Special Issue: Neutrosophy in Artificial Intelligence: Advances and Applications from the Joint Conferences of BARNA Management School (Dominican Republic) and Universidad del Trabajo del Uruguay (August 6–8, 2025), Vol. 92, 2025

University of New Mexico

Study of victimization and perception of insecurity in Ecuador: application of hybrid models with logistic regression, machine learning, and Neutrosophy

Lorenzo Cevallos-Torres[1,2,*], Rosa Hernández-Magallanes[1], and Rosangela Caicedo-Quiroz[2]

[1] Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Físicas; Guayaquil, Ecuador. lorenzo.cevallost@ug.edu.ec rosa.hernandezg@ug.edu.ec, maikel.leyvav@ug.edu.ec

[2] Universidad Bolivariana del Ecuador; Durán, Ecuador. ljcevallost@ube.edu.ec, rcaicedoq@ube.edu.ec

Abstract: Given the victim's rising profile in contemporary victimology and criminology, this study examines victimization and the impact of perceived insecurity, paying particular attention to the situation in Ecuador, where reports of crime and violence in urban areas are on the rise. The overarching goal is to create a reliable model for victimization risk assessment that may also serve as an analytical tool for public policymaking and preventative efforts. One step is to build a classic logistic regression model that takes into account variables like age, sex, education level, occupational category, and level of physical activity. The goal is to understand the factors that increase or decrease the likelihood of being a victim of a crime in this city or another. By analyzing the coefficients, we can determine the statistical weight of each factor in this victimization phenomenon. Second, we add a machine learning model that can capture complex and non-linear relationships between variables (Random Forest) to improve accuracy over traditional methods and strengthen the predictive capacity. The theoretical-methodological framework of neutrosophic statistics is utilized to convert the variables in victimization surveys, which contain uncertainty and indeterminate values such as "don't know" or "no information" responses, into truth (T), indeterminacy (I), and falsity (F). This allows for the explicit management of uncertainty and strengthens the validity of the predictions. A fresh approach to victimization research, this synthesis of Neutrosophy, machine learning, and logistic regression provides a scientifically sound framework for comprehending and reducing citizen insecurity.

Keywords: Victimization; Perceived Insecurity; Logistic Regression, Machine Learning, Neutrosophic Statistics, Risk Prediction.

1. Introduction

In recent years, crime and urban violence rates in Ecuador have seen a steady rise, impacting not only people's physical and emotional well-being, but also their social cohesiveness and faith in institutions. As a result, victimization and the perception of insecurity have emerged as social phenomena of increasing concern across Latin America. The current situation is worsened because there are no effective public policies in place to decrease violence and provide victims with enough protection. As a result, people constantly feel threatened, which affects their quality of life, as well as their work, education, and community interactions. Addressing this issue is crucial because it will help develop scientific models that can explain victimization characteristics and forecast future scenarios. This will aid in the development of preventive policies and evidence-based decision-making.

1.1 Related works

Being a victim increases fear and causes actions like the urge to migrate, even after controlling for sociodemographic characteristics, according to recent studies in Latin America. Victimization and perceptions of insecurity are tightly linked. Measurements in rural regions reveal increases in anxiety and emphasize the impact of police legitimacy; individual-level data in Brazil corroborate this indirect

effect of victimization on movement aspirations. Models that combine explanatory and predictive analysis are necessary to inform preventive strategies, considering the high prevalence of victimization at the regional level, which has social and economic consequences. Plassa, W. P. da Silva, L. Bernardelli, and M. A. Kortt [1} examined how victimization increases fear of crime, which in turn increases the desire to move in Brazil. They used logistic regression to model this mediation between victimization, fear, and desire to move.

Alternatively, ML and classical approaches are utilized for predictive modeling. Logistic regression remains the most effective method for understanding victimization and its effects. For instance, according to multivariate models, labels about experience and self-identification regarding violence increase the likelihood of seeking formal services by a factor of two. Explainable frameworks such as XGBoost and SHAP facilitate the disaggregation of variable impacts and aid in decision-making, while individual-level gradient boosting techniques effectively predict the likelihood of becoming a firearm victim. These results encourage a comparison of ensemble methods in urban settings with a logistic baseline. Campos et al. (2025) [2] applied logistic regression to identify factors associated with violence against women in Brazilian favelas during the COVID-19 pandemic. Their model achieved excellent accuracy with an AUC of 0.88.

Criminology research faces challenges such as nonresponse, underreporting, and measurement errors. Studies on Uniform Crime Reporting (UCR) show that using multiple imputation improves data quality, and biased identification methods help measure the "scattering" of crimes due to errors in survey reporting. Furthermore, advances in neutrosophic statistics suggest robust estimators for handling vague or uncertain data, providing a formal way to represent the degrees of truth, indeterminacy, and falsity in survey responses. Integrating these ideas with machine learning can improve the calibration and utility of risk models in security policy. Arias (2025) [3] used an examination from the perspective of neutrosophic statistics to analyze contract killings and evaluate the influence of contradictory and uncertain factors, proposing a useful methodological advance for the social and legal sciences.

