Disentangling Victim–Perpetrator Overlap: A Dual-Classifier Study Based on Global Predictors from a 347-Item Violence Survey

Abstract

This study investigates the overlap between adolescent victimisation and perpetration by introducing a dual-classifier framework based exclusively on global predictors. The goal is to provide an interpretable, recall-oriented screening tool capable of identifying youths at risk of occupying both roles.

Our methodological design decomposes the highly imbalanced victim–perpetrator label into two balanced binary tasks: (i) a cost-complexity pruned decision tree for victims and (ii) a compact feed-forward neural network for perpetrators. Both classifiers are trained on orthogonalised features obtained through principal-components analysis (PCA) after a structured pipeline of feature selection, engineering, and scaling. An ensemble rule then predicts dual-role status only when both models concur.

Empirical results show that the victim classifier achieves a recall of 91.1 % with $F_1 = 0.68$, while the perpetrator network reaches 90.0 % recall with $F_1 = 0.45$. The ensemble, though stricter, still attains 93.8 % recall and a balanced accuracy of 70.3 %, substantially improving specificity compared to single models.

We conclude that the dual-classifier decomposition offers a transparent, high-sensitivity approach to violence screening. Practically, the decision tree enables rapid identification of likely victims, while the neural network captures behavioural volatility linked to offending. Scientifically, the results highlight shared global predictors—particularly household structure, substance use, and coping resources—as cross-cutting mechanisms of victim–perpetrator overlap.

Keywords: victim–perpetrator overlap; dual classifier; decision tree; neural network; PCA; recall; balanced accuracy

Contents

1 Introduction and Motivation **5**

2 Statistical Analysis **6**

3 Data Processing **7**
3.1 Feature Selection . . . 7
3.1.1 Conceptual Filtering via Four-Way Typology . . . 7
3.1.2 Retained Global Predictors (35 variables) . . . 8
3.1.3 Why a Dual-Model Decomposition? . . . 8
3.1.4 Analytical Uses of Excluded Items . . . 8
3.2 Feature Engineering . . . 9
3.2.1 Engineered Global Predictors (27 variables) . . . 10
3.2.2 Descriptive tables . . . 11
3.3 Feature Pre-processing . . . 12
3.3.1 Normalisation . . . 12
3.3.2 Dimension Reduction via PCA . . . 12
3.3.3 Explained Variance by Principal Component . . . 13

4 Model Building **14**
4.1 Victim Classifier . . . 14
4.2 Perpetrator Classifier . . . 15
4.3 Summary of Over-fitting Safeguards . . . 15

5 Model Performance **16**
5.1 Victim Classifier Performance . . . 17
5.2 Perpetrator Classifier Performance . . . 18
5.3 Ensemble Model Overview . . . 19
5.4 Ensemble Classifier Performance . . . 19
5.5 Discussion of Performance . . . 20

6 Discussion of Results **21**
6.1 Decision Tree Feature Importance . . . 21
6.1.1 How to Recover the Raw-Feature Weights . . . 21
6.1.2 Complete Raw-Feature Coefficient Table . . . 22
6.2 Neural-Network Feature Importance . . . 23
6.2.1 Methodology Applied . . . 23
6.2.2 Global Importance of Principal Components . . . 24
6.2.3 Importance of Original Variables . . . 25
6.3 Ensembling model Feature-Importance . . . 26
6.4 Key Observations & Cross-Model Comparison . . . 27

7 Limitations and Future Implications **29**

List of Tables

1 Four-way typology used for conceptual filtering 7
2 List of retained predictors after conceptual filtering 8
3 Engineered predictor set used for modelling 10
4 Binary variables 11
5 Ordinal variables 11
6 Continuous variables 11
7 Explained and cumulative variance by principal component, feature importance. 13
8 Over-fitting control measures for the victim classifier 14
9 Regularisation mechanisms for the perpetrator classifier 15
10 Interpretation of evaluation metrics 16
11 Confusion matrix for the victim-detection tree. 17
12 Confusion matrix for the perpetrator-detection network. 18
13 Confusion matrix of the ensembled victim–perpetrator classifier. 19
14 Performance metrics for victim, perpetrator, and overlap classifiers 21
15 Decision-tree raw feature weights 22
16 SHAP importance of principal components 24
17 SHAP-derived importance of engineered predictors 25
18 Ensemble importance 26
19 Extending cluster insights toward predictive applications. 30
20 Key formulas used in the project 31

List of Acronyms

CCP Cost-Complexity Pruning. 14, 15

CV Cross-Validation. 14

Min-Max Min-Max Scaler. 6, 12, 21

NN Neural Network. 15

PCA principal-components analysis. 6, 7

SHAP Shapley additive explanations. 23

List of Equations

1 Scale-level mean, median and variance 9
2 Min–Max rescaling of feature j 12
3 Mean-centering of rescaled feature 12
4 Sample covariance matrix $\mathbf{C}$ 12
5 Eigen-decomposition of $\mathbf{C}$ 12
6 Cost–complexity criterion $R_\alpha(T)$ 14
7 Forward-pass equations of the neural network 15
8 Definitions of evaluation metrics 16
9 Split rule in PCA space 21
10 Inequality after reversing PCA 21
11 Hyperplane in raw feature space 21
12 Normalised feature-importance weights 21
13 SHAP definition for PC k 23
14 Raw SHAP importance of PC k 23
15 Normalised importance of PC k 23
16 SHAP importance for original variable x_i 23

1 Introduction and Motivation

Interpersonal-violence research increasingly recognises that a single individual may act as both victim and perpetrator. Nevertheless, most quantitative studies model these outcomes separately, obscuring shared mechanisms and limiting the reach of integrated prevention efforts. *Grounding our analysis, all behavioural, environmental and psychosocial indicators are drawn from the most recent complete calendar year of observation available for each adolescent, ensuring a consistent temporal frame across victim and perpetrator roles.*

To address this challenge, we introduce a *dual-classifier framework* that decomposes the victim–perpetrator overlap into two balanced binary tasks. Our methodological pipeline comprises (i) a three-stage feature-preparation procedure that systematically selects, engineers and orthogonalises predictors; (ii) two lightweight, high-recall classifiers—a cost complexity pruned decision tree for victims and a compact feed-forward neural network for perpetrators; and (iii) an ensemble logic that labels an adolescent as a victim–perpetrator only when *both* base models concur.

The study contributes (i) a principled decomposition of a tri-label problem into two tractable binaries without sacrificing sensitivity; (ii) an empirical assessment of global predictors that simultaneously shape victim and perpetrator roles; and (iii) an end-to-end implementation suitable for evidence-based screening and triage.

The remainder of the article is organised as follows. Section 3 details the data-processing workflow, Section 3 describes model construction, and Section 4 reports results and implications.

2 Statistical Analysis

Study design

The present study adopts a *cross-sectional, classificatory, and non-causal* design. All indicators refer to the same 12-month reference window and are analysed contemporaneously. The aim is to detect statistical regularities that predict victimisation, perpetration, and their overlap, without attempting to infer temporal ordering or causal effects.

Classifier choice

Two complementary algorithms were selected in line with the study goals of high recall and interpretability: (i) a *cost-complexity pruned decision tree*, which offers transparent if–then rules while suppressing overfitting via α-regularisation; (ii) a compact *feed-forward neural network* with dropout regularisation, chosen for its ability to capture non-linear relations while keeping the architecture small enough for explanatory analysis. The logical AND ensemble of these two base models yields a recall-oriented yet parsimonious tool for detecting victim–perpetrator overlap.

