International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 39 No. 1 Mar. 2023, pp. 173-183

http://www.ijias.issr-journals.org/

Modèle ensembliste pour la prédiction des maladies cardiaques dans des milieux insécurisés: Cas de la Province du Nord-Kivu, RD Congo

[Ensemble model for predicting heart diseases in insecure areas: The case of North-Kivu Province, DR Congo]

Zawadi Sirisombola Corinne, Héritier Nsenge Mpia, and Julien Kabuyahia

Département d'Informatique de Gestion, Université de l'Assomption au Congo, B.P. 104, Butembo, Nord-Kivu, RD Congo

ABSTRACT: Heart diseases are the leading cause of death worldwide. Nowadays, heart diseases are killing more people than ever before. Thus, the authors of this study designed this project to analyze data on heart diseases prediction. The project uses raw data in the form of a.csv file as data set. The authors collected the used dataset from the cardiology department of the Graben University Clinic (Butembo/DR Congo) that included 389 records and 25 variables including age, employment, pulse rate, blood pressure and clinical symptoms. The aim was to compare Machine Learning (ML) ensemble methods such as Boosting type (AdaBoosting, GradientBoosting and XGBoosting) with single ML models (KNN, Stochastic gradient descent (SGD), Decision Tree) to see which of the models predict better heart diseases in unstable and insecure areas. Thus, the results showed that the XGBoost model performed better with accuracy, precision and recall of 85% respectively. In this research the authors concluded that Boosting as ensemble method classifies accurately heart diseases data in an insecure area such as Butembo, in the province of North-Kivu, DR Congo.

KEYWORDS: Ensemble methods, Boosting, Machine Learning, Heart disease, Insecure zone, Butembo.

RESUME: Les maladies cardiaques sont restées la principale cause de décès au niveau mondial. Cependant, elles tuent maintenant plus que jamais auparavant. C'est ainsi que les auteurs de cette étude ont conçu ce projet d'analyse des données sur la prédiction des maladies cardiaques. Le projet utilise des données brutes sous la forme d'un fichier.csv comme jeu des données. Les auteurs de cette étude ont recueilli cet ensemble de données au sein du service de cardiologie de la clinique universitaire du graben (Butembo/RD Congo) qui comprenait 389 enregistrements, après nettoyage des données, et 25 variables dont l'âge, l'emploi, le pouls, la pression artérielle et les symptômes cliniques. Le but a été de comparer les modèles ensemblistes Machine Learning de type Boosting (AdaBoosting, GradientBoosting et XGBoosting) avec ceux non ensemblistes (KNN, SGD, Arbre de décision) afin de voir lequel des modèles prédit mieux les maladies cardiaques dans des régions d'insécurités, telle que la Ville de Butembo, en RD Congo. Ainsi, les résultats ont montré que l'algorithme XGBoost a obtenu la meilleure performance de classement avec accuracy, precision et recall de 85% pour toutes ces mesures respectives. Dans cette recherche les auteurs ont montré que Boosting comme modèle d'apprentissage de type ensembliste pouvait surmonter le problème de classification d'un ensemble de données sur les maladies cardiaques dans une zone insécurisée comme Butembo.

MOTS-CLEFS: Méthodes ensemblistes, Boosting, Machine Learning, Maladies cardiaques, Zone insécurisée, Butembo.

1 INTRODUCTION

Les maladies cardiovasculaires sont une cause majeure d'incapacité et de décès prématurés dans le monde entier [1]. Il est fréquent que les événements coronariens aigus (crises cardiaques) et les accidents vasculaires cérébraux (AVC) surviennent brutalement et provoquent la mort de la victime avant qu'on ait pu lui prodiguer des soins. En agissant sur les facteurs de risque, il est possible de réduire la fréquence des événements cliniques et la mortalité prématurée chez les personnes présentant déjà une pathologie cardiovasculaire établie, ainsi que chez celles dont le risque cardiovasculaire est majoré par la présence d'un ou plusieurs facteurs de risque supplémentaires [1]. Il est possible de prévenir la plupart des maladies cardiovasculaires en s'attaquant aux facteurs de risque comportementaux comme, le tabagisme, la mauvaise alimentation et l'obésité, la sédentarité et l'utilisation nocive de l'alcool à l'aide de stratégies à l'échelle de la population.

Il a été constaté en 2015 que, bien que les urgences cardiovasculaires fussent les plus fréquentes en Afrique Sub-saharienne, les cas d'urgences cardiovasculaires sont peu nombreux à Kinshasa, en RD Congo [2]. Par contre, toujours en RD Congo, dans des zones fortement insécurisées, le taux des maladies cardiovasculaires est en augmentation exponentielle. Plusieurs facteurs justifient cela. Dans ces régions d'instabilité, environ 57.5% des personnes souffrant des maladies cardiovasculaires étaient ignorant de leurs maladies et seulement 13.6% avaient un suivi médical approprié [3]. En fait, la situation de conflits armés permanents dans ces régions serait à la base de plusieurs conséquences aussi bien économique, physique que physiologique dont le trauma et les hypertensions [4]. Malheureusement, peu de recherches sont menées dans cette ligne pour pallier ce type des maladies. Soulignons qu'en Occident, il existe déjà des technologies pouvant surveiller à domicile en transmettant des données électrocardiogrammes (ECG) du patient, ainsi que des informations sur la tension artérielle et l'oxymétrie du pouls par téléphone ou même sans fil afin de faciliter titration ambulatoire des médicaments anti arythmiques [5], en Afrique Sub-saharienne, spécialement en RD Congo, il y a un manque de technologies capables d'un tel suivi médical. Plus encore, dans les régions de l'Est de la RD Congo où les crises sociopolitiques ne cessent de faire naitre des conséquences sanitaires et psychologiques, il se constate un manque total des recherches pouvant prédire et même prévenir ces maladies [6]. Sur ce, l'analyse de données ainsi qu'une approche d'apprentissage automatique parait être dans cette recherche nécessaire pour identifier les causes des maladies cardiaques, en utilisant la méthode d'ensemble Machine Learning (ML) [7].

