Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté des Sciences
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option: Réseaux et Systèmes Distribués (R.S.D)

Thème

L'Estimation du risque de Césarienne à l'aide d'IoT basé sur l'Apprentissage Supervisé

Réalisé par :

- **MAHI Ghizlene**
- **LOKBANI Souad**

Présenté le 2 juillet 2025 devant le jury composé de :

- **LEHSAINI Mohamed** **(Président)**
- **BENMAHDI Meryem Bochra** **(Encadrante)**
- **MANA Mohamed** **(Examinateur)**

Année universitaire : 2024-2025

Remerciements

Avant toute chose, nous tenons à remercier Allah, le Grand Dieu qui nous a offert la santé, la patience, la force et l'endurance requises pour accomplir ce travail avec succès. C'est grâce à sa volonté que nous avons pu surmonter les difficultés et réaliser nos objectifs.

À notre encadrante madame Benmahdi Meriem Bouchra, nous tenons à exprimer toute notre reconnaissance pour votre patience, votre écoute et vos conseils avisés. Votre accompagnement bienveillant a été essentiel dans la réalisation de ce travail.

Nous adressons aussi nos vifs remerciements aux membres des jurys Mr **LEHSAINI Mohamed** et Mr **MANA Mohamed** pour avoir bien voulu examiner et juger ce travail.

À ma cousine, **Dr Lokbani Manel**, pour son accompagnement constant, ses conseils avisés et son aide précieuse qui ont grandement contribué à la réalisation de ce projet.

À nos amis, Pour leur présence fidèle, leur amitié sincère, leur écoute attentive et leur bonne humeur qui ont illuminé notre parcours. Un remerciement tout particulier à **Mohammed** et **Mouad**, dont le soutien, la disponibilité et les mots réconfortants ont été précieux tout au long de ce parcours.

Enfin, Nous tenons aussi à remercier tous ce qui nous a aidés de près ou de loin pour réaliser ce travail.

Merci, du fond du cœur.

Dédicace

À Allah, pour nous avoir guidées, donné la force, la patience et la persévérance tout au long de ce parcours.

À moi-même, pour avoir cru en ce chemin, même lorsque tout semblait flou ou trop lourd. Pour avoir trouvé la force de continuer, malgré la fatigue et les incertitudes.

Une pensée spéciale à ma très chère grand-mère Chiali Fadèla, qui nous a quittés récemment. Son amour, sa tendresse et sa sagesse resteront à jamais gravés dans mon cœur, et dans chaque pas de mon parcours. Elle a été, et reste encore aujourd'hui, une source de force et de lumière dans ma vie. Ce travail, je le lui dédie avec émotion, gratitude et un profond respect.

À mon père, Lokbani Fayçal pour sa force, ses précieux conseils, ses encouragements constants et les nombreux sacrifices qu'il a faits pour m'offrir les meilleures conditions possibles.

À ma mère, Rahmoun Nawel pour son amour inconditionnel, ses prières sincères, sa tendresse et son soutien sans faille dans chaque étape de notre vie. Elle a toujours été mon piliers, celles qui m'a portée dans les moments de doute et qui m'a encouragée à ne jamais baisser les bras.

À mes frères Wahib et Oussama et ma belle sœurs Manel, pour leur affection, leur présence bienveillante et leur soutien discret mais si précieux.Je vous porte dans mon cœur avec une immense gratitude.

À toute ma famille, Pour votre présence, vos prières, votre bienveillance et votre réconfort dans les moments difficiles. Vous avez été une source de motivation et de stabilité.

À mon binôme et amie Ghizlene, merci pour ta complicité, ton soutien inébranlable et ta bonne humeur tout au long de cette aventure, entre les fous rires et les moments de stress, nous avons traversé bien des hauts et des bas, toujours avec courage et solidarité.Je suis profondément reconnaissante d'avoir vécu ce parcours avec toi.

À mes enseignants, votre savoir nous a guidés vers la lumière.

LOKBANI SOUAD

Dédicace

Je dédie ce modeste travail :

À moi-même,
Pour avoir cru en ce chemin, même lorsque tout semblait flou ou trop lourd. Pour avoir trouvé la force de continuer nuit après nuit malgré la fatigue et les incertitudes.

À mes chers parents,
Pour avoir toujours cru en moi, pour leur amour inconditionnel, leurs innombrables sacrifices, leur tendresse silencieuse et leurs prières qui m'ont accompagné à chaque étape.

À mes sœurs,
Amira, **Djihane** et ma petite **Dodos** que j'aime énormément.
Vous êtes mes trésors, ma plus grande source de joie, de motivation et de réconfort au quotidien. Je vous souhaite une vie remplie de bonheur, de succès et d'amour.
Et un doux mot à ma nièce **Chahrazed**, notre petit rayon de soleil récemment arrivée dans nos vies.
Que Dieu la garde pour nous et pour ses parents.

À mes grands-parents maternelle,
Pour leur amour, leurs prières et leur présence précieuse dans ma vie. Que Dieu vous garde pour nous.

À mon amie de parcours **Souad**,
Pour ta présence réconfortante, ton soutien indéfectible, et toutes ces jours passées à discuter, à rire, à douter… à traverser les hauts et les bas ensemble.
Merci d'avoir été là, dans le bon comme dans le moins bon. Ton amitié a rendu ce chemin plus doux.

MAHI GHIZLENE

Résumé

Cette recherche s'inscrit dans l'intégration des technologies IoT dans le domaine médical, notamment pour la surveillance prénatale. Il propose une méthode de prévision du taux de probabilité d'accouchement par césarienne, à partir de données cliniques collectées grâce à des capteurs connectés. Les paramètres considérés incluent l'âge, la pression artérielle et le taux de glycémie. Cinq modèles de machine learning ont été utilisés (Régression Logistique, SVM, Random Forest, Gradient Boosting, XGBoost) pour estimer le risque selon trois niveaux : bas, moyen et haut. Les résultats obtenus soulignent le potentiel de l'intelligence artificielle en tant qu'outil d'assistance à la prise de décision médicale, notamment pour prévoir les complications associées à l'accouchement.

Mots clés : IoT, Régression Logistique, SVM, Random Forest, Gradient Boosting, XGBoost, apprentissage supervisè.

Abstract

This research falls within explores the integration of Internet of Things (IoT) technologies in the medical field, with a particular focus on prenatal monitoring. It presents a method for predicting the probability of delivering by cesarean section based on clinical data collected through connected sensors. The considered parameters include age, blood pressure, and blood glucose level. Five machine learning models were applied (Logistic Regression, SVM, Random Forest, Gradient Boosting, and XGBoost) to estimate the risk at three levels: low, medium, and high. The results highlight the potential of artificial intelligence as a decision-support tool in medical practice, particularly in anticipating childbirth-related complications.

Keywords: IoT, Logistic Regression, SVM, Random Forest, Gradient Boosting, XGBoost, supervised learning.

الملخص

يندرج هذا البحث ضمن سياق دمج تقنيات إنترنت الأشياء في المجال الطبي، لاسيما في مجال مراقبة فترة ما قبل الولادة. يقترح منهجية للتنبؤ بمعدل احتمالية الولادة القيصرية، اعتمادًا على بيانات سريرية بواسطة أجهزة استشعار متصلة. تشمل المعايير المعتمدة: العمر، ضغط الدم، ومستويات السكر في الدم. تم تطبيق خمسة نماذج من التعلم الآلي) الانحدار اللوجستي، وآلة المتجهات الداعمة SVM ، والغابة العشوائية، والتعزيز التدريجي، و (XGBoostلتقدير مستوى الخطر وفقًا لثلاث درجات: منخفض، متوسط، وعالٍ. تُبرز النتائج المحصّلة قدرة الذكاء الاصطناعي على دعم اتخاذ القرار الطبي، لاسيما في التنبؤ بالمضاعفات المرتبطة بالولادة.

الكلمات المفتاحية: IoT، الانحدار اللوجستي، آلة الدعم المتجه، الغابة العشوائية، التعزيز المتدرج، XGBoost، التعلم الخاضع للإشراف.

Table des matières

Liste des figures

Liste des tableaux

Acronymes

IoT Internet of Things.

BS Blood Sugar.

SVM Support Vector Machine.

KNN K-plus proche voisins.

RF Random Forest.

IF Isolation Forest.

MIT Massachusetts Institute of Technology.

RFID Radio-Frequency Identification.

IPv6 Internet Protocol version 6.

RCSFs Réseaux de Capteurs Sans Fil.

RF Radio fréquence.

SO Smart Object.

IdO Internet des Objets.

MCU Microcontroller Unit.

SS Smart Services.

NFC Near Field Communication.

QR Code Quick Response Code.

LAN Local Area Network.

W3C World Wide Web Consortium.

IETF Internet Engineering Task Force.

IEEE Institute of Electrical and Electronics Engineers.

ETSI European Telecommunications Standards Institute.

RNR Réseaux de Neurone Récurrent.

RNN Recurrent Neural Network.

XGBoost Extreme Gradient Boosting.

ARIMA AutoRegressive Integrated Moving Average.

SARIMA Seasonal AutoRegressive Integrated Moving Average.

K-Means Partitionnement en K clusters autour de leurs moyennes.

DBSCAN Density-based algorithm for discovering clusters in large spatial databases with noise.

BSTS Bayesian Structural Times Series.

LSTM Long Short-Term Memory.

AUC Area Under the Curve.

RMSE Root Mean Squared Error.

CNN Convolutional Neural Network.

NARX Nonlinear AutoRegressive model with eXogenous inputs.

PM2.5 Particules fines en suspension dans l'air dont le diamètre est inférieur ou égal à 2,5 micromètres (µm).