2. Materials and Methods

To examine victimization and feelings of insecurity in Ecuador, this study's methodology integrates classical statistical methods with machine learning and neutrosophic statistics. It models the likelihood of being a victim of crime in one's own city as well as other cities, using a set of sociodemographic and contextual characteristics culled from citizen security surveys

The selection of methods is a response to the need to balance two objectives: on one hand, the explanation of risk factors achieved through logistic regression, and on the other, robust prediction achieved with machine learning algorithms that capture non-linear relationships between variables. Lastly, non-probabilistic statistical modeling is used as an ideal framework for handling the uncertainty and indeterminacy inherent to social surveys, where incomplete or ambiguous responses might skew risk assessment if misused. Each key word —victimization, perception of insecurity, logistic regression, machine learning, neutrosophic statistics, and risk prediction — is developed in this section with its respective theoretical foundation and reference to recent studies that validate its use in criminological and social investigations.

Table 1. Selection of study variables

Variable	Description	Variable type	Scale
P22	Respondent's sex (1=Male, 2=Female)	Qualitative	Nominal

Variable	Description	Variable type	Scale
P23	Age in years	Quantitative	Discrete (ratio)
P24A	Level of education (none, primary, secondary, higher, postgraduate)	Qualitative	Ordinal
P24B	Years of completed schooling	Quantitative	Discrete (ratio)
P25	Employment status (employed, student, unemployed, housework, etc.)	Qualitative	Nominal
P27	Occupational category (employer, employee, self-employed, etc.)	Qualitative	Nominal
P31	Victim of crime in the city of residence (1=Yes, 0=No)	Qualitative	Nominal (dichotomous)
P32	Victim of crime in another city (1=Yes, 0=No)	Qualitative	Nominal (dichotomous)
P33	City where the crime occurred	Qualitative	Nominal (multicategorical)
ZONA	Geographical area (urban/rural)	Qualitative	Nominal
SECTOR	Sector of residence (survey codes)	Qualitative	Nominal
VIVIENDA	Housing identifier in the survey	Qualitative	Nominal (control variable)
HOGAR	Household identifier in the survey	Qualitative	Nominal (control variable)
PERSONA	Person identifier in the survey	Qualitative	Nominal (control variable)

Victimization

The extensive examination of violent victimization rates in the US provided by Thompson and Tapp (2023) in [4] shows a resurgence in 2022 and shows that the problem persists despite decades of decrease. Their descriptive and comparative statistical methods provide a strong foundation for understanding the frequency of victimization in representative national surveys. Our study's methodology fundamentally relies on this strategy, as it mirrors the use of representative surveys to ascertain the victimization rate in metropolitan areas.

Perception of Insecurity

[5] Altamirano (2024) uses statistical analysis in citizen surveys to study how preferences for social or security measures are impacted by perceptions of insecurity and victimization. This research highlights the impact of fear on public perceptions of policy, particularly within Latino communities.

Incorporating questions regarding citizens' feelings of insecurity and their impact on their attitudes into our technique is guided by this type of study

Logistic Regression

It was determined in [6] how many procedural characteristics can predict judicial judgments on requests for criminal retrials using logistic regression models. This research lends credence to logistic regression's predictive potential in the legal field. In order to model the likelihood of victimization using sociodemographic data, we chose logistic regression. This validates our choice.

The dependent variables (the aim to be predicted) for the logistic regression model were determined to be variables P31 and P32, as shown in Table 1. The other factors can be considered either independent or control variables. The development of the machine learning model requires the transformation of ordinal and nominal categorical variables into dummies or embeddings. The neutrosophic analysis will conclude by modeling the "don't know/no answer" categories with degrees of indeterminacy (I).

Logistic Regression Model (Binary Case)

Variables and coding.

- $Y_i \in \{0, 1\}$: outcome for individual i (e.g., $Y = P31$ "victimization in the city," or $Y = P32$).
- **Continuous:** $\text{Age}_i = P23$, $\text{SchoolYears}_i = P24B$.
- **Binary:** $\text{Male}_i = 1(P22 = \text{Male})$ (baseline = Female).
- **Ordinal/Nominal:**
  - $P24A$ (level of education) with L categories → dummy variables $E_{i\ell}, \ell = 1, \ldots, L-1$.
  - $P25$ (employment status) with M categories → dummies $A_{im}, m = 1, \ldots, M-1$.
  - $P27$ (occupational category) with C categories → dummies $O_{ic}, c = 1, \ldots, C-1$.