De ce fait, la prédiction des maladies cardiovasculaires est ainsi considérée comme l'un des sujets les plus importants dans la section de l'analyse des données cliniques. Cela étant, le tableau ci-dessous recapitule les questions de recherche et les objectifs spécifiques poursuivis dans cette étude.

Tableau 1. Mapping des objectifs spécifiques avec les questions de recherche

Objectifs spécifiques	Questions de recherche
Identifier les facteurs qui prédisent les risques des maladies cardiaques en régions d'insécurités	Quels sont les facteurs les plus récurrents qui prédisent les maladies cardiaques en Ville de Butembo?
Comparer les résultats des modèles individuels ML et celui de la méthode d'ensemble Boosting	Quel algorithme ML peut-on retenir afin de prédire avec exactitude et précision les maladies cardiaques en Ville de Butembo, dans la province du Nord-Kivu?
Prédire les maladies cardiaques dans la Province du Nord-Kivu en utilisant la méthode d'ensemble ML le plus performant	Le modèle d'ensemble peut-il prédire mieux les facteurs qui influencent les maladies cardiaques en Ville de Butembo?

2 REVUE DE LITTERATURE

2.1 REVUE DE LITTERATURE THEORIQUE

2.1.1 INTELLIGENCE ARTIFICIELLE

L'intelligence Artificielle (IA) est la science dont le but est de concevoir des systèmes capables de reproduire le comportement de l'humain dans ses activités de raisonnement [8]. Des technologies comme le ML et le traitement automatique du langage naturel (NLP) font partie de l'IA [9].

2.1.2 MACHINE LEARNING

ML également connu sous l'appellation de l'apprentissage automatique, est l'une des sous-catégories de l'IA qui permet à des systèmes de construire des modèles à partir de données, et d'améliorer leurs performances ultérieures en conséquence [9]. Par définition, ML est une science moderne qui permet d'établir des prévisions à partir des motifs récurrents identifiés dans les flux de données [9].

Dans le domaine médical, le ML impacte de nombreux domaines et l'on y voit émerger de nombreux usages. Il n'est pas une technologie nouvelle mais elle prend une tout autre ampleur avec l'émergence de l'IA. L'objectif n'est pas de remplacer le médecin par la machine, mais de l'accompagner dans l'analyse et l'interprétation d'énormes volumes de données biométriques collectées [10]. Il existe plusieurs techniques ML, à savoir: l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage par renforcement. A cela s'ajoute la méthode d'ensemble. Cette dernière comprend à son tour quelques techniques très populaires, telles que Bagging, Boosting et stacking [11]. Sur ce, cette étude s'est focalisée sur la technique boosting.

2.1.3 METHODE D'ENSEMBLE BOOSTING

L'idée de l'apprentissage d'ensemble est d'employer plusieurs apprenants et de combiner leurs prédictions [12]. Par définition, la méthode d'ensemble boosting, initialement nommé Hypothesis Boosting, consiste à filtrer ou à pondérer les données utilisées pour former une équipe d'apprenants faibles, de sorte que chaque nouvel apprenant donne plus de poids ou ne soit formé qu'avec des observations qui ont été mal classées par les apprenants précédents. Ainsi, pour atteindre les objectifs de cette étude, les auteurs ne se sont focalisés que sur les algorithmes de Boosting les plus célèbres : Ada Boost (Adaptive Boosting), Gradient Boosting, XGBoost (Xtreme Gradient Boosting) [12].

ADABOOST

AdaBoost est adaptif dans le sens où les apprenants faibles suivants sont modifiés en faveur des instances mal classées par les classificateurs précédents [13]. Aussi, les weak learners (apprenants faibles) d'AdaBoost sont généralement des arbres décisionnels à seulement deux branches et deux feuilles (aussi appelés souches) [14].

GRADIENTBOOST

La particularité de Gradient Boosting est qu'il essaye de prédire à chaque étape non pas les données elles-mêmes mais les résidus. Ainsi, les seconds weak learners sont ensuite multipliés par un facteur inférieur à 1. L'idée derrière cette multiplication est que plusieurs petits pas sont plus précis que quelques grands pas. La multiplication réduit donc la taille des pas pour augmenter la précision. Pour demander la prédiction de gradient boosting sur une observation, il suffit d'interroger chaque weak learner. La prédiction du strong learner sera donc la somme de toutes les réponses des weak learner [14].

XGBOOST

XGBoost est une bibliothèque distribuée optimisée de boosting de gradient conçu pour être hautement efficace, flexible et portable [15]. A d'autres termes, XGBoost est en fait une version particulière de l'algorithme de Gradient Boost. Sa particularité réside dans le type de weak learner utilisé. Les weak learners sont des arbres décisionnels. Les arbres qui ne sont pas assez bons sont élagués. Cette méthode est appelée le pruning en anglais ou élagage en français [15].

2.1.4 MALADIE CARDIAQUE

Une maladie cardiaque désigne plusieurs types d'affections cardiaques causées par une accumulation de plaque dans les vaisseaux sanguins, ou artères. L'accumulation de cette plaque collante peut provoquer une maladie coronarienne, des douleurs thoraciques, des crises cardiaques et des AVCs [16].