Software environment

All analyses were conducted in `Python 3.12.0`. The main libraries were:

- `pandas 2.3.0, numpy==2.1.3` for data wrangling and tabulation;
- `scikit-learn 1.7.0` for preprocessing, PCA, decision trees, and evaluation metrics;
- `TensorFlow 2.19.0` with `Keras 3.11.0` for neural-network training under CPU/GPU/TPU strategies;
- `joblib 1.5.1, shap 0.48.0` for model persistence and feature-importance analysis.

Predictor inclusion criteria

Predictors entered the models only if they satisfied the following conditions: (i) *Global* in nature (demographic, psychosocial, lifestyle), excluding any items that directly defined victimisation or offending to prevent label leakage; (ii) free of extreme sparsity or redundancy; (iii) transformed to ensure monotonic alignment with risk (higher values always denoting greater vulnerability or propensity). After selection and engineering, a final set of 27 predictors was scaled via Min-Max normalisation, mean-centred, and orthogonalised by PCA.

3 Data Processing

Dataset. The raw corpus consists of a cross-sectional survey matrix $D \in \mathbb{R}^{n \times p}$ with $n = 4024$ adolescent respondents and $p = 347$ items spanning demographic, socio-cultural and behavioural domains. Listwise deletion of 6.4 % incomplete cases yields a working matrix of $n = 3767$ complete observations, each represented by a feature vector $\mathbf{x}_i$. Three binary outcome variables are derived: victimisation (y_V), offending (y_P) and their conjunction $y_{VP} = y_V \wedge y_P$. Owing to extreme imbalance in y_{VP} (719 positives vs. 3048 negatives), the subsequent modelling strategy decomposes the overlap into the two better-balanced binaries y_V (1861/1906) and y_P (885/2882).

Overview. The raw survey matrix undergoes a three-stage pipeline before any predictive modelling is attempted:

(i) *Feature Selection* prunes the original $p = 347$ items to a conceptually safe subset, eliminating target leakage, extreme sparsity and multicollinearity.

(ii) *Feature Engineering* recodes, aggregates and merges the selected items so that each respondent is described by a dense, risk-aligned descriptor vector.

(iii) *Feature Pre-processing* scales the engineered predictors and orthogonalises them via principal-components analysis (PCA), creating an input matrix that is dimension-reduced, collinearity-free and numerically comparable across variables.

Each subsection below details the operations and underlying rationale for its respective stage.

Throughout this section the raw feature vector for respondent i is denoted $\mathbf{x}_i \in \mathbb{R}^{347}$; all derived quantities are written as $\varphi(\mathbf{x}_i)$.

3.1 Feature Selection

3.1.1 Conceptual Filtering via Four-Way Typology

To avoid target leakage and multicollinearity, each survey item was assigned to exactly one of the mutually exclusive categories listed in Table 1. Only *Global* items were retained for modelling.

Table 1: Four-way typology used for conceptual filtering

Category	**Definition**	**Kept?**	**Rationale**
Global	Demographic, lifestyle and psychosocial variables	Yes	Contextual information untainted by the targets.
Perpetrator-defining	Items whose *yes* response directly implies offending	No	Would allow the model to read the label.
Victim-defining	Items that explicitly mark victimisation	No	Same leakage risk as above.
Derived targets	Composite/transformed outcomes (e.g. `VICTIMA_PERPETRADOR`)	No	Reserved exclusively for evaluation.

3.1.2 Retained Global Predictors (35 variables)

Table 2: List of retained predictors after conceptual filtering

Feature	Brief description	Type
PAÍS	Country of birth (1 = Spain, 2 = Other).	Global
ETNIA.BN	Binary ethnicity: European only vs. minority group.	Global
GENERO.BN	Binary gender: 0 = male, 1 = female.	Global
ORIENTSEX.BN	Binary orientation: hetero vs. non-hetero.	Global
EDAD	Age in years (13–18).	Global
FUGAS	Freq. of escape from home/centre.	Global
ABUSOSUBS1	Alcohol consumption freq.	Global
ABUSOSUBS2	Binge drinking ($\geq$5 drinks) freq.	Global
CONVIVEN1-7	Household co-residence flags (parents, step-parents, siblings, others, residential care).	Global
IMPULS1-8	Impulsivity scale items (planning, acting without thinking, etc.).	Global
AUTOEFIC1-5	Self-efficacy items (coping with difficulties).	Global
APOYO1-7	Social-support items (trusted adults, friends).	Global
MORAL1-5	Moral-emotion items (guilt, regret, shame, pride).	Global
PORNO.T	Combined frequency of pornography consumption in the last year (PORNO1+PORNO2).	Global

3.1.3 Why a Dual-Model Decomposition?

Extreme class skew (719:3048) in y_{VP} made a direct classifier impractical. Decomposing into two balanced binaries— $f_V : \mathbf{x} \mapsto y_V$ and $f_P : \mathbf{x} \mapsto y_P$—leverages higher base rates (49 %, 32 %) and isolates dual-role cases through a logical *AND*.

3.1.4 Analytical Uses of Excluded Items

Perpetrator- and victim-defining variables—as well as other items removed during screening—can still inform descriptive profiling. They are suitable for unsupervised methods such as k-means or hierarchical clustering to delineate behavioural subgroups among victims, perpetrators and dual-role adolescents.

3.2 Feature Engineering

Starting from the 35 *Global* items, we produced a compact vector $\varphi(\mathbf{x}_i) \in \mathbb{R}^{27}$ through the seven operations below. Each step is motivated by the need for (i) conceptual coherence —every larger value must denote *more risk*—and (ii) numerical comparability across predictors.

1. One-hot encoding `GENERO.BN` and `ORIENTSEX.BN` are expanded into two binary columns each to avoid spurious ordinality.

2. Age clipping $\text{EDAD}_{\text{norm}} = \min(\max(\text{EDAD}, 14), 17)$ confines outliers to the 14–17 window used in the questionnaire.

3. Co-residence merge `CONVIVEN.6` = `CONVIVEN6` $\vee$ `CONVIVEN7` pools rare residential-care flags so that every binary reflects a meaningfully sized subgroup.

4. Scale condensation For each multi-item construct—*Impulsivity* (8 items), *Self-efficacy* (5), *Social support* (7), *Moral emotions* (5)—we compute

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \tilde{x} = \text{median}\{x_i\}, \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2. \tag{1}$$

Items phrased protectively (e.g. `AUTOEFIC1-5`, `IMPULS1,4,5,6`) are reverse-coded via $x_i' = 1 + x_{\max} - x_i$ so that higher scores consistently signal higher risk. The result is twelve continuous predictors (μ, $\tilde{x}$ where $m \geq 7$, and σ^2 per scale). *(This drastically reduces dimensionality, mitigates multicollinearity, ensures conceptual coherence, captures both location and spread, and facilitates subsequent PCA/scaling.)*

5. Item aggregation `PORNO.T` = `PORNO1` + `PORNO2` records the maximum observed frequency of pornography exposure.

6. Informative missingness Unanswered substance-use questions are recoded as $\texttt{ABUSOSUBS}_j = 1$ (risk assumed present); all other remaining NaN values are set to 1.

