

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITÉ DE 8 MAI 1945 – GUELMA -
FACULTÉ DES MATHÉMATIQUES, D'INFORMATIQUE ET DES SCIENCES DE LA
MATIÈRE

Département d'Informatique



Mémoire de Fin d'études Master

Filière : Informatique

Option : Systèmes Informatiques

Thème _____

**Le traitement des données manquantes dans le « Big
Data » médical**

Encadré Par :

DR. BENHAMZA Karima

Présenté par :

BOUDJEHEM Bilal

Juin 2022

Remerciements

Alhamdoulilah , Je remercie dieu le tout puissant de m'avoir donné la santé et la volonté d'entamer ce mémoire et de le terminer. Tout d'abord, ce travail ne serait pas aussi riche et n'aurais pas vu le jour sans l'aide et l'encadrement De Mme Benhamza Karima, je la remercie pour la qualité de son encadrement exceptionnel, pour ses précieux conseils, pour sa patience, sa rigueur et sa disponibilité durant la préparation de ce mémoire. Mes remerciements s'adressent également à tous mes professeurs pour la qualité de l'enseignement qui m'ont prodigué au cours de mes années passées à l'université. Je remercie toute personne qui a contribué d'une manière ou d'une autre à la réalisation de ce mémoire

Dédécace

Je dédié se travail à mon très chers défunt père, celui qui a toujours été là pour moi, celui qui m'a tant incité à terminer mes études, et sans qui je ne serais jamais arrivé là où je suis aujourd'hui, que le bon Dieu lui accorde sa clémence et sa miséricorde, et l'accueil bien à ses côtés. Je dédié se travail à ma chère mère que le bon dieu la garde pour moi, et lui offre une longue vie pleine de bonheur et de bonne santé. Je dédié ce travail aussi a mes collègues et proches qui m'ont encouragé et aider tout au long de la route et spécialement « Hazem Bensalah ».

RÉSUMÉ

Face à l'explosion des données qui a connu le monde ces dernières années. Tous les domaines ont été envahi par le « Big Data » et se sont retrouver face à ses défis. Le domaine médicale n'a pas eu d'exception et s'est retrouver face au plus grand défi, qui est les données manquantes ou « Missing Data ».

Les données manquantes posent un gros problème, leurs traitements dans le domaine médicale est dangereux vu que la vie des gens en dépendait. Le but de ce travail est de montrer l'importance des méthodes analytiques dans le traitement des données manquantes dans le domaine médicale. Tout l'intérêt et de récupérer ces données manquantes ou les prédire .

Le résultat de la combinaison des méthodes analytiques utilisés sur un Dataset Médicale nous ont permet de souligné l'importance de ces méthode dans le traitement des données manquantes.

Mots clés : Big Data, Méthodes Analytiques, Machine Learning, K-means, k-NN, Feature Selection, Dataset Médical.

ABSTRACT

In front of the explosion of data that has experienced the world in recent years. All domains have been invaded by "Big Data" and have found themselves faced with his challenges. The medical domain was no exception and found itself facing the greatest challenge, which is the "Missing Data".

Missing data poses a big problem, their treatment in the medical domain is dangerous because people's lives depended on it. The purpose of this work is to show the importance of analytical methods in the treatment of missing data in the medical field. All the interest is to recover these missing data or predict them.

The result of the combination of analytical methods applied on a Dataset allowed us to underline the importance of these methods in the treatment of missing data.

Keywords : Big Data, Analytical Methods, Machine Learning, K-means, k-NN, Feature Selection, Medical Dataset.

ملخص

في مواجهة انفجار البيانات التي عرفت العالم في السنوات الأخيرة. تعرضت جميع المجالات للغزو من قبل البيانات الضخمة ووجدت نفسها تواجه تحدياتها. لم يكن لدى المجال الطبي استثناءات ووجد نفسه يواجه التحدي الأكبر، وهو البيانات المفقودة أو «البيانات المفقودة».

البيانات المفقودة مشكلة كبيرة، وعلاجهم في المجال الطبي خطير لأن حياة الناس تعتمد عليه. الغرض من هذا العمل هو إظهار أهمية الأساليب التحليلية في معالجة البيانات المفقودة في المجال الطبي. كل الاهتمام واستعادة هذه البيانات المفقودة أو التنبؤ بها.

سمحت لنا نتيجة الجمع بين الأساليب التحليلية المستخدمة في مجموعة البيانات الطبية بالتأكد على أهمية هذه الأساليب في علاج البيانات المفقودة.

الكلمات المفتاحية: البيانات الضخمة، الأساليب التحليلية، K-mean، k-NN، اختيار المواصفات، مجموعة البيانات الطبية

TABLE DES MATIÈRES

Liste des figures		xi
Liste des tableaux		xii
Introduction générale		1
1 BIG DATA		3
1.1 Introduction		3
1.2 Définitions		4
1.3 Caractéristiques du Big Data		5
1.3.1 Volume		5
1.3.2 Variété		5
1.3.3 Vitesse		6
1.3.4 Véracité		6
1.3.5 Valeur		6
1.4 Big Data Médical		6
1.4.1 « Big Data » dans la littérature médicale		7
1.4.2 Caractéristiques uniques du « Big Data » Medical		7
1.4.3 Hétérogénéité		7
1.4.4 Incomplétude		8

1.4.5	Ponctualité et longévité	8
1.4.6	Confidentialité des données	9
1.5	Applications Big data Medical	9
1.5.1	Surveillance régionale de la santé	9
1.5.2	Amélioration du développement de médicaments pharmaceu- tiques	9
1.5.3	Amélioration des systèmes de soutien aux soins de santé	10
1.6	Défis du « Big Data » médical	10
1.7	Missing Data dans les « Big Data »	11
1.8	Big Data Analytique	12
1.8.1	Hadoop	12
1.8.2	MapReduce	12
1.9	Apache Spark	14
1.9.1	Ensemble de Données Distribué Résilient (RDD)	14
1.9.2	Bibliothèque d'apprentissage automatique Spark (MLlib)	15
1.10	Conclusion	15
2	SYNTHESE DES TRAVAUX	17
2.1	Introduction	17
2.2	Méthodes existantes pour le traitement des données médicales man- quantes	17
2.2.1	Méthodes de suppression des données (Delete data)	18
2.2.2	Méthodes d'imputation simple (Single Imputation)	18
2.2.2.1	Moyenne, Médiane et Mode	18
2.2.2.2	Dernière observation reportée	19
2.2.2.3	Interpolation Linéaire	19
2.2.2.4	Imputation par point commun	19
2.2.2.5	Imputation Par ajout d'une catégorie	20
2.2.2.6	Imputation par catégorie fréquente	20

2.2.2.7	Imputation Par Valeurs Arbitraires	20
2.2.2.8	Imputation Par échantillonnage aléatoire	20
2.2.3	Imputation Multiple	21
2.2.4	Méthodes d'Intelligence artificielle (Model-Based Methods) . .	21
2.2.4.1	Régression linéaire	22
2.2.4.2	Random Forest	22
2.2.4.3	k-NN (k Nearest Neighbour)	22
2.2.4.4	Maximum likelihood	23
2.2.4.5	Expectation-Maximization	23
2.2.4.6	Analyse de sensibilité	23
2.2.4.7	Réseau de neurones multicouches (Multi-layer percep- tron MLP)	24
2.3	Table de comparaison et Discussion	24
2.4	Conclusion	30
3	CONCEPTION ET IMPLÉMENTATION	31
3.1	Introduction	31
3.2	Conception	31
3.2.1	Architecture proposée	31
3.2.2	Acquisition des données	32
3.2.3	Prétraitement	33
3.2.4	Modèle de classification	33
3.2.4.1	Feature selection	33
3.2.4.2	Méthode améliorée du K-means	34
3.2.5	Modèle d'Imputation	36
3.2.5.1	Algorithme k-NN	37
3.3	Implémentation	38
3.3.1	Matériels utilisés :	38
3.3.2	Logiciels utilisés :	38

3.3.3	Dataset Utilisé	38
3.3.4	Modélisation d'exécution avec Plateforme Spark	39
3.3.4.1	Lire CSV en RDD	39
3.3.4.2	Création du maitre	41
3.3.4.3	Chargement du dataset	41
3.3.5	Modèle de classification	42
3.3.5.1	Feature selection	42
3.3.5.2	Elbow Méthode	43
3.3.5.3	K-means	44
3.3.6	Modèle d'imputation	45
3.3.6.1	Imputation par K-nn	47
3.3.7	Le RMSE (Root Mean Square Error)	49
3.4	Conclusion	51
	Conclusion générale	53
	Bibliographie	54

TABLE DES FIGURES

1.1	Volume annuel des données [6]	4
1.2	Big Data structuré et non structuré [14]	8
1.3	Architecture du MapReduce[16]	13
1.4	Apache Spark Composant[17]	14
1.5	les flux de données en RDD[18]	15
3.1	Architecture proposée	32
3.2	architecture K-means Elbow	36
3.3	Dataset Utilisé	39
3.4	Lecture du Dataset en RDD	40
3.5	Fonctionnement De Spark	40
3.6	Création du Maitre	41
3.7	Chargement du dataset	41
3.8	Dataset après l'utilisation de 'Feature Selection'	42
3.9	Matrice de corrélation	43
3.10	Elbow Méthode	44
3.11	Représentation de la Population dans les clusters	45
3.12	DataTest avant l'imputation	46

3.13 Classification des individus avec des données manquantes dans les clusters	47
3.14 DataTest après l'imputation	48
3.15 Histogramme des résultats obtenus	51

LISTE DES TABLEAUX

2.1	Table de comparaison des méthodes existantes	25
2.1	Table de comparaison des méthodes existantes	26
2.1	Table de comparaison des méthodes existantes	27
2.1	Table de comparaison des méthodes existantes	28
2.1	Table de comparaison des méthodes existantes	29
2.1	Table de comparaison des méthodes existantes	30
3.1	Table excel représentant les valeurs réelles et prédites	49
3.2	Table de comparaison des résultats	50

INTRODUCTION GÉNÉRALE

Ces dernières années, le monde de l'information a été confronté à une explosion des données. Ces données numériques proviennent de différentes sources (Internet, réseaux de capteurs, trajectoires GPS, etc.). Des masses de données massives qui dépassent les limites des technologies traditionnelles ont poussées plusieurs chercheurs dans tout les domaines à concevoir de nouvelles technologies pour le stockage, le traitement et l'analyse de ces données massives.

Cette révolution phénoménale qui a envahi le monde, a donné naissance à une nouvelle technologie dont le nom est « Big Data ». Cette technologie a imposé des nouveaux défis aux différents chercheurs pour stocker, analyser et gérer correctement ces énormes ensembles de données.

Le « Big Data » s'applique pleinement dans le domaine médical. La récupération des données massives dans ce domaine se présente comme une tâche relativement facile, contrairement à son traitement d'une façon précise et efficace à cause de son incomplétude. Ce problème a obligé les chercheurs à relever le grand défi dans cet axe de recherche difficile qui est le traitement des données manquantes.

Notre projet de fin d'étude a pour but d'étudier les technologies et les méthodes

du « Big Data », Nous nous intéresserons particulièrement aux méthodes de traitement de données manquantes dans le domaine médical. L'intérêt de cette recherche est d'analyser et de comparer les méthodes utilisés afin de proposer un nouveau modèle.

Ce mémoire est divisé en trois chapitres :

Dans le premier chapitre, on présentera la technologie « Big Data » avec ses concepts et ses caractéristiques ainsi que le « Big Data » dans le domaine médical .

Puis, dans le deuxième chapitre, les méthodes utilisées pour le traitement des données manquantes dans les « Big Data » médical seront détaillées.

Le troisième chapitre traitera la conception et l'implémentation du modèle proposé pour imputer les données manquantes dans les « Big Data » médicales. La plateforme Spark est utilisée pour le traitement des données et la présentation des résultats obtenus.

Finalement, ce mémoire est clôturé par une conclusion générale, des perspectives de ce travail et les références bibliographiques utilisée.

CHAPITRE 1

BIG DATA

1.1 Introduction

Avec l'énorme développement que nous avons réalisé ces dernières années, l'industrie technologique subit des changements spectaculaires dans la quantité de données, qui doit être gérée et l'emplacement où ces actifs sont stockés. Cette grande quantité de données se définit sous le terme de « Big Data » [1]. Le « Big Data » est devenu aussi un mot à la mode dans le monde de l'innovation médicale. Le développement rapide des techniques d'apprentissage automatique, en particulier l'intelligence artificielle, promet de révolutionner la pratique médicale : de l'allocation des ressources au diagnostic des maladies complexes. Cependant les « Big Data » présentent aussi de grands défis et risques [2]. Dans ce chapitre, nous aborderons les bases du « Big Data » et leurs caractéristiques et rôles dans le domaine médical.

1.2 Définitions

Le « Big Data » (ou mégadonnées) fait référence à des ensembles de données d'une taille, d'une vitesse et d'une variété telles que leur conversion en valeur utilisable nécessite l'utilisation des techniques et des méthodes analytiques spécifiques[2]. Aussi, la définition du « Big Data », selon Japtec et al [3], englobe plus que le volume massif, la vitesse et la variété. Cela implique également la création de données en premier lieu, la création de nouveaux types de données et le développement de nouveaux processus pour analyser et traiter ces données générées. Selon Vivekanand et Vidyavathi [4], les mégadonnées sont définies aussi comme de grandes quantités d'informations numériques que les entreprises et les gouvernements collectent sur les personnes et leur environnement, avec des tailles de données de 2500 pétaoctets (10^3 Tera-octets) ou plus[4]. D'autre part Padmapriya et al. [5] ont défini le « Big Data » comme une collection d'ensembles de données volumineuses et compliquées qui s'adaptent à des processus complexes en utilisant un ensemble de méthodes requises.