2.2 REVUE DE LITTERATURE EMPIRIQUE

Tout travail scientifique étant une complémentarité, les auteurs de cette étude confirment que chaque chercheur ayant abordé le sujet similaire au leur, a tenté de proposer des solutions pour pallier les risques des maladies cardiaques. Asma Baccouche et al., par exemple dans leur article intitulé « Ensemble Deep Learning Models for Heart Disease Classification: A

Case Study from Mexico », affirment que les maladies cardiaques sont classées parmi les principales causes de mortalité dans le monde [17]. Selon ces chercheurs, la distribution des dossiers recueillis était très déséquilibrée sur les différents types de maladies cardiaques, où 17 % des dossiers des patients avaient une maladie cardiaque hypertensive, 16 % des dossiers avaient une cardiopathie ischémique, 7 % des dossiers avaient une maladie cardiaque mixte et 8 % des dossiers avaient une maladie cardiaque valvulaire. Sur ce, ces auteurs avaient proposé un cadre d'apprentissage d'ensemble de différents modèles de réseaux neuronaux et une méthode d'agrégation du sous-échantillonnage aléatoire [17]. Ainsi, ils avaient mené des expériences avec des modèles de réseaux neuronaux unidirectionnels et bidirectionnels et les résultats avaient montré qu'un classificateur d'ensemble avec un BiLSTM ou BiGRU avec un modèle CNN avait la meilleure performance de classification avec précision et F1-score entre 0.91 et 0.96 respectivement pour les différents types de maladies cardiaques [17]. Ces résultats sont compétitifs et prometteurs pour l'ensemble de données sur les maladies cardiaques. Pour ce faire, ils avaient montré qu'un cadre d'apprentissage d'ensemble basé sur des modèles profonds pouvait surmonter le problème de la classification d'un ensemble de données sur les maladies cardiaques déséquilibrées [17].

Dans la même perspective, Nico Dragan et al., dans leur recherche portant sur "Effort–Reward Imbalanceat Work and Incident Coronary Heart Disease", ont insisté sur le fait que les preuves épidémiologiques du stress au travail en tant que facteur de risque de maladie coronarienne sont principalement basées sur une seule mesure du stress au travail, appelée job strain [18]. Sur ce, parmi les 90164 participants inclus dans leur recherche, l'âge moyen à l'entrée dans l'étude était de 45 ans, avec 60.8% des femmes et 39.2% des hommes. La proportion d'individus présentant un déséquilibre effort-récompense variait entre 8% et 51% selon les études. Ce qui pouvait refléter à la fois des différences réelles de prévalence et des différences dans les mesures du déséquilibre effort-récompense malgré l'harmonisation. Cette dernière était de 31.7 % de la population totale [18].

Sims et al. dans l'article « Importance of Housing and Cardiovascular Health and Well-Being » soulignent que les disparités en matière de maladies cardiovasculaires sont façonnées par les différences de facteurs de risque entre les groupes raciaux et ethniques. Sur ce, ils affirment que le logement demeure un déterminant social important de la santé. L'objectif de cette déclaration était d'examiner et de résumer la recherche qui a examiné les associations entre l'état de logement et la santé cardiovasculaire et la santé globale [19]. Selon eux, les efforts visant à éliminer les disparités en matière de maladies cardiovasculaires ont récemment mis l'accent sur l'importance des déterminants sociaux de la santé. Dans ce sens, ils définissent le logement comme étant un élément important déterminant social de la santé et du bien-être cardiovasculaires et qui devrait être pris en compte dans l'évaluation des efforts de prévention visant à réduire et à éliminer les disparités raciales/ethniques et socioéconomiques [19]. Partant des études antérieures, certains auteurs susmentionnés se sont plus focalisés sur la caractéristique sociale, les auteurs de la présente étude ont compris que le facteur social n'est pas le seul critère à prendre en considération dans la prédiction des maladies cardiaques mais qu'il faut aussi tenir compte des facteurs cliniques. Comme: échographie cardiaque, tension artérielle, difficulté respiratoire, perte d'appétit, pouls, ictère. Cette présente étude menée dans une zone insécurisée, Butembo, a révélé également que Age, Emploi, échographie cardiaque, tension artérielle, difficulté respiratoire, échographie cardiaque générale, perte d'appétit, pouls, ictère ainsi que la zone où réside un patient se sont avérés être les facteurs les plus importants dans la prédiction des maladies cardiaques dans une zone comme Butembo.

2.3 CADRE CONCEPTUEL

Partant des recherches antérieures, les auteurs de cette étude ont maintenu trois types de facteurs comme prédisant les maladies cardiaques en Ville de Butembo, à savoir les facteurs démographiques, psycho-sociaux et cliniques.

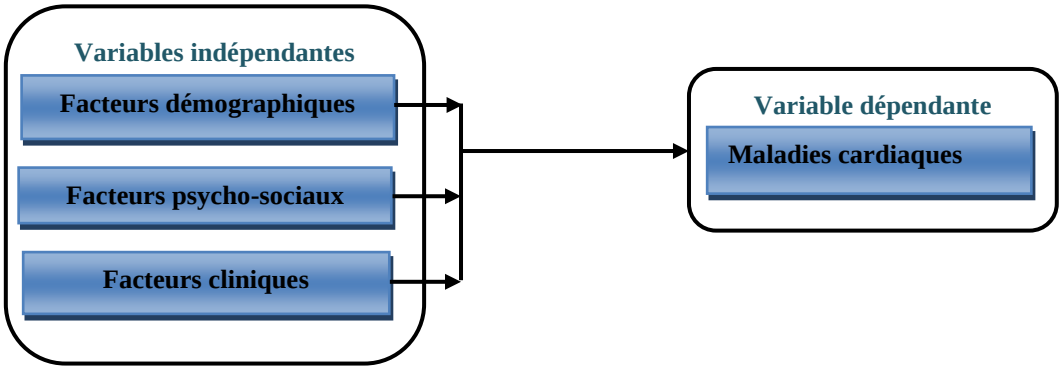

Fig. 1. Cadre conceptuel graphique de la recherche

3 METHODOLOGIE

Le concept *Méthodologie* se rapporte à un ensemble d'étapes permettant de chercher, d'identifier et de trouver des documents relatifs à un sujet par l'élaboration d'une stratégie de recherche [20]. Autrement dit, la méthodologie est une procédure logique en science. C'est-à-dire l'ensemble des pratiques particulières qu'elle met en œuvre pour que le cheminement de ses démonstrations et de ses théorisations soit clair, évident et irréfutable [21].