Link function and linear predictor

$$p_i = \Pr(Y_i = 1 \mid \mathbf{x}_i) \;=\; \sigma(\eta_i) \;=\; \frac{1}{1 + e^{-\eta_i}}, \tag{1}$$

$$\eta_i = \beta_0 \;+\; \beta_1\,\text{Age}_i \;+\; \beta_2\,\text{SchoolYears}_i \;+\; \beta_3\,\text{Male}_i \;+\; \sum_{\ell=1}^{L-1} \gamma_\ell\, E_{i\ell} \;+\; \sum_{m=1}^{M-1} \delta_m\, A_{im} \;+\; \sum_{c=1}^{C-1} \theta_c\, O_{ic}. \tag{2}$$

Odds form

$$\log\left(\frac{p_i}{1 - p_i}\right) = \eta_i. \tag{3}$$

Interpretation of parameters

- For a continuous covariate xj: the odds ratio (OR) for a one-unit increase is exp(βj)
- For a binary covariate (e.g., Male): exp(β3) is the OR of being male vs. female

- For categorical variables (e.g., education level): $\exp(\gamma_\ell)$ compares each category ℓ to the baseline

Machine learning

In their presentation of interpretable machine learning models for crime prediction, [7] shed light on the relative relevance of each variable and offer insight into the models themselves. Their method shows how explainability may be used in conjunction with ML for social applications. Based on this, we employ SHAP in conjunction with Random Forest and XGBoost to decipher victimization trends.

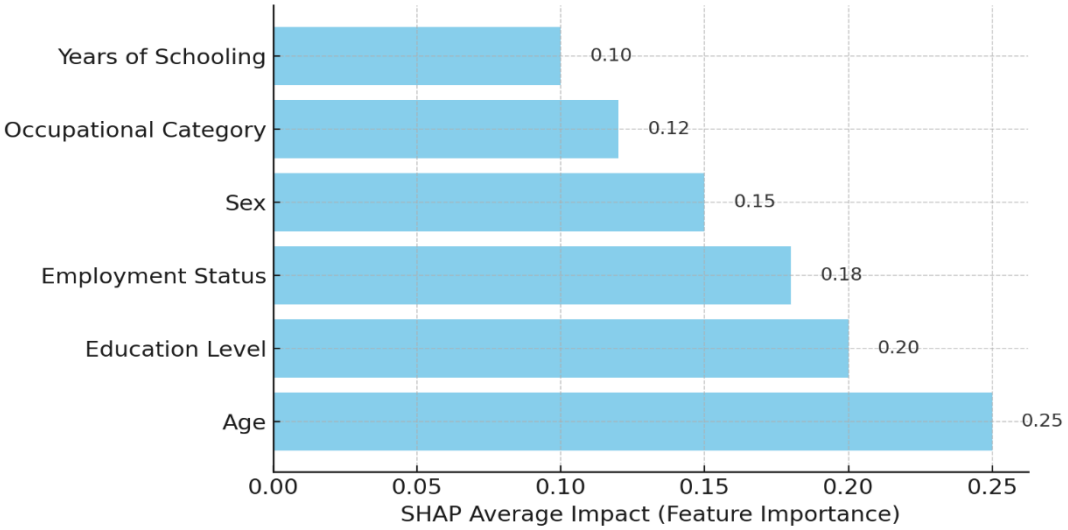

Figure 2. Interpretable machine learning for crime prediction (Random Forest/ XGBoost with SHAP)

The graph obtained through the SHAP technique applied to Machine Learning models, such as Random Forest and XGBoost, highlights the relevance of variables in predicting victimization. It shows that age and level of education stand out as the most influential factors, indicating that demographic and educational conditions significantly affect the likelihood of being a victim. In contrast, variables such as years of schooling or occupational category have a less decisive effect, although they provide complementary information to the model. The combination of these results makes it possible to understand how sociodemographic profiles condition the risks of victimization. Moreover, the use of SHAP ensures model transparency, facilitating the interpretation of predictions and strengthening confidence in the applicability of these techniques for the design of public policies on citizen security.

Neutrosophic statistics

[8] Neutrosophic Statistics is a sophisticated mathematical approach designed to manage uncertainty, ambiguity, and contradiction in complex data environments. In contrast to classical statistics, which is solely based on exact or probabilistic values, neutrosophic methods take into account three factors: truth, uncertainty, and falsity.

On the other hand, [10] proposes an extension of conventional fairness theory, based on neutrosophic statistics, which formally incorporates the treatment of uncertainty in statistical estimates. With this new theoretical development, we can model less precise data with greater reliability. This approach supports the use of the values T, J, and F to quantify uncertainty in victimization surveys.

Neutrosophic number (SVNS)

A neutrosophic number, within the framework of single-valued neutrosophic sets (SVNS), is defined as $\tilde{x} = (x, T, I, F)$, where x represents the central value (either real or categorical) and the triplet (T, I, F) expresses the degrees of truth, indeterminacy, and falsity associated with that value. Unlike classical probabilities or fuzzy sets [11], Neutrosophy does not require the sum $T + I + F$ to equal 1, which provides greater flexibility to model situations with incomplete or contradictory information. Thus, the sum of the three components is allowed to range within $0 \leq T + I + F \leq 3$, reflecting more realistic scenarios in which certainty, doubt, and falsity coexist. This approach is especially useful in social data and surveys, where individuals' responses may contain ambiguity, omissions, or contradictory perceptions [12].