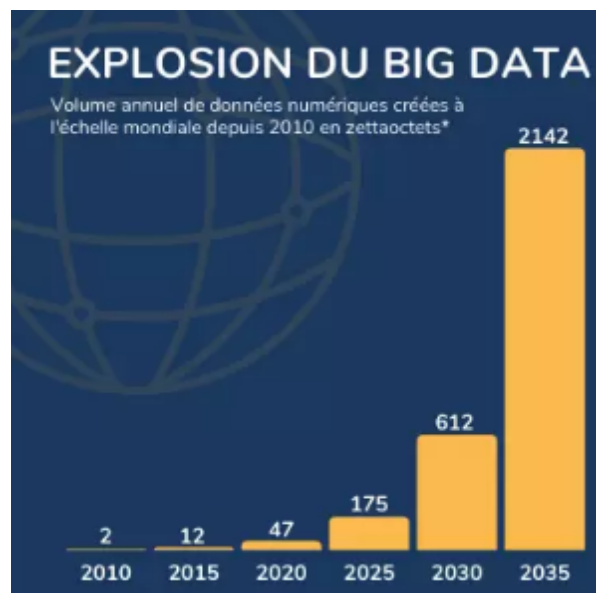


FIGURE 1.1 – Volume annuel des données [6]

1.3 Caractéristiques du Big Data

Un « Big Data » est souvent défini par le nombre de « V ». Nizam et Hassan [7] l'ont défini sur la base de « 5V » : Volume, Variété, Vitesse, Valeur et Véracité.

1.3.1 Volume

La grande quantité de données qui peut être collectée, transférée, stockée et analysée avec des outils et des applications de stockage de données et de mise en réseau avancés est appelée volume [7]. Il fait référence aussi à la quantité totale de données qui a été générée à partir de diverses sources et qui continue de croître jusqu'à 2,5 exaoctets (10^3 pétaoctet) ou plus[8]. Selon Ranjan [9], certains experts considèrent que les volumes de données massives sont supérieurs à 5 Pétaoctets, et la croissance des données volumineuses devrait atteindre 35 zettaoctets (10^3 exaoctets) de données en 2022.

1.3.2 Variété

La variété fait référence à un large éventail de représentations de données (par exemple, texte, documents, photographies, vidéo, audio, etc.) qui sont acquises à partir de diverses sources et représentées à l'aide de divers supports et collecteurs tels que des capteurs, des téléphones portables et d'autres appareils. pour représenter des ensembles de données hétérogènes [8]. Les données structurées, avec des champs fixes intégrés, les données non structurées, qui sont produites de manière aléatoire et difficiles à analyser, et les données semi-structurées, qui ne peuvent pas tenir dans des champs définis mais contiennent des balises pour séparer les éléments de données, sont les trois catégories de données du Big Data [8].

1.3.3 Vélacité

Yadranjiaghdam et al. [10] ont souligné que des quantités massives de grands ensembles de données sont produites et générées quotidiennement. Dans les Big Data, la vélocité est une mesure de la vitesse à laquelle les données sont créées, diffusées, agrégées et transmises à partir de diverses sources [8].

1.3.4 Véracité

Sans cette caractéristique, les modèles construits sur les données n'auront pas de vraie valeur, compte tenu du volume, de la variété et de la vélocité que permet le « Big Data ». La qualité des données dérivées après traitement est déterminée par la véracité des données sources. Les biais de données, les anomalies ou les incohérences, la volatilité et la duplication doivent être atténués par le système. [11]

1.3.5 Valeur

En termes d'entreprise, le « V » le plus important est la Valeur. Le « Big Data » devrait théoriquement fournir de la "Valeur". Les équipes d'analyse et de recherche doivent considérer, concevoir et fournir la taille et la portée de cette valeur. Dans le domaine des affaires, la valeur est l'une des premières propriétés discutées, et une certaine quantité de valeur sera projetée au début d'un projet « Big Data ». Le Big Data aide au développement de l'infrastructure sur laquelle l'apprentissage automatique et l'intelligence artificielle peuvent être construits[11].

1.4 Big Data Médical

Les « Big Data » ont un large éventail d'applications, mais l'une des plus importantes est leur potentiel pour faire avancer la recherche médicale, améliorant ainsi la qualité de la vie humaine. L'utilisation des « Big Data » dans la recherche et l'avancement médicaux est d'une importance primordiale. L'intelligence artificielle est à

l'avant-garde de la collecte de données médicales. En analysant les problèmes de santé publique en temps réel, les mégadonnées peuvent faire progresser la recherche médicale dans de multiples domaines. Le « Big Data » peut également améliorer les soins aux patients et prévenir la propagation de maladies mortelles [12].

1.4.1 « Big Data » dans la littérature médicale

Avec le développement du domaine médical (appelé aussi clinique), de nombreux articles de recherche et des connaissances structurées sont produits à grande vitesse. De plus, beaucoup de matériel ancien persiste dans ce domaine. Cette avancée dans la littérature a apporté des contributions significatives au domaine des « Big Data » Medicals [13].

1.4.2 Caractéristiques uniques du « Big Data » Medical

Le « Big Data » dans le domaine de la santé a ses propres caractéristiques en plus des caractéristiques '5V' du Big Data. Ces dernières sont l'hétérogénéité, l'incomplétude, l'actualité et la longévité, la confidentialité des données et la propriété.

1.4.3 Hétérogénéité

Le « Big Data » Médical a souvent des formats incompatibles, qui peuvent être classés en structurés et non structurés(Figure 1.2). Cependant la majorité des « Big Data » médicales sont non structurés vu la différence de leurs provenances exemple de source : la tomodensitométrie, de l'IRM ,rayons X, la surveillance Holter, l'angiographie et les laboratoires [13].

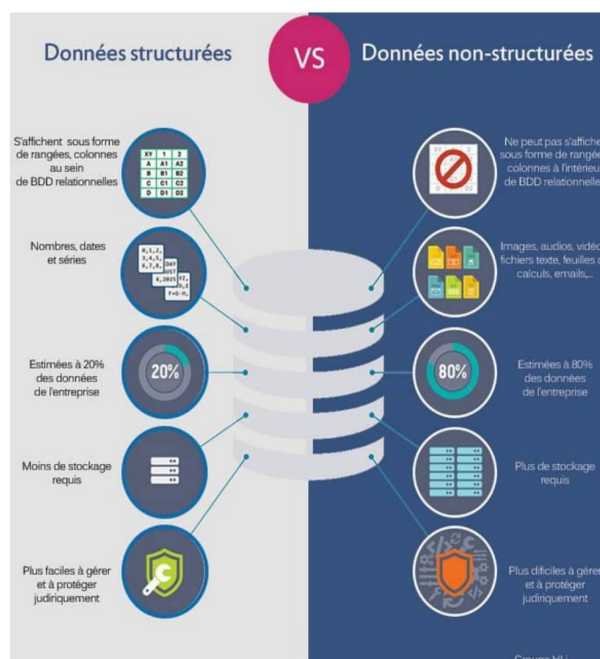


FIGURE 1.2 – Big Data structuré et non structuré [14]

1.4.4 Incomplétude

Le système hospitalier exige des médecins ou des infirmiers à enregistrer les informations sur les maladies des patients (médicaments, allergies,..). En effet, les dossiers médicaux soutiennent non seulement les soins directs aux patients, mais également l'audit clinique, l'épidémiologie, la recherche médicale et l'allocation des ressources[13]. Il est donc trop coûteux de stocker toutes ces mégadonnées de santé. Cette situation a conduit à l'incomplétude des données qui est le défi majeur du «Big Data» médical.

1.4.5 Ponctualité et longévité

Pour certaines maladies familiales ou génétiques, il est utile de connaître les antécédents familiaux afin d'aider à la prise de décision médicale. À ce stade, il n'y a aucun lien entre le dossier médical du patient et celui des membres de sa famille[13] et la durée de conservation des dossiers médicaux diffère aussi d'un lieu à un autre.

1.4.6 Confidentialité des données

En raison de la centralisation d'une grande partie des informations de santé, les données sont très vulnérables aux attaques, et vu la sensibilité de ces données, il existe des préoccupations importantes concernant la confidentialité et la sécurité. Des précautions extrêmes sont généralement prises pour protéger la vie privée des patients[13].

1.5 Applications Big data Medical

Voici quelques-unes des applications dont les mégadonnées contribuent à faire progresser la recherche médicale et les soins de santé[12].

1.5.1 Surveillance régionale de la santé

Les données peuvent être utilisées pour la recherche médicale prédictive, aidant à prévenir la propagation des maladies. Par exemple, une façon de faire progresser les soins préventifs et la recherche consiste à suivre les populations de patients et leurs besoins en matière de soins de santé en suivant les questions médicales qu'ils recherchent ainsi que les informations qu'ils fournissent sur les sites Web médicaux. Ces informations ont le potentiel d'aider à prévoir les épidémies régionales de diverses maladies et les problèmes de santé publique actuels.

1.5.2 Amélioration du développement de médicaments pharmaceutiques

La production de médicaments est essentielle à la survie de l'humanité. Cependant, la découverte et le développement de médicaments ne sont que la première étape. Un médicament pharmaceutique doit subir plusieurs séries de tests rigoureux avant d'être approuvé pour la vente. Plutôt que de mener de véritables expériences

en laboratoire, les « Big Data » permettent aux créateurs d'utiliser des applications de données dans le processus de développement de médicaments, ce qui facilite la découverte de médicaments par calcul. Les sociétés pharmaceutiques peuvent utiliser ces informations pour créer des modèles et des simulations réalistes afin de tester leurs produits.

1.5.3 Amélioration des systèmes de soutien aux soins de santé

L'une des principales avancées de la technologie médicale est la robotique des soins de santé, qui devrait atteindre 2,8 milliards de dollars de revenus en 2022. Les spécialités comprennent la formation de robots chirurgicaux, les infirmières robotiques et les prothèses intelligentes, ainsi que l'assistance en thérapie, la télé-présence, et la logistique. L'utilisation de la robotique alimentée par le «Big Data» a le potentiel d'améliorer considérablement la qualité du soutien aux soins de santé.

1.6 Défis du « Big Data » médical

La structure des données, la sécurité, la normalisation des données, le stockage et les transferts, ainsi que les compétences managériales telles que la gouvernance des données, étaient les principaux défis. L'amélioration de la qualité, la gestion et la santé de la population, la détection précoce des maladies, l'amélioration de la prise de décision et la réduction des coûts ont été les principales opportunités révélées. Il y a une augmentation exponentielle des grands ensembles de données électroniques sur la santé, qui ne peuvent être gérés par les techniques conventionnelles. L'utilisation optimale de ces ensembles de données dans l'informatique de la santé fait face à de nombreux défis, notamment le volume et la vitesse des données générées, la variété des types de données, la véracité des données et la confidentialité des informations médicales du patient. Le domaine de l'informatique de la santé devrait bénéficier du développement rapide d'outils d'analyse de données volumineuses pour résoudre des problèmes critiques tels que la représentation des connaissances, le diagnostic

des maladies et l'aide à la décision clinique [7].

S'intégrant sous ces problèmes, le traitement des données manquantes (Missing Data) reste le plus grand déficit pour le «Big Data» médical.

1.7 Missing Data dans les « Big Data »

Il existe trois mécanismes majeurs de perte de données qui peuvent être décrites selon la relation de dépendance entre la donnée observée (existante) et la donnée non-observée (manquante).

- Missing Completely at Random (MCAR) : Lorsque les observations manquantes dépendent des mesures observées et non observées. Dans ce cas, la probabilité qu'une observation manque ne dépend que d'elle-même. Il n'y a pas de mécanisme caché lié à une variable et il ne dépend d'aucune caractéristique des patients [15]. À titre d'exemple, lorsqu'un médecin oublie d'enregistrer le sexe de tout les patients qui entrent dans l'unité de soins intensifs.
- Missing At Random (MAR) : Dans ce cas, la probabilité qu'une valeur soit manquante est liée uniquement aux données observables, c'est-à-dire que les données observées sont statistiquement liées aux variables manquantes et il est possible d'estimer les valeurs manquantes à partir des données observées. Autrement dit, la probabilité que certaines données soient manquantes pour une variable particulière ne dépend pas des valeurs de cette variable, après ajustement pour les valeurs observées. Par exemple, si les personnes âgées sont moins susceptibles d'informer le médecin qu'elles ont déjà eu une pneumonie, le taux de réponse de la variable pneumonie dépendra de la variable âge [15].
- Missing Not At Random (MNAR) : Il s'agit du cas où ni MCAR ni MAR ne sont maintenus. Les données manquantes dépendent à la fois des valeurs manquantes et observées. Déterminer le mécanisme manquant est généralement impossible, car cela dépend de données invisibles. De là découle l'importance d'effectuer des analyses de sensibilité et de tester la validité des inférences sous

différentes hypothèses. Par exemple : les patients souffrant d'hypotension artérielle sont plus susceptibles de faire mesurer leur tension artérielle moins fréquemment (les données manquantes pour la variable « tension artérielle » dépendent en partie des valeurs de la tension artérielle) [15].

1.8 Big Data Analytique

L'objectif de l'analyse du Big Data est de découvrir des modèles intéressants, des relations cachées et des informations précieuses. Le marché compte de nombreux outils d'analyse qui peuvent être classés selon les tâches suivantes : stockage, gestion, analyse, intégration de données, exploration de données et visualisation de données. Parmi ces outils, on peut citer les plus utilisés : Hadoop, Spark,...