3.1 CONCEPTION DE LA RECHERCHE

Pour mener à bien cette recherche, les auteurs se sont servis de la méthode quantitative. Celle-ci leur a permis d'obtenir des données secondaires et concrètes, principalement sous forme numérique auprès de la clinique universitaire du Graben, en Ville de Butembo [21]. Sur ce, ils ont proposé le processus ci-après qui les a aidés à recueillir, à nettoyer, à analyser les données, et à implémenter le modèle prédictif.

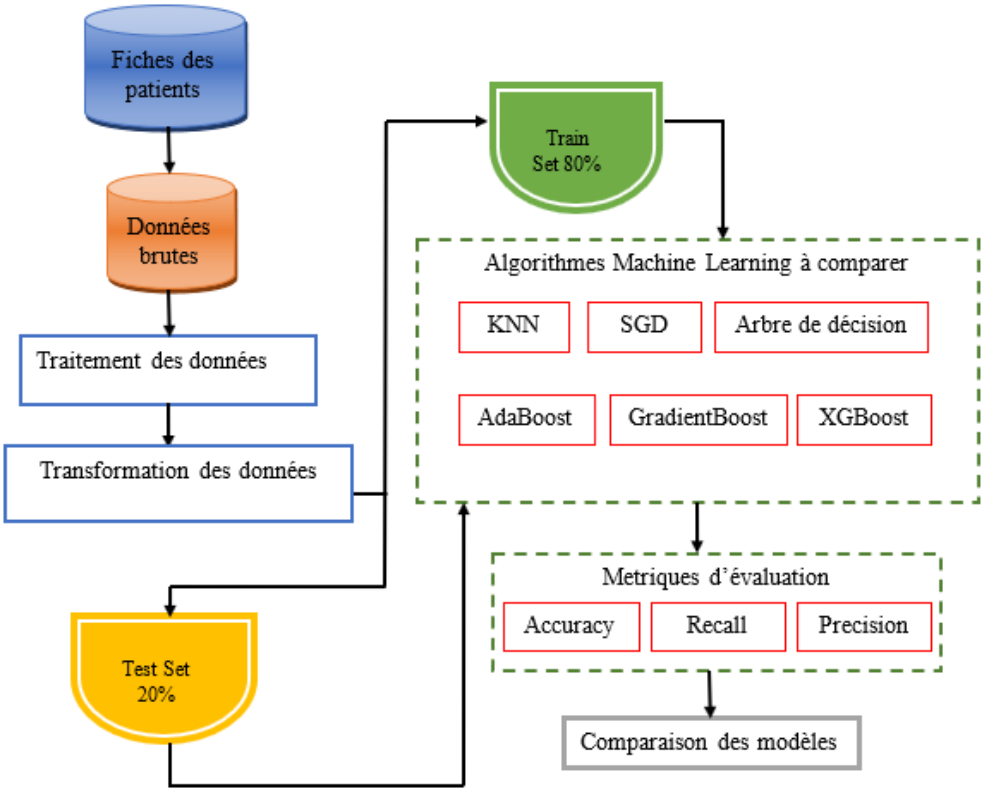

Fig. 2. Processus de mise en place du projet

Comme indiqué sur la Figure 2, ci-haut, les auteurs ont tout d'abord organisé les données recueillies sur les fiches de différents patients en données brutes en un fichier.csv. Ensuite, ils ont procédé au prétraitement et à l'analyse des données brutes. Aussi, pour que ces données soient alimentées aux différents modèles ML, les auteurs ont fait la transformation (encodage) de celles-ci et ainsi les ont subdivisées en deux parties, à savoir le train set et le test set. Le train set est la base d'apprentissage qui leur a permis de réaliser la phase d'entraînement et le test Set a été réservé au test ou à l'évaluation des modèles.

Algorithmes Machine Learning à comparer comme phase a consisté à construire différents modèles ML de type boosting et ceux de type individuel. C'est-à-dire AdaBoost, GradientBoost et XGBoost pour le modèles ensemblistes Boosting et KNN, Stochastic gradient descent (SGD), et l'arbre de décision comme modèles individuels. Ajoutons que, pour mesurer la performance de ces différents modèles, les auteurs ont utilisé l'accuracy, le Recall et la Précision comme mesures d'évaluation afin d'effectuer un jugement de performance sur ces algorithmes. L'accuracy indique le pourcentage de bonnes prédictions. Il s'agit simplement du rapport entre le nombre d'observations correctement prédites et le nombre total d'observations [22].

3.2 POPULATION CIBLE

Dans le cadre d'un projet de recherche, il est indispensable de préciser une population cible qui fera l'objet de l'intervention [23]. Cette dernière représente le total des unités de collecte au sujet desquelles les résultats seront applicables. La population cible dans cette recherche a été composée des patients du service de cardiologie des Cliniques Universitaires du Graben/Butembo, connu sous le nom de « Horizon ».

3.3 PROCEDURE DE COLLECTE DES DONNEES

Les auteurs ont fait usage des données secondaires, en recueillant les données à partir des fiches des consultations des différents patients du service de cardiologie Horizon/Butembo. De ce fait, pour atteindre les objectifs de cette recherche, les auteurs ont jugé mieux de délimiter la période à laquelle ils comptaient récolter ces données. Cette période partait de Janvier 2021 à Décembre 2021. Ainsi, ils avaient recueilli 389 données prélevées à partir des fiches des patients du service de cardiologie des Cliniques Universitaires du Graben/Butembo. Le tableau ci-dessous illustre les différentes variables retenues après le feature selection et la façon dont leurs valeurs catégorielles ont été encodé pour les rendre adéquates dans la phase de ML:

