Prédiction de la première violence physique subie avant 18 ans

NOAH OBAMA LUC-LILIAN DARRYL

Superviseur : Pr NGUEFACK

Mai 2025

INTRODUCTION

La violence physique subie dès le plus jeune âge représente une atteinte grave aux droits fondamentaux et a des conséquences à long terme sur la santé mentale, émotionnelle et physique des victimes. Ce phénomène est particulièrement préoccupant dans les contextes où les violences domestiques sont banalisées ou peu rapportées. L'analyse prédictive fondée sur les données peut jouer un rôle crucial dans l'identification précoce des facteurs de risque, afin de mieux orienter les politiques de prévention.

Ce projet s'inscrit dans une perspective de data science appliquée aux questions sociales, avec pour ambition de soutenir la lutte contre la violence basée sur le genre par des méthodes statistiques et d'apprentissage automatique.

REVUE DE LA LITTÉRATURE

La violence sexuelle à l'encontre des femmes demeure un problème de santé publique majeur à l'échelle mondiale. Selon l'Organisation mondiale de la santé (OMS), environ 1 femme sur 3 dans le monde a subi des violences physiques ou sexuelles au cours de sa vie. Cette réalité est encore plus préoccupante dans certaines régions d'Afrique subsaharienne, où les normes socioculturelles patriarcales, la stigmatisation des survivantes et la faible application des lois renforcent le silence autour de ces violences.

Parmi les différentes formes de violences sexuelles, les violences perpétrées dans l'enfance ont des répercussions particulièrement graves sur le développement psychologique, émotionnel et social des survivantes. L'âge de la première agression sexuelle forcée (d126) est ainsi un indicateur clé dans l'étude des violences précoces. Les recherches montrent qu'être exposé tôt à la violence augmente le risque de revictimisation à l'âge adulte, y compris dans le cadre conjugal.

Plusieurs études ont mis en évidence le lien entre les violences sexuelles subies par un partenaire intime (comme d105h, d105i, d105k) et les antécédents de violence dans l'enfance. Les femmes ayant été contraintes d'avoir des rapports sexuels ou d'autres actes non désirés par leur partenaire présentent souvent un historique de violences passées. Les violences sexuelles récentes par des auteurs autres que le partenaire (d124) ou sous des formes variées (d125) sont aussi liées à des antécédents de victimisation.

Enfin, la violence sexuelle exercée par un ex-partenaire (d130b) illustre la persistance du risque au-delà de la relation actuelle. Ces différentes dimensions, bien que distinctes, s'entrecroisent souvent dans la trajectoire des survivantes, rendant indispensable une approche analytique intégrée pour mieux comprendre les facteurs prédictifs d'une première violence avant l'âge de 18 ans.

OBJECTIF ET QUESTION DE RECHERCHE

L'objectif principal est de construire un modèle de classification permettant de prédire si une femme a subi sa première violence physique avant l'âge de 18 ans, en utilisant des caractéristiques individuelles et familiales issues des enquêtes DHS. L'analyse permettra également de déterminer quels sont les facteurs les plus fortement associés à ce phénomène.

Question : Quels facteurs liés à d'autres formes de violences sexuelles sont significativement associés à une première agression sexuelle précoce chez les femmes ?

MÉTHODOLOGIE

Description du jeu de données

a) Source : Les données sont extraites du module de violence domestique de l'enquête DHS Cameroun (CMIR71FL).

b) Cible : La variable cible de cette étude vise à prédire si une femme a subi sa première violence physique avant l'âge de 18 ans. Cette information n'était pas directement disponible dans les données brutes ; elle a donc été créée à partir de la variable existante indiquant l'âge au moment de la première violence physique (`age_at_first_violence`).

Nous avons procédé comme suit :

— Si `age_at_first_violence` $<$ 18 $\rightarrow$ la variable cible prend la valeur 1 (violence précoce)

— Si `age_at_first_violence` $\geq$ 18 $\rightarrow$ la variable cible prend la valeur 0 (pas de violence précoce)

Ainsi, la variable binaire obtenue prend les significations suivantes :

— 1 = la femme a subi une première violence physique avant 18 ans

— 0 = la femme a subi une première violence physique à 18 ans ou plus

Variables explicatives :

— `d105h` : Viol par mari/partenaire

— `d105i` : Forcée à d'autres actes sexuels par mari/partenaire

— `d105k` : Forcée à faire des actes sexuels qu'elle ne voulait pas (par le partenaire)

— `d124` : Viol par une autre personne (dans les 12 derniers mois)

— `d125` : Forcée à réaliser des actes sexuels non désirés

— `d130b` : Forcée par un ex-partenaire à avoir des rapports ou actes sexuels

Processus d'analyse

a) Prétraitement :

— Nettoyage des données (suppression des valeurs manquantes)

— Séparation du jeu de données en un ensemble d'entraînement (80 %) et un ensemble de test (20 %)

— Standardisation des variables continues si nécessaire

— Entraînement de plusieurs modèles de classification supervisée

b) Modèles évalués :

— Régression Logistique

— Random Forest

— Gradient Boosting

— SVM (Support Vector Machine)

— k-Nearest Neighbors (KNN)

— Naive Bayes

c) Évaluation :

— Accuracy (exactitude)

— Précision, rappel et F1-score pour la classe minoritaire (target = 1)

— Matrice de confusion

RÉSULTATS

La régression logistique est le modèle qui offre les meilleures performances globales, tout en permettant une interprétation directe des coefficients.