1.8.1 Hadoop

Hadoop est un Framework logiciel (open source) qui distribue le traitement de grands ensembles de données sur des clusters d'ordinateurs en utilisant des modèles de programmation agiles. Il offre aussi un haut niveau d'évolutivité en mettant à disposition des centaines, voire des milliers de machines. Il combine des stockages et des calculs locaux plutôt que de dépendre de serveurs centralisés[7]. Hadoop utilise "MapReduce" pour traiter rapidement les données massives.

1.8.2 MapReduce

MapReduce est un modèle de programmation qui facilite le développement d'applications pour le traitement de grands ensembles de données (téraoctets ou pétaoctets) en parallèle sur un grand cluster (nombre de nœuds) de serveurs de base, de manière fiable et tolérante aux pannes. Les composantes principales de MapReduce sont [7] :

— **Mapper Function**

La fonction Mapper analyse les données sources brutes et fait correspondre les valeurs à des clés uniques sous la forme de paires <key, value>. La fonction Mapper est décrite comme l'opération d'extraction et d'organisation des informations.

— **Shuffle and Sort Function**

Il s'agit d'une fonction intermédiaire qui fonctionne après l'achèvement de tous les Mappers et avant le début des Réducteurs pour mettre la sortie des "Mapers" dans une forme appropriée pour l'exécution des Réducteurs.

— **Reducer**

La fonction de réduction opère sur les données intermédiaires mélangées et triées générées par les fonctions de mappage afin de produire le résultat final.

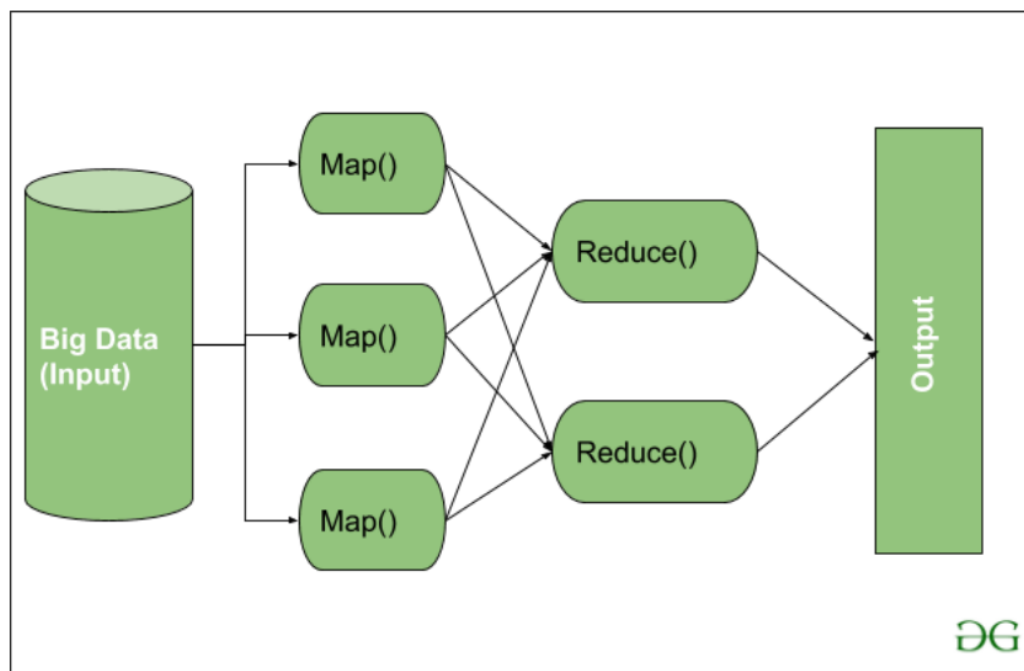


FIGURE 1.3 – Architecture du MapReduce[16]

1.9 Apache Spark

Spark est un moteur à usage général qui peut traiter rapidement de grands ensembles de données. Il est cent fois plus rapide en mémoire et dix fois plus rapide sur disque que MapReduce. Le flux de travail est optimisé par le moteur de Graphe Acyclique Dirigé (DAG). Spark peut être utilisé avec Elastic MapReduce, HDFS (Hadoop Distributed File System), S3, HBase, Sequoia DB, MongoDB et d'autres bases de données utilisant les plateformes de programmation : Python, Java, R et Scala. Le Resilient Distributed Dataset (RDD) est le concept central qui sous-tend la fonctionnalité de Spark. Spark core, Spark Structured Query Language (SQL), MLib, Spark Streaming et GraphX sont les principaux composants.

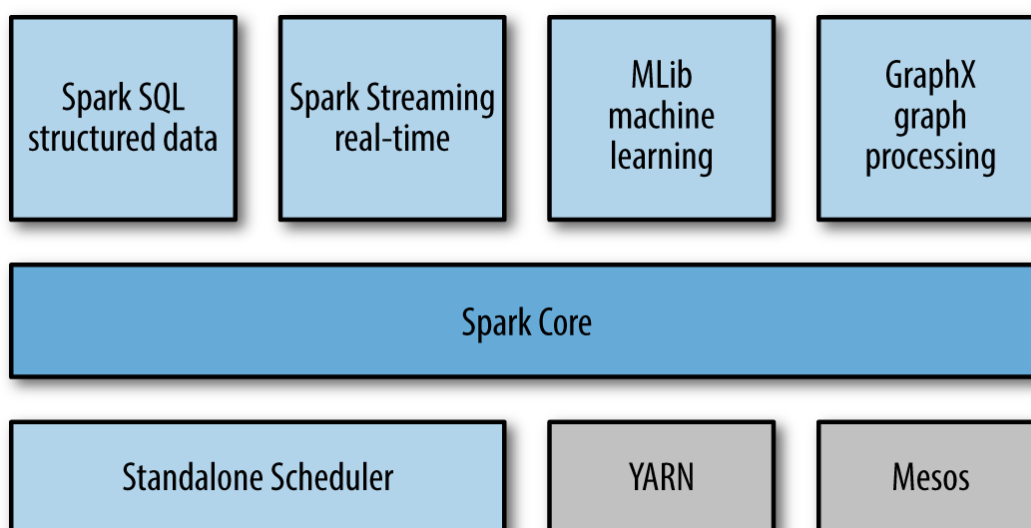


FIGURE 1.4 – Apache Spark Composant[17]

1.9.1 Ensemble de Données Distribué Résilient (RDD)

Le point faible de MapReduce réside dans les calculs distribués itératifs nécessaires à de nombreux algorithmes. Spark a résolu ce problème en effectuant ces calculs en mémoire, et non dans HDFS (Hadoop Distributed File System). C'est pourquoi Spark est jugé supérieur à Mapreduce en termes de rapidité de traitement.

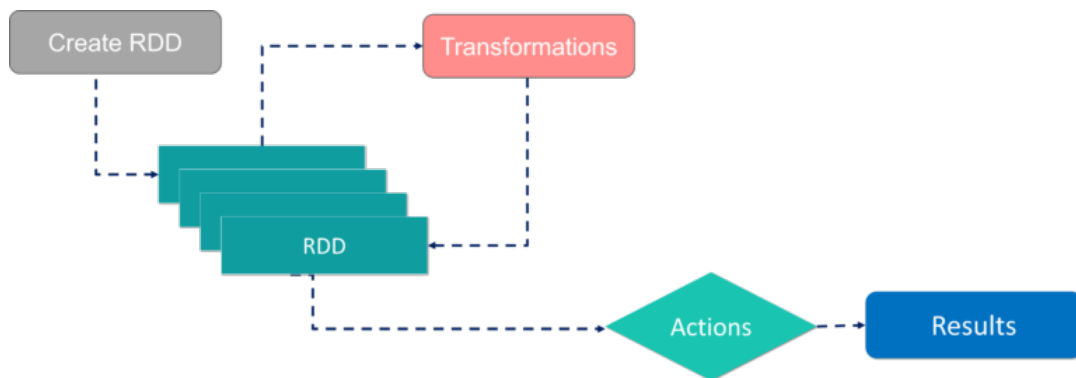


FIGURE 1.5 – les flux de données en RDD[18]

1.9.2 Bibliothèque d'apprentissage automatique Spark (MLlib)

Spark MLlib est la principale raison du succès de Spark, Ce dernier fournit une bibliothèque d'algorithmes qui peuvent être utilisés pour résoudre la plupart des problèmes d'apprentissage automatique. La classification, la régression, le clustering, la recommandation, la modélisation de sujet, l'ensemble d'éléments fréquents et d'autres algorithmes sont disponibles dans Spark MLlib. Spark MLlib est simple, évolutif et fonctionne avec un large éventail d'outils de science des données, notamment R, Python, Weka et autres. [7]

1.10 Conclusion

Les « Big Data » peuvent être utilisés pour faire avancer la recherche médicale et améliorer la vie humaine. Google, par exemple, partage les données des utilisateurs avec les chercheurs. Les « Big Data » peuvent également accélérer et améliorer la faisabilité des tests de dépistage de drogues grâce à des algorithmes et à l'apprentissage automatique. Ces mégadonnées permettent aux créateurs de médicaments d'utiliser des applications de données dans le processus de développement de médicaments. Le « Big Data » peut ainsi améliorer les soins aux patients et prévenir la propagation de maladies mortelles. Le domaine de l'informatique de la santé devrait bénéficier du développement rapide d'outils d'analyse de données massives pour résoudre les

problèmes critiques tels que le diagnostic de maladies et l'aide à la décision clinique, mais pour avoir des diagnostics et des décisions cliniques précises il faut que les données du « Big Data » soient complètes. C'est pour cela que les données manquantes sont le défi numéro un des «Big Data» médicaux.

CHAPITRE 2

SYNTHESE DES TRAVAUX

2.1 Introduction

Nous avons établis des recherches, dans le moteur de recherche académique « Google Scholar » en utilisant l'équation de recherche « Missing Data in medical Big-Data », tout en fixant la date entre 2014 et 2022 (Période de pic de travail de recherche). Plusieurs approches ont été abordées (des méthodes analytiques aux .méthodes de L'Intelligence Artificielle) pour gérer les données manquantes mais peu de méthodes ont été utilisées dans le domaine médical.

2.2 Méthodes existantes pour le traitement des données médicales manquantes

Les méthodes de traitement des données médicales manquantes précédemment collectées ont été synthétisés comme suit :

2.2.1 Méthodes de suppression des données (Delete data)

L'approche la plus courante pour les données manquantes consiste à omettre les cas avec les données manquantes et à analyser les données restantes [19]. Cette approche est connue sous le nom d'analyse complète (ou de cas disponible) ou de suppression par liste (List-wise deletion). Une autre approche existante connue sous le nom de Suppression par paire (Pairwise Deletion) où seules les observations manquantes sont ignorées, et l'analyse est effectuée sur les variables présentes[20]. Étant donné qu'une suppression par paires utilise toutes les informations observées, elle conserve plus d'informations que la suppression par liste. Néanmoins, s'il manque beaucoup d'observations, l'analyse sera déficiente. D'autre part; s'il manque trop de données pour une variable, il peut être possible de supprimer la variable ou la colonne du jeu de données. Cette méthode est connue sous le nom de "Suppression des variables" (ou Dropping Variables)[21]. Une analyse appropriée des données est nécessaire avant que la variable ne soit complètement abandonnée (vérifier les performances du modèle).

2.2.2 Méthodes d'imputation simple (Single Imputation)

Toute technique d'imputation vise à produire un ensemble de données complet qui peut ensuite être utilisé pour l'apprentissage automatique.

2.2.2.1 Moyenne, Médiane et Mode

Dans cette technique d'imputation, l'objectif est de remplacer les données manquantes par des estimations statistiques des valeurs manquantes. La moyenne, médiane ou mode peuvent être utilisés comme valeur d'imputation.

A. Dans une substitution moyenne, la valeur moyenne d'une variable est utilisée à la place de la valeur de données manquante pour cette même variable[22]. Cela a l'avantage de ne pas modifier la moyenne de l'échantillon pour cette variable. La distorsion de la variance(mesure de la dispersion des valeurs de l'échantillon) d'origine et la

distorsion de la covariance (l'évaluer du sens de variation de deux variables) avec les variables restantes dans l'ensemble de données sont deux inconvénients majeurs de cette méthode.

B. La médiane peut être utilisée lorsque la variable a une distribution asymétrique [23].

C. L'intérêt de l'estimation du mode est de remplacer la population de valeurs manquantes par la valeur la plus fréquente, car il s'agit de l'occurrence la plus probable [22].

2.2.2.2 Dernière observation reportée

Si les données sont des données chronologiques, l'une des méthodes d'imputation les plus utilisées est la dernière observation reportée (Last Observation Carried Forward LOCF). Chaque fois qu'une valeur est manquante, elle est remplacée par la dernière valeur observée [24]. Cette méthode est avantageuse car elle est facile à comprendre et à communiquer. Bien que simple, cette méthode suppose fortement que la valeur du résultat reste inchangée par les données manquantes, ce qui n'est pas toujours correct.

2.2.2.3 Interpolation Linéaire

L'interpolation est une méthode mathématique qui ajuste une fonction aux données et utilise cette fonction pour extrapoler les données manquantes. Le type d'interpolation le plus simple est l'interpolation linéaire, c'est-à-dire entre les valeurs avant les données manquantes et les valeurs sans données manquantes [25]. Néanmoins, le modèle peut être très complexe et l'interpolation linéaire ne pourrait pas suffire pour l'imputation.

2.2.2.4 Imputation par point commun

Pour une échelle d'évaluation, en utilisant le point médian ou la valeur la plus couramment choisie [26]. Similaire à l'imputation par la valeur moyenne mais cette

méthode est plus appropriée pour les valeurs ordinales.