LightGBM Light Gradient Boosting Machine.

macOS Macintosh Operating System.

Matplotlib Mathematical Plotting Library.

NumP Numerical Python.

SciPy Scientific Python.

IPython Interactive Python.

PAN Personal Area Network.

Wi-Fi Wireless Fidelity.

DiastolicBp Diastolic Blood Pressure.

SystolicBp Systolic Blood Pressure.

Scikit-learn Scikit for Machine Learning.

RBF Radial Basis Function.

SVC Support Vector Classification.

Log-loss Logarithmic Loss.

max_depth Profondeur maximale de l'arbre.

Introduction Générale

Introduction Générale

Lors de ces dernières décennies l'Internet des Objets (IoT) a vu le jour. Aujourd'hui l'IoT s'impose comme un levier majeur d'innovation dans de nombreux domaines. Cette technologie permet à des dispositifs physiques de collecter, de transmettre et d'analyser des données en temps réel. Dans le secteur médical, l'IoT permet le développement de systèmes connectés capables d'assurer un suivi précis et personnalisé de l'état de santé des patients, en temps réel et à distance [1].

Dans le domaine du suivi prénatal, les technologies IoT permettent d'automatiser la collecte de données cliniques importantes soit de l'état du fœtus ou l'état de la maman. L'état de la maman comme la pression artérielle, la fréquence cardiaque ou le taux de glycémie, l'état du fœtus comme le rythme cardiaque ou les mouvements. Grâce à ces dispositifs connectés, les données peuvent être envoyées instantanément aux professionnels de santé, ce qui permet un suivi individualisé de l'état de santé de chaque patiente. Lorsque ces paramètres présentent des valeurs anormales, des signes d'alerte sont déclenchés par exemple, si la pression artérielle dépasse 140 mmHg une alerte est générée automatiquement afin de prévenir un éventuel risque d'accouchement par césarienne. Une dérive significative par rapport aux seuils habituels peut signaler un risque de complications pendant la grossesse ou lors de l'accouchement. Il est donc essentiel de détecter ces anomalies à temps afin de prévenir les situations critiques.

Ce projet de fin d'étude s'inscrit dans cette dynamique en exploitant les données collectées par des dispositifs IoT afin de prédire le risque d'un accouchement par césarienne. À partir de facteurs cliniques de l'état de la maman tels que l'âge, la pression artérielle systolique et diastolique, ainsi que le taux de glycémie, nous proposons une méthode de prédiction du taux de probabilité d'accouchement par césarienne (haut, moyen ou bas). Dans cette optique, nous avons impliqué et comparé plusieurs modèles d'apprentissage automatique comme la régression logistique, le SVM (Support Vector Machine), le Random Forest, GradientBoosting, et XGBoost_model.

L'objectif de notre contribution est de développer un système de prédiction du taux de la probabilité d'accoucher par césarienne. Ce système est basé sur des données cliniques de l'état de santé de la maman collectées via des dispositifs IoT, afin d'aider les professionnels de santé d'anticiper les complications possibles et d'intervenir à temps. Nous avons supposé que ces données ont été collectées sur des femmes enceintes à terme et que leur fœtus est en bonne santé.

Ce mémoire est structuré comme suit :

- Le premier chapitre présente les concepts généraux de l'Internet des objets, son fonctionnement, ses domaines d'application, et la prévision des données dans l'IoT et ses défis.
- Le deuxième chapitre propose un état de l'art sur les approches de prédiction dans l'Internet des Objets, en mettant l'accent sur les modèles intelligents utilisés, le cycle de vie d'une prédiction des données, ainsi que les principales applications de ces méthodes dans le domaine de la santé connectée.
- Le troisième chapitre présente l'implémentation concrète du système de prédiction, en détaillant les étapes de traitement des données, l'application des modèles choisis, ainsi que l'évaluation et l'analyse des résultats obtenus.

Enfin, une conclusion récapitule avec des perspectives d'amélioration pour de futurs travaux.

Chapitre I

Les réseaux de l'Internet des objets : généralités et concepts

I.1 Introduction

L'IoT est une technologie en pleine évolution qui modifie la manière de collecter, d'analyser et d'exploiter les données issues du monde réel. Dans le domaine médical, cette innovation prend une dimension particulièrement stratégique, car elle permet une surveillance continue et en temps réel des paramètres vitaux des patients grâce à des dispositifs connectés.

Ce chapitre propose une vue globale sur les réseaux de l'IoT. Il retrace leur historique, précise leurs définitions, décrit leur fonctionnement général et présente leurs types d'architectures. Il aborde également leurs domaines d'application, ainsi que les principaux avantages et limites de cette technologie. Enfin, une attention particulière est portée à la prévision des données dans IoT**.**

I.2 Historique et évolution de l'IoT

Le concept d'Internet des objets (IoT) a été introduit en 1999 par Kevin Ashton, chercheur au MIT, dans le but de connecter les objets physiques à Internet à l'aide de la technologie RFID. Cette idée a pu améliorer la traçabilité et la communication entre les objets sans intervention humaine. L'apparition du protocole IPv6 a ensuite permis de dépasser les limitations liées au nombre d'adresses IP, favorisant ainsi le développement de l'IoT dans divers domaines, comme l'aéronautique. Ce n'est qu'à partir de 2007 que l'IoT commence réellement à se populariser à l'échelle mondiale [2].

L'évolution de l'IoT remonte aux années 1990, avec les premières expériences de connexion d'objets simples comme des grille-pains ou cafetières. En 2000, des entreprises comme LG évoquent déjà l'idée d'un électroménager intelligent. À cette époque, le nombre d'appareils connectés reste faible : en 2003, on ne comptait que 500 millions de dispositifs pour une population mondiale de 6,3 milliards. C'est en 2010, avec l'essor des smartphones et tablettes, que le nombre de dispositifs connectés atteint 12,5 milliards, dépassant pour la première fois le nombre d'habitants. Depuis, cette croissance s'est accélérée, avec des prévisions estimant plus de 50 milliards d'objets connectés dans les années à venir [3]. Dix ans plus tard, en 2020, le nombre d'objets connectés a été estimé entre 22 et 30 milliards, selon différentes sources, confirmant ainsi l'essor rapide et continu des technologies IoT.

La technologie d'IoT est en progrès continu depuis ces dernières décennies dans plusieurs domaines. Les deux technologies prédécesseurs qui ont abouti à ce progrès se sont :

- **RCSFs :** sont composés de plusieurs nœuds capteurs limités en termes de calcul, de stockage et d'énergie. Ces nœuds sont capables de capter des grandeurs physique de les traiter et de les envoyer à une ou plusieurs stations de base [4].
- **RFID** : La technologie RFID est une méthode de communication sans fil qui utilise un champ électromagnétique (onde radio) pour transférer des données entre une étiquette RFID et un lecteur RFID. Le lecteur est composé d'une station de base connectée à une base de données et d'une antenne émettrice-réceptrice. Il émet une onde radiofréquence (RF) pour alimenter le ou les étiquettes se trouvant dans sa zone de lecture [5].

I.3 Internet des objets (IoT)

En anglais appelé Internet of Things (IoT). Dans la littérature, il existe plusieurs définitions de l'IoT. Les plus pertinentes sont :

- Les auteurs dans [6] définissent IoT comme l'ensemble des objets physiques équipés de capteurs ou de dispositifs intelligents capables de se connecter à Internet pour échanger des données comme montrer dans la figure I.1. Ces objets interconnectés permettent de collecter en continu des informations utiles, qui peuvent ensuite être exploitées par divers acteurs tels que les entreprises, les administrations, les collectivités locales, les établissements de santé ou encore les citoyens. Il repose sur une infrastructure unifiée combinant des technologies d'identification normalisées et des systèmes de communication sans fil, offrant ainsi la possibilité d'identifier avec précision aussi bien des objets physiques que des entités numériques. Cela permet d'assurer une circulation fluide des données entre le monde réel et le monde virtuel, en facilitant leur collecte, leur stockage, leur transmission et leur traitement [7].
- Dans [8], les auteurs ont décrit l'IoT comme "*une infrastructure de réseau mondial dynamique avec des capacités d'auto-configuration basées sur des protocoles de communication standards et interopérables où les 'objets' physiques et virtuels ont des identités, des attributs physiques et des personnalités virtuelles en utilisant des interfaces intelligentes qui sont parfaitement intégrées au réseau d'information*".
- Une autre explication par [9] : "*l'IoT permet aux personnes et aux objets d'être connectés à tout moment et en tout lieu en utilisant idéalement n'importe quel chemin/réseau et tout service*".

- En addition à ces explications les auteurs dans [10], ont défini l'IoT dans différents contextes principalement dans : "L'intégration des appareils minuscules appelés objets intelligents' (Smart Object (SO)), généralement alimentés par des batteries équipés d'un microcontrôleur (MCU) et d'émetteurs-récepteurs. Les services offerts par ces objets intelligents sont appelés services intelligents (SS) [11, 12]".

Figure I.1 : Internet des objets [13]

I.4 Fonctionnement de l'IoT

Afin de garantir le bon fonctionnement d'IoT, il est nécessaire de connaitre ses éléments clés et de s'appuyer sur une architecture bien structurée capable de décrire les échanges entre les équipements connectés, les systèmes de traitement et les applications. Ce type de représentation permet non seulement de mieux comprendre l'ensemble, mais aussi de localiser les étapes clés du traitement des données et de faciliter l'intégration des technologies issues de différents fabricants. Dans cette partie, nous allons voire les éléments clés d'IoT, ses fonctionnalités et son architecture.

