)
    for i in range(1, n):
        terms += s(i)
    return terms



def make_plot(path: Path, size=(8, 6)):
    fig, ax = plt.subplots(figsize=size)
    return fig, ax, lambda: (fig.savefig(path, dpi=300, bbox_inches="tight"),
plt.close(fig))



if SITE not in SITE_FEATURES:
    raise ValueError(f"Invalid SITE value: {SITE}")


features = SITE_FEATURES[SITE]
required = features + [TARGET]


df = pd.read_excel(DATA_PATH)
missing = [c for c in required if c not in df.columns]
if missing:
```

```
    raise ValueError("Missing required columns: " + ", ".join(missing))


df = df.dropna(subset=required).copy()
X, y = df[features], df[TARGET]


out = Path(f"GAM_{SITE}_Final_Reduced_Model")
out.mkdir(parents=True, exist_ok=True)


X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=TEST_SIZE, random_state=RANDOM_STATE
)


gam = LinearGAM(build_terms(X.shape[1])).fit(X_train.values, y_train.values)
y_pred = gam.predict(X_test.values)
residuals = y_test.values - y_pred


mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)


cv = KFold(n_splits=N_SPLITS_CV, shuffle=True,
random_state=RANDOM_STATE)
cv_scores = [
    LinearGAM(build_terms(X.shape[1])).fit(X.values[tr],
y.values[tr]).score(X.values[val], y.values[val])
    for tr, val in cv.split(X)
]
cv_mean = float(np.mean(cv_scores))


tables = {
```

```
    "model_summary.csv": pd.DataFrame({
        "Term": ["Intercept"] + features,
        "p-value": gam.statistics_["p_values"],
    }),
    "cross_validation_scores.csv": pd.DataFrame({
        "Fold": range(1, N_SPLITS_CV + 1),
        "R2": cv_scores,
    }),
    "model_evaluation.csv": pd.DataFrame({
        "Metric": [
            "Mean Squared Error",
            "Root Mean Squared Error",
            "R-squared",
            "CV R-squared (Mean)",
        ],
        "Value": [mse, rmse, r2, cv_mean],
    }),
}

for name, table in tables.items():
    table.to_csv(out / name, index=False)

fig, axes = plt.subplots(2, 4, figsize=(20, 10))
for i, feature in enumerate(features):
    ax = axes.flat[i]
    XX = gam.generate_X_grid(term=i)
    pdp = gam.partial_dependence(term=i, X=XX)
    ci = gam.partial_dependence(term=i, X=XX, width=0.95)
    ax.plot(XX[:, i], pdp)
```

```
    ax.plot(XX[:, i], ci[1], linestyle="--")
    ax.set_title(feature)
    ax.set_xlabel(feature)
    ax.set_ylabel("Partial effect")
    ax.grid(True)
fig.tight_layout()
fig.savefig(out / "partial_dependence_plots.png", dpi=300, bbox_inches="tight")
plt.close(fig)

fig, ax, save = make_plot(out / "observed_vs_predicted.png", (6, 6))
ax.scatter(y_test, y_pred, alpha=0.6)
ax.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], linestyle="--")
ax.set(xlabel="Observed", ylabel="Predicted", title=f"Observed vs Predicted –
{SITE}")
ax.grid(True)
save()

fig, ax, save = make_plot(out / "residual_histogram.png")
ax.hist(residuals, bins=30, edgecolor="black")
ax.set(title=f"Residual Histogram – {SITE}", xlabel="Residual", ylabel="Frequency")
ax.grid(True)
save()

for filename, x, xlabel, title in [
    ("residuals_vs_predicted.png", y_pred, "Predicted", f"Residuals vs Predicted –
{SITE}"),
    ("residuals_vs_actual.png", y_test, "Actual", f"Residuals vs Actual – {SITE}"),
]:
    fig, ax, save = make_plot(out / filename)
    ax.scatter(x, residuals, alpha=0.6)
```

```
    ax.axhline(0, linestyle="--")
    ax.set(title=title, xlabel=xlabel, ylabel="Residual")
    ax.grid(True)
    save()

fig, ax, save = make_plot(out / "qq_plot.png")
stats.probplot(residuals, dist="norm", plot=ax)
ax.set_title(f"QQ Plot of Residuals – {SITE}")
ax.grid(True)
save()

print(f"Site: {SITE}")
print(f"Dataset used: {DATA_PATH}")
print(f"Selected features: {features}")
print(f"Output folder: {out}")
print(f"Mean Squared Error: {mse:.6f}")
print(f"Root Mean Squared Error: {rmse:.6f}")
print(f"R-squared: {r2:.6f}")
print(f"CV R-squared (Mean): {cv_mean:.6f}")
```

Appendix D.2: CatBoost Python Scripts

```
from __future__ import annotations

from pathlib import Path

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from catboost import CatBoostClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, f1_score, log_loss
from sklearn.model_selection import train_test_split

SITE = "Kamalpokhari"   # "Kamalpokhari" or "Mitrapark"
DATA_PATH = "your_dataset.xlsx"
TARGET = "Crossing Path"
TEST_SIZE = 0.30
RANDOM_STATE = 42

SITE_CONFIG = {
    "Kamalpokhari": {
        "selected_vars": [
            "Vehicle Yield", "Pedestrian Speed Change", "Pedestrian Speed",
            "Waiting Time", "Average Rejected Gap", "Safety Distance 1",
            "Safety Distance 2", "Speed 1",
        ],
        "categorical_vars": ["Vehicle Yield", "Pedestrian Speed Change"],
    },
    "Mitrapark": {
        "selected_vars": [
            "Pedestrian Speed Change", "Vehicle Yield", "Average Rejected Gap",
            "Safety Distance 2", "Accepted Gap", "Pedestrian Size",
            "Speed 2", "Waiting Time",
        ],
```