2.2.2.5 Imputation Par ajout d'une catégorie

C'est la méthode la plus largement utilisée pour l'imputation des données manquantes pour les variables catégorielles. Cette méthode consiste à traiter les données manquantes comme une catégorie supplémentaire de la variable. Toutes les observations manquantes sont regroupées dans une étiquette 'Manquant' nouvellement créée [27]. Cette imputation est bien adaptée lorsque le nombre de données manquantes est élevé.

2.2.2.6 Imputation par catégorie fréquente

Le remplacement des valeurs manquantes par la catégorie la plus fréquente équivaut à une imputation moyenne/médiane. Elle consiste à remplacer toutes les occurrences de valeurs manquantes dans une variable par l'étiquette ou la catégorie la plus fréquente de la variable[28].

2.2.2.7 Imputation Par Valeurs Arbitraires

L'imputation de valeur arbitraire consiste à remplacer toutes les occurrences de valeurs manquantes dans une variable par une valeur arbitraire (différente de la médiane/moyenne/mode). Les valeurs arbitraires généralement utilisées sont 0, 999, -999 (ou d'autres combinaisons de 9) ou -1 [29]. Il s'agissait d'une méthode courante lorsque les bibliothèques et les algorithmes d'apprentissage automatique prêts à l'emploi n'étaient pas très habiles à travailler avec les données manquantes.

2.2.2.8 Imputation Par échantillonnage aléatoire

L'imputation par échantillonnage aléatoire est en principe similaire à l'imputation moyenne/médiane car elle vise à préserver les paramètres statistiques de la variable d'origine, pour laquelle des données sont manquantes. L'échantillonnage aléatoire

consiste à prendre une observation aléatoire à partir d'observations disponibles et à utiliser cette valeur extraite au hasard pour remplir le manque[30]. Généralement, on prend autant d'observations aléatoires que les valeurs manquantes sont présentes dans la variable.

2.2.3 Imputation Multiple

L'Imputation Multiple (IM) est une technique statistique permettant de traiter les données manquantes. Le concept clé de l'IM est d'utiliser la distribution des données observées pour estimer un ensemble de valeurs plausibles pour les données manquantes. Des composantes aléatoires sont incorporées à ces valeurs estimées pour montrer leur incertitude. Plusieurs ensembles de données sont créés, puis analysés individuellement mais de manière identique pour obtenir un ensemble d'estimations de paramètres. L'avantage des multiples imputations est que la restauration des valeurs manquantes intègre l'incertitude due aux données manquantes, ce qui se traduit par une inférence statistique valide [31]. La méthode d'Imputation multiple la plus connue est la méthode MICE (Multiple Imputation using Chained Equations). Le traitement des données à l'aide de MICE comprend trois étapes principales : génération d'imputations multiples, analyse des données imputées et mise en commun des résultats d'analyse[31].

2.2.4 Méthodes d'Intelligence artificielle (Model-Based Methods)

L'imputation des données manquantes basé sur des modèles, consiste à utiliser des modèles entraînés pour prédire les valeurs des données manquantes. Il existe de nombreuses options pour un tel modèle prédictif, y compris le réseau de neurones [32]. On énumère ci-dessous les modèles les plus utilisés dans le domaine médical.

2.2.4.1 Régression linéaire

Dans l'imputation par régression, les variables existantes sont utilisées pour prédire, puis la valeur prédite est substituée comme si elle était une valeur réellement obtenue[33]. Cette approche présente plusieurs avantages car l'imputation conserve une grande quantité de données par rapport à la suppression par liste ou par paires et évite de modifier considérablement l'écart-type (i.e. dispersion des données autour de la moyenne). Cependant, comme dans une substitution moyenne, aucune information nouvelle n'est ajoutée, alors que la taille de l'échantillon peut être considérablement augmentée et que l'erreur-type (La mesure de l'erreur d'estimation) est réduite.

2.2.4.2 Random Forest

Random Forest est une méthode d'imputation non paramétrique applicable à divers types de variables qui a été aussi appliquées aux données manquantes. Random Forest utilise plusieurs arbres de décision pour estimer les valeurs manquantes et les sorties OOB (out of the bag) qui sont des estimations des erreurs d'imputation. La méthode Random Forest fonctionne mieux avec les grands ensembles de données, et l'utilisation de Random forest sur les petits ensembles de données peut conduire à des sur-ajustements [34].

2.2.4.3 k-NN (k Nearest Neighbour)

k-NN impute les valeurs d'attribut manquantes en fonction du nombre des voisins K les plus proches. Les voisins sont déterminés en fonction d'une mesure de distance. Une fois que K voisins sont déterminés, la valeur manquante est imputée en prenant la moyenne/médiane ou le mode des valeurs d'attribut connues de l'attribut manquant [25].

2.2.4.4 Maximum likelihood

Plusieurs stratégies utilisent la méthode de la probabilité maximale pour traiter les données manquantes[35]. L'hypothèse est que les données observées sont un échantillon tiré d'une distribution normale multivariée. Une fois les paramètres estimés à l'aide des données disponibles, les données manquantes sont estimées sur la base des paramètres qui viennent d'être estimés.

2.2.4.5 Expectation-Maximization

La maximisation des attentes (EM) est la méthode de probabilité maximale utilisée pour créer un nouvel ensemble de données. Toutes les valeurs manquantes sont imputées avec des valeurs estimées par les méthodes de probabilité maximale. Cette approche commence par l'étape d'attente, au cours de laquelle les paramètres (variances, covariances et moyennes) sont estimés. Ces estimations sont ensuite utilisées pour créer une équation de régression afin de prédire les données manquantes. L'étape de maximisation utilise ces équations pour remplir les données manquantes. L'étape d'attente est ensuite répétée avec les nouveaux paramètres, où les nouvelles équations de régression sont déterminées pour « remplir » les données manquantes. Ces étapes sont répétées jusqu'à ce que le système se stabilise [36].

2.2.4.6 Analyse de sensibilité

L'analyse de sensibilité est définie comme l'étude de l'incertitude de la sortie d'un modèle et qui peut être attribuée aux différentes sources d'incertitude dans ses intrants. Lors de l'analyse des données manquantes, des hypothèses supplémentaires sur les données manquantes sont faites, et ces hypothèses sont souvent applicables à une analyse primaire[37]. Cependant, l'exactitude des hypothèses ne peut pas être définitivement validée.

2.2.4.7 Réseau de neurones multicouches (Multi-layer perceptron MLP)

Un MLP (Multi-layer perceptron) se compose de plusieurs couches d'unités de calcul interconnectées de manière anticipée. Chaque unité dans une couche est directement connectée aux neurones de la couche suivante[38]. L'architecture MLP standard est composée de deux couches de poids. Les poids de la première couche relient les variables de données d'entrée aux unités cachées (neurones) et les poids de la deuxième couche connectent ces neurones cachés aux unités de sortie[38]. Les réseaux MLP ont été utilisés pour estimer les valeurs manquantes en entraînant un MLP pour apprendre les fonctionnalités incomplètes (utilisées comme sorties), en utilisant les caractéristiques complètes restantes comme entrées[38]. L'approche MLP peut être un outil utile pour reconstruire les valeurs manquantes. Cependant, son principal inconvénient est que lorsque les éléments manquants apparaissent dans plusieurs combinaisons d'attributs dans un problème de grande dimension, de nombreux modèles MLP doivent être construits[38].

2.3 Table de comparaison et Discussion

TABLE 2.1 – Table de comparaison des méthodes existantes

Méthode	Auteur	Avantages	Inconvénients
Suppression Par Listes (List-wise Deletion)	Strobl, E.,et al (2018)	-Une stratégie raisonnable s'il y a un échantillon suffisamment grand.	-N'est pas la stratégie optimale lorsqu'il n'y a pas un grand échantillon.
Suppression par paire (Pairwise Deletion)	Weaver, B.,et al (2014)	-S'il manque des données ailleurs dans l'ensemble de données, les valeurs existantes sont utilisées. - Utilise toutes les informations observées. -Conserve plus d'informations que la suppression par liste.	-S'il manque beaucoup d'observations, l'analyse sera déficiente. -On ne peut pas comparer les analyses car elles sont différentes à chaque fois.
Suppression des Variables (Dropping Variables)	Janssen,K. J et al (2010)	-Il n'y a pas de règle empirique.	-Peut dégrader la performance du modèle.

TABLE 2.1 – Table de comparaison des méthodes existantes

Méthode	Auteur	Avantages	Inconvénients
Moyenne	P, Kumar et. al (2021)	- Simple et facile à réaliser -Ne modifier pas la moyenne de la variable dans l'échantillon.	-Peut conduire a une incohérence. -La distortion de la variance d'origine.
Médiane	Madhu, G.et al. (2020)	-Simple et facile à réaliser	-L adistortion de la covariance avec les variables restans
Mode	Rani,P et al (2021)		
Dernière observation reportée (Last Observation Carried Forward (LOCF))	Dimitrakopoulou, V.et al (2015)	-Facile à comprendre et à communiquer. -Simple à réaliser. -Suppose fortement que la valeur du résultat reste inchangée par les données manquantes.	dans l'ensemble de données -Ne fonctionne pas dans de nombreux contextes.

TABLE 2.1 – Table de comparaison des méthodes existantes

Méthode	Auteur	Avantages	Inconvénients
Interpolation Linéaire	Daberdaku,S.et al. (2020)	-Donne des bons résultats avec des modèles qui ont des données non-complexes.	-Pourrait ne pas fonctionner dans un modèle assez complexe.
Imputation par point commun	Choi, J.et al. (2018)	-Simple et facile à réaliser -Plus approprié pour les valeurs ordinales.	-La distorsion de la variance d'origine. -La distorsion de la covariance avec les variables restantes dans l'ensemble de données.
Imputation par Catégorie Fréquente	Kunzmann, K., et al (2021)	-Facile à communiquer et simple à réaliser.	
Ajout d'une catégorie	Andrew Jet al (2018).	- Donne des bons résultats lorsque le nombre des données manquantes est élevé.	-Donne des mauvais résultats quand les données manquantes ne sont pas nombreuses.

TABLE 2.1 – Table de comparaison des méthodes existantes

Méthode	Auteur	Avantages	Inconvénients
Imputation Par Valeurs Arbitraires	Zhang, Z. (2016)	-Fonctionne raisonnablement bien pour les caractéristiques numériques principalement positives en valeur et pour les modèles arborescents.	-Peut affecter la performance d'un modèle si les données manquantes sont nombreuses.
Imputation Par Échantillonnage aléatoire	Giganti, M. J. et al. (2020)	-Préserve les paramètres statistiques de la variable d'origine	-Peut conduire à des incohérences.
MICE	Bartlett, J. W. et al (2020)	-Très flexible et peut gérer des données manquantes de différents types. -Crée des Datasets complètes.	-Difficile à communiquer. -Prend les imputations incertaines en compte.
Régression Linéaire	Fedushko, S. et al (2019)	-Conserve une grande quantité de données. - Évite de modifier considérablement l'écart-type ou la forme de la distribution.	-La véritable distribution du prédicteur est généralement inconnue et nécessite des hypothèses.

TABLE 2.1 – Table de comparaison des méthodes existantes

Méthode	Auteur	Avantages	Inconvénients
Random Forest	Yang, B.et al (2018)	-Fonctionne mieux avec les grands ensembles de données,	-L'utilisation sur les petits ensembles de données donne un sur-ajustement
k-NN (k Nearest Neighbour)	Daberdaku,S.et al. (2020)	-Peut prédire les attributs qualitatifs et les attributs quantitatifs. -Il n'est pas nécessaire de créer un modèle prédictif pour chaque attribut avec des données manquantes.	-La performance est affectée quand la valeur de 'k' est grande et pour les grands ensembles des données.
Maximum likelihood	Tomita, H.et al (2018)	-Facile à réaliser	-Un peu compliquer à interpréter
Expectation-Maximization	Huang, S. F.et al (2020)		-Peut donner des estimations hors rang

TABLE 2.1 – Table de comparaison des méthodes existantes

Méthode	Auteur	Avantages	Inconvénients
Analyse des sensibilités	Cro,S.Morris et al (2020)	-Conserve l'intégralité de la Dataset	-Difficile à réaliser - L'exactitude des hypothèses ne peut pas être définitivement validée
Réseau de Neurones Multicouches (MLP)	Cheng, C. Y.et al (2020).	-Fournir des prédictions rapides après avoir été entraîné. -Donne des bonnes précisions.	-Lorsque les données manquantes sont plusieurs dans un grand Dataset, de nombreux modèles MLP doivent être construits. - Temps de réponse élevé.

2.4 Conclusion

Après la synthèse des travaux dans le domaine (Missing Medical Data) on remarque que les recherches les plus importantes s'insère dans l'axes des méthodes analytiques qui présentes plusieurs atouts. C'est pourquoi notre modèle proposé s'insère dans cet axe et sera présenter dans le chapitre suivant .

CHAPITRE 3

CONCEPTION ET IMPLÉMENTATION

3.1 Introduction

L'objectif de ce chapitre est de présenter les étapes de conception de notre modèle d'imputation des données manquantes dans les « Big Data » Médicales. En fait, on a proposé une approche, qui est basée sur l'algorithme d'apprentissage automatique (K-means) combiné avec une méthode de « Feature Selection » pour choisir les meilleurs caractéristiques afin de rendre le modèle plus performant . Finalement l'algorithme k-NN a été utilisé pour l'imputation des données manquantes . Dans la section suivante, on expose l'architecture du modèle proposé suivi par l'implémentation et les résultats obtenus.