I.4.1 Éléments clés de l'IoT

D'après les auteurs dans [14] , Un système IoT repose sur plusieurs éléments techniques indispensables à son bon fonctionnement. Une architecture IoT standard comprend généralement cinq composantes principales :

A) Les objets connectés : Ces objets, actifs ou passifs, sont chargés de collecter des données utiles selon le domaine d'application. Ils comprennent principalement des capteurs, chargés de mesurer des données (température, humidité, pression, etc.), et des actionneurs, qui exécutent des actions en réponse à des commandes (par exemple, activer un moteur ou déclencher une alarme).

B) Le réseau : Il permet le transport des données entre les objets et les serveurs ou plateformes de traitement. Il doit garantir une couverture adaptée à la zone ciblée (locale, urbaine ou mondiale).

C) Les données : Les objets génèrent des données brutes qui doivent être stockées dans des bases bien structurées. Cela permet leur exploitation ultérieure pour produire des résultats fiables [14].

D) Les informations : Issues du traitement et de l'analyse des données, les informations obtenues doivent également être sauvegardées pour garantir leur intégrité et leur disponibilité à tout moment.

E) Les applications d'exploitation : Ces interfaces, souvent visuelles (tableaux de bord, graphiques), facilitent l'interaction entre l'utilisateur et le système. Elles permettent une compréhension rapide des données et une prise de décisions claire [15].

I.4.2 Les fonctionnalités de l'IoT

Le bon fonctionnement d'un système IoT repose sur plusieurs fonctionnalités techniques majeures [16], chacune jouant un rôle spécifique dans la chaîne de traitement des données. Ces fonctions sont décrites comme suit :

A) Identification : Chaque objet IoT reçoit un identifiant unique, ce qui permet de le distinguer dans le réseau et d'assurer la traçabilité des données échangées.

B) Sensing (capture) : Les capteurs jouent un rôle central dans la collecte d'informations physiques ou environnementales, utilisées ensuite pour des analyses ou des décisions automatisées.

C) Communication : Pour que l'IoT fonctionne, les données collectées par les capteurs doivent être transmises à des dispositifs capables de les analyser et de les traiter. Cela repose sur diverses technologies de communication, filaires ou sans fil, telles que le Wi-Fi, le Bluetooth, le Zigbee, le NFC ou encore les réseaux à bande étroite [17].

D) Computation (traitement) : Le traitement des données peut se faire localement via l'edgecomputing, réduisant ainsi la latence et la charge du réseau. Les données peuvent ensuite être envoyées vers le cloud pour des analyses plus poussées [16].

E) Services : Les services IoT correspondent aux fonctionnalités concrètes effectuer par les systèmes : surveillance, détection d'anomalie, sécurité, prédiction des données, contrôle à distance, etc. Ces services transforment les données en actions.

F) Sémantique : Il s'agit ici de donner du sens aux données pour qu'elles soient comprises et interprétables par différents systèmes. Cela nécessite l'usage de modèles et de standards permettant une intégration et une interopérabilité efficaces [16].

Figure I.2: Les fonctions d'IoT [17]

I.4.3 Architecture IoT

Parmi les modèles les plus répandus, on retrouve une organisation en trois principale couches. Cette approche était adoptée dès l'apparition d'IoT. Les niveaux sont :

- **La couche perception**, qui représente la partie physique du système. Elle intègre les capteurs chargés de collecter des données depuis l'environnement ou d'interagir avec d'autres objets intelligents.
- **La couche réseau**, Elle assure la connexion avec d'autres objets intelligents, équipements réseau et serveurs, tout en jouant un rôle essentiel dans la transmission et le traitement des données issues des capteurs [18].
- **La couche application**, qui correspond aux services proposés à l'utilisateur selon le domaine d'usage : domotique, santé connectée, ville intelligente, etc. [18]

Cependant, cette structure s'est révélée insuffisante pour répondre à des problématiques plus avancées. Ainsi, un modèle plus complet, à cinq couches, a été introduit. Il reprend les trois précédentes tout en y ajoutant :

- **La couche de transport**, est responsable du transfert des informations entre la couche de perception et celle de traitement, en utilisant divers moyens de communication telle que les réseaux sans fil, la 3 G, les réseaux locaux (LAN), le Bluetooth, le RFID ou le NFC [19].
- **La couche de traitement**, parfois appelée middleware, qui permet de stocker, analyser et exploiter les données, notamment grâce au cloud computing ou aux technologies Big Data.

On peut aussi mentionner une **couche métier** dans certaines approches, qui traite des aspects liés à la stratégie, à la gestion des services et à la confidentialité des données [18].

I.5 Domaines d'utilisation

L'Internet des objets (IoT) trouve des applications variées dans de nombreux domaines de la vie quotidienne comme montré dans la figure I.3. Les principaux domaines concernés sont les suivants :

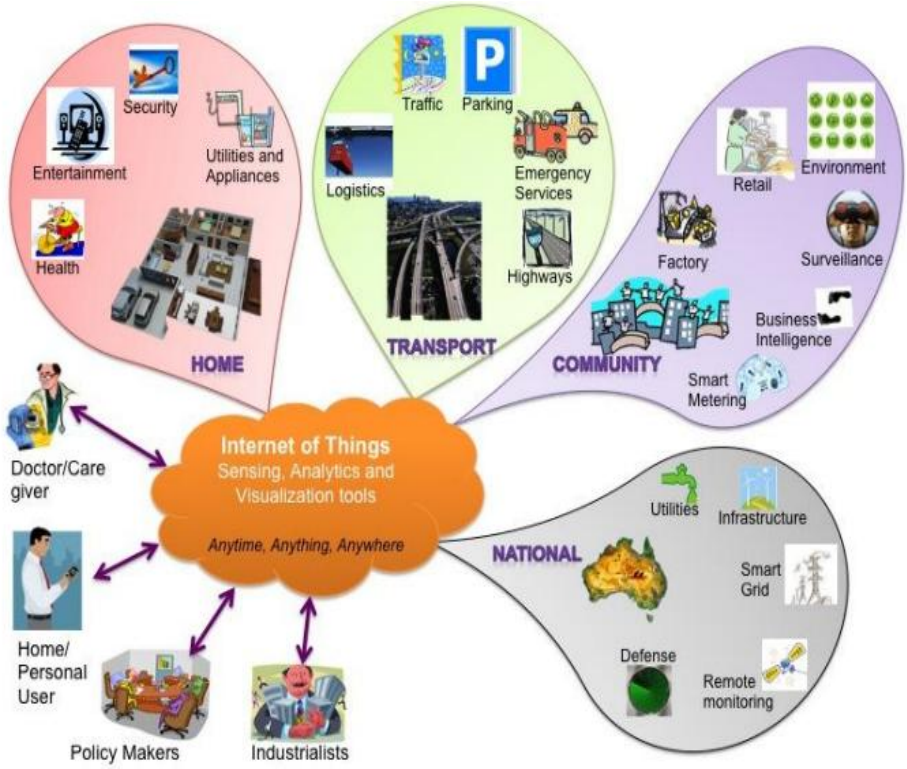

Figure I.3: Domaines d'application d'IoT [20]

I.5.1 Les villes intelligentes

De nombreuses grandes métropoles, telles que Séoul, New York, Tokyo, Shanghai, Singapour, Amsterdam ou encore Dubaï, ont initié des projets en lien avec le concept de ville intelligente (voir Figure I.4). Considérées comme les villes du futur, elles intègrent progressivement des technologies innovantes, notamment grâce à l'Internet des objets (IoT). Le rythme actuel d'innovation rend l'intégration de ces technologies de plus en plus accessible dans le processus de développement urbain [18]. La réussite de ces initiatives repose cependant sur une planification rigoureuse à chaque étape, ainsi que sur la coopération des gouvernements et des citoyens pour soutenir l'adoption des solutions IoT dans tous les domaines. Grâce à l'Internet des objets, il est possible d'optimiser différents aspects de la ville, notamment en modernisant les infrastructures et en améliorant l'efficacité des systèmes de transport.

Figure I.4: Exemple d'une ville intelligente [21]

I.5.2 L'agriculture

Dans le secteur agricole, l'Internet des objets permet d'exploiter des réseaux de capteurs interconnectés pour surveiller en temps réel les conditions environnementales des cultures comme illustré dans la figure I.5. Cette technologie favorise une meilleure gestion de l'irrigation, l'optimisation de l'utilisation des intrants, ainsi qu'une planification plus efficace des activités agricoles. Elle contribue également à la prévention des dégâts liés aux aléas climatiques et à l'amélioration globale de la qualité de l'environnement [22].

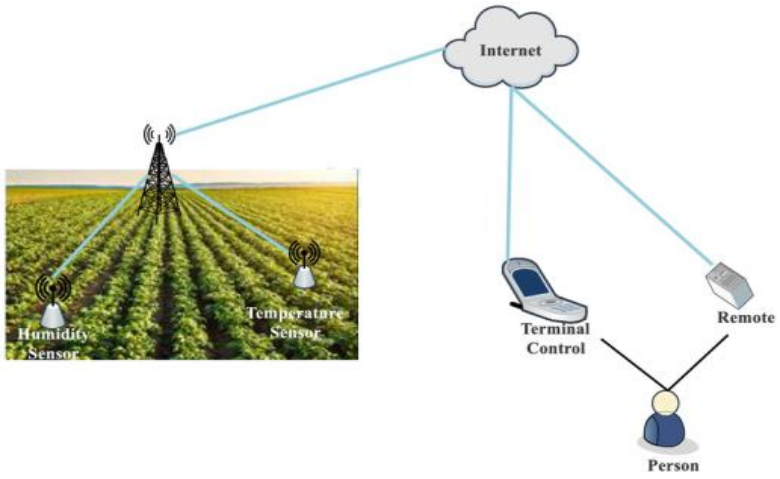

Figure I.5: Un modèle de réseau IoT agricole avec des capteurs [23]

I.5.3 Le transport

Internet des objets joue un rôle clé dans le développement des véhicules intelligents en renforçant la sécurité routière et en facilitant l'assistance à la conduite. Grâce à la communication entre les véhicules et les infrastructures, l'IoT permet d'optimiser la gestion du trafic, d'améliorer le confort des usagers, et de réaliser des gains en temps, en énergie et en efficacité globale [24].