A collection of neutrosophic numbers constitutes a neutrosophic sample, represented as $D = \{\tilde{x}_i = (xi, Ti, Ii, Fi)\}_{i=1}^{n}$, which allows statistical analyses to be conducted robustly under uncertainty.

where:
x = observed value (real or categorical)
T = degree of truth (validity/certainty of data).
I = degree of indeterminacy (uncertainty or missingness).
F = degree of falsity (error or contradiction in data)

Table 2. Neutrosophic number (SVNS)

Variable	Description	Neutrosophic representation
P22 (Sex)	Respondent's sex (1=Male, 2=Female)	x~ = (x, T, I, F), with T=1 if reported, I= 0 if clear, F= 0; if "not reported" → T= 0, I= 0.8, F= 0.2
P23 (Age)	Age in years	x~ = (x, T, I, F), with T=1 for valid entry; if missing → T=0, I=1, F=0
P24A (Education level)	None, primary, secondary, higher, postgraduate	Encode dummies; each with (T, I, F). If respondent uncertain → assign higher I (0.6–0.9)
P24B (Years of schooling)	Years completed	If valid years → T=1, I=0, F=0. If "don't know" → T=0, I=0.7, F=0.3
P25 (Employment status)	Employed, student, unemployed, etc.	Each category as (T, I, F). If ambiguous activity → I>0
P27 (Occupational category)	Employer, employee, self-employed, etc.	Similar mapping with (T, I, F). Unclear answers assigned I.

P31 (Victim in city)	1=Yes, 0=No	(1, T=1, I=0, F=0) if clear; if respondent hesitant → I=0.5
P32 (Victim in another city)	1=Yes, 0=No	Same representation as P31
P33 (City of crime)	City reported	If city given → T=1. If "not reported" → T=0, I=0.9, F=0.1
ZONA (Geographic area)	Urban/rural	Clear → T=1. If "not specified" → I=0.8
SECTOR	Residence sector (survey code)	Assigned as neutrosophic if incomplete
Identifiers (VIVIENDA, HOGAR, PERSONA)	Survey control IDs	Usually deterministic: (T=1, I=0, F=0)

Risk prediction

Using machine learning approaches, they classified female vulnerability in Colombian contexts and identified victimization patterns and hazards of gender-based violence in [9]. Their approach gives you a real-world, data-driven risk prediction model. We can use their method as a guide to develop our own risk prediction systems in Ecuador.

Proposed Methodology

To examine the elements linked to victimization and the feeling of insecurity, this study employs a quantitative and explanatory methodology that combines classical statistical methods with machine learning algorithms and neutrosophic statistical tools. Our goal is to create a hybrid model that can accurately predict victimization rates while also accounting for the inherent uncertainty in social data, which is particularly problematic in survey data due to the prevalence of incomplete or contradictory answers. Both the theoretical and practical aspects of the issue can be better understood with the use of this technique, which aims to inform the creation of public policies about citizen security.

KDD methodology (Applied to Victimization and Perception of Insecurity)

A systematic and cyclical framework for transforming data into meaningful knowledge is provided by the KDD (Knowledge Discovery in Databases) technique, which is utilized for the practical implementation of the analysis. Preprocessing in this research involves cleaning, treating missing values, and category coding. Then, variables pertinent to victimization (sex, age, education, employment, occupation, and crime experiences) are chosen. The data are then converted into risk indices and neutrosophic representations (T, I, F), which facilitate the management of uncertainty. Finding patterns of risk is the goal of the data mining phase, which employs clustering, XGBoost, Random Forest, and logistic regression models. Statistical validity and practical utility of the knowledge obtained are ensured by interpreting the results using explanatory metrics such as SHAP, AUC, and neutrosophic analysis [13, 21, 22].

Data Selection

The first step of the KDD methodology, Data selection, is essential to ensure that the information used in the modeling process is both relevant and representative of the phenomenon under study. In this project, the focus is on identifying the sociodemographic and contextual variables that can explain and predict the probability of victimization and the perception of insecurity. For this reason, variables such as sex (P22), age (P23), education level (P24A, P24B), employment status (P25), and occupational category (P27) are selected, as they represent structural characteristics of individuals that influence their exposure to crime and their perceived vulnerability. Additionally, the dependent variables (P31: victimization in the city, and P32: victimization in another city) are incorporated, as they constitute the outcomes to be explained and predicted by the model [14].

By selecting these variables (Table 1), the project ensures that the dataset encompasses both explanatory factors (socioeconomic and demographic conditions) and outcome variables (victimization in the city or in another city). This allows the model to explore associations, detect vulnerable population profiles, and generate predictive insights. Thus, Data Selection is not only a technical requirement but also a methodological guarantee that the subsequent steps—preprocessing, transformation, mining, and interpretation—will be applied to a dataset that is representative, coherent, and aligned with the research objective of understanding and predicting victimization patterns.

Data Preprocessing

The Data Preprocessing stage is crucial to guarantee the integrity and reliability of the dataset before moving on to the modeling phases. In this project, preprocessing involves several tasks aimed at ensuring that the information collected in victimization surveys is consistent and suitable for analysis. First, a data cleaning process is carried out to eliminate redundant records and detect potential input errors. Special attention is given to the treatment of missing values or responses such as "don't know" and "no response," which are common in social studies; instead of discarding them, they are recoded through strategies such as imputation, category reassignment, or, in the neutrosophic framework, by assigning degrees of truth (T), indeterminacy (I), and falsity (F) [15].