3.2 Conception

3.2.1 Architecture proposée

Pour imputer les données manquantes, plusieurs étapes ont été suivies : initialement, on fait entrer les individus dans la fonction de classification générée par le modèle (K-means) pour affecté chaque individu qui contient des données manquantes

au cluster auquel il appartient. Tous les individus avec des données manquantes seront imputés en utilisant le modèle k-NN proposé. L'architecture du modèle proposé est représenté dans la figure suivante puis détaillée dans la section suivante.

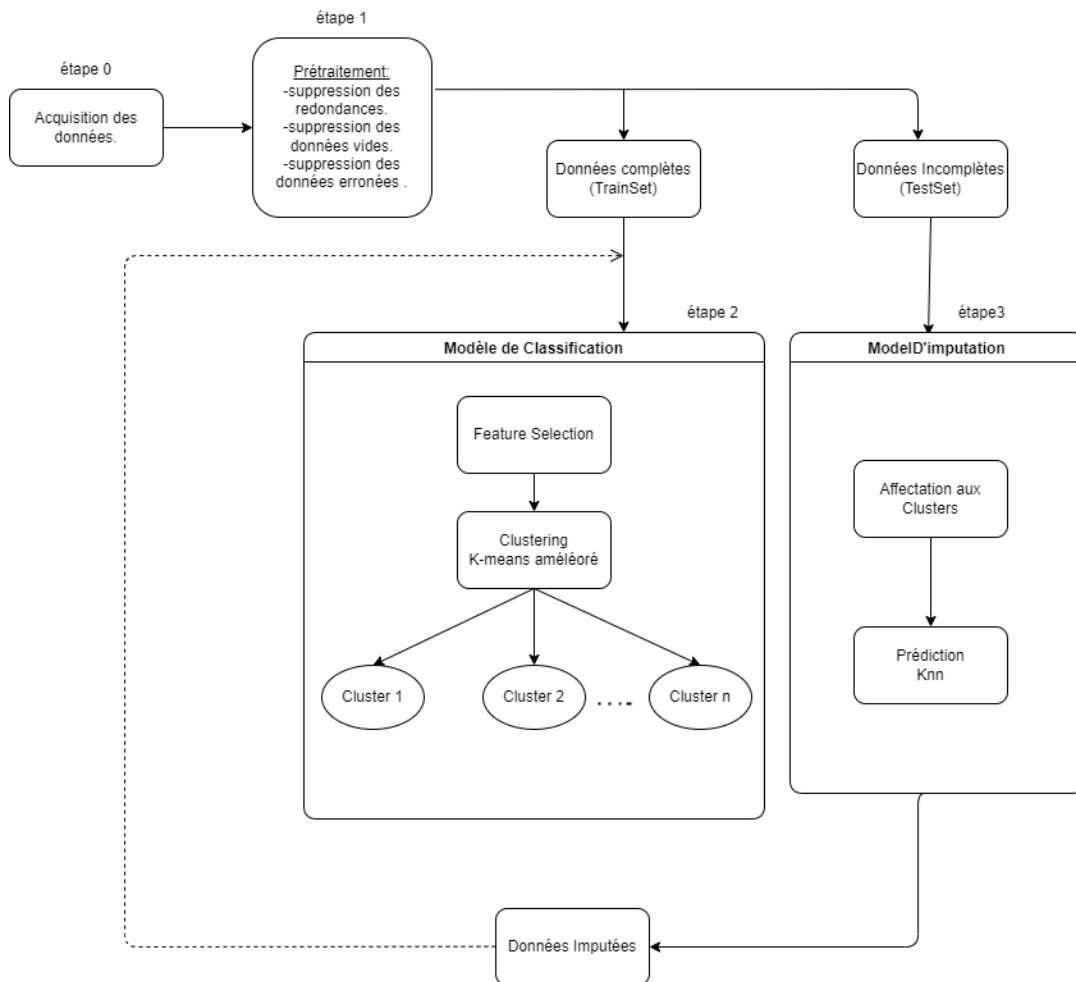


FIGURE 3.1 – Architecture proposée

3.2.2 Acquisition des données

L'acquisition des données est la première étape essentielle de l'accès au soins de la santé. Elle se fait avec des dispositifs médicaux, qui englobe tout article, instruments, appareils ou équipements utilisés pour prévenir, diagnostiquer ou traiter une affection ou une maladie[39].

3.2.3 Prétraitement

L'étape de prétraitement (nettoyage de données) est effectuée comme suit :

- Phase 1 : Suppression des lignes qui contiennent beaucoup de données manquantes (80%) en utilisant la méthode de Suppression par listes(list-wise deletion) .
- Phase 2 : Suppression des redondances (données qui se répètent).
- Phase 3 :Suppression des données erronées (Valeurs dépassant le seuil reconnu pour les variables).

Ce prétraitement améliorera amplement la qualité des données utilisées dans le modèle. Une fois le prétraitement terminé, le dataset obtenu est divisé en deux parties : Partie apprentissage(Training Set : 80%) et Partie Test (Test set :20%).

3.2.4 Modèle de classification

Le modèle de classification que nous avons proposé comporte deux étapes La sélection des caractéristiques (Feature Selection) et l'étape du K-means :

3.2.4.1 Feature selection

La sélection des caractéristiques ou (Feature selection) peut être définie comme le processus de sélection d'un sous-ensemble d'entités (ou de variables) pertinentes à partir d'un ensemble de données [40].

En d'autres termes, la sélection des caractéristiques peut être considérée comme un cas particulier de réduction de la dimensionnalité, visant à réduire le nombre de variables aléatoires : Réduire le nombre de variables aléatoires est utile pour donner un ensemble de variables primaires [40].

La méthode la plus utilisée de « Feature Selection » est la méthode de filtrage[40]. Elle est basée sur des techniques de classement. Plus précisément, les caractéristiques d'entrée sont classés à l'aide de « La variance » où les caractéristiques supérieures à un certain seuil sont supprimées. De nombreuses techniques statistiques appartiennent

au type de méthodes de filtrage, y compris le gain d'information et la régression pas à pas [40].

La méthode que nous avons utilisé (la plus connue) est « Variance_Inflation_Factor (VIF) ». Le facteur d'inflation de la variance est une mesure de l'augmentation de la variance des estimations des paramètres (mesure servant à caractériser la dispersion d'une distribution ou d'un échantillon.) [41]

Algorithm 1 Variance Inflation Factor

Entrées :

1. Ensemble de N données, noté par x ;

Sortie :

2. Ensemble de N données réduit;

Début :

3. Calculer la variance de tous les axes;

4. Supprimer les axes qui ont les valeurs les plus élevés;

5. Garder les axes qui ont les moins de variances;

Fin.

3.2.4.2 Méthode améliorée du K-means

Le clustering K-means est l'un des simples algorithmes d'apprentissage automatique non supervisés. Les algorithmes non supervisés font des inférences à partir d'ensemble de données en utilisant uniquement des vecteurs d'entrée sans se référer à des résultats connus ou étiquetés. Le principe du K-means est le suivant : L'algorithme identifie K nombres de centroïdes, puis alloue chaque point de données au cluster le plus proche, tout en gardant les centroïdes aussi petits que possible. Les « moyennes » dans l'algorithme présenté ci-dessous font référence à la moyenne des données, c'est-à-dire à la recherche du centroïde [42].

Algorithm 2 K-means**Entrées :**

1. Ensemble de N données, noté par x
2. Nombre de groupes souhaité, noté par k

Sortie :

3. Une partition de K groupes (C_1, C_2, \dots, C_k)

Début :

4. Initialisation aléatoire des centres C_K ;

Répéter

4. Affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche;

$$X_i \in C_k \text{ si } \forall j |x_i - \mu_k| = \min_j |x_i - \mu_j| \quad (1)$$

Avec μ_k le centre de la classe K;

5. Représentation : Calculer les centres associée à la nouvelle partition;

$$\mu_k = \frac{1}{N} \sum_{x_i \in C_k} x_i \quad (2)$$

Jusqu'à convergence de l'algorithme vers une partition stable;

Fin.

La faiblesse de l'algorithme K-means reste à trouver le nombre initial du cluster « K » qui conditionne le résultat final. Pour y remédier, nous avons utilisé la méthode Elbow [43] pour générer la valeur optimale du nombre des clusters « K ». La méthode Elbow est utilisée avec une variable, WCSS (Within-Cluster Sum-of-Squares), qui mesure la variance au sein de chaque cluster [44].

La variance des clusters se calcule comme suit :

$$V = \sum_j \sum_{x_i \rightarrow C_j} D(c_j, x_i)^2 \quad (3.1)$$

Avec :

c_j : Le centre du cluster.

x_i : La i ème observation dans le cluster

$D(c_j, x_i)$: La distance euclidienne entre le centre du cluster c_j et x_i le point Après avoir utilisé la méthode Elbow, le nombre de clusters est obtenu et leurs centres sont calculés. La figure 3.2 montre l'architecture du K-means améliorée par la méthode Elbow.

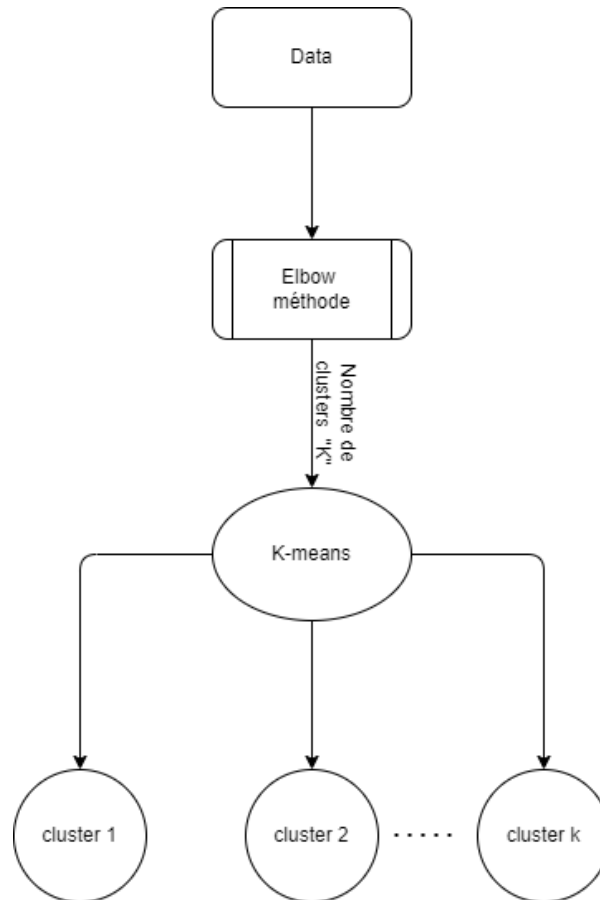


FIGURE 3.2 – architecture K-means Elbow

Le modèle de classification est suivi du modèle d'imputation proposé

3.2.5 Modèle d'Imputation

Notre modèle proposé d'imputation des données manquantes comporte 3 étapes : En premier, on affecte chaque individu contenant des données manquantes au cluster approprié selon la fonction de classification proposée. Puis, on impute les données

manquantes par la moyenne des données appartenant au « K » plus proche voisins en utilisant l'algorithme k-NN.

3.2.5.1 Algorithme k-NN

En classification, l'algorithme supervisé du « k » plus proche voisin (ou k-NN) est une technique de classification d'objets basée sur l'exemple d'apprentissage le plus proche dans l'espace du problème. k-NN est un type d'apprentissage basé sur les instances ou d'apprentissage paresseux où la fonction n'est approchée que localement et tous les calculs sont différés jusqu'à la classification. L'algorithme du « k » plus proche voisin est le plus simple de tous les algorithmes d'apprentissage automatique : L'objet est affecté à la classe la plus courante parmi ses k plus proches voisins (k est un entier positif, généralement petit)

Algorithm 3 Imputation par k-NN

Entrées :

1. Un ensemble de données
2. Un nombre entier k

Sortie :

3. Données manquantes imputées

Début :

Pour une nouvelle observation dont on veut prédire sa variable de sortie Faire :

4. Calculer toutes les distances de cette observation avec les valeurs de l'ensemble des données ;
5. Retenir les k valeurs du jeu de données les plus proches de cette observation ;
6. Choisir la moyenne des valeurs des k valeurs précédentes ;
7. Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par k-NN pour la donnée manquante ;

Fin.

3.3 Implémentation

Dans cette étape, on décrit en premier le matériel, logiciel et DataSet utilisés dans l'implémentation du modèle proposé. Les résultats obtenus seront aussi présentés et discutés dans les sections suivantes.

3.3.1 Matériels utilisés :

L'implémentation de notre système a été réalisée sur une machine possédant les caractéristiques suivantes avec les logiciels présentés ci-dessous :

Processeur : I5

Mémoire :4.00 Go

Disque dur :500 GB

3.3.2 Logiciels utilisés :

a) Système d'exploitation : Linux Ubuntu

b) Outils de développement :

Python version 3.7 : Python est un langage de programmation de haut niveau avec une syntaxe simple et une puissance remarquable.

Les bibliothèques connexes sont : pyspark, pandas, matplotlib, statsmodels ,sklearn,numpy

c) Apache Spark version 2.4.6 : il permet d'effectuer des traitements sur de large volume de données.