7. Risk alignment & final check A final sweep verifies that every transformed variable is monotonically increasing with the hypothesised propensity to victimise or offend, yielding a coherent 27-element vector.

3.2.1 Engineered Global Predictors (27 variables)

Table 3: Engineered predictor set used for modelling

Feature	Brief description	Type
PAÍS	Country of birth (1 = Spain, 2 = Other).	Binary
ETNIA.BN	Ethnicity: majority vs. minority.	Binary
EDAD	Age, clipped to the 14–17 range.	Continuous
FUGAS.BN	Any runaway episode in the last 12 months.	Binary
ABUSOSUBS1	Alcohol use frequency (1–5; missing → 1).	Ordinal
ABUSOSUBS2	Binge-drinking frequency (1–5; missing → 1).	Ordinal
CONVIVEN.1-6	Six household co-residence flags (parents, step-parents, siblings, others, shared care, residential care).	Binary
IMPULS.MEAN	Mean impulsivity score (reverse-coded so higher = riskier).	Continuous
IMPULS.MEDIAN	Median impulsivity score.	Continuous
IMPULS.VAR	Within-scale variance of impulsivity responses.	Continuous
AUTOEFIC.MEAN	Mean self-efficacy (reverse-coded).	Continuous
AUTOEFIC.VAR	Variance of self-efficacy responses.	Continuous
APOYO.MEAN	Mean perceived social support (reverse-coded).	Continuous
APOYO.MEDIAN	Median social-support score.	Continuous
APOYO.VAR	Variance of social-support responses.	Continuous
MORAL.MEAN	Mean moral-emotion score (reverse-coded).	Continuous
MORAL.VAR	Variance of moral-emotion responses.	Continuous
PORNO.T	Maximum reported frequency of pornography consumption.	Ordinal
GENERO_BIN_0	Male indicator.	Binary
GENERO_BIN_1	Female indicator.	Binary
ORIENTSEX.BN_1	Heterosexual indicator.	Binary
ORIENTSEX.BN_2	Non-heterosexual indicator.	Binary

3.2.2 Descriptive tables

Table 4: Binary variables

Variable	Value	Count	Percent
PAÍS	1	3432	91.1%
ETNIA.BN	1	751	19.9%
FUGAS.BN	1	519	13.8%
CONVIVEN.1	1	3468	92.1%
CONVIVEN.2	1	2841	75.4%
CONVIVEN.3	1	296	7.9%
CONVIVEN.4	1	114	3.0%
CONVIVEN.5	1	2382	63.2%
CONVIVEN.6	1	354	9.4%
GENERO_BIN_0	1	1805	47.9%
GENERO_BIN_1	1	1962	52.1%
ORIENTSEX.BN_1	1	3264	86.6%
ORIENTSEX.BN_2	1	503	13.4%

Table 5: Ordinal variables

Variable	Value	Count	Percent
ABUSOSUBS1	0	1540	40.9%
	1	8	0.2%
	2	1183	31.4%
	3	668	17.7%
	4	347	9.2%
	5	21	0.6%
ABUSOSUBS2	0	754	20.0%
	1	1487	39.5%
	2	883	23.4%
	3	419	11.1%
	4	201	5.3%
	5	23	0.6%
PORNO.T	1	1791	47.5%
	2	727	19.3%
	3	362	9.6%
	4	482	12.8%
	5	405	10.8%

Table 6: Continuous variables

Variable	Mean	SD	Median	Min	Max
EDAD	15.43	1.04	15.00	14.0	17.00
IMPULS.MEAN	2.40	0.35	2.38	1.0	4.00
IMPULS.MEDIAN	2.37	0.53	2.50	1.0	4.00
IMPULS.VAR	0.82	0.43	0.73	0.0	2.25
AUTOEFIC.MEAN	2.93	0.58	3.00	1.0	4.00
AUTOEFIC.VAR	0.44	0.39	0.24	0.0	2.16
APOYO.MEAN	3.27	0.58	3.43	1.0	4.00
APOYO.MEDIAN	3.46	0.76	4.00	1.0	4.00
APOYO.VAR	0.61	0.55	0.49	0.0	2.20
MORAL.MEAN	3.96	0.77	4.00	1.0	5.00
MORAL.VAR	0.62	0.67	0.40	0.0	3.84

3.3 Feature Pre-processing

The engineered feature matrix of 27 variables was subjected to a two-step pipeline to ensure numerical comparability, eliminate redundancy, and concentrate predictive signal into a smaller set of uncorrelated factors. These transformations not only reduce computational cost—by lowering dimensionality and speeding up convergence—but also improve prediction quality by removing noise and multicollinearity.

3.3.1 Normalisation

Before any multivariate analysis, all inputs were placed on a common numerical scale via Min-Max rescaling followed by mean-centering:

$$x_{ij}^{\text{scaled}} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}, \tag{2}$$

$$\tilde{x}_{ij} = x_{ij}^{\text{scaled}} - \frac{1}{n}\sum_{i=1}^{n} x_{ij}^{\text{scaled}}, \tag{3}$$

By mapping every feature into $[0, 1]$ and centering at zero, we prevent any one variable's range from dominating the analysis, improve numerical stability, and accelerate optimization in downstream modelling.

3.3.2 Dimension Reduction via PCA

Even after normalisation, inter-feature correlations can inflate computational burden and obscure latent structures. We therefore performed Principal-Components Analysis (PCA):

$$\mathbf{C} = \frac{1}{n-1}\,\mathbf{X}_{\text{centered}}^{\top}\mathbf{X}_{\text{centered}}, \tag{4}$$

$$\mathbf{C}\,\mathbf{v}_k = \lambda_k\,\mathbf{v}_k, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{27} \geq 0, \tag{5}$$

$$z_{ik} = \mathbf{x}_i^{(\text{centered})} \cdot \mathbf{v}_k, \quad k = 1, \ldots, 27.$$

Here, $\mathbf{v}_k$ are orthonormal loadings and λ_k their associated variances. Projecting onto the first $K \ll 27$ components retains most variability in a compact form, removes multicollinearity, reduces model size for faster training, and often yields more robust predictions.

3.3.3 Explained Variance by Principal Component

Table 7: Explained and cumulative variance by principal component, feature importance.