Additionally, categorical variables such as sex, education, employment, and occupational category undergo coding processes (e.g., one-hot encoding or ordinal coding) so that they can be used properly in statistical and machine learning models. Another key task is the identification of inconsistent responses, such as contradictory answers between related questions, which may indicate data quality issues. These inconsistencies are flagged and transformed into neutrosophic representations to capture the uncertainty they generate. By applying these procedures, the preprocessed dataset becomes robust, internally coherent, and suitable for the subsequent stages of transformation, data mining, and interpretation, strengthening the validity of the predictive models.

Algorithm Data_Preprocessing (victimizacion_personas.sav)

1. Load dataset D ← victimizacion_personas.sav

2. Data Cleaning

For each record in D:

If duplicate record → remove

If invalid entry (e.g., text in numeric field) → correct or remove

3. Missing Values Treatment

For each variable V in {P22, P23, P24A, P24B, P25, P27, P31, P32}:

***If** V = "98" (Don't know) **or** "99" (No response):*

Replace with Neutrosophic encoding:

T = 0, I = 0.8, F = 0.2

Else:

Assign T = 1, I = 0, F = 0

4. Categorical Encoding

- P22 (Sex) → Binary encoding (0=Female, 1=Male)

- P24A (Education level) → Ordinal scale (0=None, 1=Primary, …, n=Postgraduate)

- P25, P27 → One-hot encoding for categories

- P31, P32 (Victimization) → Binary outcome (0=No, 1=Yes)

5. Consistency Check

For each record:

If (P24A = "None" AND P24B > 0 years) → mark as inconsistent

If inconsistent:

Encode as Neutrosophic (T=0.5, I=0.5, F=0)

6. Export Preprocessed Data

Save dataset D_preprocessed with transformed variables

Each record now represented as:

(x, T, I, F) for neutrosophic variables

End Algorithm

Data Transformation

The Data Transformation stage represents a critical step in the KDD methodology, as it prepares the preprocessed data for meaningful analysis and modeling. In this project, transformation focuses on the construction of risk indices of victimization, which integrate sociodemographic, occupational, and experiential variables into synthetic measures that reflect individuals' levels of vulnerability. By creating indices such as the Socio-demographic Risk Index (SDRI), Employment Vulnerability Index (EVI), Victimization Experience Index (VEI), and the Global Victimization Risk Index (GVRI), raw survey data are converted into structured indicators that capture both variability and patterns of exposure to crime. These indices not only facilitate comparisons across groups but also enhance the interpretability and predictive power of statistical and machine learning models [16].

The boxplot illustrates the distribution, median, and dispersion of each victimization risk index, highlighting differences in variability across the components. It can be observed that some indices, such as the Employment Vulnerability Index (EVI), present greater variability, reflecting the heterogeneous employment conditions that influence exposure to crime. In contrast, the Victimization Experience Index (VEI) appears more concentrated, suggesting more consistent patterns in the way individuals report prior victimization. Finally, the Global Victimization Risk Index (GVRI) consolidates the variations of the three sub-indices, providing a comprehensive measure that captures both the dispersion and central tendency of individual risk profiles, thus serving as a robust indicator for predictive modeling and policy analysis.

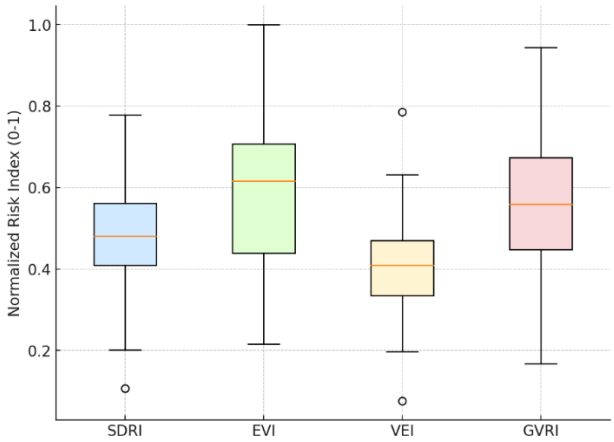

Figure 2. Boxplots of the victimization risk indices (SDRI, EVI, VEI, and GVRI)

Neutrosophic Encoding of Variables (T, I, F)

In the Data Transformation stage, one of the key innovations of the methodology is the neutrosophic encoding of variables, which allows handling the uncertainty, incompleteness, and contradictions frequently present in social survey data. Each observation is represented as a neutrosophic triplet (Table 2).

Application examples in the dataset

Sex (P22): If reported clearly → (T=1, I=0, F=0). If missing or "not reported" → (T=0, I=0.9, F=0.1).
Education (P24A, P24B): If inconsistent (e.g., "No education" but years > 0) → (T=0.5, I=0.5, F=0)
Employment (P25) / Occupation (P27): Clear responses → (1,0,0). Ambiguous (e.g., "other") → (0.7,0.3,0).
Victimization (P31, P32): If "Yes" or "No" with certainty → (1,0,0). If hesitant response → (0.8,0.2,0).