3.3.3 Dataset Utilisé

Nous avons utilisé un Dataset médical « Pima » qui représente un ensemble de données de « Diabète », provenant du « National Institute of Diabetes and Digestive

and Kidney Diseases » Ce Dataset contient des informations sur des femmes issues d'une population proche de Phoenix, en Arizona, aux États-Unis et disponible sur :¹

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	4	129	70	18	122	29.43	1.17	45
1	1	205	76	36	249	37.28	0.92	29
2	8	97	82	0	0	37.82	0.59	68
3	7	141	90	41	0	34.25	0.40	39
4	4	120	72	0	0	29.12	0.39	46
...
77563	3	91	58	11	54	26.26	0.27	22
77564	2	112	62	32	56	26.40	0.13	21
77565	4	128	68	0	0	36.47	0.40	29
77566	1	101	68	21	0	28.56	1.11	22
77567	9	169	88	0	0	31.13	0.32	49

77568 rows x 8 columns

FIGURE 3.3 – Dataset Utilisé

3.3.4 Modélisation d'exécution avec Plateforme Spark

3.3.4.1 Lire CSV en RDD

Après avoir installé tout les logiciels requis, Spark procède à la lecture du Dataset en utilisant l'outil RDD tel qu'il est présenté dans la figures suivante :

1. <https://www.kaggle.com/datasets/pradeepgurav>

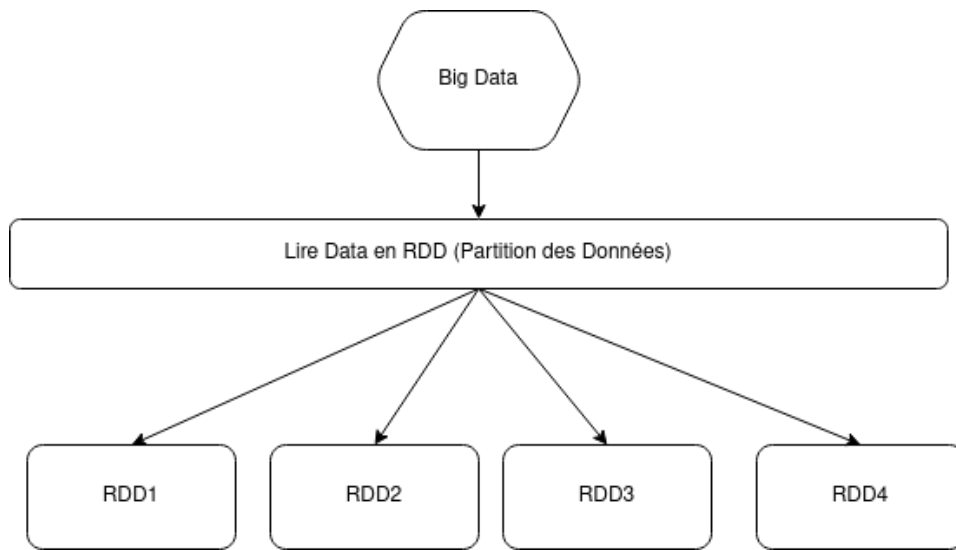


FIGURE 3.4 – Lecture du Dataset en RDD

Spark est généralement composé de deux programmes suivants :

- Le programme "Pilot" (SparkContext).
- Le programme "Esclaves".

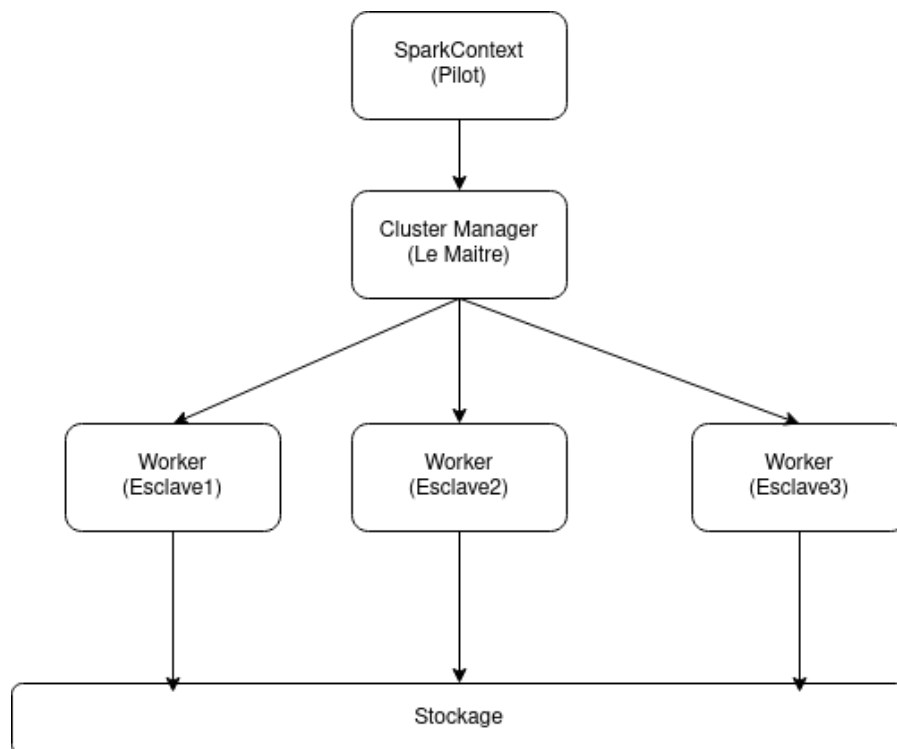


FIGURE 3.5 – Fonctionnement De Spark

3.3.4.2 Création du maitre

Au départ, il est nécessaire de créer un "Maitre" pour gérer les "Esclaves". La Création du "Maitre" se fait par le "pilote", ce dernier se connecte au "cluster manager" pour créer un "Maitre" et avoir ainsi une session Spark.



FIGURE 3.6 – Création du Maitre

3.3.4.3 Chargement du dataset

Après sa création, le "Maître" ordonne aux "Esclaves" de lire le Dataset ,en utilisant la lecture distribuée sur les différents nœuds des clusters .Les "Esclaves" créent des tâches pour lire le fichier. Chaque "Esclave" a accès à la mémoire du nœud et attribue une partition de mémoire à la tâche.

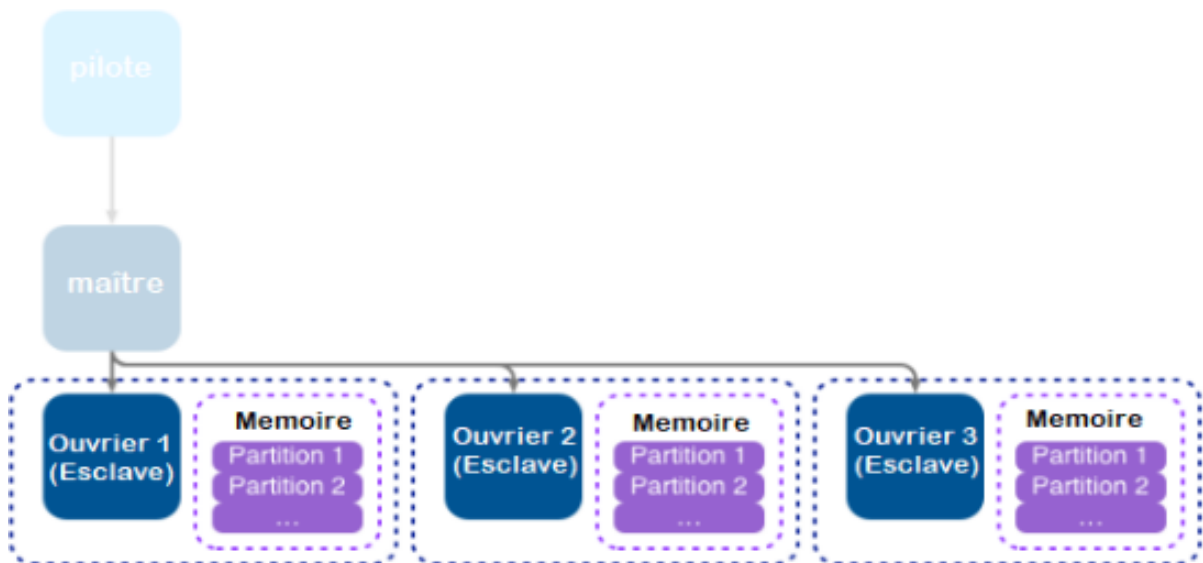


FIGURE 3.7 – Chargement du dataset

3.3.5 Modèle de classification

3.3.5.1 Feature selection

La méthode « Variance Inflation Factor VIF », utilisée sur le dataset lu, calcule la variance selon la formule suivante

$$V = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (3.2)$$

avec :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (3.3)$$

Cette étape est suivie par la suppression des axes qui ont une variance supérieur a 5 (valeur empirique par défaut) Les axes retenus(Pregnancies,SkinThickness,Insulin,DiabetsPedigreeFunction) seront les variables les plus importantes (pour la prédiction) et qui seront utilisé dans le modèle proposé (Figure 3.8)

	Pregnancies	SkinThickness	Insulin	DiabetesPedigreeFunction	cluster_pred
Unnamed: 0					
0	4.0	18.0	122.0	1.17	2
1	1.0	36.0	249.0	0.92	3
5	5.0	41.0	42.0	0.16	2
6	1.0	23.0	94.0	0.17	2
12	1.0	14.0	415.0	0.41	3
...
75288	1.0	32.0	156.0	0.70	0
75290	4.0	17.0	49.0	0.35	2
75295	1.0	39.0	110.0	1.09	2
75300	2.0	38.0	360.0	0.35	3
75301	7.0	40.0	105.0	0.21	2

32999 rows × 5 columns

FIGURE 3.8 – Dataset après l'utilisation de 'Feature Selection'

La figure 3.9 montre qu'il ne reste aucune corrélation entre les axes après l'utilisation de la méthode de sélection de caractéristiques ('Feature Selection').

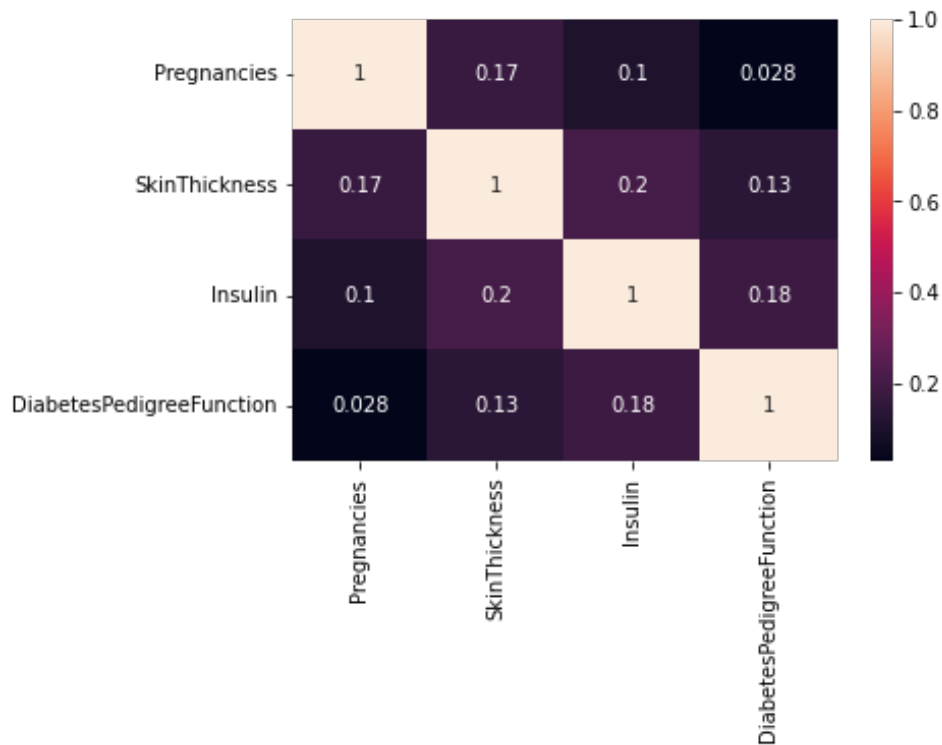


FIGURE 3.9 – Matrice de corrélation

Après avoir appliqué la méthode de sélection de caractéristiques (VIF) sur notre Dataset nous l'avons divisé en deux : 80% pour entraîner le modèle (Training set) et 20% pour tester les résultats (Test set). Le Test set est généré en éliminant aléatoirement des variables.

3.3.5.2 Elbow Méthode

La méthode Elbow a été appliquée sur le Dataset d'entraînement (Training set) pour générer le nombre de cluster K utile au modèle K-means présenté ci-dessous,

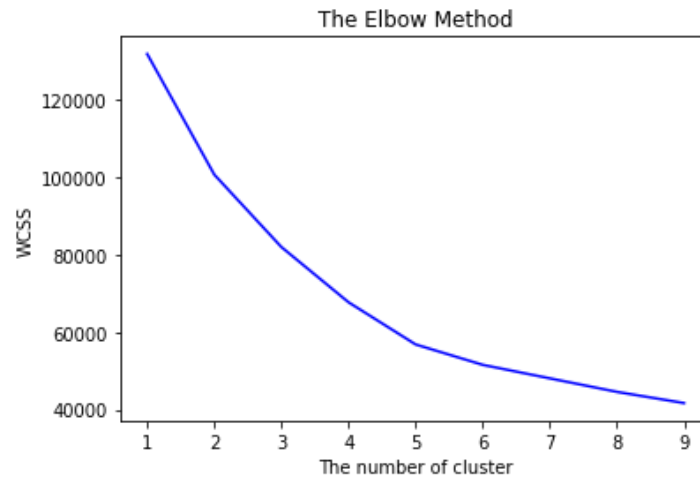


FIGURE 3.10 – Elbow Méthode

Comme le montre la figure 3.10 le nombre K est égale à 5 ($K=5$); représentant le coude (Elbow) de la courbe obtenue,

3.3.5.3 K-means

Nous avons regroupé les individu du Dataset d'entraînement avec le K-means en utilisant la valeur du K obtenue précédemment. La figure 3.11 représente les Clusters obtenus (5) après application du K-means.