```
        "categorical_vars": ["Pedestrian Speed Change", "Vehicle Yield"],
    },
}



def make_plot(path: Path, size=(8, 6)):
    fig, ax = plt.subplots(figsize=size)
    return fig, ax, lambda: (fig.savefig(path, dpi=300, bbox_inches="tight"), plt.close(fig))



if SITE not in SITE_CONFIG:
    raise ValueError(f"Invalid SITE value: {SITE}")

selected_vars = SITE_CONFIG[SITE]["selected_vars"]
categorical_vars = SITE_CONFIG[SITE]["categorical_vars"]
required_cols = [TARGET] + selected_vars

df = pd.read_excel(DATA_PATH)
missing = [c for c in required_cols if c not in df.columns]
if missing:
    raise ValueError("Missing required columns: " + ", ".join(missing))

df = df[required_cols].dropna().copy()
X, y = df.drop(columns=[TARGET]), df[TARGET]
cat_features = [X.columns.get_loc(c) for c in categorical_vars]

out = Path(f"CatBoost_{SITE}_Final_Reduced_Model")
out.mkdir(parents=True, exist_ok=True)

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=TEST_SIZE, random_state=RANDOM_STATE, stratify=y
)

model = CatBoostClassifier(
    iterations=1000,
```

```
    learning_rate=0.05,
    depth=6,
    eval_metric="Accuracy",
    random_seed=RANDOM_STATE,
    verbose=False,
)

model.fit(
    X_train, y_train,
    cat_features=cat_features,
    eval_set=(X_test, y_test),
    use_best_model=True,
)

y_pred = np.array(model.predict(X_test)).ravel()
try:
    y_pred = y_pred.astype(y_test.dtype)
except Exception:
    pass

classes = model.classes_
probs = model.predict_proba(X_test)

base_dist = y_train.value_counts(normalize=True).reindex(classes, fill_value=0).values
null_probs = np.tile(base_dist, (len(y_test), 1))

pseudo_r2 = 1 - (
    log_loss(y_test, probs, labels=classes) /
    log_loss(y_test, null_probs, labels=classes)
)

acc = accuracy_score(y_test, y_pred)
macro_f1 = f1_score(y_test, y_pred, average="macro")
weighted_f1 = f1_score(y_test, y_pred, average="weighted")
```

```
summary_df = pd.DataFrame({
    "Metric": [
        "Total Samples", "Training Samples", "Testing Samples", "Number of Classes",
        "Pseudo R² (McFadden-like)", "Accuracy", "Macro F1-score", "Weighted F1-score",
    ],
    "Value": [
        len(df), len(X_train), len(X_test), y.nunique(),
        round(float(pseudo_r2), 5), round(float(acc), 5),
        round(float(macro_f1), 5), round(float(weighted_f1), 5),
    ],
})

report_df = pd.DataFrame(
    classification_report(y_test, y_pred, output_dict=True, zero_division=0)
).transpose()

feature_importance = model.get_feature_importance(prettified=True)
if "Feature Id" in feature_importance.columns:
    feature_importance = feature_importance.rename(columns={"Feature Id": "Feature"})

params_df = pd.DataFrame(model.get_params().items(), columns=["Parameter", "Value"])

mis_mask = y_test.to_numpy() != y_pred
misclassified = X_test.loc[mis_mask].copy()
misclassified["Actual"] = y_test.loc[mis_mask].values
misclassified["Predicted"] = y_pred[mis_mask]
misclassified = misclassified[["Actual", "Predicted"] + selected_vars].head(10)

tables = {
    "model_summary.csv": summary_df,
    "classification_report.csv": report_df,
    "feature_importance.csv": feature_importance,
    "model_parameters.csv": params_df,
```

```
    "misclassifications_sample.csv": misclassified,
}
for name, table in tables.items():
    table.to_csv(out / name, index=False)

cm = confusion_matrix(y_test, y_pred, labels=classes)

fig, ax, save = make_plot(out / "confusion_matrix.png")
im = ax.imshow(cm, aspect="auto")
ax.set_xticks(np.arange(len(classes)))
ax.set_yticks(np.arange(len(classes)))
ax.set_xticklabels(classes)
ax.set_yticklabels(classes)
ax.set_xlabel("Predicted")
ax.set_ylabel("Actual")
for i in range(cm.shape[0]):
    for j in range(cm.shape[1]):
        ax.text(j, i, str(cm[i, j]), ha="center", va="center")
fig.colorbar(im, ax=ax)
save()

fig, ax, save = make_plot(out / "feature_importance.png", (10, 6))
top = feature_importance.head(len(selected_vars))
ax.barh(top["Feature"], top["Importances"])
ax.set_xlabel("Importance")
ax.set_ylabel("Feature")
save()

actual_dist = y_test.value_counts().reindex(classes, fill_value=0)
pred_dist = pd.Series(y_pred).value_counts().reindex(classes, fill_value=0)
x = np.arange(len(classes))

fig, ax, save = make_plot(out / "actual_vs_predicted_distribution.png", (10, 6))
ax.bar(x - 0.2, actual_dist.values, width=0.4, label="Actual")
ax.bar(x + 0.2, pred_dist.values, width=0.4, label="Predicted")
```

```
ax.set_xticks(x)
ax.set_xticklabels(classes)
ax.set_ylabel("Frequency")
ax.legend()
save()

y_pred_s = pd.Series(y_pred, index=y_test.index)
class_accuracy = {
    label: (((y_test == label) & (y_pred_s == label)).sum() / (y_test == label).sum())
    if (y_test == label).sum() > 0 else 0.0
    for label in classes
}

fig, ax, save = make_plot(out / "class_wise_accuracy.png", (10, 6))
ax.bar(list(class_accuracy.keys()), list(class_accuracy.values()))
ax.set_ylabel("Accuracy")
ax.set_ylim(0, 1)
save()