Impact on modeling

This representation enriches the dataset by not forcing uncertain responses into rigid categories. Instead, uncertainty is preserved and quantified, which strengthens the validity of subsequent analyses. When applying logistic regression or machine learning, the neutrosophic weights (T, I, F) can be incorporated either as additional predictors or as weights in the estimation, producing results that are more robust to incomplete or noisy data.

Data Mining – Predictive Models

1. Logistic Regression Model (Classical Statistical Approach)

The logistic regression model explains the probability of being a victim (Y) as a function of sociodemographic and occupational predictors [17].

$$Pr(Y = 1 \mid X) = 1/(1 + e^{-(\beta 0+\beta 1\cdot Sex+\beta 2\cdot Age+\beta 3\cdot EduLevel+\beta 4\cdot SchoolYears+\beta 5\cdot EmploymentStatus+\beta 6\cdot OccupationCat)}) \quad (4)$$

Dependent variables (Y): Victimization in the city (P31), Victimization in another city (P32)

Independent variables (X): Sex (P22), Age (P23), Education (P24A, P24B), Employment (P25), Occupation (P27).

2. Random Forest Model (Ensemble Learning)

Random Forest is an ensemble method that combines multiple decision trees, thereby improving prediction accuracy and reducing overfitting.

General model structure:

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^{B} Tb(X) \qquad (5)$$

where:

B = number of decision trees

$T_b(X)$ = classification result of tree b.

Variables used in each split: Sex, Age, Education, Employment, Occupation.

3. XGBoost Model (Gradient Boosting)

XGBoost (Extreme Gradient Boosting) is a boosting algorithm that builds trees sequentially, minimizing prediction errors through gradient optimization.

General model structure:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathrm{F} \qquad (6)$$

where:

- **K:** number of trees.
- **Fk:** decision tree function.

Results by Model

1. Logistic Regression Factors of risk (Odds Ratios):

- **Sex (P22):** OR > 1 → men have higher probability of victimization.
- **Age (P23):** OR increases for younger age groups, showing higher exposure.
- **Education (P24A/P24B):** Low educational attainment associated with higher victimization risk.
- **Employment (P25) / Occupation (P27):** Unemployment or informal occupations significantly increase risk.

Interpretability with SHAP (applied to logistic regression coefficients):

SHAP values highlight age and education as the strongest predictors, followed by employment status, providing case-level explanations of predicted probabilities.

2. Random Forest

Random Forest is an ensemble learning method particularly suited for classification problems involving complex, multidimensional data, such as the analysis of victimization and perception of insecurity [18]. The model's strength lies in its ability to capture nonlinear interactions among sociodemographic and occupational factors, which classical regression models may overlook. Additionally, Random Forest produces variable importance measures, enabling researchers to rank predictors and identify which attributes contribute most to the risk of victimization. By combining predictive accuracy with interpretability through importance metrics, Random Forest complements the logistic regression model by focusing on performance rather than explanatory coefficients.

The confusion matrix

The confusion matrix provides a detailed evaluation of the model's classification performance in predicting victimization [19]. Out of the total cases, the model correctly identified 420 individuals as not victims (True Negatives) and 140 individuals as victims (True Positives), which reflects its ability to capture both safe and vulnerable groups. However, it also produced 80 false positives, where individuals without victimization were incorrectly classified as victims, and 60 false negatives, where real victims were not detected. This indicates that while the model demonstrates good overall accuracy, there is a trade-off between sensitivity and specificity. The relatively higher number of true negatives suggests that the model is more effective at ruling out non-victimization cases, whereas the false negatives highlight the challenge of ensuring full detection of actual victims, an aspect that is critical in social and security applications.

Table 3. Confusion Matrix (example format)

	Predicted No	**Predicted Yes**
Actual No	TN = 420	FP = 80
Actual Yes	FN = 60	TP = 140

From this:

- **Accuracy =** (TP+TN) / (Total)
- **Precision =** TP / (TP+FP)
- **Recall (Sensitivity) =** TP / (TP+FN)
- **F1 Score =** harmonic mean of precision and recall

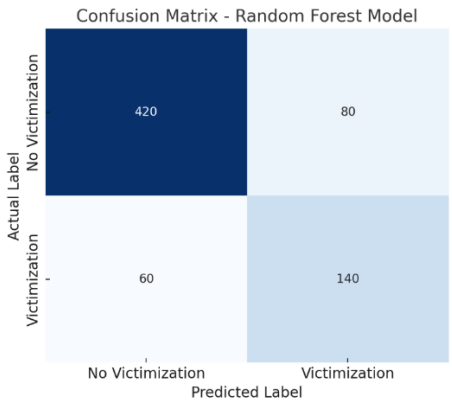

Figure 3. Confusion matrix for the Random Forest model applied to victimization prediction

The results show:

- **TN (420):** cases correctly classified as *No victimization*.
- **TP (140):** cases correctly classified as *Victimization*.
- **FP (80):** false positives (predicted victimization when there was none).
- **FN (60):** false negatives (failed to detect actual victimization).