Figure I.6 : Un modèle de réseau de transport avec IoT [25]

I.5.4 Sécurité et surveillance

Internet des objets contribue à renforcer la sécurité dans divers espaces tels que les bâtiments d'entreprise, les centres commerciaux, les usines, les parkings ou encore les lieux publics, tout en respectant la vie privée des utilisateurs. Pour garantir un haut niveau de sécurité de manière efficace, différents types de capteurs sont utilisés. Par exemple, des capteurs multimédias pour la surveillance vidéo tels que montré dans la figure I.7 ou des capteurs environnementaux permettent de détecter la présence de substances chimiques dangereuses, tandis que d'autres dispositifs sont capables d'analyser le comportement des individus afin d'identifier des actions potentiellement suspectes [2].

Figure I.7: La surveillance vidéo intelligente basée sur l'IoT [26]

I.5.5 Industrie et entreprises

Dans le cadre professionnel, l'Internet des objets est souvent déployé comme une solution interne destinée à répondre aux besoins spécifiques de l'entreprise. Les données collectées par ces réseaux sont utilisées exclusivement par les responsables concernés, et leur exploitation se fait de manière ciblée et sécurisée. L'une des principales applications concerne le contrôle de l'environnement de travail, notamment pour assurer la traçabilité du personnel et la gestion automatisée de services tels que le chauffage, la ventilation, la climatisation ou encore l'éclairage.

Dans les milieux industriels, les capteurs jouent depuis longtemps un rôle essentiel en matière de sécurité, de régulation climatique et d'automatisation. Aujourd'hui, ces dispositifs filaires sont progressivement remplacés par des systèmes sans fil intégrés à l'IoT, permettant des ajustements dynamiques selon les besoins. Ce type de réseau localisé constitue un sous-système dédié à la maintenance et au bon fonctionnement des installations. L'introduction de l'IoT dans le secteur industriel s'inscrit dans une tendance plus large vers l'Industrie 4.0, (comme montré dans la figure I.8) marquant une nouvelle ère d'optimisation intelligente des processus industriels [27].

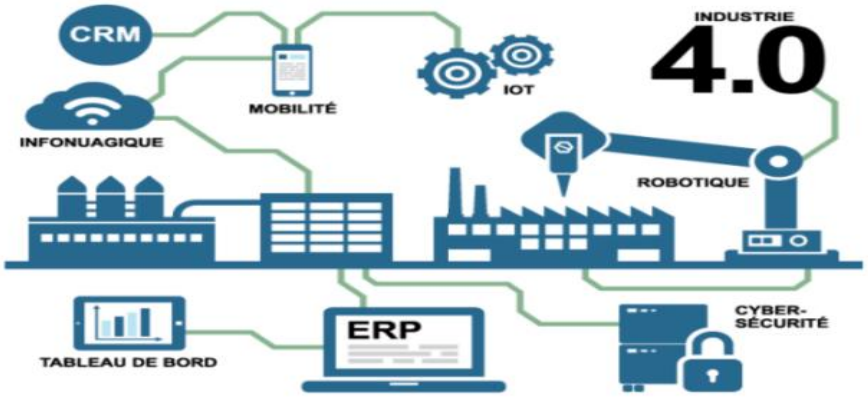

Figure I.8: Schéma des composantes technologiques de l'Industrie 4.0 [28]

I.5.6 La santé

Dans le domaine médical, l'Internet des objets permet une surveillance continue des signes vitaux des patients grâce à des réseaux personnels composés de capteurs médicaux. Ces dispositifs mesurent diverses constantes biologiques telles que la température corporelle, la pression artérielle, fréquence cardiaque ou encore la fréquence respiratoire. Ce système connecté s'avère particulièrement utile pour le suivi des personnes à mobilité réduite, malade ou des grossesses à risque, dont les activités quotidiennes peuvent être observées à l'aide de capteurs portables (comme des accéléromètres ou des gyroscopes) comme montré dans la figure I.9 ou installés dans leur environnement, grâce à la technologie IoT, il devient possible d'assurer une prise en charge plus efficace et personnalisée des patients [29].

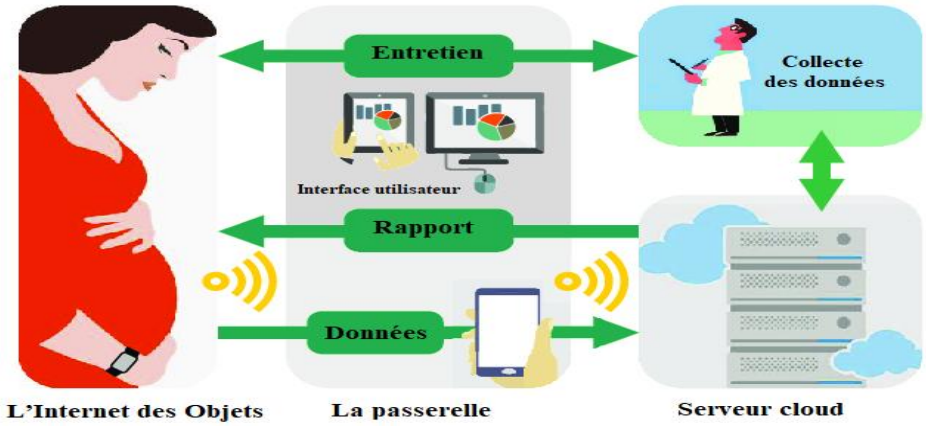

Figure I.9: L'IoT dans la santé

I.6 Les avantages et les inconvénients de l'Internet des objets

L'Internet des objets comporte plusieurs avantages et inconvénients en voici quelques-uns :

I.6.1 Avantages

- Favorise la transmission des informations et améliore la communication. Contribue à une meilleure efficacité et performance dans les entreprises [30]
- Contribue à l'amélioration des conditions de la vie en facilitant le confort au quotidien.
- Optimise la productivité et l'expérience client : les objets connectés transmettent des données aux fabricants concernant les préférences et comportements des utilisateurs, permettant ainsi aux entreprises d'adopter une approche proactive et personnalisée, mieux alignée sur les attentes et besoins de la clientèle [17].
- Simplifie le quotidien dans plusieurs domaines clés comme la santé**,** exemple les objets connectés permettent aux patients de limiter certains déplacements vers les établissements médicaux, en transmettant à distance des données utiles au diagnostic de leur état de santé [31].

- Dans certains cas d'usage, IoT permet d'optimiser les dépenses et de réaliser des économies, en favorisant une consommation basée sur les besoins réels, que ce soit pour les achats, l'énergie (éclairage, climatisation), ou d'autres ressources [17].

I.6.2 Inconvénients

- L'intégration de l'IoT dans les entreprises et les systèmes de santé soulève d'importants défis en matière de sécurité de l'information. En cas de faille, les conséquences peuvent être graves : atteinte à la réputation de l'organisation, pertes financières, interruption des activités, voire implications juridiques. De plus, la collecte massive de données par les objets connectés représente une menace sérieuse pour la vie privée des utilisateurs, leurs informations personnelles pouvant être exploitées sans consentement. Ces risques rendent la sécurité et la protection des données essentielles dans tout déploiement d'une solution IoT.
- L'absence de normes de compatibilité universelles limite l'interopérabilité entre les dispositifs connectés [17].

- L'automatisation des tâches manuelles entraîne la suppression de certains emplois, ce qui conduit un nombre important de personnes à se retrouver au chômage.

I.7 La prédiction dans l'IoT et ses défis

IoT est une technologie qui a changé le monde, cependant son déploiement s'accompagne de défis importants liés à la sécurité, à l'interopérabilité et à la gestion des données massives. La prévision des données apparaît alors comme une solution essentielle pour anticiper les événements, optimiser les décisions et garantir la fiabilité des systèmes. Pour que l'IoT tienne ses promesses, il est indispensable d'exploiter intelligemment les données qu'il génère. C'est à cette condition que l'IoT pourra réellement soutenir le développement de technologies intelligentes et durables.

La prédiction des données dans IoT consiste à anticiper des événements futurs en exploitant les données recueillies par des capteurs et divers dispositifs connectés. Cette capacité à prédire s'avère particulièrement précieuse dans des secteurs comme la santé, où elle permet de surveiller en continu les paramètres vitaux des patients et de détecter à l'avance d'éventuelles complications [32]. Cependant, pour que ces prévisions soient efficaces, il est indispensable de gérer des volumes très importants de données souvent hétérogènes, issues de multiples sources, ce qui complique leur traitement en temps réel. De plus, la diversité des protocoles de communication et la multiplicité des formats de données rendent leur intégration et leur analyse encore plus complexes.

Les défis spécifiques liés à la prévision dans les systèmes IoT sont multiples. La qualité des données collectées représente un enjeu majeur, car elles peuvent être incomplètes, bruitées ou erronées, ce qui affecte directement la fiabilité des modèles prédictifs. Par ailleurs, la gestion de la sécurité et de la confidentialité des informations est essentielle, notamment lorsque les données sont sensibles, comme dans le domaine médical [33]. Enfin, la scalabilité des solutions analytiques est mise à rude épreuve par l'augmentation continue du nombre d'objets connectés et le volume croissant de données générées, nécessitant des infrastructures capables de s'adapter sans dégrader les performances [33].