PC	Expl. var.	Cumul. var.	Direct var.	Direct corr.	Inv. var.	Inv. corr.
PC1	0.2198	0.2198	GENERO_BIN_0	0.6612	GENERO_BIN_1	-0.6612
PC2	0.1103	0.3300	CONVIVEN.5	0.6375	ETNIA.BN	-0.2142
PC3	0.0903	0.4204	ORIENTSEX.BN_2	0.6174	ORIENTSEX.BN_1	-0.6174
PC4	0.0842	0.5046	CONVIVEN.5	0.6152	CONVIVEN.2	-0.2919
PC5	0.0679	0.5725	EDAD	0.4543	CONVIVEN.2	-0.3091
PC6	0.0576	0.6300	CONVIVEN.2	0.5891	CONVIVEN.3	-0.2501
PC7	0.0486	0.6786	FUGAS.BN	0.5380	EDAD	-0.5117
PC8	0.0402	0.7188	APOYO.MEDIAN	0.4604	EDAD	-0.3984
PC9	0.0337	0.7525	CONVIVEN.6	0.8586	CONVIVEN.1	-0.3753
PC10	0.0320	0.7845	PORNO.T	0.7411	FUGAS.BN	-0.3696
PC11	0.0289	0.8133	CONVIVEN.1	0.6759	CONVIVEN.3	-0.5137
PC12	0.0248	0.8382	ABUSOSUBS1	0.5635	PORNO.T	-0.4473
PC13	0.0227	0.8609	PAÍS	0.8077	ETNIA.BN	-0.3673
PC14	0.0213	0.8821	CONVIVEN.3	0.6095	PAÍS	-0.3597
PC15	0.0200	0.9021	APOYO.VAR	0.5633	IMPULS.MEDIAN	-0.3460
PC16	0.0176	0.9197	MORAL.MEAN	0.5787	IMPULS.MEDIAN	-0.4630
PC17	0.0165	0.9362	AUTOEFIC.MEAN	0.4594	AUTOEFIC.VAR	-0.4345
PC18	0.0142	0.9503	APOYO.VAR	0.5400	IMPULS.VAR	-0.4925
PC19	0.0135	0.9638	IMPULS.VAR	0.7165	AUTOEFIC.VAR	-0.2834
PC20	0.0097	0.9735	ABUSOSUBS2	0.7991	ABUSOSUBS1	-0.5722
PC21	0.0094	0.9829	CONVIVEN.4	0.9313	CONVIVEN.3	-0.2522
PC22	0.0079	0.9908	AUTOEFIC.MEAN	0.6398	IMPULS.VAR	-0.2323
PC23	0.0065	0.9973	MORAL.VAR	0.7459	AUTOEFIC.VAR	-0.2182
PC24	0.0016	0.9989	APOYO.MEAN	0.8379	APOYO.MEDIAN	-0.4792
PC25	0.0011	1.0000	IMPULS.MEAN	0.8527	IMPULS.MEDIAN	-0.5137
PC26	0.0000	1.0000	GENERO_BIN_0	0.7071	IMPULS.MEAN	-0.0000
PC27	0.0000	1.0000	ORIENTSEX.BN_2	0.7071	GENERO_BIN_0	-0.0005

4 Model Building

The modelling strategy pursued two complementary goals: (i) maximise sensitivity to both victimisation and offending (*high recall* for the positive class) and (ii) keep model complexity low enough for frontline interpretability.[1]

4.1 Victim Classifier

A cost-complexity–pruned decision tree was trained on the first $K = 18$ principal components (95% cumulative variance), yielding an extremely sparse, depth-1 model.

Table 8: Over-fitting control measures for the victim classifier

Phase	**Purpose**	**Mechanism**
Stratified 80/20 split	Preserve class balance between train and test	Unbiased hold-out estimate.
Depth sweep (1–20)	Diagnose variance via train vs. CV recall	Soft regularisation by depth.
Full growth & CCP	Generate candidate sub-trees	Removes spurious splits.
α-grid search	Optimise recall_1 on unseen data	Direct bias–variance tuning.
Artefact export	Persist final tree, rule, confusion template	Ensures reproducibility.

Cost-complexity pruning (CCP) minimises

$$R_\alpha(T) \;=\; R(T) + \alpha\,|T|, \tag{6}$$

where $R(T)$ is the empirical error, $|T|$ the number of leaves, and $\alpha \geq 0$ the complexity penalty. The grid search identified an optimal value

$$\alpha^* = 0.01495 \qquad (\textit{Mejor ccp\_alpha})$$

which retains a single split on PC_2 (threshold 0.54). The resulting rule achieves 91 % recall with maximal transparency.

[1] Details of over-fitting control procedures are provided in the supplementary material.

4.2 Perpetrator Classifier

A lightweight feed-forward neural network (NN) was trained on the first 22 principal components (≈ 99 % cumulative variance; $\approx 2\,000$ parameters). Let $\mathbf{z} \in \mathbb{R}^{22}$ denote the PCA input. The network computes

$$\begin{aligned} \mathbf{a}_1 &= \sigma_1\big(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1\big), && |\mathbf{a}_1| = 64, \\ \tilde{\mathbf{a}}_1 &= \mathrm{Drop}_{0.10}(\mathbf{a}_1), && \text{(training only)}, \\ \mathbf{a}_2 &= \sigma_2\big(\mathbf{W}_2\tilde{\mathbf{a}}_1 + \mathbf{b}_2\big), && |\mathbf{a}_2| = 8, \\ \hat{y} &= \sigma_3\big(\mathbf{w}_3^\top \mathbf{a}_2 + b_3\big), && \hat{y} \in [0,1], \end{aligned} \tag{7}$$

where σ_1 is the identity (linear), σ_2 and σ_3 are element-wise sigmoids, and Drop_p denotes dropout with keep-probability $1-p$. Equation (7) encapsulates the full forward pass referenced in Table 9.

Table 9: Regularisation mechanisms for the perpetrator classifier

Design choice	**Mechanism**
Class weighting (inverse prevalence)	Emphasises minority class during optimisation.
Dropout layers (10%, then 0%)	Prevents neuron co-adaptation.
Early stopping on validation recall	Halts training after five epochs without improvement.
Custom over-fit gap detector	Stops if train–val recall gap > 0.10.
Grid search (648 combos)	Finds simplest architecture meeting recall target.
Mixed-precision & TPU strategy	Accelerates exhaustive search.

4.3 Summary of Over-fitting Safeguards

- **Data protocol:** Stratified splits and untouched hold-out sets.
- **Structural regularization:** Depth-1 tree via CCP; compact neural net with dropout.
- **Validation-driven tuning:** Depth sweeps, α grids, early stopping, custom gap detector.
- **Recall-centric selection:** Hyper-parameters chosen solely for positive-class recall.
- **Post-hoc transparency:** Decision rule and network equations fully documented.

5 Model Performance

This section presents evaluation results for both base classifiers and the ensemble, emphasising metrics that prioritise recall. Table 10 gives concise guidance on how to read each metric in our violence-screening context.

Metric definitions

$$
\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\
\text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\
\text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\
F_1 &= 2\,\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \\
\text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\
\text{NPV} &= \frac{\text{TN}}{\text{TN} + \text{FN}}, \\
\text{Balanced Accuracy} &= \frac{\text{Recall} + \text{Specificity}}{2}.
\end{aligned}
\tag{8}
$$

Table 10: Interpretation of evaluation metrics

Metric	**Description**	**Utility in this study**
Accuracy	Overall fraction of correct predictions	Can be misleading under class imbalance; used only as a rough check.
Precision	Share of positive calls that are truly positive	Reflects investigation workload: lower precision → more false alarms frontline teams must review.
Recall	Share of true positives that are detected	*Primary goal*: missing a victim or perpetrator imposes high social cost.
F_1	Harmonic mean of precision and recall	Convenient single number for model ranking when recall and precision both matter.
Specificity	Share of true negatives correctly dismissed	Indicates how many low-risk youths are correctly left unflagged, balancing recall.
NPV	Share of negative calls that are truly negative	Confidence that a "safe" label indeed corresponds to low risk.
Balanced Accuracy	Mean of recall and specificity	Adjusts for imbalance; gauges overall discrimination while keeping recall central.

5.1 Victim Classifier Performance

Table 11 reproduces the 2×2 confusion matrix for the depth-1, cost-complexity–pruned decision tree. The associated evaluation metrics follow directly from Eqs. (8).

	Pred. Non-Victim	**Pred. Victim**
Actual Non-Victim	TN = 95	FP = 287
Actual Victim	FN = 33	TP = 339

Table 11: Confusion matrix for the victim-detection tree.