Tableau 2. Les features utilisés comme prédicteurs et leur encodage

Feature	Détail sur le feature	Format d'encodage
Age	L'âge du patient	Les âges des patients sont des valeurs numériques
Emploi	Emploi du patient	Cultivateur=0, Couturier=1, Ménagère=2, Agent de l'Etat=3, Enseignant=4, Comptable=5, Commerçant=6, Elève=7, Ecolière=7, Enfant=8, Morgateur=9, Autre=10, Religieux=11, Etudiant=7, Ingénieur=12, Agronome=13, Chauffeur=14, Hôtelier=15, Médecin=16, Caissier=17, Convoilleur=18, Déclarant=19, Menuisier=20, Magasinier=21, Humanitaire=22, Vétérinaire=23, Vendeur ambulant=24, Electricien=25, Coiffeur=26, Commissionnaire=27, Inspecteur=28, Taximen=29
ECG	Résultats des données électrocardiogrammes du patient	Normal=0, Anormal=1
EC	Echographie cardiaque générale	Normale=0, Anormale=1
TA	Tension artérielle du patient	Normale=0, Anormale=1
Pouls	Pulsation du flux sanguin du patient ressentie en palpant son artère	Normale=0, Anormale=1
DR	Le patient éprouve ou non la difficulté respiratoire	Oui=0, Non=1
PA	Le patient a perdu ou l'appétit	Oui=0, Non=1
Ictère	Le patient présente une encéphalopathie hépatique ou non	Oui=0, Non=1
Zone	La zone d'habitation du patient	Sécurisée=0, Moyennement sécurisée=1, Insécurisée=2

3.4 OUTILS DE TRAITEMENT ET D'ANALYSE DES DONNEES UTILISES

Les auteurs ont utilisé les bibliothèques importantes de python pour effectuer le traitement des données mais aussi développer les modèles ML. Parmi ces différentes bibliothèques et librairies il y a matplotlib, seaborn, skeatlearn et tant d'autres.

4 ANALYSE DES DONNEES, RESULTATS ET DISCUSSIONS

4.1 STATISTIQUES DESCRIPTIVES

L'analyse descriptive (AD) est une perspective pour comprendre les données et leurs différents modèles existants. Fondamentalement, elle fait partie des quatre types de concepts d'analyse de données [24]. L'analyse descriptive tend principalement vers l'apprentissage non supervisé pour décrire, classer et extraire des informations afin d'obtenir des réponses sur ce qui s'est passé dans le passé. Elle aide également les gens à comprendre correctement certains scénarios et leurs résultats [25]. De ce fait, ci-dessous on peut voir les différentes représentations graphiques illustrant les statistiques des certaines variables ayant fait objet de cette recherche.

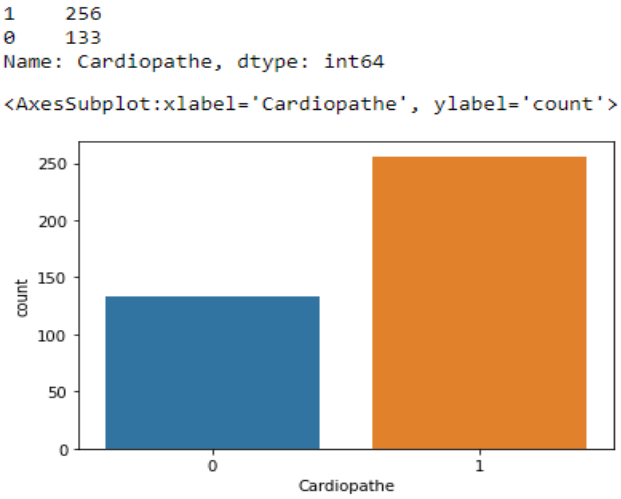

Fig. 3. Etat en cardiopathie à Butembo

La figure ci-dessus indique qu'environ 34.19%, soit 133, des patients dans le dataset utilisé ont été atteints d'une maladie cardiaque et 65.81%, soit 256, des patients ne le sont pas.

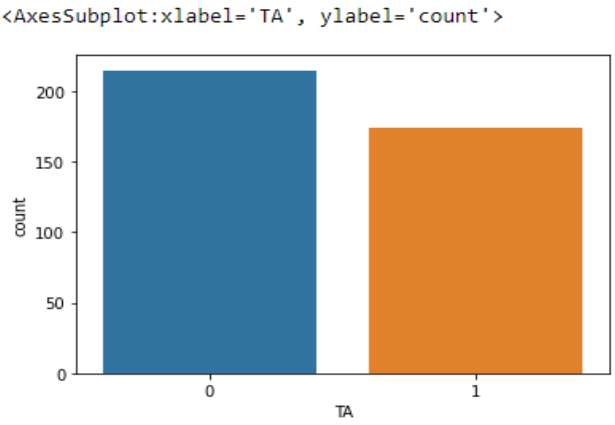

Fig. 4. Cardiopathie par zone

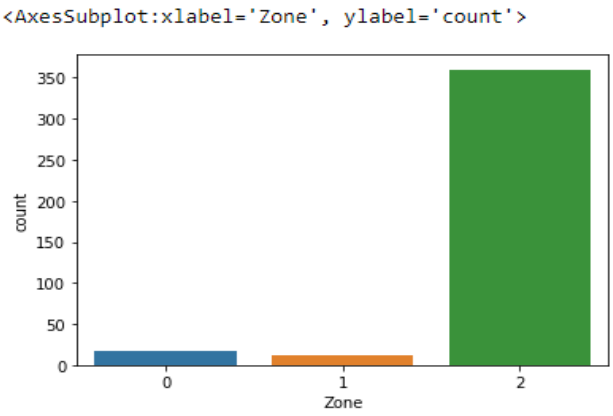

Fig. 5. Fréquence de la TA

La figure 4, indique que statistiquement 92.54% de la population habitant dans la zone insécurisée de Butembo se figurant dans le dataset étaient plus vulnérables aux différents types de maladies cardiaques comparativement aux personnes œuvrant dans des zones moyennement sécurisées et sécurisées. Par contre, la figure 5 indique qu'environ 55.27% de patients ayant une tension artérielle (TA) normale avaient plus de chance de contracter une maladie cardiaque; ce qui ne pas le cas pour les patients présentant une tension artérielle anormale.