I.8 Conclusion

Dans ce premier chapitre, nous avons vu les concepts de base d'IoT. Cette technologie joue un rôle clé dans la transformation numérique des secteurs comme la santé, en permettant une collecte de données efficace et en temps réel, leur traitement et la prévision des complications. Dans le prochain chapitre, nous allons voir un état de l'art sur la prévision des données dans IoT.

Chapitre II

État de l'art sur la prédiction des données dans les réseaux IoT

II.1 Introduction

Dans un environnement de transition numérique accélérée, IoT occupe une position significative dans divers secteurs, en particulier celui de la santé. Les objets connectés permettent de collecter instantanément des informations cruciales sur les personnes, comme des mesures physiologiques, dans le but de prévoir certains dangers et d'optimiser leur gestion. Cette aptitude à anticiper des circonstances critiques offre la possibilité de préparer les interventions à l'avance et d'améliorer la sécurité des individus impliqués.

Dans ce chapitre, nous allons présenter un état de l'art sur la prédiction des données dans IoT qui contient une classification des modèles intelligents de prédiction des données, le cycle de vie d'une prédiction des données et quelques applications de prédiction des données dans IoT.

II.2 Catégories des modèles intelligents de prédiction des données dans IoT

La prédiction des donnés dans IoT est très utile dans certains cas comme dans la gestion de l'énergie, la détection des anomalies ou la maintenance prédictive. Il existe plusieurs méthodes impliquées dans la prédiction des données. Les méthodes les plus répandues ont été classées en plusieurs catégories, représentées dans la figure II.1.

II.2.1 Prédiction basée sur les méthodes d'apprentissage automatique supervisé

Dans le cadre de l'apprentissage automatique supervisé pour la prédiction des données, le modèle est entraîné sur un jeu de données déjà étiquetées [34]. Il est conçu pour apprendre à effectuer des prédictions en se basant sur une liste d'exemples étiquetés, c'est-à-dire associés à la valeur qu'il faut prédire. Selon les auteurs dans [35], l'apprentissage supervisé est décrit comme suit : « L'apprentissage supervisé c'est l'apprentissage d'une mise en correspondance entre un ensemble de variables d'entrée X et une variable de sortie Y et en appliquant cette correspondance pour prédire les sorties pour de nouvelles données non vues au paravent » [35].

Dans ce qui suit nous présentons les principales méthodes d'apprentissage supervisé :

A) K-plus proche voisins

En anglais appelé K-nearest neighbors (KNN). Un algorithme couramment employé pour traiter les problèmes de classification dans le secteur industriel, grâce à sa simplicité d'interprétation de la variable cible, mais aussi à sa capacité fiable de prédiction du résultat avec un temps de calcul

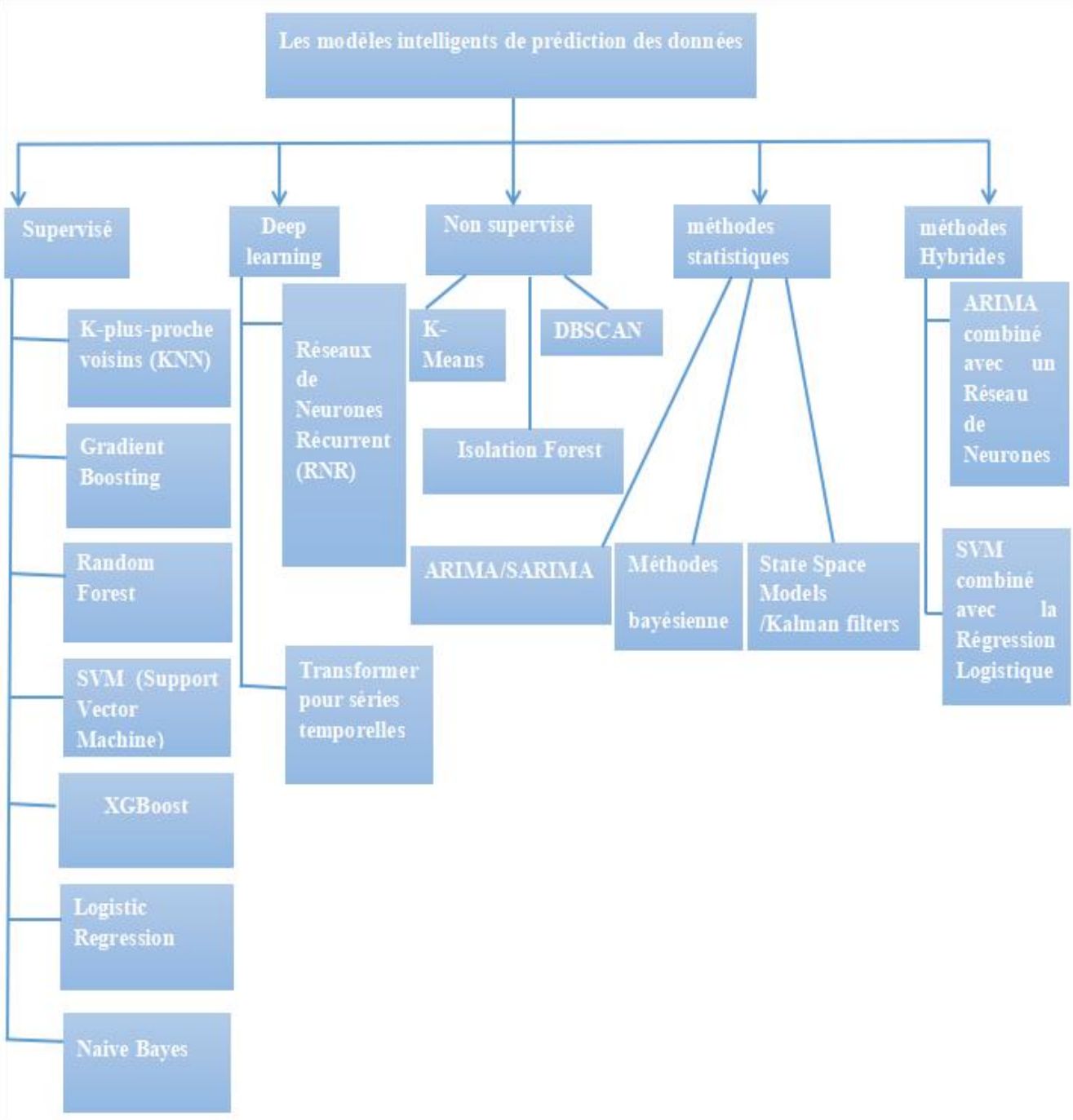

Figure II.1: Classification des modèles intelligents de prédiction des données dans IoT

réduit [36]. L'algorithme des k plus proches voisins est décrit par « Chloé-Agathe Azencott » comme une méthode qui attribue un label à une nouvelle observation x en se basant sur les labels des k points les plus proches du jeu de données d'entraînement [37]. Il est performant pour des jeux de données de petite taille à moyenne et pour des données linéaires mais il peut devenir non performant sur des données bruitées ou non linéaires.
Dans [38], les auteurs ont appliqué KNN sur le célèbre jeu de données Pima Indians Diabète [39] pour prédire si une personne est atteinte de diabète à partir de caractéristiques médicales. Les résultats ont montré que KNN, avec un choix de **k optimale** et une normalisation des données, à atteint une précision acceptable dans les tâches de prédiction médicale.

B) Gradient Boosting

Pour un cas de classification binaire basique, l'algorithme du gradient boosting a été présenté par M. Kuhn et Johnson en 2013 [40]. Le Gradient Boosting est une technique d'apprentissage supervisé basée sur le boosting, dont l'objectif est de combiner plusieurs modèles de faible performance afin de créer un modèle avec d'excellente performance [41]. Il vise à minimiser la fonction objectif en intégrant des apprenants faibles, tout en recourant à une optimisation par descente de gradient [40].

Dans [42], les auteurs ont appliqué Gradient Boosting sur le jeu de données Cleveland Heart Disease Dataset [43]. pour prédire la maladie cardiaque à partir de caractéristiques cliniques. Et les résultats ont montré que G**radient Boosting** donne de meilleurs résultats que KNN en termes de précision, **rappel**, et **AUC-ROC**, grâce à sa capacité à gérer les relations complexes et interactions non linéaires entre les variables.

C) Random Forest (RF)

(Random Forest) est une méthode d'apprentissage automatique supervisé décrit par Leo Breiman (2001) [44]. La méthode d'apprentissage par ensembles la plus célèbre, « bagging » ou « bootstrap aggregating », consiste à générer plusieurs instances identiques d'un modèle (plusieurs arbres de décision) et à former chacun de ces arbres sur une fraction aléatoire d'un ensemble de données [45]. Il est performant pour améliorer la précision de la classification et minimiser le danger du surapprentissage grâce à la variété des arbres et la compilation de leurs prévisions [46].

Dans [47], les auteurs ont utilisé random forest pour prédire le diabète sur Pima Indians Diabet [48]. Dans l'article [47], plusieurs algorithmes ont été comparés (KNN, SVM, ANN, Random Forest, etc.). Les résultats ont montré que la **Random Forest est stable et performante** même avec peu de réglage.