Metric values

$$\begin{aligned} \text{Accuracy} &= 0.576 \quad (57.6\%), \\ \text{Precision} &= 0.542 \quad (54.2\%), \\ \text{Recall} &= 0.911 \quad (91.1\%), \\ F_1 &= 0.678 \quad (67.8\%), \\ \text{Specificity} &= 0.249 \quad (24.9\%), \\ \text{NPV} &= 0.742 \quad (74.2\%). \end{aligned}$$

Interpretation

- *Recall (91.1%)*—our primary objective—shows that the tree misses fewer than one in ten true victims, satisfying the design goal of maximal sensitivity.
- *Precision (54.2%)* indicates that roughly half of the youth flagged as victims are confirmed positive, translating into a manageable follow-up workload for practitioners.
- *Specificity (24.9%)* is deliberately sacrificed: three quarters of non-victims are falsely flagged. This trade-off is acceptable in an early-warning context where the cost of a missed victim outweighs that of an unnecessary assessment.
- *Accuracy (57.6%)* is low because the minority class is privileged; under strong imbalance, accuracy is not a reliable quality signal.
- *Negative Predictive Value (74.2%)* offers reassurance that three out of four adolescents labelled "safe" truly are, an important comfort for front-line staff.
- F_1 *(67.8%)* summarises the precision–recall balance; it confirms that the model remains practical despite its aggressive recall tuning.

5.2 Perpetrator Classifier Performance

The feed-forward neural network was evaluated on the same hold-out set; its confusion matrix appears in Table 12. Metrics are computed analogously.

	Pred. Non-Perp.	**Pred. Perp.**
Actual Non-Perp.	TN = 245	FP = 476
Actual Perp.	FN = 22	TP = 199

Table 12: Confusion matrix for the perpetrator-detection network.

Metric values

$$\begin{aligned} \text{Accuracy} &= 0.471 \quad (47.1\%), \\ \text{Precision} &= 0.295 \quad (29.5\%), \\ \text{Recall} &= 0.900 \quad (90.0\%), \\ F_1 &= 0.445 \quad (44.5\%), \\ \text{Specificity} &= 0.340 \quad (34.0\%), \\ \text{NPV} &= 0.918 \quad (91.8\%). \end{aligned}$$

Interpretation

- *Recall (90.0%)* meets the same sensitivity threshold as the victim model, ensuring that nine in ten true perpetrators are detected.
- *Precision (29.5%)* is lower, meaning roughly one in three positive calls is correct. This reflects the greater behavioural volatility of the perpetrator class and is acceptable for preventive screening.
- *Specificity (34.0%)* and *accuracy (47.1%)* are again intentionally modest; the classifier is biased toward catching the hard-to-find positive cases.
- A notably high *NPV (91.8%)* implies strong confidence that those dismissed as non-perpetrators rarely offend, which is vital for triage safety.
- The reduced F_1 *(44.5%)* captures the imbalance between high recall and low precision; nevertheless, the model clears the pre-defined recall bar while keeping the false-alarm rate within operational limits.

5.3 Ensemble Model Overview

We construct an ensemble classifier that combines:

- A decision tree trained to detect *victims*, optimized for very high recall on the "victim" class.
- A feed-forward neural network trained to detect *perpetrators*, likewise optimized for high recall on the "perpetrator" class.

Each base model's target is well balanced within its own training set, allowing us to push thresholds aggressively for recall without collapsing to trivial predictions. The ensemble's final rule is:

$$\text{Predict "victim–perpetrator" (label 1)} \iff \big(\text{Decision tree predicts victim}\big) \wedge \big(\text{Neural network predicts perpetrator}\big).$$

This logical AND ensures that only those cases flagged by both high-recall models are called positive in the combined task.

We did not train a single model directly on the merged `victim_perpetrator` label because that target is extremely imbalanced in the raw data, leading to unstable training and poor recall if one simply optimized a one-step classifier. By splitting into two balanced subproblems and ensembling by conjunction, we retain high sensitivity on both dimensions.

5.4 Ensemble Classifier Performance

Table 13 shows the confusion matrix for the ensembled victim–perpetrator classifier.

	Predicted 0	**Predicted 1**	**Total Actual**
Actual 0	TN = 76	FP = 86	162
Actual 1	FN = 2	TP = 30	32
Total Pred.	78	116	194

Table 13: Confusion matrix of the ensembled victim–perpetrator classifier.

Using the definitions from Equations **??**–**??**, we compute:
The ensemble achieves Accuracy ≈ 0.5464, Precision ≈ 0.2586, Recall 0.9375, Specificity ≈ 0.4691, F_1-score ≈ 0.4056, False Pos. Rate ≈ 0.5309, False Neg. Rate 0.0625, and Balanced Accuracy ≈ 0.7033.

5.5 Discussion of Performance

Single–task models. The cost–complexity–pruned decision tree (*victim task*) and the compact feed-forward neural network (*perpetrator task*) both achieve recall above 90 % (91.1 % and 90.0 %, respectively). In absolute terms this means that, on the held-out sets, the tree retrieves 339 of 372 victims and the network captures 199 of 221 perpetrators. Although precision is moderate (54.2 % for victims, 29.5 % for perpetrators), the deliberate bias toward recall is justified for an early-warning tool: the marginal cost of investigating a false alarm is lower than the social cost of overlooking a genuine case of violence.

Overlap (victim–perpetrator) model. When the two high-recall classifiers are combined through a logical AND, the ensemble flags an adolescent as dual-role only if *both* base models fire. This seemingly severe rule still attains **93.8 % recall** on the dual-role group, missing just 2 of the 32 known victim–perpetrators (Table 13). Importantly, the ensemble eliminates a large share of the false positives produced by either single model in isolation: of 763 unique alarms emitted by the tree and the network, only 86 survive the conjunction yet prove to be non–dual-role (overall precision = 25.9 %). While far from perfect, this marks a substantive gain over the naïve baseline where every adolescent would be labelled positive (precision = 18.9 %).

Balanced view of errors. The ensemble's balanced accuracy (70.3 %) underscores that the method does more than "guess positive"; its specificity (46.9 %) is nearly double that of the victim tree and 12 points higher than the perpetrator net. Qualitative inspection of the 86 false positives shows that most were either victims-only or perpetrators-only—profiles that still merit preventive attention even if they do not satisfy the strict dual-role definition.

Practical implications. Taken together, the results support a two-tier deployment strategy:

(a) **Universal screening:** Use the individual victim and perpetrator models to cast a wide net, ensuring that almost no high-risk youth slips through.

(b) **Targeted triage:** Apply the ensemble rule to prioritise the subset most likely to occupy both roles, thereby concentrating resources on cases with the highest propensity for mutual violence escalation.

In resource-constrained settings, investigators may start with tier (b); in comprehensive programmes, tier (a) can serve as a front-line filter followed by contextual assessment. Either way, the dual-model architecture offers a transparent, data-driven compromise between maximal sensitivity and tolerable workload.