print(f"Site: {SITE}")
print(f"Dataset used: {DATA_PATH}")
print(f"Selected features: {selected_vars}")
print(f"Categorical features: {categorical_vars}")
print(f"Output folder: {out}")
print(f"Pseudo R² (McFadden-like): {pseudo_r2:.6f}")
print(f"Accuracy: {acc:.6f}")
print(f"Macro F1-score: {macro_f1:.6f}")
print(f"Weighted F1-score: {weighted_f1:.6f}")
```

त्रिभुवन विश्वविद्यालय
TRIBHUVAN UNIVERSITY
इन्जिनियरिङ्ग अध्ययन संस्थान
INSTITUTE OF ENGINEERING
पुल्चोक क्याम्पस
PULCHOWK CAMPUS

Accredited by University Grants Commission (UGC) Nepal 2020

5-521260
5-521611
5-522104
5-522809

पुल्चोक, ललितपुर।
Pulchowk, Lalitpur

Date: May 7, 2026

To Whom It May Concern:

This is to certify that the paper titled "*A Statistical Approach to Modeling Pedestrian Crossing Paths Using Multinomial Logit in Kathmandu's Uncontrolled Midblock Crossings*" (Submission ID #1002), with **Sudarshan Tamang** as the first author, was accepted through the peer-review process and has been presented at the 18th IOE Graduate Conference, organized at Pulchowk Campus, Lalitpur, Nepal, from May 7 to 9, 2026.

Please note that inclusion of the accepted manuscript in the conference proceedings is contingent upon timely compliance with any further editorial requirements during the publication process.

Prof. Sangeeta Singh
Convener
18th IOE Graduate Conference

A Statistical Approach to Modeling Pedestrian Crossing Paths Using Multinomial Logit in Kathmandu's Uncontrolled Midblock Crossings

Sudarshan Tamang [a], Pradeep Kumar Shrestha [b]

[a] *Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Trbhuwan University, Lalitpur, Nepal*
[b] *Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Trbhuwan University, Lalitpur, Nepal*
[a] 079mstre023.sudarshan@pcampus.edu.np, [b] pradeep.shrestha@pcampus.edu.np

Abstract
Pedestrian safety remains a critical concern in urban areas, particularly at uncontrolled midblock crossings where pedestrian and vehicle interactions occur without signal regulation. This study aims to develop a statistical model to predict pedestrian crossing path choices; *Perpendicular*, *Oblique* or *Irregular* using the Multinomial Logit (MNL) model in SPSS. Field data were collected from two major midblock locations in Kathmandu: Kamalpokhari and Mitrapark, and analyzed to extract behavioral, traffic and environmental variables such as pedestrian speed change, running behavior, safety distance, vehicle yielding, pedestrian size, roadside obstructions and other contextual factors. The results show that factors such as pedestrian speed change, vehicle yield and safety distance were statistically significant determinants for the choice of crossing path. Model fit statistics including -2 Log Likelihood and McFadden's Pseudo R^2 (**0.36-0.39**) indicate moderate explanatory power for behavioral data at both sites. The model accuracies were between **15%-20%** indicating limited predictive performance in capturing the variability of pedestrian behavior under simplifying model assumptions. Nonetheless, the results underscore the value of behavioral and contextual variables in explaining pedestrian path selection. The study demonstrates that despite these limitations, statistical models such as MNL remain useful tools for interpreting pedestrian decision making and offer actionable insights for safer midblock design in mixed traffic urban contexts.

Keywords
pedestrian behavior, multinomial logit (MNL) model, midblock crossing, Kathmandu, statistical modeling, pedestrian safety

1. Introduction

Pedestrian safety has become one of the major concerns in today's rapidly urbanizing cities, particularly in developing countries where pedestrian related crashes account for a significant portion of total road traffic accidents. Midblock crossings which occur away from intersections are especially vulnerable due to the lack of traffic control and the unpredictable interactions between pedestrians and vehicles. In Nepal, particularly in dense urban areas such as Kathmandu, the situation is more critical due to mixed traffic conditions, high pedestrian volumes and inadequate crossing facilities. Therefore, understanding pedestrian crossing behavior at uncontrolled midblock locations is essential for developing appropriate safety measures and improving overall traffic management. While previous studies have addressed crossing frequency and gap acceptance, limited attention has been given to crossing path choice, which this study directly investigates.

Several studies have analyzed pedestrian behavior under varying traffic and environmental conditions. Previous research has employed a range of modeling approaches, from conventional statistical techniques to advanced machine learning algorithms to understand factors influencing pedestrian behavior including accepted gap, crossing speed and decision-making characteristics. However, most studies have focused on signalized intersections or controlled environments while fewer have examined uncontrolled midblock crossings. This creates a significant research gap, particularly in developing countries where pedestrian behavior is strongly influenced by mixed traffic conditions and inadequate or inoperable pedestrian infrastructure. Similar efforts have examined pedestrian path choice behavior at uncontrolled midblock locations. [1] classified pedestrian crossings as perpendicular, oblique or mixed and related path deviation to traffic conditions. Similarly, [2] used discrete choice modeling to examine pedestrian route selection at midblock locations without crossing facilities, confirming that path choice is influenced by perceived safety and vehicle approach speed.

In this context, this study evaluates pedestrian crossing path behavior at uncontrolled midblock locations using the Multinomial Logit (MNL) model. Two study sites, Kamalpokhari and Mitrapark were selected as case study locations in Kathmandu to compare behavioral differences under mixed traffic conditions. The findings aim to identify key behavioral and contextual factors influencing crossing path choice, providing insights for safer urban design and data driven policy formulation. Accordingly, the MNL model is applied to quantify pedestrian crossing path choices and identify the behavioral and contextual factors influencing these decisions.

2. Literature Review

Pedestrian behavior has been a subject of increasing research interest due to its critical role in ensuring road safety and improving traffic efficiency. Several studies have focused on