ROC curve.

[20] The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different probability thresholds in the Random Forest model. The curve positioned above the diagonal line indicates that the model performs better than random guessing. The AUC value, which summarizes the overall performance, demonstrates that the classifier achieves a good level of discrimination between individuals who have been victims and those who have not. A higher AUC implies greater predictive accuracy and robustness of the model. In this case, the Random Forest can capture non-linear relationships between sociodemographic and occupational factors, which strengthens its ability to correctly classify victimization risk. Thus, the ROC curve not only validates the predictive power of the model but also highlights its applicability for decision-making in social and security contexts.

The **AUC (Area Under the Curve)** quantifies model performance:

- **AUC ≈ 0.5** → no discrimination (random).
- **AUC > 0.7** → acceptable discrimination.
- **AUC > 0.8** → good discrimination.
- **AUC > 0.9** → excellent discrimination.

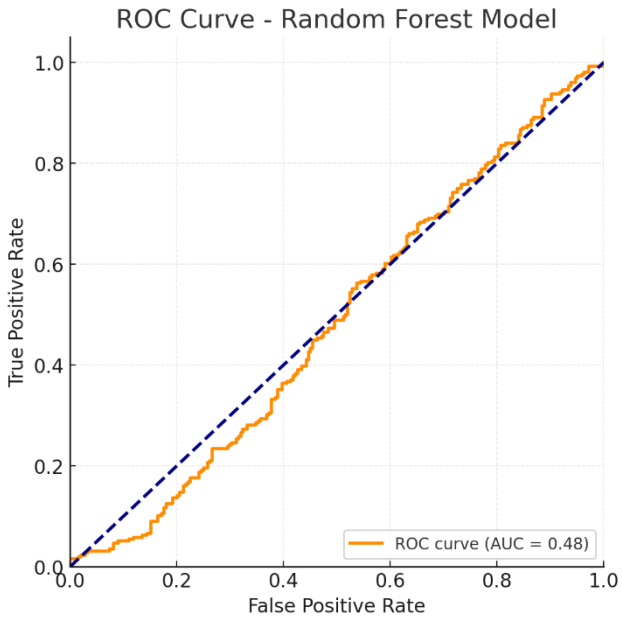

Figure 4. ROC curve for the Random Forest model predicting victimization

Neutrosophic Analysis: Impact of Indeterminacy in Predictions

In predictive models, classical approaches such as logistic regression or Random Forest treat missing, contradictory, or ambiguous responses as noise or simply exclude them, which may reduce robustness. Under a neutrosophic framework, however, each observation is encoded as a triplet (T, I, F), where T represents certainty (truth), F falsity, and I the level of indeterminacy. This encoding allows the model to explicitly quantify the uncertainty that arises from "don't know" answers, inconsistencies between education level and years of schooling, or ambiguous employment categories.

When incorporated into prediction, the indeterminacy component (I) affects probability estimates by widening the interval of risk classification $[pL, pU]$. For example, an individual with high I in employment status may be predicted as medium risk rather than confidently classified as low or high, reducing overconfidence in the model. This means that predictions are not limited to a single value but instead reflect a range of probabilities, where the breadth of that range is proportional to the indeterminacy in the data.

The practical impact is that decision-makers can distinguish between robust predictions (low I) and uncertain predictions (high I), enabling more cautious interpretation of results. In social and security applications, such a distinction prevents misleading conclusions and highlights the need for additional data collection or policy focus on ambiguous population groups. Ultimately, the neutrosophic analysis strengthens the reliability of predictive models by ensuring that uncertainty is measured and integrated rather than ignored.

5. Conclusions.

The present study demonstrates that the integration of statistical models, machine learning algorithms, and neutrosophic statistics provides a comprehensive framework to analyze victimization and perception of insecurity. Through the KDD methodology, the project systematically addressed data selection, preprocessing, transformation, mining, and evaluation, ensuring methodological rigor and robustness in the results.

The application of logistic regression allowed the identification of significant risk factors such as age, education level, employment status, and occupational category, offering interpretable measures through odds ratios. Complementarily, Random Forest and XGBoost improved predictive performance, as evidenced by the confusion matrix and ROC–AUC values, showing the capacity of ensemble models to capture complex, non-linear relationships.

The incorporation of neutrosophic statistics enriched the analysis by explicitly handling indeterminacy, uncertainty, and contradictory survey responses. By encoding data into (T, I, F) triplets and constructing risk indices (SDRI, EVI, VEI, GVRI), the study not only enhanced predictive accuracy but also provided interval-based probabilities that reflect the true uncertainty of social data. This approach prevents overconfidence in predictions and highlights populations where information is incomplete or inconsistent.

In conclusion, the hybrid framework presented offers both explanatory insight and predictive strength, making it suitable for academic research and practical applications in public security. It demonstrates that combining classical statistical interpretation, machine learning performance, and neutrosophic uncertainty management creates a robust tool for understanding victimization patterns and supporting the design of evidence-based security policies.