Table 14: Performance metrics for victim, perpetrator, and overlap classifiers

Metric	Victim_DT	Perpetrator_NN	Overlap_DT_NN
Accuracy	0.576	0.471	0.546
Precision	0.542	0.295	0.259
Recall	0.911	0.900	0.938
F1-score	0.678	0.445	0.406
Specificity	0.249	0.340	0.469
NPV	0.742	0.918	0.974
Balanced Accuracy	0.580	0.620	0.703

6 Discussion of Results

6.1 Decision Tree Feature Importance

For our pruned tree we have a single internal node:

$$\mathrm{PC}_2 \;\le\; t, \qquad t = 0.54. \tag{9}$$

6.1.1 How to Recover the Raw-Feature Weights

The decision tree splits on PC_2 with the threshold in Eq. 9. To express that rule as a linear inequality in the *original* (*un-scaled*) variables we proceed in three algebraic steps:

1. Undo the PCA projection. Replace PC_2 by its definition:

$$\mathbf{w}_2^\top\big(\mathbf{z} - \mu\big) \;\le\; t. \tag{10}$$

2. Undo the Min-Max scaling. Substitute $\mathbf{z} = \mathbf{S}(\mathbf{x} - \mathbf{d}^{\min})$ and regroup all constants:

$$\underbrace{\mathbf{w}_2^\top \mathbf{S}}_{\beta^\top}\mathbf{x} \;\le\; t + \mathbf{w}_2^\top\big(\mathbf{S}\mathbf{d}^{\min} + \mu\big) \;=\; c. \tag{11}$$

3. Normalise to interpret "importance". Transform the raw coefficients into unit-free weights:

$$w_j \;=\; \frac{|\beta_j|}{\sum_{k=1}^{27}|\beta_k|}, \qquad j = 1, \ldots, 27. \tag{12}$$

6.1.2 Complete Raw-Feature Coefficient Table

Table 15: Raw-space coefficients that reproduce the decision-tree rule $\mathrm{PC}_2 \leq t$. "Weight" is the normalised absolute value; "Effect" shows whether an increase in the variable moves a respondent towards (+) or away from (–) the victim class.

#	**Feature**	β_j	Weight	Effect
1	CONVIVEN.5	0.637478	0.229	+
2	CONVIVEN.2	0.546152	0.196	+
3	ETNIA.BN	-0.214185	0.077	–
4	ORIENTSEX.BN_1	0.203226	0.073	+
5	ORIENTSEX.BN_2	-0.203226	0.073	–
6	FUGAS.BN	-0.172597	0.062	–
7	CONVIVEN.1	0.156661	0.056	+
8	PAÍS	-0.130422	0.047	–
9	CONVIVEN.3	-0.122660	0.044	–
10	CONVIVEN.6	-0.106702	0.038	–
11	APOYO.MEDIAN	0.040205	0.014	+
12	APOYO.VAR	-0.045495	0.016	–
13	EDAD	-0.031849	0.011	–
14	ABUSOSUBS1	-0.017070	0.006	–
15	ABUSOSUBS2	-0.013711	0.005	–
16	CONVIVEN.4	-0.013151	0.005	–
17	AUTOEFIC.MEAN	0.018166	0.007	+
18	AUTOEFIC.VAR	-0.021052	0.008	–
19	APOYO.MEAN	0.034071	0.012	+
20	PORNO.T	-0.016569	0.006	–
21	IMPULS.MEDIAN	0.000463	0.000	+
22	IMPULS.MEAN	-0.000006	0.000	–
23	IMPULS.VAR	-0.005535	0.002	–
24	MORAL.MEAN	0.010998	0.004	+
25	MORAL.VAR	-0.008597	0.003	–
26	GENERO_BIN_0	0.008369	0.003	+
27	GENERO_BIN_1	-0.008369	0.003	–

6.2 Neural-Network Feature Importance

6.2.1 Methodology Applied

1. (SHAP) value computation for PCs. Let M be the number of input principal components (PCs) and $f(\mathbf{x})$ the network output for a sample $\mathbf{x}$. The SHAP value of component k for $\mathbf{x}$ is the Shapley value

$$\phi_k(\mathbf{x}) \;=\; \sum_{S\subseteq\{1,\dots,M\}\setminus\{k\}} \frac{|S|!\,(M-|S|-1)!}{M!}\Big[f_{S\cup\{k\}}\big(\mathbf{x}_{S\cup\{k\}}\big) - f_S\big(\mathbf{x}_S\big)\Big], \tag{13}$$

where $f_S\big(\mathbf{x}_S\big)$ is the prediction when only the features in subset S are present.

2. Aggregation and normalisation. For N observations, the global (mean-absolute) importance of PC k is

$$I_k \;=\; \frac{1}{N}\sum_{j=1}^{N}\big|\phi_k(\mathbf{x}^{(j)})\big|, \tag{14}$$

which we convert to a percentage of total importance via

$$\widehat{I}_k \;=\; 100\times\frac{I_k}{\sum_{\ell=1}^{M} I_\ell} \quad \big(\%\text{ of total}\big), \tag{15}$$

3. Propagation to original variables. Each engineered predictor x_i is a linear combination of PCs, $x_i = \sum_{k=1}^{M} w_{k,i}\,\mathrm{PC}_k$, where $W = (w_{k,i})$ is the loading matrix. We distribute the SHAP importance of every PC onto the original variables:

$$\psi_i \;=\; \sum_{k=1}^{M} w_{k,i}\,I_k, \qquad \Psi_i \;=\; 100\times\frac{|\psi_i|}{\sum_j|\psi_j|}, \tag{16}$$

where ψ_i is the signed contribution and Ψ_i its percentage share of the total.

6.2.2 Global Importance of Principal Components

Table 16: SHAP-based importance of each principal component in the perpetrator neural network.

Component	**Importance (%)**
PC2	18.69
PC4	14.06
PC5	10.01
PC7	9.27
PC1	8.73
PC6	8.60
PC16	5.65
PC10	5.46
PC21	5.26
PC17	4.02
PC9	3.54
PC18	2.68
PC22	1.53
PC11	1.41
PC3	0.44
PC20	0.39
PC14	0.15
PC19	0.08
PC18	0.02
PC15	0.01
PC13	0.002
PC12	0.001

6.2.3 Importance of Original Variables

Table 17: SHAP-propagated coefficients for the neural-network rule. "Weight" is the normalised absolute value; "Effect" indicates whether an increase in the variable pushes the prediction towards (+) or away from (–) the perpetrator class.

#	**Feature**	ψ_i	Weight	Effect
1	CONVIVEN.5	22.02	0.201	+
2	CONVIVEN.2	9.46	0.086	+
3	FUGAS.BN	9.02	0.082	+
4	ORIENTSEX.BN_2	-7.74	0.071	–
5	ORIENTSEX.BN_1	7.74	0.071	+
6	ABUSOSUBS2	7.34	0.067	+
7	APOYO.VAR	6.92	0.063	+
8	PORNO.T	6.35	0.058	+
9	GENERO_BIN_1	-4.45	0.041	–
10	GENERO_BIN_0	4.45	0.041	+
11	CONVIVEN.6	4.19	0.038	+
12	AUTOEFIC.MEAN	3.33	0.030	+
13	IMPULS.MEDIAN	-2.65	0.024	–
14	EDAD	-2.40	0.022	–
15	ETNIA.BN	1.98	0.018	+
16	IMPULS.MEAN	-1.71	0.016	–
17	CONVIVEN.3	-1.28	0.012	–
18	APOYO.MEAN	-1.10	0.010	–
19	AUTOEFIC.VAR	1.07	0.010	+
20	CONVIVEN.4	1.00	0.009	+
21	ABUSOSUBS1	0.83	0.008	+
22	APOYO.MEDIAN	0.81	0.007	+
23	MORAL.VAR	0.66	0.006	+
24	PAÍS	0.64	0.006	+
25	IMPULS.VAR	-0.57	0.005	–
26	CONVIVEN.1	-0.09	0.001	–
27	MORAL.MEAN	0.03	0.000	+

6.3 Ensembling model Feature-Importance

Table 18: Ensemble feature-importance breakdown with condensed column names.