4.2 RESULTAT DE LA RECHERCHE

Ce point consiste à présenter les résultats obtenus en se référents aux trois différents objectifs assignés dans cette recherche.

PRESULTATS DU PREMIER OBJECTIF SPECIFIQUE DE L'ETUDE

Rappelons que notre le premier objectif de cette étude a consisté à identifier les facteurs qui prédisent les risques des maladies cardiaques. Pour ce faire, les auteurs ont appliqué la technique SelectKBest de la librairie feature selection pour identifier les features qui prédisent mieux les maladies cardiaques partant des 25 variables du dataset collecté [25]. Sur ce, pour garantir une meilleure performance pour les modèles ML développés, les résultats de feature selection a révélé que l'âge du patient, le type d'emploi du patient, les résultats de l'électrocardiogramme, les résultats de l'échographie cardiaque, la tension artérielle du patient, les résultats de l'électrocardiogramme, pouls, difficulté à respirer, perte d'appétit, ictère et la zone d'habitation du patient (sécurisée ou insécurisée) sont des variables qui prédisent le mieux les maladies cardiaques chez les patients vivant en Ville de Butembo.

PRESULTATS DU DEUXIEME OBJECTIF SPECIFIQUE DE L'ETUDE

Pour prédire les maladies cardiaques dans la Province du Nord-Kivu dans la ville de Butembo, les auteurs ont appliqué un des algorithmes de boosting, le XGBoost. Le code ci-dessous illustre l'exemple des données d'un patient X dont on a prédit s'il souffre ou non d'une maladie cardiaque. AgeF représente l'âge du patient qui est 78 ans, Emploi égal à 0 signifie que ce patient est un cultivateur. ECGF fait référence aux données électrocardiogrammes du patient qui sont 0, c'est-à-dire le test ECG est normal, Letère (Ictère) de 0 indique ce patient souffre d'encéphalopathie hépatique. ECF est le résultat de l'échographie générale du patient qui est à 0 signifiant que c'est normal. TAF égale à 1 signifie que la tension artérielle de ce patient est anormale. PoulsF de 0 indique que la pulsation du flux sanguin de ce patient est normale, tandis que le DRF de 0 signifie que ce patient a des difficultés respiratoires. 0 comme valeur de PAF fait référence à la perte d'appétit. ZoneF égale à 0 indique que ce patient X habite dans une zone sécurisée. Précisons que les encodages des données sont illustrés dans le tableau 2.

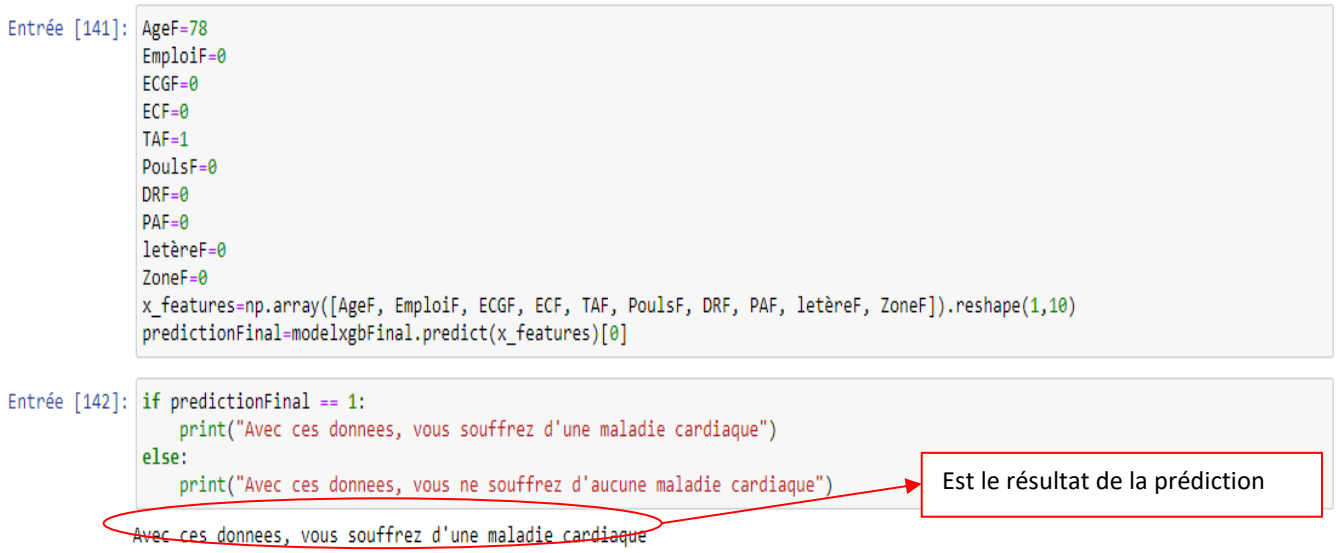

Fig. 6. Résultat de la prédiction de maladie cardiaque pour le Patient X

Le code python se trouvant dans le premier bloc (cellule 141 du notebook de Jupyter) de la figure 4, a permis de définir les entrées relatives aux données d'un patient X comme illustration du fonctionnement du modèle dans la phase de son utilisation. Dans le deuxième bloc (cellule 142), se trouve le code qui définit deux conditions qui permettent de savoir si ce patient X est ou non sujet d'une maladie cardiaque grâce aux données de test introduites dans le premier bloc. Sur ce, le modèle XGBoost (modelxgbFinal) a prédit que ce patient soufre d'une maladie cardiaque.