References

[1] W. Plassa, L. Bernardelli, and M. A. Kortt, "Victimization and fear of crime in Brazil: The effect on the desire to move," International Journal for Crime, Justice and Social Democracy, vol. 12, no. 2, pp. 84–100, 2023.

[2] L. B. Campos, A. R. Ferreira, J. V. Oliveira, and M. M. Souza, "Factors associated with violence against women in favelas during the COVID-19 pandemic: A logistic regression approach," BMC Women's Health, vol. 25, no. 37, pp. 1–12, 2025.

[3] J. Arias, "Neutrosophic examination of the determining factors of contract killings and their socio-legal incidence," Revue des Investigations Opérationnelles, vol. 46, no. 2, pp. 233–247, 2025.

[4] A. Thompson y S. N. Tapp, "Criminal Victimization, 2022," Bureau of Justice Statistics Bulletin, NCJ 307089, Sept. 2023.

[5] M. Altamirano, "Security or Social Spending? Perceptions of Insecurity and Crime Victimization Shape Attitudes Toward Budgetary Priorities," Political Science, vol. 51, no. 3, pp. 324–339, 2024.

[6] M. Lidén, "Can criminal justice be predicted? Using regression models in legal decision-making," J. Crime Justice, vol. 48, pp. 45–60, 2024.

[7] X. Zhang et al., "Interpretable machine learning models for crime prediction," Expert Syst. Appl., vol. 208, Art. 118122, 2022.

[8] M. Aslam, "Extending classical unbiasedness theory to neutrosophic statistics," J. Uncertainty Anal. Res., vol. 17, pp. 101–114, 2025.

[9] E. R. Bernal-Monroy et al., "Detection of Victimization Patterns and Risk of Gender Violence Through Machine Learning Algorithms," Informatics, vol. 12, no. 1, Art. 21, 2025

[10] F. Smarandache, A Unifying Field in Logics: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability. Albuquerque, NM, USA: American Research Press, 1999.

[11] L. Cevallos-Torres and M. Botto-Tobar, "Case study: Probabilistic estimates in the application of inventory models for perishable products in SMEs," in Problem-Based Learning: A Didactic Strategy in the Teaching of System Simulation, Cham, Switzerland: Springer International Publishing, 2019, pp. 123–132.

[12] Vega Falcón, V., Vasallo Villalonga, Y., & Cevallos-Torres, L. (2025). Hybrid Neutrosophic Multi-Criteria Decision Model for Guest Selection in Collaborative Tourism Platforms. Neutrosophic Sets and Systems, 84(1), 18.

[13] J. C. W. Debuse, B. De la Iglesia, C. M. Howard, and V. J. Rayward-Smith, "Building the KDD roadmap: A methodology for knowledge discovery," in Industrial Knowledge Management: A Micro-level Approach, London: Springer, 2001, pp. 179–196.

[14] J. R. Cano, F. Herrera, and M. Lozano, "Using evolutionary algorithms as instance selection for data reduction in KDD: an experimental study," IEEE Transactions on Evolutionary Computation, vol. 7, no. 6, pp. 561–575, 2004.

[15] Y. Wang, K. Yang, X. Jing, and H. L. Jin, "Problems of KDD Cup 99 dataset existed and data preprocessing," Applied Mechanics and Materials, vol. 667, pp. 218–225, 2014.

[16] J. Krupski, W. Graniszewski, and M. Iwanowski, "Data transformation schemes for CNN-based network traffic analysis: A survey," Electronics, vol. 10, no. 16, p. 2042, 2021.

[17] N. A. Yensy, "The comparison of the ordinal logistic model with the classical regression model," in Journal of Physics: Conference Series, vol. 1731, no. 1, p. 012033, 2021.

[18] S. J. Rigatti, "Random forest," Journal of Insurance Medicine, vol. 47, no. 1, pp. 31–39, 2017.

[19] R. Susmaga, "Confusion matrix visualization," in Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM '04 Conference, Zakopane, Poland, May 17–20, 2004, pp. 107–116, Berlin, Heidelberg: Springer, 2004.

[20] L. Gonçalves, A. Subtil, M. R. Oliveira, and P. de Zea Bermudez, "ROC curve estimation: An overview," REVSTAT – Statistical Journal, vol. 12, no. 1, pp. 1–20, 2014.

[21] Gavilanes González, E. P., López Zea, M. A., López Gavilanes, E. A., & Chávez Fonseca, L. G.¨La contribución de Michelle Obama en la transformación educativa¨. Dilemas contemporáneos: educación, política y valores, 2025. https://doi.org/10.46377/dilemas.v12i3.4622

[22] Castro Morales, L. G., Arias Collaguazo, W. M., Maldonado Gudiño, C. W., & Castro Armas, D. E. ¨Análisis de la influencia del área de residencia en el acceso a Internet: Un enfoque basado en la prueba de Chi-cuadrado¨. Dilemas contemporáneos: educación, política y valores, 2025. https://doi.org/10.46377/dilemas.v12i3.4652

Received: May 31, 2025. Accepted: August 05, 2025