	Feature	PW	Dir	w_DT	d_DT	w_NN	d_NN
1	CONVIVEN.5	0.262997	+	0.229	+	0.201	+
2	CONVIVEN.2	0.172477	+	0.196	+	0.086	+
3	ORIENTSEX.BN_2	0.088073	−	0.073	−	0.071	−
4	ORIENTSEX.BN_1	0.088073	+	0.073	+	0.071	+
5	ABUSOSUBS2	0.037920	+	0.005	−	0.067	+
6	ETNIA.BN	0.036086	−	0.077	−	0.018	+
7	CONVIVEN.3	0.034251	−	0.044	−	0.012	−
8	CONVIVEN.1	0.033639	+	0.056	+	0.001	−
9	PORNO.T	0.031804	+	0.006	−	0.058	+
10	APOYO.VAR	0.028746	+	0.016	−	0.063	+
11	GENERO_BIN_1	0.026911	−	0.003	−	0.041	−
12	GENERO_BIN_0	0.026911	+	0.003	+	0.041	+
13	PAÍS	0.025076	−	0.047	−	0.006	+
14	AUTOEFIC.MEAN	0.022630	+	0.007	+	0.030	+
15	EDAD	0.020183	−	0.011	−	0.022	−
16	IMPULS.MEDIAN	0.014679	−	0.000	+	0.024	−
17	APOYO.MEDIAN	0.012844	+	0.014	+	0.007	+
18	FUGAS.BN	0.012232	+	0.062	−	0.082	+
19	IMPULS.MEAN	0.009786	−	0.000	−	0.016	−
20	IMPULS.VAR	0.004281	−	0.002	−	0.005	−
21	MORAL.MEAN	0.002446	+	0.004	+	0.000	+
22	CONVIVEN.4	0.002446	+	0.005	−	0.009	+
23	MORAL.VAR	0.001835	+	0.003	−	0.006	+
24	AUTOEFIC.VAR	0.001223	+	0.008	−	0.010	+
25	APOYO.MEAN	0.001223	+	0.012	+	0.010	−
26	ABUSOSUBS1	0.001223	+	0.006	−	0.008	+
27	CONVIVEN.6	0.000000	+	0.038	−	0.038	+

To understand which predictors drive the ensemble decision, we merge the feature-importance weights from the victim decision tree and the perpetrator neural network according to the following process:

1. **Combine weights per feature.**

   - For each feature present in both models, compare their `proportional_effect` ("directly" vs. "inversely"):
     - If they agree, sum their `ponderated_weight` and retain that direction.
     - If they disagree, subtract the inverse weight from the direct weight; the resulting sign determines the net effect (non-negative → directly, negative → inversely).

2. **Normalize combined weights.**

   - Take the absolute value of each combined weight.
   - Divide by the sum of all absolute combined weights so that the new `ponderated_weight` values sum to 1.

6.4 Key Observations & Cross-Model Comparison

1. Overlap in the strongest signals. Both models converge on the same *household-structure* variables as the single biggest risk markers. `CONVIVEN.5` (shared care) and `CONVIVEN.2` (step-parent presence) jointly explain almost half of the decision-tree hyperplane weight (42.5%) and one-third of the neural-network SHAP mass (28.7%). This agreement suggests that instability in living arrangements plays a pivotal, role-independent part in violent involvement.

2. Breadth versus parsimony. The pruned tree needs only nine materially non-zero coefficients (all others carry $< 2\%$ weight), whereas the network distributes importance across $\sim$20 variables. Hence the tree offers *transparent sufficiency*: a high-risk victim profile can be read off a single inequality combining a handful of predictors. The neural net, by contrast, builds a *redundant safety net*: it tempers misclassification risk by letting several weaker factors (e.g. `APOYO.VAR`, `PORNO.T`, `AUTOEFIC.MEAN`) nudge the score when the dominant ones are ambiguous.

3. Directional consistency. Where the same feature appears in both models, the sign of its effect is identical. For example, higher `FUGAS.BN` and higher binge drinking (`ABUSOSUBS2`) increase the likelihood of being predicted positive in each task, while higher age (`EDAD`) and greater social support (`APOYO.MEAN`) pull the score down. The echo across tasks reinforces the substantive interpretation that runaway episodes, substance misuse and weak support networks jointly escalate both victimisation and offending risk.

4. Model-specific accents.

- **Decision tree (victim task).** Immigration status (`PAÍS`) and minority ethnicity (`ETNIA.BN`) carry negative weights, indicating that majority-group adolescents are more likely to cross the split threshold. The tree thus surfaces a *majority-victim paradox*: after controlling for the other covariates, belonging to the dominant group predicts victimisation rather than protection.
- **Neural network (perpetrator task).** Behavioural measures dominate beyond household factors: binge drinking, pornography exposure and self-efficacy variance register sizeable positive SHAP shares. This pattern implies that the offending pathway is less about static demographics and more about *volatile conduct and coping resources*.

5. Implications for practice. *Front-line screening* can rely on the tree's handful of questions to flag likely victims quickly. *Secondary assessment* benefits from the network's richer feature palette to tease out perpetrator risk in borderline cases. Because both models spotlight the same living-arrangement markers, intervention programmes should prioritise family-stability measures; additional modules on impulse control, substance use and digital consumption appear essential when the goal is to curb offending behaviour.

6. Complementarity in ensemble use. The logical-AND ensemble retains the tree's parsimony for victim detection while leveraging the network's nuanced behavioural cues for perpetrator identification. The near-identical ranking of the top two features ensures that the combined rule remains interpretable, yet the wider tail of network-specific variables raises ensemble recall to 93.8%—a gain that would be unattainable with either model alone.

7 Limitations and Future Implications

Although the present study demonstrates that a dual-classifier framework can achieve high sensitivity in identifying adolescents at risk of occupying both victim and perpetrator roles, several limitations must be acknowledged. First, the dataset is drawn from a single cross-sectional survey of school-attending adolescents. While relatively large in size, it may not fully represent young people outside the educational system or those living in different cultural and regional contexts, thereby limiting the generalisability of the findings. Second, the exclusive reliance on self-report data introduces inherent risks of bias. Responses to sensitive questions may be subject to under- or over-reporting, social desirability effects, and recall errors, which could distort the measured associations between global predictors and outcomes. A third limitation concerns the extreme imbalance of the victim–perpetrator category relative to other groups. Although the decomposition-and-ensemble strategy alleviates this challenge by redistributing the task into two better balanced binaries, residual imbalance remains and may compromise the stability and calibration of estimates. Finally, despite the application of safeguards such as cost-complexity pruning, cross-validation, regularisation, and early stopping, the exploration of extensive hyperparameter grids carries the risk of overfitting the models to the particularities of this dataset. These caveats should be borne in mind when interpreting the results and before transferring the proposed classifiers to real-world screening contexts.