PRESULTATS DU TROISIEME OBJECTIF SPECIFIQUE DE L'ETUDE

Eu égard à ce qui précède, les méthodes ensemblistes sont sans doute les meilleures techniques de prédiction qui puissent exister dans le monde des Big Data. De ce fait, voici ci-dessous le tableau synthèse qui visualise les différents résultats de performance des modèles individuels ML et ceux des méthodes ensemblistes.

Tableau 3. Affichage des mesures de performance de chaque modèle ML développé

	Modèle	Accuracy (%)	Recall	Precision
1	Ada Boosting	80.0	0.80	0.80
2	Gradient boosting	82.0	0.82	0.82
3	XGBoosting	85.0	0.85	0.85
4	KNN	56.0	0.56	0.56
5	SGD	77.0	0.77	0.77
6	Arbre de décision	79.0	0.79	0.79

Lorsqu'il faut prendre une décision importante, il vaut souvent mieux recueillir plusieurs avis que de se fier à un seul. Cela dit, le tableau 4 ci-dessus indique que les méthodes ensemblistes ont été meilleures que les modèles non ensemblistes ML (KNN, SDG, Arbre de décision) dans la prédiction des maladies cardiaques en Ville de Butembo, u Nord-Kivu. Sur ce, le modèle le plus performant est sans doute le XGBoosting car en tenant compte des mesures de performance retenues dans cette recherche (Accuray, Recall et Precision), ce modèle a présenté des meilleurs scores. Par conséquent, les auteurs l'ont validé comme modèle adéquat et performant dans le cas des prédictions des telles maladies en des telles zones.

5 CONCLUSIONS ET RECOMMANDATIONS

Rappelons que cette étude a utilisé des données brutes sous la forme d'un fichier.csv. Ces données ont été utiles dans l'analyse pour la prédiction des maladies cardiaques à l'aide de la méthode ensembliste Boosting et de l'analyse de données en code python. Les maladies cardiaques sont l'une des principales causes de mortalité à Butembo. C'est ainsi que les auteurs

se sont posés les questions principales suivantes pour réaliser cette étude: Quels sont les facteurs les plus récurrents qui prédisent les maladies cardiaques en ville de Butembo? Le modèle ensembliste peut-il prédire mieux les facteurs qui influencent les maladies cardiaques en Ville de Butembo, dans la province du Nord-Kivu ? Quel algorithme ML peut-on retenir afin de prédire avec exactitude et précision les maladies cardiaques en Ville de Butembo, dans la Province du Nord-Kivu ?

Il est à reconnaitre qu'il est difficile d'identifier les maladies cardiaques en raison de plusieurs facteurs de risque tels que les yeux hémorragiques, l'hypertension artérielle, douleur articulaire, un pouls anormal et de nombreux autres facteurs. En raison de ces facteurs, les scientifiques se sont tournés vers des approches modernes telles que l'exploration de données et l'apprentissage automatique pour prédire la maladie. C'est ainsi que les auteurs ont pensé qu'une analyse basée sur les techniques ensemblistes Boosting et non ensemblistes ML serait une meilleure solution susceptible de servir d'outil d'analyse des données permettant aux cardiologues d'établir un bon choix lorsqu'il s'agit des maladies cardiaques et ainsi, offrir à la population de la Ville de Butembo la chance de se mettre à l'abri de ce fléau sanitaire.

Trois modèles ensemblistes Boosting et trois autres non ensemblistes, à savoir respectivement: AdaBoosting, GradientBoosting, XGBoosting, KNN Classifier, SGD et arbre de décision ont été développé dans cette recherche. Le modèle ensembliste XGBoosting a prédit le mieux avec une performance de 85% pour toutes les métriques d'évaluation utilisées dans cette étude. De ce fait, les objectifs fixés à l'introduction ont été atteints et cette recherche a répondu non seulement aux attentes des cardiologues mais aussi à celles des patients locaux. Certes, plusieurs chercheurs, tels qu'indiqués dans la littérature empirique, ont déjà mené des investigations sur les maladies cardiaques dans différents domaines scientifiques, à l'instar de l'informatique, la médecine, etc. Malheureusement, aucun d'entre eux n'a fait allusion à un contexte similaire au nôtre: la prédiction des maladies cardiaques dans des zones insécurisées, à l'occurrence en Ville de Butembo, une région de la RD Congo marquée par des insécurités incessantes. Ceci constitue l'un des points de démarcation de cette étude avec celles antérieures en ce sens que les variables retenues comme prédicteurs ont été d'un contexte socio-sécuritaire.

Contrairement aux travaux antérieurs, pour la construction du modèle d'ensemble ML, les auteurs ont fait recours à la méthode boosting dont le rôle est de pondérer les données utilisées pour former une équipe d'apprenants faibles, de sorte que chaque nouvel apprenant donne plus de poids ou ne soit formé qu'avec des observations qui ont été mal classées par les apprenants précédents [12]. D'où la motivation du choix de cette méthode d'ensemble qui a permis aux auteurs d'obtenir un ensemble des modèles parfaitement complémentaires dont les faiblesses des uns ont été compensées par les forces des autres et cela a donné la chance d'obtenir un meilleur score de performance. Afin de déterminer la performance de cette méthode d'ensemble, les auteurs ont procédé par sa comparaison avec des modèles non ensemblistes ML. Ce qui a amené à une conclusion selon laquelle les modèles ensemblistes donnent des très bonnes performances que ceux non ensemblistes pris séparément [26].

Eu égard à ce qui précède, les auteurs recommandent aux cardiologues de la Ville de Butembo de prendre ce système comme outil d'aide à l'identification certaine des facteurs clés liés à la possibilité d'être atteint d'une maladie cardiaque. Ainsi les cardiologues auront des idées éclairées et probables concernant leurs patients grâce à la prédiction que donnera ledit système. Aussi au ministère de la santé de la RD Congo de mettre à sa disponibilité un système national qui serait à mesure de prédire les maladies cardiaques dans le but de faciliter la tâche aux cardiologues dans l'exercice de leur fonction. En plus, aux futurs chercheurs, les auteurs recommandent d'améliorer la performance de ce système en tenant compte de l'aspect d'overfitting et celui d'underfitting et ainsi le déployer en mobile si possible.