D) SVM (Support Vector Machine)

Les Machines à Vecteurs de Support (SVM) constituent une technique d'apprentissage supervisé mise au point par Vladimir Vapnik en 1990 [49]. SVM est un algorithme d'apprentissage automatique très robuste et flexible, apte à réaliser une classification linéaire ou non, une régression, ainsi qu'une détection des valeurs extrêmes. Les SVM sont particulièrement appropriés pour classifier des ensembles de données complexes, mais de taille petite ou moyenne. SVM est parmi les modèles les plus utilisés en apprentissage automatique [50].

Dans [51], les auteurs ont utilisé SVM sur le jeu de données Wisconsin Breast Cancer Dataset (WBCD) [52] pour prédire le cancer du sein à partir des données cliniques des cellules mammaires. Les résultats ont montré que la précision était de **97.85%** et que SVM est performant pour les tâches de classification médicale.

E) XGBoost

C'est une version améliorée de gradient boosting grâce à plusieurs ajouts [53], permettant une exécution rapide, une gestion automatique des valeurs manquantes et excellente précision. XGBoost est utilisée pour la prédiction du diabète [54], et les résultats ont montré que les autres classifieurs (KNN, SVM, RF) en précision, robustesse et efficacité pour la détection du diabète dans des systèmes intelligents de santé.

F) Logistic Regression

Hosmer et Lemeshow ont présenté et popularisé la régression logistique en 1989 [55]. La régression logistique opère de façon très comparable à la régression linéaire, mais elle s'applique à une variable de réponse binaire. L'avantage majeur de cette technique est la possibilité d'utiliser des variables explicatives continues et de traiter simultanément plus de deux variables explicatives. Et surtout lorsqu'il s'agit d'examiner l'influence de différentes variables explicatives sur la variable dépendante [56]. La régression logistique a été appliquée pour prédire la

probabilité qu'un patient soit atteint de diabète dans [57]. L'étude a montré que **La régression logistique**, bien qu'elle est interprétable mais elle présente une précision inférieure.

G) Naive Bayes

La technique Naive Bayes se fonde sur le théorème de Bayes, établi par Thomas Bayes (environ 1763, publié après sa mort) [58]. Naive Bayes est un classificateur probabiliste supervisé en apprentissage automatique, à la fois simple et performant. Il est spécialement utile pour les enjeux de classification textuelle [59]. Dans [60], les chercheurs ont utilisé Naive Bayes pour prédire le diabète et il a donné une **précision d'environ 76 %**, montrant qu'il est **compétitif malgré sa simplicité**, surtout pour des premières analyses rapides.

Le tableau II.1 présente une synthèse comparative des modèles d'apprentissage supervisé.

II.2.2 Les méthodes basées sur les réseaux de neurones (deep learning)

Ce sont des méthodes basées sur une catégorie de l'apprentissage automatique avancé, plus exactement les réseaux de neurones artificiels profonds pour la prédiction des données. Nous allons citer deux types de ces méthodes :

A) Réseaux de neurones récurrents (RNR)

En anglais appelé Recurrent Neural Network (RNN). Dans le années 90, Jeffrey L. Elman a présenté les Réseaux de Neurones Récurrents (RNN) [61]. (RNN) sont spécifiquement élaborés pour traiter les données séquentielles telles que les séries temporelles, les textes ou les signaux vocaux [62]. L'idée fondamentale est de définir des unités basiques nommées neurones, chacune pouvant effectuer quelques opérations élémentaires sur des données numériques. Ces neurones sont agencés de façon à maintenir un état interne, ce qui permet de retenir les informations précédentes et de saisir les dépendances dans le temps. En connectant une multitude de ces unités, on crée donc un outil de calcul performant, approprié à l'examen des données complexes en séquence [63].

Les chercheurs ont employé les RNNs pour la prédiction de température intérieure dans un bâtiment intelligent à partir de capteurs IoT [64]. Les RNNs fournissent des prédictions plus précises que les méthodes classiques (ARIMA, régression linéaire)

Modèle	Description	Avantages	Inconvénients
Gradient Boosting	Un modèle d'ensemble utilisant de nombreux arbres de décision de faible performa	Performance élevée, gestion de relations non linéaires.	Plus lent à entraîner, sensible au sur apprentissage
Random Forest	Formation sur des sous-échantillons à l'aide de la technique d'arbres de décision.	Résistant au bruit, nécessite peu d'ajustements.	Moins performant que le boosting dans certaines situations.
SVM (Support Vector Machine)	Augmentation de l'écart entre les classes.	Idéal pour les petits datasets, performant avec des marges bien définies.	Moins approprié pour les données bruyantes ou les grands ensembles de données.
XGBoost	Version perfectionnée du Gradient Boosting, intégrant la régularisation et la parallélisation.	Extrêmement rapide, d'une grande précision, avec une régularisation intégrée pour éviter le surajustement.	Plus compliqué à configurer, vulnérable aux données déséquilibrées.
Logistic Regression	Classification par modèle linéaire.	Rapide et compréhensible.	Moins performant sur des relations complexes.
K-Nearest Neighbors (KNN)	Attribue à une entrée l'un des k profils les plus similaires dans les données passées.	Facile à comprendre, sans nécessité de formation, efficace pour les petits ensembles de données.	Ralentissement avec de grands volumes, sensible aux variables non standardisées.
Naive Bayes	Estimez la probabilité d'appartenir à une catégorie en présumant que les attributs sont indépendants.	Rapide, efficace pour les petits ensembles de données, performant face aux données bruitées.	L'hypothèse d'indépendance est rarement vérifiée, ce qui peut entraîner une sous-performance en présence de corrélations entre les variables.

Tableau II.1 Analyse comparative des algorithmes de classification supervisée

B) Transformer pour séries temporelles

Le modèle Transformer a été présenté par Vaswani et ses collaborateurs en 2017 [65]. La prédiction des séries temporelles a été largement utilisée dans diverses applications [65]. Les Transformers pour les séries temporelles sont des modèles qui s'appuient sur le mécanisme d'attention pour gérer de manière optimale des données multivariées dans le temps. Les Transformers saisissent en même temps les relations globales dans une séquence, ce qui les rend appropriés pour des tâches telles que la prévision de séries temporelles, la détection d'anomalies et la prévision sur plusieurs horizons [66].

L'article [67] présente un modèle Transformer pour séries temporelles multivariées, et le teste sur des données IoT industrielles. Les résultats ont été meilleurs que les méthodes classiques (LSTM,…). Afin de mieux comprendre les spécificités des modèles utilisés pour le traitement des séries temporelles, voir ci-dessous la Table II.2, qui compare les avantages et les inconvénients des modèles **RNN** et **Transformer** pour le traitement des séries temporelles.

Modèle	Avantages	Inconvénients
Réseaux de Neurones (RNN)	- Bien appropriés pour les données séquentielles - Possèdent une mémoire interne (états cachés) - Moins coûteux pour des séquences de courte durée - Faciles à mettre en œuvre	- Difficulté à saisir des dépendances sur le long terme - Problèmes de gradient qui disparaît/explose - Entraînement peu rapide (séquentiel) - Moins adaptable au parallélisme
Transformer pour séries temporelles	- Aptitude à saisir des dépendances globales - Totalement parallélisables (accélération de l'entraînement) - Attention compréhensible - Plus efficaces pour les séquences longues	- Grande exigence en mémoire (particulièrement pour les séquences de longue durée) - Nécessite un plus grand volume de données pour une meilleure généralisation. - Traitements préliminaires parfois sophistiqués (codage de position, masquage)

Table II.2: Analyse comparative des avantages et inconvénients entre RNN et Transformer pour séries temporelles

II.2. 3 Prédiction basée sur les méthodes statistiques intelligentes

Les méthodes statistiques intelligentes sont réputées pour leur performance dans l'analyse et la prédiction de données chronologiques. Ils font partie des méthodes statistiques avancées. Ces techniques offrent la possibilité de modéliser les tendances, les variations saisonnières et les éléments aléatoires, tout en demeurant assez peu onéreuses du point de vue exécution.

A) Modèle ARIMA/SARIMA

Le modèle ARIMA (Moyenne Mobile Intégrée Autorégressive) fusionne les éléments d'auto régression, de différenciation (intégration) et de moyenne mobile pour représenter des séries temporelles stationnaires ou transformées en stationnaires. Le modèle SARIMA (ARIMA saisonnier) est une extension d'ARIMA qui prend en compte explicitement un élément saisonnier, le rendant ainsi approprié pour les séries présentant des fluctuations périodiques [68].

Les chercheurs dans [69] ont présenté comment prédire la consommation d'énergie dans les prochaines heures ou jours, en se basant sur des capteurs IoT qui mesurent la consommation d'énergie (électricité, chauffage, etc.) en temps réel dans les bâtiments intelligents (smart buildings) avec le modèle SARIMA. Afin d'éviter les pics de consommation, et d'optimiser l'usage des ressources pour ainsi réduire les coûts énergétiques et régler automatiquement les systèmes de chauffage/climatisation. Le modèle SARIMA a plusieurs avantages entre autres c'est un modèle statistique robuste pour les séries temporelles linéaire, il est facile à mettre en œuvre pour les données stationnaires et saisonnières et il n'a pas besoin de jeu de données énorme.

B) Model State Space Models /Kalman filters

Les modèles d'espace d'état décrivent les séries temporelles en utilisant des variables latentes qui progressent dans le temps. L'algorithme de Kalman est un outil récursif qui sert à évaluer ces variables invisibles et à réaliser des projections malgré la présence de bruits ou d'incertitudes. Ces techniques se révèlent particulièrement bénéfiques pour les séries non stationnaires et dans des situations où les données sont incomplètes ou entachées de bruit [70].