Nous avons utilisé des couleurs pour visualiser les regroupements obtenus :

Rouge : Cluster 1

Noire : Cluster 2

Vert : Cluster 3

Bleu : Cluster 4

Jaune : Cluster 5

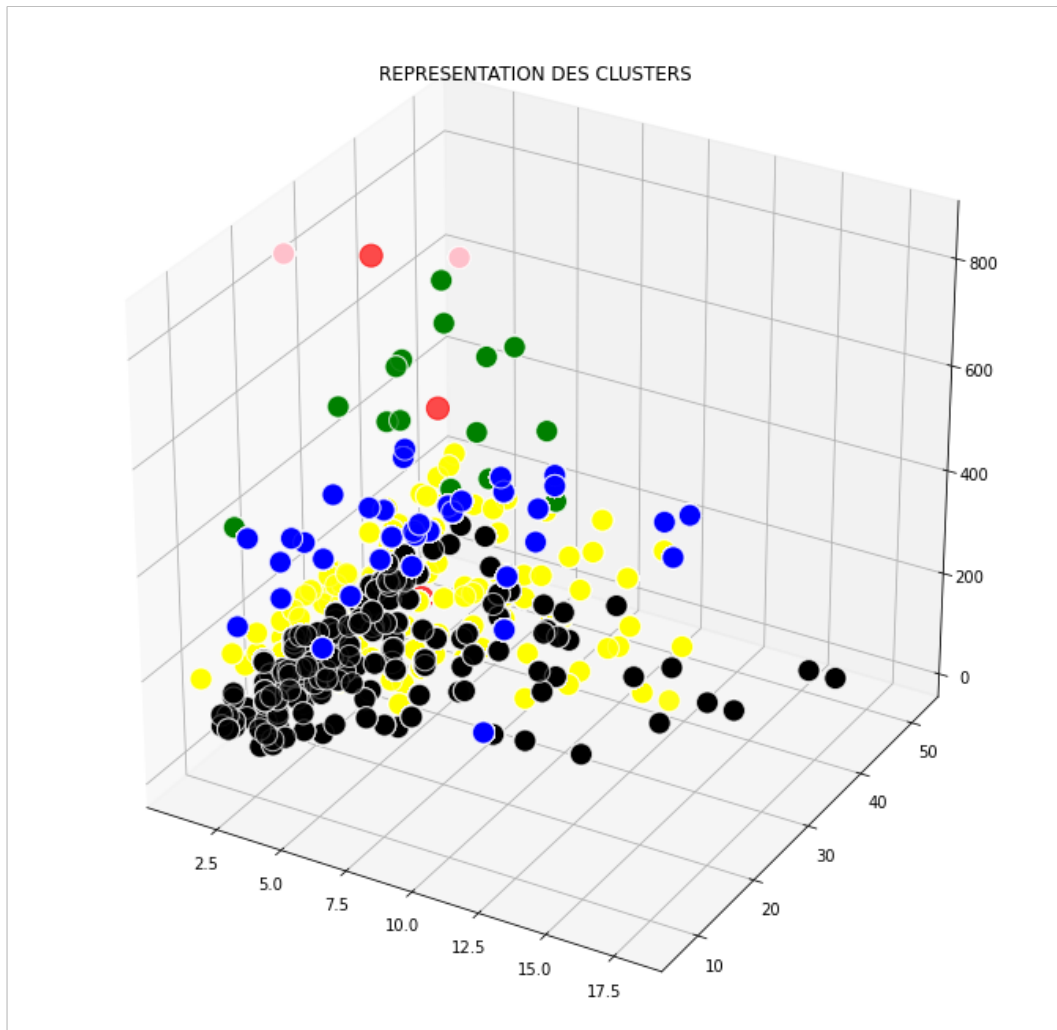
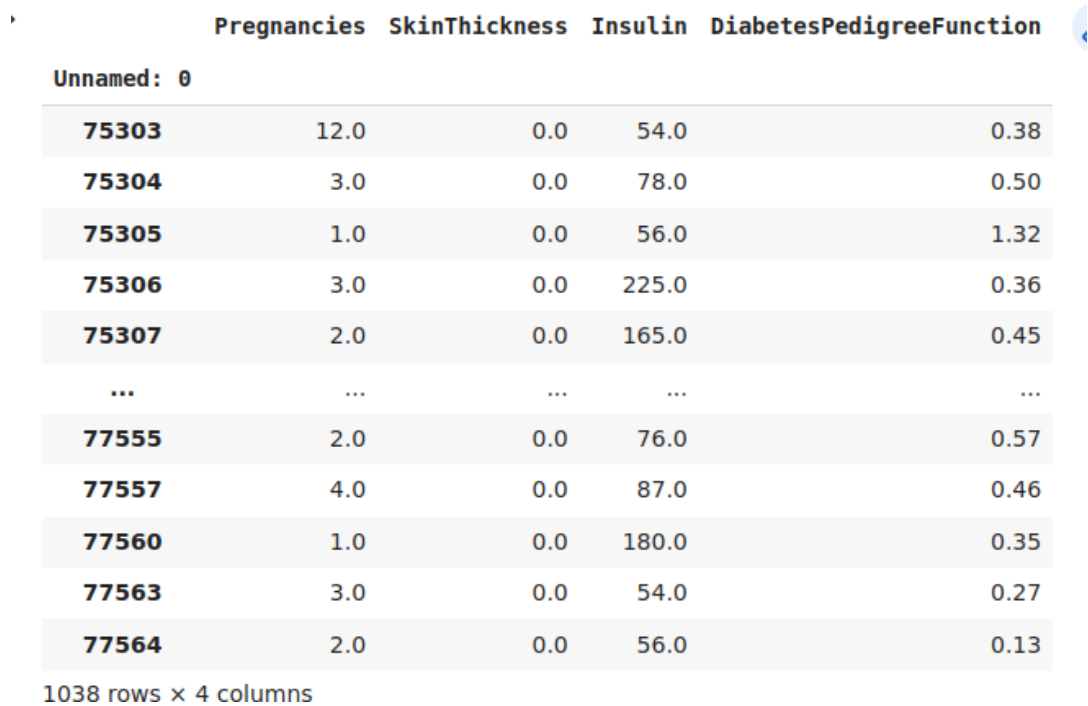


FIGURE 3.11 – Représentation de la Population dans les clusters

3.3.6 Modèle d'imputation

Cette étape présente les données générées après imputation avec la méthode k-NN dans le dataset de test (20% du dataset), La figure 3.13 présente l'affectation des individus du TestSet aux clusters appropriés (Couleur : gris), Les figures (3.12, 3.14) présentent le Test Set avant et après imputation.



	Pregnancies	SkinThickness	Insulin	DiabetesPedigreeFunction
Unnamed: 0				
75303	12.0	0.0	54.0	0.38
75304	3.0	0.0	78.0	0.50
75305	1.0	0.0	56.0	1.32
75306	3.0	0.0	225.0	0.36
75307	2.0	0.0	165.0	0.45
...
77555	2.0	0.0	76.0	0.57
77557	4.0	0.0	87.0	0.46
77560	1.0	0.0	180.0	0.35
77563	3.0	0.0	54.0	0.27
77564	2.0	0.0	56.0	0.13

1038 rows x 4 columns

FIGURE 3.12 – DataTest avant l'imputation

Après nous avons introduit les individus avec les individus qui ont des données complètes.

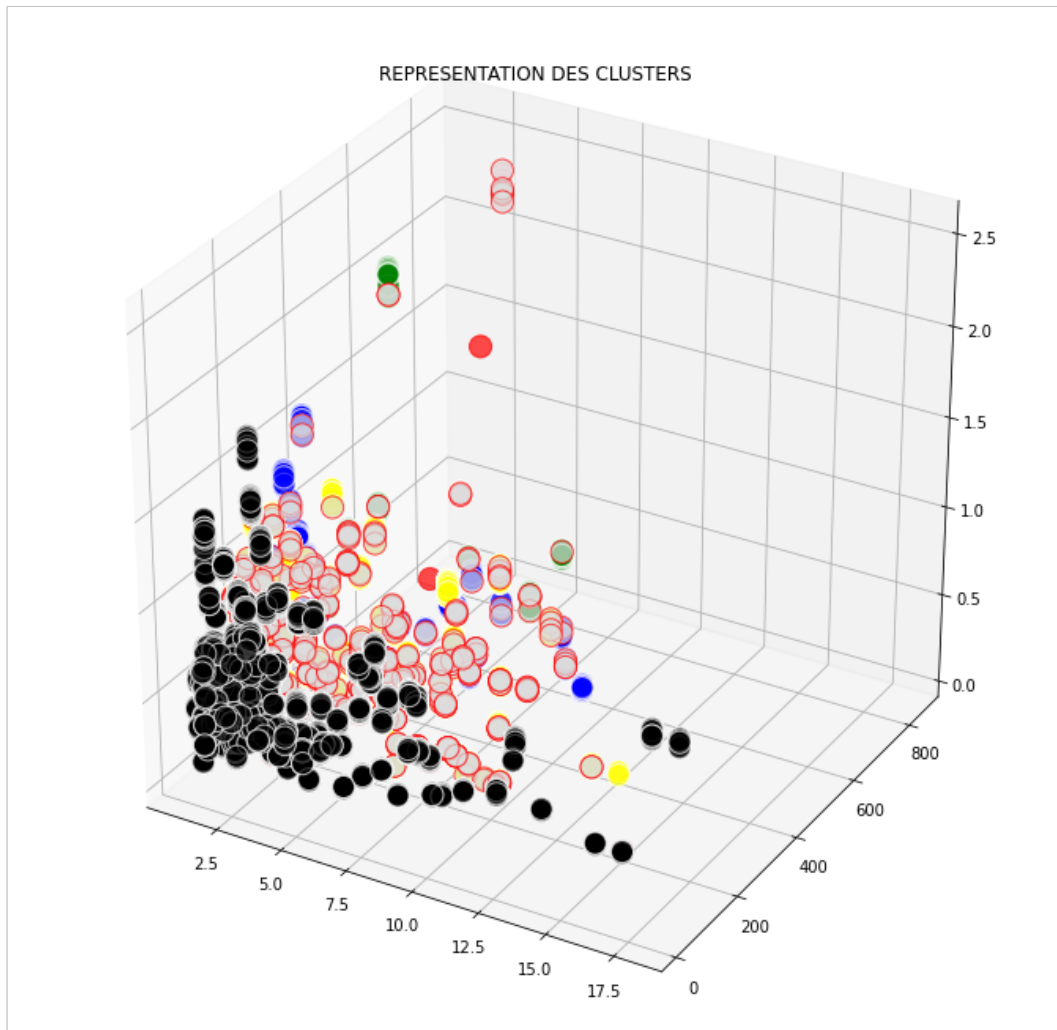


FIGURE 3.13 – Classification des individus avec des données manquantes dans les clusters

3.3.6.1 Imputation par K-nn

Après avoir classer les individus du dataset du test dans les custers appropriés, nous avons imputer les valeurs manquantes par la moyenne des valeurs des plus proches voisins.

↳

	Pregnancies	SkinThickness	Insulin	DiabetesPedigreeFunction
Unnamed: 0				
75303	12.0	40.0	54.0	0.38
75304	3.0	12.0	78.0	0.50
75305	1.0	30.0	56.0	1.32
75306	3.0	37.0	225.0	0.36
75307	2.0	27.0	165.0	0.45
...
77555	2.0	16.5	76.0	0.57
77557	4.0	12.0	87.0	0.46
77560	1.0	38.0	180.0	0.35
77563	3.0	11.0	54.0	0.27
77564	2.0	32.0	56.0	0.13

1038 rows x 4 columns

FIGURE 3.14 – DataTest après l'imputation

Comparativement aux données réel notre modèle permet de générer des valeurs très proches des valeurs réelles avec un écart jugé très satisfaisant

Une table contenant les valeurs réelles et imputées sur 500 individus (après suppression d'une colonne a la fois) permet de valoriser les resultats présentés dans le tableau 3.1

TABLE 3.1 – Table excel représentant les valeurs réelles et prédites

Index	Pregnancies	Skin Thickness	Insulin	Diabetes Pedigree Function	Pred Pregnancies	Pred SkinThickness
76471	1.0	11.0	60.0	0.54	1.0	11.0
76475	4.0	32.0	88.0	0.48	4.0	32.0
76476	1.0	18.0	58.0	0.26	1.0	18.0
76478	8.0	35.0	225.0	0.43	8.0	35.0
76480	2.0	23.0	50.0	0.53	2.0	23.0
76482	6.0	41.0	140.0	0.61	6.0	41.0
76483	1.0	15.0	140.0	0.49	1.0	15.0
76486	9.0	30.0	100.0	0.17	9.0	30.0
76487	6.0	32.0	190.0	0.34	6.0	32.0
76488	1.0	19.0	82.0	0.32	1.0	19.0
76490	2.0	52.0	57.0	0.7	2.0	52.0
76492	2.0	19.0	53.0	0.23	2.0	19.0
76495	6.0	30.0	120.0	0.49	6.0	30.0
76499	3.0	19.0	86.0	0.16	3.0	19.0

3.3.7 Le RMSE (Root Mean Square Error)

Le RMSE est la racine carrée de la variance des résidus. Cette valeur indique l'ajustement absolu du modèle aux données, c.à.d à quel point les points de données observés sont proches des valeurs prédites du modèle. La formule du RMSE est la suivante :

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (p_i - o_i)^2}{n}}$$

ou,

p est la valeur prédite.