REFERENCES

[1] Mohamed M.S. et al. La mort subite de l'adulte, particularités en Afrique, à propos de 476 cas. *The Pan African Medical Journal,* 16, 2013, pp. 16-125. https://doi.org/10.11604/pamj.2013.16.125.2490.

[2] Mboliasa I. et al. Profil épidémiologique et clinique des urgences cardiovasculaires admises aux soins intensifs de médecine interne des Cliniques Universitaires de Kinshasa. *Ann. Afr. Med*., 8 (2), 2015, pp. 1933-1938.

[3] Katchunga P.B. et al. Hypertension artérielle chez l'adulte Congolais du Sud Kivu: résultats de l'étude Vitaraa. Presse Med, 40 (6), 2011, pp. 315-323.

[4] Tankink Marian T.A., Slegh H. Living Peace in Democratic Republic of Congo: An Impact Evaluation of an Intervention with Male Partners of Women Survivors of Conflict-Related Rape and Intimate Partner Violence, Promundo, Washington, 2017.

[5] Fahey T. et Schroeder K. Cardiology. *British Journal of General Practice,* 54 (506), 2004, pp. 695-702.

[6] International Crisis Group. RD Congo: En finir avec la violence cyclique en Ituri. *Rapport Afrique de Crisis Group,* 292, Brussels, 2020.

[7] Salhi D.E. et al. Using Machine Learning for Heart Disease Prediction. *Advances in Computing Systems and Applications,* 199, 2021, pp.70-81. https://doi.org/10.1007/978-3-030-69418-0_7.

[8] Majnarić L.T., Babič F., O'Sullivan S., Holzinger A. AI and Big Data in Healthcare: Towards a More Comprehensive Research Framework for Multimorbidity. *Journal of Clinical Medicine,* 2021, 10 (4), 766. https://doi.org/10.3390/jcm10040766.

[9] Singh B. et al. A Trade-off between ML and DL Techniques in Natural Language Processing. *International Conference on Robotics and Artificial Intelligence (RoAI),* 1831, 2021, pp.1-7. https://doi.org/10.1088/1742-6596/1831/1/012025.

[10] Li B, Jiang L, Lin D, Dong J. Registered Clinical Trials for Artificial Intelligence in Lung Disease: A Scoping Review on ClinicalTrials.gov. *Diagnostics,* 2022, 12 (12), 3046. https://doi.org/10.3390/diagnostics12123046.

[11] Zhang Y., Liu J., Shen W. A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Applied Sciences,* 2022, 12 (17), 8654. https://doi.org/10.3390/app12178654.

[12] Ferreira, A.J., Figueiredo, M.A.T. Boosting Algorithms: A Review of Methods, Theory, and Applications. Zhang, C., Ma, Y. (eds) Ensemble Machine Learning, 2012, Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-9326-7_2.

[13] Ding Y., Zhu H., Chen R., Li R. An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification. *Applied Sciences,* 2022, 12 (12), 5872. https://doi.org/10.3390/app12125872.

[14] Natras R., Soja B., Schmidt M. Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. *Remote Sensing,* 2022, 14 (15), 3547. https://doi.org/10.3390/rs14153547.

[15] Li W. et al. Gene Expression Value Prediction Based on XGBoost Algorithm. *Frontiers in Genetics,* 2019. https://doi.org/10.3389/fgene.2019.01077.

[16] Frąk W. et al. Pathophysiology of Cardiovascular Diseases: New Insights into Molecular Mechanisms of Atherosclerosis, Arterial Hypertension, and Coronary Artery Disease. *Biomedicines,* 2022, 10 (8), 1938. https://doi.org/10.3390/biomedicines10081938.

[17] Baccouche A. et al. Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico. I*nformation,* 11 (207), 2020, pp.1-28. https://doi.org/10.3390/info110402007.

[18] Gragao N. et al. Effort-Reward Imbalance at Work and Incident Coronary Heart Disease. *Epidemiology,* 28 (4), 2017, pp. 619-626.

[19] Sims M. et al. Importance of Housing and Cardiovascular Health and Well-Being. *Circ Cardiovasc Qual Outcomes,* 13 (8), 2020, pp. 596-605. https://doi.org/10.1161/HCQ/0000000000000089.

[20] Yvonne G., Alain J. Pourquoi je préfère la recherche quantitative/Pourquoi je préfère la recherche qualitative ? *Revue internationale P.M.E,* 2016, 29, pp.7-17.

[21] Denise T. Méthodes qualitatives, Quantitatives et Mixtes, L'approche Delphi: Application dans la conception d'un outil clinique en réadaptation au travail en santé mentale, Presses de l'université du Québec, Paris, 2014.

[22] Maxwell A.E., Warner T.A., Guillén L.A. Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 1: Literature Review. *Remote Sensing,* 2021, 13 (13): 2450. https://doi.org/10.3390/rs13132450.

[23] Casteel A., Bridier, N.L. Describing populations and samples in doctoral student research. *International Journal of Doctoral Studies,* 16, 2021, pp. 339-362. https://doi.org/10.28945/4766.

[24] Lawless H.T., Heymann, H. Descriptive Analysis. Sensory Evaluation of Food, 2010, pp. 227-257. https://doi.org/10.1007/978-1-4419-6488-5_10.

[25] Albattah W., Khan R.U., Alsharekh M.F., Khasawneh S.F. Feature Selection Techniques for Big Data Analytics. *Electronics,* 2022, 11 (19): 3177. https://doi.org/10.3390/electronics11193177.

[26] Mpia H.N., Mwendia S.N., Mburu L.W. Predicting Employability of Congolese Information Technology Graduates Using Contextual Factors: Towards Sustainable Employability. *Sustainability,* 2022, 14 (20), 13001. https://doi.org/10.3390/su142013001.