understanding pedestrian behavior at different types of crossings under varying traffic and environmental conditions. These studies generally examine factors such as pedestrian age, gender, walking speed, group size, waiting time and vehicle speed which significantly influence a pedestrian's crossing decision.

Many researchers have employed statistical and simulation based modeling techniques to examine pedestrian crossing behavior. Among them, discrete choice models such as the Binary Logit (BL) and Multinomial Logit (MNL) models have been widely used to understand pedestrian decision-making processes. These models estimate the probability of pedestrians choosing specific crossing behaviors based on influencing variables, making them suitable for analyzing behavioral tendencies under different conditions. For instance, [3] analyzed pedestrian road crossing behavior under mixed traffic conditions using discrete choice modeling and identified critical variables such as vehicle speed, pedestrian speed and gap size. Similarly, [4] examined pedestrian behavior at midblock crossings and observed that waiting time and accepted gap are strongly associated with crossing decisions.

Other studies have highlighted the importance of environmental and situational factors in determining pedestrian behavior. [5] studied pedestrian characteristics at uncontrolled midblock crossings in India and reported that crossing speeds vary with age and gender. [6] investigated illegal midblock crossings in China and found that higher traffic flow and shorter gaps increase the probability of pedestrians crossing irregularly. [7] found that unsignalized midblock crossings substantially affect vehicular speeds and driver yielding behavior emphasizing the interaction between pedestrian and vehicular movements in mixed traffic. Similarly, [8] identified pedestrian speed change and rolling behavior as significant indicators of midblock safety supporting the inclusion of these behavioral predictors in the present study. [9] critically assessed existing pedestrian behavior models and emphasized the importance of incorporating psychological and contextual variables to improve behavioral predictions.

Recent advancements have also seen the integration of machine learning and hybrid modeling techniques to capture non-linear relationships and unobserved diversity in pedestrian decision making. However, despite these developments, the Multinomial Logit (MNL) model remains one of the most widely used and interpretable methods for analyzing pedestrian crossing behavior, particularly in data limited environments such as developing cities. The MNL model's ability to compare multiple behavioral alternatives while maintaining interpretability makes it well suited for this research.

Although several studies have contributed valuable insights into pedestrian behavior, most have focused on signalized intersections or controlled crossings. Limited research has been conducted on uncontrolled midblock crossings, particularly in the context of developing countries where traffic is diverse and pedestrian facilities are inadequate. This study addresses this research gap by applying the MNL model to analyze pedestrian crossing path behavior at uncontrolled midblock crossings in Kathmandu.

3. Methodology

This study aims to analyze pedestrian crossing behavior at uncontrolled midblock crossings in Kathmandu by developing statistical models that explain pedestrians' path choices based on observed behavioral and traffic characteristics. The methodological framework adopted in this research involves several key steps including site selection, data collection, variable identification and model development using the Multinomial Logit (MNL) model.

3.1 Study Area Selection

The study was conducted at two uncontrolled midblock crossings in Kathmandu: Kamalpokhari and Mitrapark. These locations were selected because they represent typical midblock conditions in the city characterized by high pedestrian activity, mixed traffic flow and an absence of traffic control devices. Both sites have commercial surroundings that attract large numbers of pedestrians throughout the day. Observations were carried out during peak traffic periods to capture realistic pedestrian-vehicle interactions.

A preliminary pilot study was conducted to assess the suitability of potential sites and to determine the feasibility of data collection through video recording. The pilot helped identify camera placement points, field-of-view limitations and the optimal observation duration required to record sufficient pedestrian samples. Based on the pilot findings, Kamalpokhari and Mitrapark were finalized as the study sites due to their high pedestrian volumes and suitability for observing uncontrolled midblock crossing behavior.

3.2 Data Collection

Data was collected through video recording using cameras positioned at elevated vantage points to ensure complete visibility of the crossing areas. Recordings were conducted during morning peak hours between 9 and 12 to capture variations in traffic and pedestrian flow. Each video was later analyzed manually using Kinovea software which allows for precise frame-by-frame observation of pedestrian and vehicular movement. The software has been widely validated in previous studies for accuracy in measuring speed, distance and time-based variables from video footage [10].

From the recorded videos, pedestrian, traffic and environmental characteristics were extracted including waiting time, pedestrian speed, pedestrian size, running behavior, group size, gap acceptance, vehicle speed, yielding behavior, roadside obstructions, etc. Each pedestrian crossing instance was treated as an individual observation resulting in a dataset suitable for statistical Multinomial Logit (MNL) modeling, ensuring consistency across both locations.

The dataset consisted of 460 observations at Kamalpokhari and 490 at Mitrapark. As the total pedestrian population is large and not precisely defined, an infinite population assumption was adopted. Using a standard sample size formula with a 95% confidence level and 5% margin of error,