L'article [71] explore une application de suivi et de prédiction de la position d'un véhicule autonome dans une smart city en se basant sur filtre de Kalman. Une voiture connectée utilise le filtre de Kalman pour estimer sa position exacte à partir de données GPS bruitées, et ainsi planifier sa trajectoire dans une zone urbaine dense. Parmi les avantages de ce modèle

l'adaptation à des modèles réelles, la fiabilité avec des jeu de données incomplets ou bruités et la précision dans la fusion de capteurs IoT.

II.2.4 Les méthodes basées sur l'apprentissage non supervisé et clustering

La prédiction des données par apprentissage non supervisé se fait sans recourir à des étiquettes de données [72]. L'objectif consiste à pousser la machine à produire sa propre image à partir de l'environnement, avant d'élaborer une solution. L'auto organisation est une caractéristique de ce genre, contrairement à l'apprentissage supervisé où les données sont annotées par des personnes [73].

1. K-Means

Le K-Means, même s'il est bénéfique pour identifier certaines consommations irrégulières, ne se révèle pas toujours être un outil solide pour la prédiction des données, en particulier lorsque les clusters ne sont pas clairement établis ou sont impurs [74]. C'est un algorithme élémentaire d'agrégation des observations autour de centres en déplacement. Les paramètres essentiels du K-Means incluent le nombre de clusters et les points sélectionnés en tant que centroides. L'inconvénient majeur de cet algorithme est le choix du nombre de classes car il peut définir la qualité du clustering [75].

Dans l'article [76], les chercheurs ont utilisé k-Means pour identifier automatiquement les anomalies ou les comportements inhabituels dans la consommation d'eau pour la sensibilisation et la maintenance préventive. Parmi les avantages de ce modèle : mise en œuvre facile, exécution rapide, efficacité sur les jeux de données volumineux. Il peut aussi être une référence pour des systèmes de prévision indirecte basée sur les similarités et de détection d'anomalies.

2. Density-based algorithm for discovering clusters in large spatial databases with noise (DBSCAN)

DBSCAN [77] est le premier algorithme de regroupement fondé sur la densité. Les chercheurs ont proposé ce concept en 1996, visant à organiser des données de formes diverses en présence de bruit dans des bases de données, qu'elles soient spatiales ou non, de grande taille [78]. Il se distingue grâce à sa faculté d'isoler les anomalies en tant que bruit, cependant, il peut passer à côté des erreurs artificielles fortement dépendantes de la variable temporelle [74].

L'article [79] donne la possibilité de détecter automatiquement les anomalies médicales dans les données physiologiques (ex. : fréquence cardiaque, saturation en oxygène, température) en se basant sur l'algorithme DBSCAN afin de détecter des changements soudains ou dangereux dans l'état de santé du patient. Les résultats ont montré une excellente précision de détection d'anomalies.

3. Isolation Forest (IF)

Isolation Forest est un algorithme proposé par Fei Tony Liu, Kai Ming Ting et Zhi-Hua Zhou en 2008 [80]. Il emploie une collection d'arbres décisionnels pour identifier les anomalies plus rapidement que les observations standards. Il évalue la normalité d'une observation en déterminant le nombre moyen de divisions d'arbres nécessaires à son isolement [80]. Il est généralement employé pour prédire des comportements anormaux ou défaillants dans un système IoT, ce qui permet gérer les pannes, les intrusions, ou les conditions dangereuses.

L'article [81] présente un système pour détecter de manière précoce les signes d'une future défaillance d'un équipement avant qu'elle ne survienne afin d'éviter les arrêts coûteux en utilisant IF. Les résultats ont montré que IF est une méthode robuste, rapide, et précise pour la détection d'anomalies dans les flux de données capteurs issus de l'IoT industriel.

II.2.5 Les méthodes Hybrides

Les stratégies hybrides cherchent à combiner les avantages de divers algorithmes d'apprentissage pour mieux refléter la complexité des données. Le but premier est de saisir à la fois les relations linéaires, généralement bien représentées par des méthodes statistiques telles que l'ARIMA ou la régression linéaire, et les relations non linéaires qui sont efficacement modélisées par les modèles d'apprentissage automatique ou profond, tels que les réseaux de neurones ou les arbres de décision.

1. ARIMA combiné avec un Réseau de Neurones

Un modèle hybride intègre ici ARIMA, qui saisit les tendances ainsi que les éléments linéaires, associé à un réseau de neurones chargé d'apprendre les résidus non linéaires restants. Cette association optimise l'exactitude globale des prédictions, particulièrement dans le cas de séries temporelles compliquées [82].

L'article [83] présente une application de surveillance de la qualité de l'air de stations IoT basé sur l'hybridation d'ARIMA et de LSTM. Les résultats ont montré que cette hybridation est supérieure à chaque méthode utilisée séparément, car elle combine la rigueur statistique des modèles classiques avec la flexibilité des réseaux neuronaux profonds. Elle est donc très adaptée à la prévision intelligente de données environnementales complexes issues de l'IoT.

2. SVM combiné avec la Régression Logistique

Un autre cas serait l'utilisation d'un SVM pour distinguer les données non linéaires dans un espace modifié, suivi par l'application d'une régression logistique pour donner une estimation probabiliste de la classification. Ceci offre l'opportunité d'exploiter la solidité du SVM tout en fournissant une interprétation probabiliste du résultat [84].

L'article [85] présente une application de télésurveillance médicale basée sur une combinaison de SVM et la régression logistique. Cette application permet de de prédire l'hyperglycémie grave à l'avance et d'éviter l'hospitalisation. Les résultats ont montré que cette combinaison offre une précision meilleure que les deux méthodes utilisées séparément.

II.3 Cycle de vie de la prévision des données dans IoT

Dans cette partie, nous expliquons les cinq étapes pour la prédiction des données dans IoT en impliquant des modèles prédictifs :

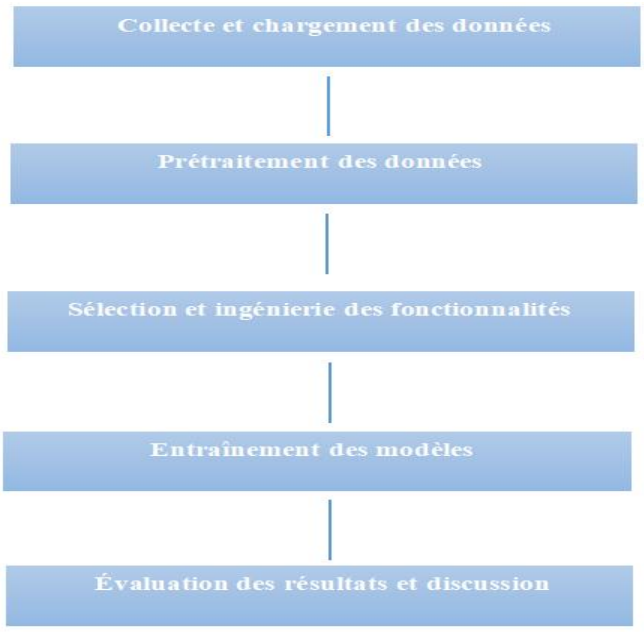

Figure II.2 : Cycle de vie de la prédiction des données dans IoT

II.3.1 Collecte des données

La collecte de données représente la première étape essentielle dans toute démarche de modélisation prédictive ou d'analyse de données. Cela signifie rassembler les données initiales indispensables pour saisir, modéliser et régler un problème spécifique en employant des méthodes de statistique ou d'apprentissage automatique.

II.3.2 Prétraitement des données

Le prétraitement des données constitue une phase essentielle dans l'analyse de données et l'apprentissage automatique. Il offre la possibilité de nettoyer et de convertir les données brutes en un format approprié pour l'entraînement des modèles. Voici les étapes de prétraitement principales :

- Chargement des données.
- Nettoyage des données (suppression des valeurs manquantes, correction des incohérences ou doublon...).
- Encodage des variables cible et catégoriel.
- Sélection des variables pertinentes(feautres).
- Conversion des variables numériques.
- Normalisation des données d'entrée.
- Division du jeu de données (entraînement et test).

II.3.3 Entraînement et évaluation des modèles

Après avoir préparé les données, l'étape suivante consiste à sélectionner et entraîner un ou plusieurs modèles de prédiction. Le processus se déroule habituellement de la manière suivante :

- Sélection d'un ou plusieurs algorithmes (régression, arbre décisionnel, SVM, LSTM, etc.).
- Formation du ou des modèles à partir des données d'apprentissage.
- Recours aux indicateurs d'évaluation (précision, rappel, F1-score, AUC, RMSE...).
- Évaluation comparative des performances des modèles.

II.4 Application de prévision des données dans IoT

Dans cette section, nous illustrons quelques exemples concrets d'utilisation des modèles prédictifs dans l'IoT :

II.4.1 Prédiction de la qualité de l'air avec des systèmes IoT

Dans [86], l'ambition première était de concevoir un système IoT en mesure de contrôler la qualité de l'air en temps réel et de prédire les taux de PM2.5 (particules fines en suspension dans l'air dont le diamètre est inférieur ou égal à 2,5 micromètres (µm)), en s'appuyant sur des appareils situés à la périphérie du réseau (computing edge) et des algorithmes d'apprentissage automatique(CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), Modèle NARX (Nonlinear AutoRegressive model with eXogenous inputs)).

II.4.2 Prédiction des maladies cardiovasculaires via des dispositifs IoT

L'objectif principal dans [87] était de concevoir un modèle capable de prédire les affections cardiovasculaires à l'avance en exploitant des informations recueillies grâce à des appareils IoT. Ce qui a permis une meilleure prise en charge des patients en termes de qualité et de rapidité. La recherche emploie divers algorithmes d'apprentissage supervisé pour effectuer des tâches de classification et de prédiction : Random Forest, Arbre de décision, Naïve Bayes, k-plus proches voisins (k-NN), Machine à vecteurs de support (SVM).