TABLE 3.2 – Table de comparaison des résultats

	Prédiction pregnancies	Prédiction Skinthikness
Imputation par Notre modèle	0,031	0,73
Imputation par la moyenne	1,48	4,52

o est valeur observée.

n est le nombre d'observations.

Des valeurs faibles de RMSE indiquent un meilleur ajustement. Le RMSE est une bonne mesure de la précision avec laquelle le modèle prédit des items, et c'est le critère le plus important pour l'ajustement si l'objectif principal du modèle est la prédiction[45].

Nous avons calculer le RMSE des prédictions de notre modèle et nous l'avons comparer avec un autre modèle. Les résultats sont représenter dans le tableau 3.2

La Figure 3.15 presente une comparaison des résultats obtenus par le modèle proposé et la methode d'imputation par la moyenne (Mean).

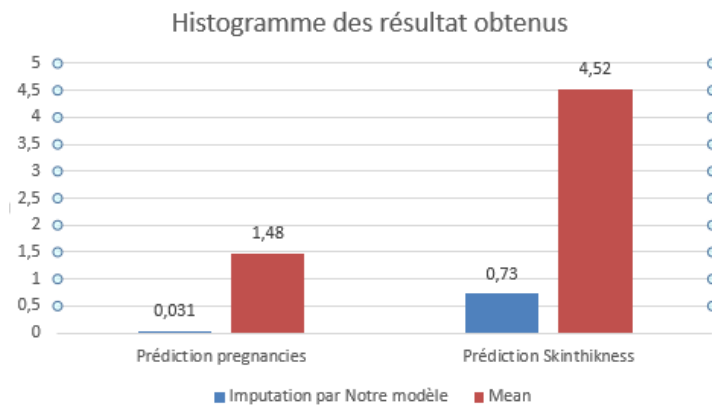


FIGURE 3.15 – Histogramme des résultats obtenus

3.4 Conclusion

Dans ce chapitre, nous avons exposé le modèle proposé pour l'imputation des données manquantes. Les étapes du modèle ont été ensuite détaillées. L'implémentation a permis de présenter les résultats obtenus qui s'avèrent très satisfaisants.

CONCLUSION GÉNÉRALE

Les méthodes analytiques sont de plus en plus utilisées pour les "Missing Data". Elles s'appuient principalement sur l'analyse statistique et l'exploration de données. Ces méthodes sont basées sur des techniques exploratoires et des algorithmes pour découvrir les relations qui relient les données afin de fournir des résultats fiables. L'application de ces méthodes nous permet de mieux comprendre les données qui nous entourent et d'améliorer les performances pour prédire des résultats.

Dans ce travail, la méthode du K-means amélioré par la méthode Elbow, a été proposé avec une méthode « Feature Selection ». Cette hybridation a permis d'apporter une amélioration aux résultats de classification. Ensuite la méthode K-nn a été utilisé pour prédire les valeurs des données manquantes.

Dans ce projet :

1. Un état de l'art sur les concepts du « Big Data » et « Big Data » médical a été présenté dans le premier chapitre.
2. Une étude des méthodes utilisées pour le traitement des données manquantes dans les « Big data » médical a été exposée dans le deuxième chapitre.

3. Dans le chapitre 3

- Une implémentation de la méthode de partitionnement K-means et « Feature Selection » ont été proposées.
- L'algorithme a été combiné avec la méthode du k-NN est validé sur des données massives réelles « Médicales ».
- L'utilisation du Framework Spark a été d'un grand apport dans le traitement de ces données massives.

En perspective, ce travail peut être complété par les point suivants :

- (1) Automatisation du choix des paramètres des méthodes utilisées.
- (2) D'autres Datasets peuvent faire aussi l'objet de bases de tests pour valider le système proposé.
- (3) Implémentation du GAN (Generative adversarial network) pour générer les données manquantes.

BIBLIOGRAPHIE

- [1] S.-T. PARK, Y.-R. KIM, S.-P. JEONG, C.-I. HONG et T.-G. KANG, « A case study on effective technique of distributed data storage for big data processing in the wireless internet environment, » *Wireless Personal Communications*, t. 86, n° 1, p. 239-253, 2016.
- [2] W. N. PRICE et I. G. COHEN, « Privacy in the age of medical big data, » *Nature medicine*, t. 25, n° 1, p. 37-43, 2019.
- [3] L. JAPEC, F. KREUTER, M. BERG et al., « Big data in survey research : AAPOR task force report, » *Public Opinion Quarterly*, t. 79, n° 4, p. 839-880, 2015.
- [4] B. VIDYAVATHI et al., « Security Challenges in Big Data, » *International Journal of Advanced Research in Computer Science*, t. 6, n° 6, 2015.
- [5] V. PADMAPRIYA, J. AMUDHAVEL, V. GOWRI, K. LAKSHMIPRIYA, S. VINOETHINI et K. P. KUMAR, « Demystifying challenges, opportunities and issues of Big data frameworks, » in *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, 2015, p. 1-5.
- [6] N. KLEIN, *Edge computing : Comprendre son impact sur l'IOT*, nov. 2021. adresse : <https://openest.io/iotsmartobjects/edge-iot/>.

- [7] M. K. HASSAN, A. I. EL DESOUKY, S. M. ELGHAMRAWY et A. M. SARHAN, « Big data challenges and opportunities in healthcare informatics and smart hospitals, » *Security in smart cities : Models, applications, and challenges*, p. 3-26, 2019.
- [8] M. AL-MEKHLAL et A. A. KHWAJA, « A Synthesis of Big Data Definition and Characteristics, » in *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, IEEE, 2019, p. 314-322.
- [9] J. RANJAN, « The 10 Vs of Big Data framework in the Context of 5 Industry Verticals., » *Productivity*, t. 59, n° 4, 2019.
- [10] B. YADRANJIAGHDAM, N. POOL et N. TABRIZI, « A survey on real-time big data analytics : applications and tools, » in *2016 international conference on computational science and computational intelligence (CSCI)*, IEEE, 2016, p. 404-409.
- [11] *The 5V of big data : A basic guide*, mars 2021. adresse : <https://www.jigsawacademy.com/blogs/big-data/5v-of-big-data/>.
- [12] *Big Data and Medical Research*, août 2020. adresse : <https://insidebigdata.com/2020/08/08/big-data-and-medical-research/>.
- [13] L. HONG, M. LUO, R. WANG, P. LU, W. LU et L. LU, « Big data in health care : Applications and challenges, » *Data and information management*, t. 2, n° 3, p. 175-197, 2018.
- [14] P. da VITO LAVECCHIA LAVECCHIA VITO INGEGNERE INFORMATICO (POLITECNICO DI BARI) EMAIL : [EMAILNOSP;PROTECTED] SITO WEB : [HTTPS://VITOLAVECCHIA.ALTERVISTA.ORG](https://vitolavecchia.altervista.org) MOSTRA ALTRI ARTICOLI, P. d. V. LAVECCHIA, L. V. I. (di BARI) EMAIL : [EMAILNOSP;PROTECTED] SITO WEB : [HTTPS://VITOLAVECCHIA.ALTERVISTA.ORG](https://vitolavecchia.altervista.org) MOSTRA ALTRI ARTICOLI et M. a. ARTICOLI, *Differenza Tra Dati Strutturati, Non Strutturati e semi-strutturati*, fév. 2021. adresse : <https://vitolavecchia.altervista.org/differenza-tra-dati-strutturati-non-strutturati-e-semi-strutturati/>.

- [15] Z. ZHANG, « Missing data exploration : highlighting graphical presentation of missing pattern, » *Annals of translational medicine*, t. 3, n° 22, 2015.
- [16] *Hadoop - Architecture*, fév. 2022. adresse : <https://www.geeksforgeeks.org/hadoop-architecture/>.
- [17] L. SFAXI, *P2 - introduction à Apache Spark*. adresse : <https://liliasfaxi.github.io/Atelier-Spark/p2-spark/>.
- [18] S. P. M. SAYS : *Apache Spark Architecture : Distributed System Architecture explained*, mars 2022. adresse : <https://www.edureka.co/blog/spark-architecture/>.
- [19] E. V. STROBL, S. VISWESWARAN et P. L. SPIRITES, « Fast causal inference with non-random missingness by test-wise deletion, » *International journal of data science and analytics*, t. 6, n° 1, p. 47-62, 2018.
- [20] B. WEAVER et R. KOOPMAN, « An SPSS macro to compute confidence intervals for Pearson's correlation, » *The Quantitative Methods for Psychology*, t. 10, n° 1, p. 29-39, 2014.
- [21] K. J. JANSSEN, A. R. T. DONDEERS, F. E. HARRELL JR et al., « Missing covariate data in medical research : to impute is better than to ignore, » *Journal of clinical epidemiology*, t. 63, n° 7, p. 721-727, 2010.
- [22] P. RANI, R. KUMAR et A. JAIN, « HIOC : a hybrid imputation method to predict missing values in medical datasets, » *International Journal of Intelligent Computing and Cybernetics*, 2021.
- [23] G. MADHU, B. LALITH BHARADWAJ, K. SAI VARDHAN et G. NAGA CHANDRIKA, « A normalized mean algorithm for imputation of missing data values in medical databases, » in *Innovations in Electronics and Communication Engineering*, Springer, 2020, p. 773-781.
- [24] V. DIMITRAKOPOULOU, O. EFTHIMIOU, S. LEUCHT et G. SALANTI, « Accounting for uncertainty due to 'last observation carried forward' outcome imputation in a meta-analysis model, » *Statistics in Medicine*, t. 34, n° 5, p. 742-752, 2015.

- [25] S. DABERDAKU, E. TAVAZZI et B. DI CAMILLO, « A combined interpolation and weighted K-nearest neighbours approach for the imputation of longitudinal ICU laboratory data, » *Journal of Healthcare Informatics Research*, t. 4, n° 2, p. 174-188, 2020.
- [26] J. CHOI, J. CHOI et H.-T. JUNG, « Applying machine-learning techniques to build self-reported depression prediction models, » *CIN : Computers, Informatics, Nursing*, t. 36, n° 7, p. 317-321, 2018.
- [27] A. J. STEELE, S. C. DENAXAS, A. D. SHAH, H. HEMINGWAY et N. M. LUSCOMBE, « Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease, » *PloS one*, t. 13, n° 8, e0202344, 2018.
- [28] K. KUNZMANN, L. WERNISCH, S. RICHARDSON et al., « Imputation of ordinal outcomes : a comparison of approaches in traumatic brain injury, » *Journal of neurotrauma*, t. 38, n° 4, p. 455-463, 2021.
- [29] Z. ZHANG, « Multiple imputation for time series data with Amelia package, » *Annals of translational medicine*, t. 4, n° 3, 2016.
- [30] M. J. GIGANTI, P. A. SHAW, G. CHEN et al., « Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples, and multiple imputation, » *The annals of applied statistics*, t. 14, n° 2, p. 1045, 2020.
- [31] J. W. BARTLETT et R. A. HUGHES, « Bootstrap inference for multiple imputation under uncongeniality and misspecification, » *Statistical methods in medical research*, t. 29, n° 12, p. 3533-3546, 2020.
- [32] Y. FATOUMATA, A. ADNANE et Z. ATAOUA, « Machine Learning Pour La Maintenance Prédictive., » 2021.

- [33] S. FEDUSHKO, T. USTYIANOVYCH et al., « Medical card data imputation and patient psychological and behavioral profile construction, » *Procedia Computer Science*, t. 160, p. 354-361, 2019.
- [34] B. YANG, G. DAI, Y. YANG et al., « Automatic text classification for label imputation of medical diagnosis notes based on random forest, » in *International Conference on Health Information Science*, Springer, 2018, p. 87-97.
- [35] H. TOMITA, H. FUJISAWA et M. HENMI, « A bias-corrected estimator in multiple imputation for missing data, » *Statistics in Medicine*, t. 37, n° 23, p. 3373-3386, 2018.
- [36] S.-F. HUANG et C.-H. CHENG, « A Safe-region imputation method for handling medical data with missing values, » *Symmetry*, t. 12, n° 11, p. 1792, 2020.
- [37] S. CRO, T. P. MORRIS, M. G. KENWARD et J. R. CARPENTER, « Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation : a practical guide, » *Statistics in medicine*, t. 39, n° 21, p. 2815-2842, 2020.
- [38] C.-Y. CHENG, W.-L. TSENG, C.-F. CHANG, C.-H. CHANG et S. S.-F. GAU, « A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder, » *Frontiers in psychiatry*, p. 673, 2020.
- [39] W. H. ORGANIZATION et al., « Processus d'acquisition : guide pratique, » 2012.
- [40] C.-H. LIU, C.-F. TSAI, K.-L. SUE et M.-W. HUANG, « The feature selection effect on missing value imputation of medical datasets, » *Applied Sciences*, t. 10, n° 7, p. 2344, 2020.
- [41] *Statsmodels.stats.outliers_influence.variance_inflation_factor*. adresse : https://www.statsmodels.org/dev/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html.

-
- [42] D. M. J. GARBADE, *Understanding K-means clustering in machine learning*, sept. 2018. adresse : <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>.
- [43] M. SYAKUR, B. KHOTIMAH, E. ROCHMAN et B. D. SATOTO, « Integration k-means clustering method and elbow method for identification of the best customer profile cluster, » in *IOP conference series : materials science and engineering*, IOP Publishing, t. 336, 2018, p. 012 017.
- [44] F. O. ISINKAYE, Y. O. FOLAJIMI et B. A. OJOKOH, « Recommendation systems : Principles, methods and evaluation, » *Egyptian informatics journal*, t. 16, n° 3, p. 261-273, 2015.
- [45] K. GRACE-MARTIN, « Assessing the fit of regression models, » *The Analysis Factor*, t. 2015, 2018.