the minimum required sample size was estimated at approximately 384 observations. The number of observations collected at both sites exceeds this requirement, providing sufficient data for reliable analysis. The distribution of crossing path categories was also examined to assess class balance. At Kamalpokhari, perpendicular, oblique and irregular crossings accounted for 27.2%, 42.2% and 30.7% of observations respectively, while at Mitrapark the corresponding proportions were 30.0%, 42.4% and 27.6% respectively. The observed proportions indicate that irregular crossings constitute a substantial share of the dataset at both sites and do not represent a severely underrepresented class. Therefore, the inability of the MNL model to predict irregular crossings in certain cases cannot be attributed solely to class imbalance but also reflects the limitations of the model in capturing complex and less structured pedestrian behaviors.

3.3 Variable Selection

The variables used in this study were selected based on previous research findings, site observations and data availability. The dependent variable was the crossing path choice categorized into perpendicular, oblique and irregular crossings. Perpendicular crossings refer to movements where pedestrians cross the roadway along a path approximately normal to the traffic flow. Oblique crossings involve angled movements across the roadway while irregular crossings include non-linear or multi-stage movements in which pedestrians deviate from a consistent path, including stopping or changes in direction during crossing. The classification of crossing paths was based on visual interpretation of video recordings and was carried out by a single observer following consistent criteria throughout the dataset and therefore, inter-rater reliability was not assessed and is acknowledged as a limitation of this study. Independent variables included pedestrian, traffic and environmental characteristics such as pedestrian speed, speed change, running behavior, safety distance, vehicle speed, vehicle yield, presence of a crosswalk nearby and roadside obstructions. Table 1 shows the different variables used in this study.

Table 1: Classification of variables

Category	Variables
Pedestrian	Pedestrian Speed, Pedestrian Speed Change, Running, Pedestrian Size, Gender, No. of Crossing Attempts, Mobile Phone Use, Age, Carrying Object
Traffic	Speed 1, Speed 2, Vehicle Yield, Accepted Gap, Average Rejected Gap, Flow Against, Vehicle Type
Environmental	Waiting Time, Safety Distance 1, Safety Distance 2, Presence of Roadside Obstructions, Presence of Crosswalk Nearby, Road Surface Conditions

To maintain consistency during data coding, variables with measurable thresholds were explicitly defined based on prior studies and on-site observations:

3.3.1 Pedestrian Speed

Measured as the average speed (m/s) during the crossing phase, calculated from the time taken to traverse the road width. Speeds between 1.2m/s and 1.5m/s were considered typical walking speeds while values exceeding 2.2m/s were identified as running behavior. These reference ranges align with prior studies of pedestrian movement in urban mixed traffic conditions [5], [4].

3.3.2 Running

Classified when pedestrian speed exceeded 2.2m/s for at least two seconds during crossing. Speeds between 0.9m/s and 2.2m/s were coded as walking, following benchmarks from mixed traffic environments [5].

3.3.3 Safety Distance

Defined as the gap between the pedestrian and the nearest approaching vehicle at the moment the pedestrian entered the first lane.

3.3.4 Pedestrian Speed Change

This variable was coded as binary where 0 indicates no meaningful speed change and 1 indicates a noticeable change during crossing. A threshold of ±25% from the average crossing speed was used to identify meaningful behavioral adjustments such as hesitation or urgency, while excluding minor normal fluctuations.

3.3.5 Waiting Time

Measured as the duration (in seconds) between arrival at the curb and initiation of crossing; values > 5s were considered high waiting, representing cautious behavior.

The remaining variables were coded directly from observed categories or recorded values and therefore did not require separate threshold-based definitions.

The threshold values were determined by combining published parameters and field evidence from the Kathmandu midblock context. The 2.2m/s running threshold distinguishes purposeful acceleration from normal walking speeds (1.2-1.5m/s). Safety distance categories reflect reaction distances at common midblock speeds, while the ±25% criterion for speed change captures meaningful behavioral adaptation without oversensitivity to normal walking variation. The operational definition of safety distance in this study aligns with the findings of [11], who observed that pedestrians judge approaching vehicle speed and distance to maintain a minimal safety margin while initiating crossings. Waiting time and yielding thresholds were grounded in local observations to differentiate assertive from cautious behavior. Establishing these limits ensured uniform data coding and reproducible behavioral interpretation across both study sites.

3.4 Model Development

The Multinomial Logit (MNL) model was used to analyze pedestrian crossing path behavior. The MNL framework has been widely applied in pedestrian route choice modeling to evaluate categorical behavioral decisions [12], supporting its use for analyzing crossing path selection. The model assumes that each pedestrian selects a crossing path that provides the highest utility among available alternatives. The general functional form of the utility equation is given by:

$$U_{ij} = \beta_{0j} + \beta_{1j} X_1 + \beta_{2j} X_2 + \cdots + \varepsilon_{ij} \quad \text{(i)}$$

The probability that a pedestrian i chooses crossing path j is expressed as:

$$P_{ij} = \frac{\exp(U_{ij})}{\sum_k \exp(U_{ik})} \quad \text{(ii)}$$

The MNL model was calibrated using SPSS software through maximum likelihood estimation. The Perpendicular crossing path was set as the reference category for analysis. Model performance was evaluated using indicators such as the -2 Log Likelihood, Likelihood Ratio Chi-Square, McFadden's Pseudo R^2 and classification accuracy for both training and testing datasets to assess explanatory power. Significance levels (p-values) of independent variables were used to identify the most influential factors for each site.

Multicollinearity among the independent variables was assessed using the Variance Inflation Factor (VIF). The VIF values ranged from 1.041 to 4.892 at Kamalpokhari and from 1.043 to 4.548 at Mitrapark, indicating that multicollinearity was within acceptable limits and was not a major concern in the model. Although a few variables showed moderate collinearity, all values remained below the commonly accepted threshold of 5.

3.5 Model Validation