Les arbres (Random Forest, Gradient Boosting) ont montré des performances proches, mais légèrement inférieures.

Le modèle Naive Bayes surclassifie les cas de violence précoce, au prix d'un très grand nombre de faux positifs.

Modèle	**Accuracy**	**Précision (classe 1)**	**Rappel (classe 1)**	**F1-score (classe**
Régression Logistique	**0.73**	**0.67**	**0.70**	**0.69**
Random Forest	0.69	0.61	0.70	0.65
Gradient Boosting	0.70	0.62	0.70	0.66
SVM	0.70	0.62	0.69	0.65
KNN	0.68	0.59	0.74	0.66
Naive Bayes	0.46	0.43	0.93	0.59

DISCUSSION

L'objectif de cette étude était de construire un modèle fiable pour prédire la probabilité qu'une femme ait subi une première violence sexuelle avant l'âge de 18 ans. Plusieurs modèles de classification supervisée ont été comparés afin d'identifier celui offrant la meilleure performance en termes de prédiction.

a) Meilleur compromis : la Régression Logistique Parmi tous les modèles testés, la régression logistique s'est révélée être le modèle le plus performant. Elle affiche une accuracy de 73

b) Modèles d'ensemble : Random Forest et Gradient Boosting Les modèles Random Forest et Gradient Boosting ont montré des performances légèrement inférieures. Bien qu'ils atteignent des rappels similaires (0.70), leurs précisions sont un peu plus faibles (0.61 et 0.62 respectivement), ce qui suggère qu'ils ont tendance à produire plus de faux positifs. Leur F1-score reste correct (0.65–0.66), ce qui en fait des alternatives valables, mais sans avantage net par rapport à la régression logistique dans ce contexte.

c) SVM et KNN : performances stables mais en retrait Les modèles SVM et KNN offrent une performance intermédiaire, avec une précision de 0.62 pour SVM et 0.59 pour KNN. Le rappel élevé de KNN (0.74) suggère une forte capacité à détecter les cas positifs, mais au prix d'une précision plus faible, ce qui peut poser problème si l'on souhaite limiter les fausses alertes. Le F1-score de 0.66 confirme qu'il s'agit d'un bon modèle de détection, mais moins équilibré que la régression logistique.

d) Naive Bayes : un rappel très élevé, mais une précision trop faible Le modèle Naive Bayes se démarque par un rappel exceptionnel (0.93), ce qui signifie qu'il détecte presque tous les cas positifs. Cependant, sa précision très basse (0.43) entraîne une forte proportion de faux positifs, ce qui rend ce modèle peu fiable en pratique. Il pourrait être utilisé dans un cadre de pré-dépistage où l'on privilégie la sensibilité, mais il nécessiterait un second filtre pour réduire les erreurs.

LIMITES DE L'ÉTUDE

Qualité des données auto-déclarées : Les données proviennent d'enquêtes déclaratives (DHS), ce qui expose les réponses à un biais de mémoire ou de désirabilité sociale, notamment pour des sujets aussi sensibles que la violence sexuelle.

Variable cible construite : La variable cible «violence sexuelle avant 18 ans» a dû être créée à partir de plusieurs variables indirectes, ce qui peut introduire une certaine subjectivité ou incertitude sur sa précision.

Portée limitée des variables explicatives : L'étude s'est concentrée uniquement sur les variables relatives à différentes formes de violence sexuelle, sans intégrer d'autres facteurs contextuels (niveau d'instruction, statut socioéconomique, environnement familial, etc.) qui pourraient enrichir les prédictions.

Modèles non optimisés : Les modèles ont été utilisés avec leurs paramètres par défaut, sans phase de tuning approfondie (optimisation d'hyperparamètres) qui pourrait améliorer leur performance.

RECOMMANDATIONS

Sur la base des observations précédentes, plusieurs recommandations peuvent être formulées :

Amélioration de la qualité de la variable cible : Utiliser une définition plus rigoureuse et validée de la variable cible pour mieux identifier les victimes de violence sexuelle avant 18 ans.

Élargissement du jeu de données : Intégrer des variables sociodémographiques, comportementales ou environnementales pour mieux comprendre les facteurs associés aux violences précoces.

Tuning des modèles : Optimiser les modèles à l'aide de techniques de Grid Search ou Randomized Search afin d'en améliorer la performance et la robustesse.

Analyse de sensibilité et validation croisée : Mettre en œuvre une validation croisée k-fold et des analyses de sensibilité pour garantir la stabilité et la généralisation des résultats.

Communication éthique des résultats : Dans toute diffusion des résultats, veiller à respecter les principes éthiques, en évitant toute stigmatisation et en valorisant une utilisation responsable des prédictions.

CONCLUSION

Cette étude visait à prédire la probabilité qu'une femme ait subi une première violence sexuelle avant l'âge de 18 ans à partir de données issues des enquêtes DHS. Grâce à la construction d'une variable cible et à l'utilisation de plusieurs algorithmes de classification, les résultats ont montré que la régression logistique constitue le modèle le plus performant, alliant précision, rappel et F1-score équilibrés.

Les variables décrivant différentes formes de violence sexuelle ont permis de dégager des tendances utiles pour la détection précoce. Toutefois, des limites subsistent, notamment liées à la qualité des données et à la portée des variables incluses.

Ce travail ouvre la voie à des recherches plus approfondies intégrant une diversité de facteurs explicatifs, des approches de modélisation avancées et une perspective éthique forte, en vue de mieux comprendre, prévenir et répondre à la violence sexuelle subie dès le plus jeune âge.