Looking forward, several avenues for enrichment emerge. One important direction is the integration of the framework with educational and preventive policies. The parsimonious structure of the decision tree and the transparency of the neural network architecture render the approach well suited for embedding into school-based early warning systems. In such settings, the models could serve as initial filters to flag adolescents at elevated risk and connect them with psychosocial or family support resources, thereby improving the efficiency of preventive interventions. Another priority for future research is replicability across datasets. Independent validation using surveys from other countries, regions, or temporal cohorts is crucial to test the robustness and generalisability of the approach. Longitudinal data would also allow the exploration of causal pathways, clarifying how global predictors evolve over time and contribute to the dynamics of victimisation and offending.

Beyond these immediate extensions, the study can be further advanced by incorporating cluster-analytic modelling. Using the retained global predictors, unsupervised methods such as k-means or hierarchical clustering can uncover more homogeneous subgroups of victims and perpetrators. Victim clusters may correspond to different abuse-type profiles, while perpetrator clusters can reveal aggression pathways, both of which reduce dimensionality and enhance interpretability. Crucially, cluster membership can itself be used as an input for predictive modelling. Victim clusters may forecast perpetrator cluster affiliation and vice versa, making explicit the latent mechanisms that link the two roles. This cross-prediction framework provides a richer perspective on the overlap and highlights the contextual factors shared across roles.

Table 19: Extending cluster insights toward predictive applications.

Cluster asset	**Current insights**	**Future model leverage**
Victim clusters	Profiles of abuse types; dimensionality reduction; improved interpretability of heterogeneous victim groups	Development of targeted questionnaires; generation of cluster-informed risk scores for early screening
Perpetrator clusters	Profiles of aggression patterns; identification of balanced behavioural subsegments; improved characterisation of offending styles	Design of tailored prevention pathways; use of cluster membership as features in advanced predictive risk models
Cross-cluster interaction (Victim $\leftrightarrow$ Perpetrator)	Identification of trajectories of risk; recognition of shared contextual factors between roles	Production of multidimensional forecasts of dual involvement; integration of findings into intervention design and policy frameworks

An additional advantage of cluster-based modelling is the ability to generate probabilistic outputs that lend themselves to qualitative interpretation. By comparing agnostic prediction probabilities with known cluster assignments, researchers can identify borderline or mixed cases, which may represent transitional trajectories or novel subtypes of risk not captured by binary labels. Such cases broaden our understanding of heterogeneity in adolescent involvement in violence and point to new opportunities for tailored interventions.

Table 19 summarises how current insights from clustering can be extended into predictive leverage. Victim clusters, currently useful for profiling abuse types and reducing dimensionality, could in the future inform targeted questionnaires and cluster-based risk scores. Perpetrator clusters, which today help describe aggression patterns and balance sample segments, could support the design of tailored prevention pathways and serve as features in advanced risk models. Finally, cross-cluster interactions between victims and perpetrators, which highlight trajectories of risk and shared contextual factors, could be leveraged for multidimensional forecasting and integrated intervention design

In conclusion, while the present study offers a transparent and recall-oriented framework for identifying dual-role adolescents, it is constrained by issues of representativeness, self-report bias, class imbalance, and potential overfitting. Future research should not only seek validation across diverse settings, but also enrich the analytic strategy through cluster-based predictive modelling and qualitative interpretation of probabilistic outputs. Such extensions promise both theoretical advances in understanding the mechanisms of victim–perpetrator overlap and practical benefits for designing early screening tools and targeted interventions.

Table 20: Key formulas used in the project

Formula Name	LaTeX Expression	Use in Project
Min–Max Scaling	$x_{ij}^{\text{scaled}} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}$	Normalises features to [0,1] to ensure comparability across variables.
Mean-Centering	$\widetilde{x_{ij}} = x_{ij}^{\text{scaled}} - \frac{1}{n}\sum_i x_{ij}^{\text{scaled}}$	Centers features around zero to stabilise optimisation and PCA.
Scale Statistics	$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \tilde{x} = \text{median}\{x_i\}, \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2$	Condenses multi-item constructs into mean, median, and variance (e.g. impulsivity, self-efficacy).
Covariance Matrix	$C = \frac{1}{n-1} X_{\text{centered}}^{\top} X_{\text{centered}}$	Captures variance and correlation structure of scaled features for PCA.
Eigen-Decomposition	$Cv_k = \lambda_k v_k, \;\; \lambda_1 \geq \cdots \geq \lambda_{27} \geq 0$	Computes eigenvectors and eigenvalues for principal components.
PCA Projection	$z_{ik} = x_i^{(\text{centered})} \cdot v_k$	Projects feature vectors into orthogonal principal component space.
Cost–Complexity–Pruned	$R_\alpha(T) = R(T) + \alpha\lvert T\rvert$	Avoid Decision Tree overfitting.
Decision Tree Rule	$\text{PC}_2 \leq t, \;\; t = 0.54$	DT classification rule on PC2.
DT Raw-Space Inequality	$\beta^{\top} x \leq c, \qquad w_j = \frac{\lvert\beta_j\rvert}{\sum_{k=1}^{27}\lvert\beta_k\rvert}$	Transforms PC split into raw variables and normalises feature-importance weights.
Neural Network Layer	$a_1 = \sigma_1(W_1 z + b_1), \;\; \lvert a_1\rvert = 64$ $\widetilde{a_1} = \text{Drop}_{0.10}(a_1)$ $a_2 = \sigma_2(W_2\widetilde{a_1} + b_2), \;\; \lvert a_2\rvert = 8$ $\hat{y} = \sigma_3(w_3^{\top} a_2 + b_3), \;\; \hat{y} \in [0,1]$	Forward pass in the perpetrator NN.
Evaluation Metrics	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ $\text{Precision} = \frac{TP}{TP+FP}$ $\text{Recall} = \frac{TP}{TP+FN}$ $\text{Specificity} = \frac{TN}{TN+FP}$ $\text{NPV} = \frac{TN}{TN+FN}$ $\text{F1} = 2 \cdot \frac{\text{Prec}\cdot\text{Rec}}{\text{Prec}+\text{Rec}}$ $\text{Bal. Acc.} = \frac{\text{Recall}+\text{Specificity}}{2}$	Defines all performance metrics used for evaluation.
Ensemble Rule	Predict VP $\iff$ (DT predicts victim) $\wedge$ (NN predicts perpetrator)	Logical AND ensemble combining victim DT and perpetrator NN.
SHAP Values	$\varphi_k(x) = \sum_{S \subseteq M \setminus \{k\}} \frac{\lvert S\rvert!(M - \lvert S\rvert - 1)!}{M!} \cdot \left[f_{S\cup\{k\}}(x_{S\cup\{k\}}) - f_S(x_S)\right]$	Explains feature importance for neural network predictions.
Global Importance	$I_k = \frac{1}{N}\sum_{j=1}^{N}\lvert\varphi_k(x^{(j)})\rvert$	Aggregates SHAP values across all individuals to rank predictors.
Normalised Importance	$\widetilde{I_k} = 100 \cdot \frac{I_k}{\sum_l I_l}$	Expresses feature importance as percentage of total importance.
Feature Contribution	$\psi_i = \sum_k w_{k,i} I_k, \;\; \Psi_i = 100 \cdot \frac{\lvert\psi_i\rvert}{\sum_j\lvert\psi_j\rvert}$	Combines tree and NN weights to compute final feature contributions.