The dataset was divided into 70:30 for training and testing subsets to validate the predictive capability of the model. The model developed using the training data was tested on the remaining data to assess its accuracy and generalizability. Consistency between the predicted and observed outcomes was measured using a confusion matrix and accuracy metrics. The results were then compared across both study locations to identify site-specific behavioral patterns and key influencing variables.

4. Results and Discussion

Table 2 presents the model fit summary for the Kamalpokhari MNL model, including -2 Log Likelihood, Likelihood Ratio Chi-Square, degrees of freedom, and McFadden's Pseudo R^2.

Table 2: Model Summary : Kamalpokhari

Metric	Value
-2 Log Likelihood (Intercept Only)	695.6673
-2 Log Likelihood (Final Model)	421.3397
Likelihood Ratio Chi-Square	274.3276
Degrees of Freedom	20
Sig. (p-value)	0
McFadden's Pseudo R^2	0.3943

Table 3 presents the estimated coefficients, odds ratios and significance levels for the Kamalpokhari MNL model, with perpendicular crossing path as the reference category. The results highlight the importance of pedestrian speed change, vehicle yield, running behavior, pedestrian speed and safety distance in influencing crossing path choice.

For the comparison between oblique and perpendicular crossings, negative coefficients for Vehicle Yield, Pedestrian Speed, Flow Against and Gender indicate that pedestrians are less likely to adopt oblique paths when drivers yield, when they walk faster and when opposing pedestrian flow is present. In contrast, Pedestrian Speed Change, Running, Speed 2 and Safety Distance 1 show positive associations, suggesting that speed adjustments, running behavior, higher far-lane vehicle speeds and greater near-lane safety distance are linked to oblique crossing movements.

For the comparison between irregular and perpendicular crossings, Vehicle Yield, Pedestrian Speed and Accepted Gap show significant negative effects, indicating that yielding behavior, faster walking speeds and larger accepted gaps reduce the likelihood of irregular crossings. Conversely, Pedestrian Speed Change, Running, Safety Distance 1 and Waiting Time are positively associated with irregular crossings, suggesting that speed adjustments, running, longer waiting times and greater near-lane safety margins are linked to more irregular movement patterns under mixed traffic conditions.

Overall, these results highlight that dynamic behavioral adjustments such as pedestrian speed change, running and waiting time, along with driver-pedestrian interaction cues such as vehicle yield, play an important role in shaping pedestrian path selection. The observed patterns are consistent with expected behavioral responses under mixed traffic conditions and support the interpretive value of the MNL model in explaining midblock crossing behavior at the Kamalpokhari site.

Table 3: MNL Coefficient : Kamalpokhari

Variable	Oblique vs Perpendicular			Irregular vs Perpendicular		
	Coefficient	Odds Ratio	p-value	Coefficient	Odds Ratio	p-value
Gender	-1.092	0.335	0.029	-0.854	0.426	0.136
Accepted Gap	-0.170	0.844	0.124	-0.273	0.761	0.024
Speed 2	0.443	1.558	0.028	0.338	1.402	0.136
Waiting Time	0.074	1.077	0.075	0.108	1.114	0.017
Flow Against	-1.946	0.143	0.022	-0.554	0.575	0.539
Running	3.995	54.337	0.041	5.031	153.112	0.015
Pedestrian Speed	-2.394	0.091	0.002	-2.828	0.059	< 0.001
Pedestrian Speed Change	3.278	26.533	< 0.001	4.989	146.724	< 0.001
Vehicle Yield	-2.437	0.087	< 0.001	-4.040	0.018	< 0.001
Safety Distance 1	0.062	1.064	0.033	0.077	1.080	0.022

The MNL model achieved an accuracy of approximately 0.146 on the training set and 0.159 on the testing set, indicating limited predictive performance for pedestrian crossing classification. The McFadden's Pseudo R^2 value of 0.3943, however, suggests moderate explanatory power for a discrete choice model of pedestrian crossing behavior.

The confusion matrix indicates that perpendicular crossings are classified with the highest accuracy, while oblique and irregular categories show substantial overlap. This reflects the inherent variability of pedestrian behavior under mixed traffic conditions, where non-perpendicular movements are less clearly separable. The model underpredicts oblique and irregular crossings, with irregular crossings not predicted in either the training or testing sets, highlighting limitations in capturing such patterns using linear decision boundaries.

Table 4 presents the model fit summary for the Mitrapark MNL model, including -2 Log Likelihood, Likelihood Ratio Chi-Square, degrees of freedom, and McFadden's Pseudo R^2.

Table 5 presents the estimated coefficients, odds ratios and significance levels for the Mitrapark MNL model, with perpendicular crossing path as the reference category. The

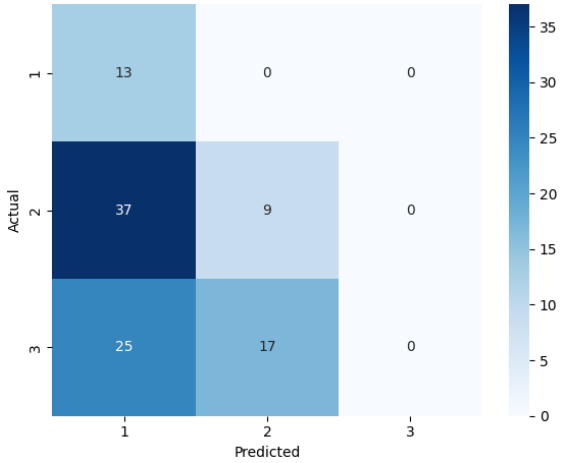


Figure 1: Confusion Matrix: Kamalpokhari

Table 4: Model Summary: Mitrapark

Metric	Value
-2 Log Likelihood (Intercept Only)	741.2576
-2 Log Likelihood (Final Model)	474.5438
Likelihood Ratio Chi-Square	266.7138
Degrees of Freedom	12
Sig. (p-value)	0
McFadden's Pseudo R^2	0.3598

results highlight the influence of waiting time, pedestrian speed change, vehicle yield, roadside obstructions and vehicle type on crossing path choice.

For the comparison between oblique and perpendicular crossings, Pedestrian Speed Change shows a strong positive effect, indicating that pedestrians who adjust their walking pattern are more likely to adopt oblique trajectories. In contrast, Vehicle Yield, Presence of Roadside Obstructions and Gender exhibit negative effects, suggesting that driver yielding, reduced visibility and gender-related characteristics are associated with a lower likelihood of oblique movement. Waiting Time shows a positive association, indicating that pedestrians who wait longer before crossing are more inclined toward oblique paths.

For the comparison between irregular and perpendicular crossings, Pedestrian Speed Change again shows a strong positive effect, reinforcing its role in shaping reactive and non-linear movement patterns. Vehicle Yield and Presence of Roadside Obstructions exhibit negative effects indicating that driver cooperation and reduced visibility discourage irregular crossings. Waiting Time shows a positive association, suggesting that longer waiting leads to more dynamic movement adjustments. Vehicle Type is also significant for this comparison, indicating that larger or heavier vehicles increase the likelihood of irregular crossing behavior.

Overall, the Mitrapark results highlight the influence of reactive behavioral adjustments such as pedestrian speed change and waiting time, along with driver-pedestrian interaction cues such as vehicle yield and environmental constraints such as roadside obstructions in shaping path choice decisions. Vehicle Type shows a marginal influence, particularly for irregular crossings. These patterns are consistent with expected behavioral tendencies in visually constrained, high-friction urban environments and support the interpretive value of the MNL model in explaining midblock crossing behavior.

Table 5: MNL coefficient estimates, odds ratios, and p-values for Mitrapark

Variable	Oblique vs Perpendicular			Irregular vs Perpendicular		
	Coefficient	Odds Ratio	p-value	Coefficient	Odds Ratio	p-value
Gender	-0.844	0.430	0.032	-0.912	0.402	0.066
Waiting Time	0.100	1.105	0.002	0.118	1.126	0.001
Pedestrian Speed Change	3.520	33.793	< 0.001	4.901	134.485	< 0.001
Vehicle Yield	-1.381	0.251	0.009	-3.624	0.027	< 0.001
Vehicle Type	-0.234	0.791	0.539	0.966	2.626	0.050
Presence of Roadside Obstructions	-1.451	0.234	0.016	-2.741	0.065	0.001

The MNL model for the Mitrapark site achieved an accuracy of approximately 0.155 on the training set and 0.197 on the testing set, indicating limited predictive performance for pedestrian crossing classification. This is consistent with the variability and overlap inherent in pedestrian movement patterns. The McFadden's Pseudo R^2 value of 0.3598 indicates moderate explanatory power for interpreting pedestrian path choice behavior within a mixed traffic environment.

The confusion matrix indicates that perpendicular crossings are classified most accurately, while oblique and especially irregular categories exhibit substantial overlap. Similar to Kamalpokhari, the model does not predict any irregular crossings in either the training or testing sets. This highlights the difficulty of distinguishing irregular trajectories using linear decision boundaries particularly in environments like Mitrapark where pedestrian movement is highly adaptive and context dependent.

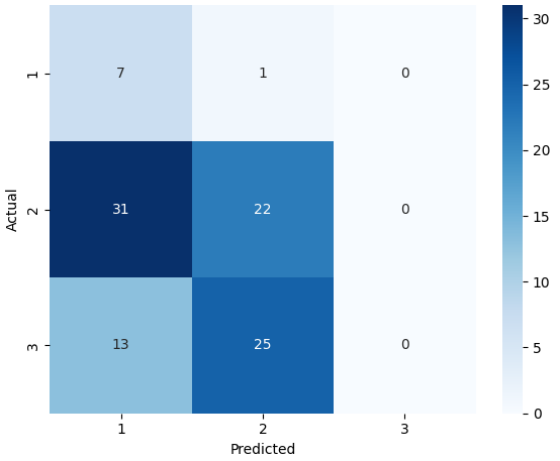


Figure 2: Confusion Matrix: Mitrapark

The MNL models developed for both study sites show low predictive accuracy with Kamalpokhari achieving 0.146 on the training set and 0.159 on the testing set, and Mitrapark achieving 0.155 and 0.197 respectively. Despite this, the models demonstrate moderate explanatory power as reflected by McFadden's Pseudo R^2 values of 0.3943 for Kamalpokhari and 0.3598 for Mitrapark. This indicates that while the models are limited in accurately predicting individual crossing path choices, they remain useful for identifying the underlying behavioral and contextual factors influencing pedestrian movement.

The differences in significant variables between Kamalpokhari and Mitrapark reflect the influence of site-specific conditions

on pedestrian crossing behavior. Variations in traffic flow, vehicle speeds and roadside environment likely contribute to these differences indicating that crossing decisions are context dependent rather than governed by a uniform set of factors.

Table 6: Summary of Both Sites

Site	Training Accuracy	Testing Accuracy	Pseudo R^2
Kamalpokhari	14.60%	15.90%	0.3943
Mitrapark	15.50%	19.70%	0.3598

These findings are consistent with previous studies [3, 13], which show that both pedestrian and traffic related factors influence crossing behavior. The reported McFadden's Pseudo R^2 values also fall within the range commonly observed in similar behavioral modeling studies. For example, [3] reported values around 0.36-0.39 for pedestrian crossing behavior under mixed traffic conditions, while [9] noted that lower values are common in discrete choice models of human behavior. This suggests that the present model provides realistic explanatory performance and useful, location specific insights for targeted safety interventions.

5. Practical Implications

The findings of this study have important implications for urban safety and infrastructure design. The influence of safety distance and vehicle yielding behavior highlights the need to improve driver awareness at midblock crossings. Measures such as refuge islands, warning signage and traffic calming can help reduce pedestrian hesitation and encourage more direct crossings. In addition, planners can use the MNL model results to identify high-risk pedestrian behaviors and adapt crossing facilities accordingly. The inclusion of variables such as pedestrian size and running behavior further suggests that safety interventions should account for behavioral diversity among road users.

The model also provides insights into how behavioral and contextual factors influence crossing type. For example, vehicle yielding is associated with more direct, perpendicular crossings, while behavioral adjustments such as speed change and running are linked to less stable crossing patterns. These findings can support the design of safer crossings by addressing high-risk behaviors and improving pedestrian-vehicle interactions.

6. Conclusion

This study analyzed pedestrian crossing path behavior at uncontrolled midblock locations in Kathmandu using the Multinomial Logit (MNL) model. Two midblock sites, Kamalpokhari and Mitrapark were selected to examine behavioral variability under different traffic and geometric conditions. The models identified pedestrian speed, vehicle yielding, running, safety distance and vehicle speed as significant predictors influencing crossing path choice.

The findings indicate that pedestrians who adjusted their speed or ran during crossing were more likely to choose oblique or irregular paths reflecting adaptive responses to short gaps or perceived traffic pressure. In contrast, driver yielding increased the likelihood of perpendicular crossings demonstrating the importance of pedestrian-vehicle interaction in shaping safer behavior.

Overall, the results confirm that both behavioral and environmental variables play key roles in determining pedestrian crossing choices at uncontrolled midblock locations. The application of the MNL model was useful in quantifying these relationships and provided interpretable insights that can support midblock safety design, enforcement and planning. The study contributes to the growing body of research on pedestrian behavior in developing cities by offering evidence based insights from two uncontrolled midblock locations in Kathmandu and highlighting the importance of integrating behavioral considerations into traffic engineering practice for safer pedestrian environments.

7. Limitations and Future Work

Despite its valuable findings, this study has certain limitations. The analysis was restricted to two selected sites in Kathmandu and therefore cannot be considered fully representative of all uncontrolled midblock crossings in the city, which may limit the general applicability of the results to other urban settings. The data were collected only during daylight hours and potential variations in nighttime pedestrian behavior were not captured. Moreover, some influencing factors such as pedestrians' psychological perceptions (e.g., risk tolerance, urgency) and environmental conditions (e.g., lighting, weather) were not explicitly measured due to practical constraints in field data collection. In addition, the classification of crossing path was based on visual interpretation by a single observer and inter-rater reliability was not assessed, which may introduce subjectivity in the categorization. The findings are also specific to the observed traffic conditions and time periods and may vary under different traffic densities or seasonal conditions. Furthermore, overlap among crossing path categories, particularly for irregular crossings may have reduced the model's ability to distinguish between crossing types.

In addition, the use of the MNL model assumes independence of irrelevant alternatives (IIA), which implies that the relative probability of choosing between any two crossing path options is unaffected by the presence of other alternatives as well as linear relationships between variables and choice probabilities. In the context of pedestrian behavior, these assumptions may not fully reflect the complexity of real-world decision making, where choices are often interdependent and influenced by situational factors. This limitation may partly explain the relatively low predictive accuracy observed in the MNL model. Although the IIA assumption was not formally tested in this study, it is acknowledged as a limitation of the modeling approach. Future studies could explore mixed logit, random parameter models or hybrid approaches where conventional statistical models such as MNL and MLR are integrated with advanced machine learning models such as XGBoost, CatBoost, GAM and Random Forest to better capture unobserved heterogeneity and non-linear effects. Combining field data with simulation based approaches could further enhance predictive accuracy and support the design of adaptive safety measures.

8. Acknowledgment

I would like to express my sincere gratitude to Dr. Pradeep Kumar Shrestha of Department of Civil Engineering, Pulchowk Campus for his invaluable guidance, constructive feedback and encouragement throughout the course of this research. Appreciation is also extended to the faculty members and staff of the Department of Civil Engineering for their assistance during data collection and analysis. Special thanks are given to the volunteers who helped with field observations at the Kamalpokhari and Mitrapark sites.

References

[1] Hüseyin Onur Tezcan, Mahmoud Elmorssy, and Göker Aksoy. Pedestrian crossing behavior at midblock crosswalks. *Journal of safety research*, 71:49–57, 2019.

[2] Ziqian Zhang, Haojie Li, NN Sze, and Gang Ren. Investigating pedestrian crossing route choice at midblocks without crossing facilities: The role of roadside environment. *Travel behaviour and society*, 32:100573, 2023.

[3] B Raghuram Kadali and P Vedagiri. Modelling pedestrian road crossing behaviour under mixed traffic condition. *European transport*, 55(3):1–17, 2013.

[4] Mohammed M Hamed. Analysis of pedestrians' behavior at pedestrian crossings. *Safety science*, 38(1):63–82, 2001.

[5] Rajat Rastogi, Ilango Thaniarasu, and Satish Chandra. Design implications of walking speed for pedestrian facilities. *Journal of transportation engineering*, 137(10):687–696, 2011.

[6] Christopher Cherry, Brian Donlon, Xuedong Yan, Samuel Elliott Moore, and Jian Xiong. Illegal mid-block pedestrian crossings in china: gap acceptance, conflict and crossing path analysis. *International journal of injury control and safety promotion*, 19(4):320–330, 2012.

[7] B Raghuram Kadali, Tadi Chiranjeevi, and Rankireddy Rajesh. Effect of pedestrians un-signalized mid-block crossing on vehicular speed. *International Journal for Traffic & Transport Engineering*, 5(2), 2015.

[8] B Raghuram Kadali and Perumal Vedagiri. Proactive pedestrian safety evaluation at unprotected mid-block crosswalk locations under mixed traffic conditions. *Safety science*, 89:94–105, 2016.

[9] Eleonora Papadimitriou, George Yannis, and John Golias. A critical assessment of pedestrian behaviour models. *Transportation research part F: traffic psychology and behaviour*, 12(3):242–255, 2009.

[10] Albert Puig-Diví, Carles Escalona-Marfil, Josep Maria Padullés-Riu, Albert Busquets, Xavier Padullés-Chando, and Daniel Marcos-Ruiz. Validity and reliability of the kinovea program in obtaining angles and distances using coordinates in 4 perspectives. *PloS one*, 14(6):e0216448, 2019.

[11] Jennifer A Oxley, Elfriede Ihsen, Brian N Fildes, Judith L Charlton, and Ross H Day. Crossing roads safely: an experimental study of age differences in gap selection by pedestrians. *Accident Analysis & Prevention*, 37(5):962–971, 2005.

[12] Thambiah Muraleetharan and Toru Hagiwara. Overall level of service of urban walking environment and its influence on pedestrian route choice behavior: analysis of pedestrian travel in sapporo, japan. *Transportation Research Record*, 2002(1):7–17, 2007.

[13] Khaled Shaaban and Anurag Pande. Evaluation of red-light camera enforcement using traffic violations. *Journal of traffic and transportation engineering (English edition)*, 5(1):66–72, 2018.

iThenticate

Similarity Report ID: oid:3117:588371757

PAPER NAME

Evaluation of Pedestrian Gap Acceptanc e and Crossing Path Choice at Uncontroll ed Midblock Crossings - A Case Study of Kamalpokhari & Mitrapark

AUTHOR

Sudarshan Tamang

WORD COUNT

21112 Words

CHARACTER COUNT

120515 Characters

PAGE COUNT

79 Pages

FILE SIZE

1.8MB

SUBMISSION DATE

May 9, 2026 10:22 PM GMT+5:45

REPORT DATE

May 9, 2026 10:23 PM GMT+5:45

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 5% Publications database
- Crossref database
- Crossref Posted Content database
- 0% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 10 words)

Summary