II.4.3 Prédiction de l'utilisation des cartes bancaires via la géolocalisation

L'objectif de la recherche dans [88] était de perfectionner la prédiction de l'usage des cartes bancaires en incorporant les informations de géolocalisation des clients. Le concept central est que les opérations bancaires (achats, retraits) sont fréquemment associées à des emplacements précis et habituels, tels que des magasins locaux, des centres commerciaux ou des machines distributrices placées à divers lieux. Des arbres de décision à gradient boosting tels que XGBoost ou LightGBM ont été employés.

II.4.4 Prédiction des risques pendant la grossesse

Le but de la recherche exposée dans [89] est d'optimiser le suivi de la santé des mères en milieu rural à l'aide de capteurs connectés (IoT) et d'algorithmes de machine learning. Les chercheurs ont fait appel à des capteurs mobiles (RFID, thermomètres, moniteurs de fréquence cardiaque,

etc.) pour recueillir en direct les informations médicales des femmes enceintes. On procède ensuite au traitement de ces informations afin d'estimer le degré du risque de mortalité (faible, moyen, élevé) associé à la grossesse. Un modèle de type arbre de décision intégrant des fonctions de régression logistique (LMT – Logistic Model Tree) a été employé, affichant une précision proche de 98 % sur la base de données Pima-Indians-diabetes, tout en présentant une bonne correspondance avec les informations réelles recueillies via l'IoT. Le système offre une surveillance à distance efficace et pourrait participer à la diminution de la mortalité maternelle dans les régions où l'accès est limité.

II.4.5 Prédiction du diabète à l'aide d'algorithmes d'apprentissage automatique

De nombreuses recherches se sont servies du jeu de données Pima Indiens Diabètes [39] pour prédire le diabète en utilisant différents algorithmes. On a effectué une comparaison entre plusieurs modèles tels que KNN, Random Forest, SVM, XGBoost, régression logistique et Naive Bayes. XGBoost est reconnu pour sa précision et sa solidité, alors que Random Forest a démontré une stabilité appréciable. KNN et Naive Bayes ont produit des résultats satisfaisants en étant faciles à mettre en œuvre. Bien que la régression logistique puisse être interprétée, elle est néanmoins moins performante. Ces méthodes attestent du potentiel de l'intelligence artificielle dans le diagnostic du diabète.

II.4.6 Prédiction des maladies cardiaques à l'aide du Gradient Boosting

Dans l'étude [43], les auteurs ont utilisé l'algorithme du Gradient Boosting Classifier sur le Cleveland Heart Disease Dataset dans le but de prédire l'existence de maladies cardiaques à partir de caractéristiques cliniques comme l'âge, la pression sanguine, le taux de cholestérol ou les résultats d'un électrocardiogramme. Ils ont effectué une comparaison avec le modèle KNN. Les performances révélées par l'analyse indiquent que le Gradient Boosting a surpassé KNN en termes de précision, de rappel et score AUC-ROC, notamment grâce à son aptitude à modéliser des relations complexes et non linéaires entre les variables. Ces résultats soulignent l'efficacité du Gradient Boosting pour les tâches liées à la classification médicale.

II.4.7 Prédiction du cancer du sein à l'aide de SVM

Dans [52], les chercheurs ont employé l'algorithme SVM sur le jeu de données du Wisconsin Breast Cancer Dataset (WBCD), dans le but de prédire l'existence du cancer du sein à partir des

caractéristiques cliniques issues des cellules mammaires. L'analyse a démontré un taux de précision élevé de 97,85 %, ce qui atteste que le SVM est un classificateur particulièrement efficace pour les tâches de classification médicale, en particulier dans le secteur de l'oncologie.

II.5 Conclusion

Ce chapitre contient les diverses méthodes de prédiction des données dans IoT. Il examine plusieurs techniques. Chaque d'entre elle présente des avantages et des contraintes en fonction du type de données ou du problème à résoudre. Nous explorons également le cycle de vie d'une prédiction des données, notamment quelques applications de la prédiction des données dans IoT. Dans le chapitre suivant, nous allons parler de notre contribution.

Chapitre III

Étude comparative des méthodes de prédiction dans l'IoT

III.1 Introduction

Ce chapitre expose la démarche employée pour estimer le risque d'accouchement par césarienne à partir des informations cliniques provenant de la base de données Maternal Health Risk Data qui se trouve sur le site de Kaggle [94]. Notre contribution se base sur l'exploitation de quatre facteurs primordiaux (l'âge, la pression artérielle systolique, la pression artérielle diastolique et le taux de glucose sanguin) et l'implémentation de divers modèles d'apprentissage automatique tels que la régression logistique, Random Forest, SVM, Gradient Boosting et XGBoost. Nous exposons le cycle de vie complet d'une prédiction données (de leur collecte à leur évaluation). La phase de prétraitement est également détaillée, ainsi que les outils utilisés, tant logiciels que matériels. En dernier lieu, les résultats des prédictions sont examinés pour déterminer le modèle offrant de meilleures performances.

III.2 Outils logiciels et matériels

Pour la réalisation de notre application de fin d'étude, nous avons employé des outils matériels et logiciels.

III.2.1 Outils logiciels

Pour analyser les modèles prédictifs appliqués à la prévision des données, nous avons utilisé les outils suivants :

A) Jupyter Notebook : Jupyter Notebook est un outil interactif qui simplifie la création et la diffusion de documents contenant du code, des graphiques et du contenu textuel. Il est employé au sein des universités et dans l'industrie dans l'analyse de données, la visualisation des données et 'apprentissage automatique [90].

B) Python : Python est un langage de programmation évolué, souvent utilisé pour l'analyse des données, la conception de sites internet, l'apprentissage automatique et de nombreuses autres applications. Il est très apprécié par la communauté scientifique pour sa simplicité et son large éventail de bibliothèques [91].

C) Spyder : autrefois appelé Pydee, c'est un environnement de développement pour le langage Python. Sous licence MIT et compatible avec diverses plateformes (Windows, macOS, Linux), ce logiciel rassemble un éventail de bibliothèques destinées à l'application scientifique : Matplotlib, NumPy, SciPy et IPython. Cet outil, conçu pour les chercheurs, ingénieurs et

analystes de données, offre des fonctionnalités avancées pour le développement scientifique en Python [92].

Nous avons utilisé Jupyter Notebook pour écrire du code, Python comme langage de programmation et Spyder pour l'interface graphique.

III.2.2 Outils matériels

Selon l'article [93] la collecte des données a été faite par des capteurs corporels IoT connectés à un réseau personnel (Personal Area Network – PAN), connectés avec WIFI/Bluetooth/ZigBee comme montré dans la figure III.1 [93]. Comme pour chaque application, nous avons fait recours à un ordinateur portable de marque Dell, équipé d'un processeur Intel Core i7 et de 8 Go de mémoire RAM, fonctionnant sous Windows 10 pour analyser les données et ainsi prédire la probabilité d'accoucher par césarienne.

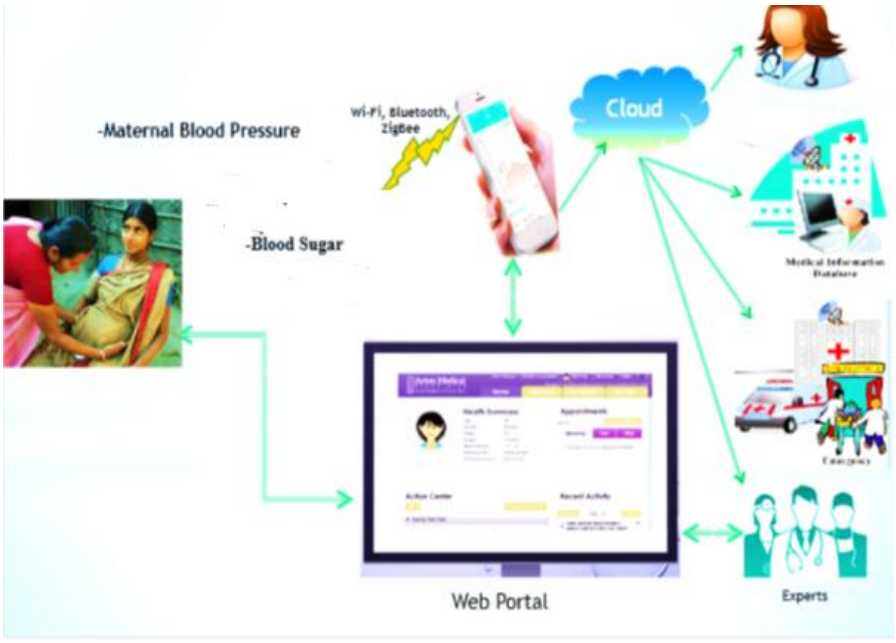

Figure III.1: Collecte des données [93]

III.3 Présentation de la contribution

La réalisation de notre projet de fin d'étude s'est déroulée en trois étapes. Au départ, notre thématique portait sur la prévision des données dans IoT. Comme nous sommes passionnées par le domaine médical, nous avons décidé d'orienter notre travail vers une application concrète de l'IoT dans le secteur de la santé.

La première étape était donc été le choix d'une base de données appropriée sur le site Kaggle [94]. Nous avons recherché un jeu de données médicales comportant des mesures physiologiques à l'aide d'équipements IoT. Nous avons choisi une base de données nommée « Maternal Health Risk Data » [95]. Elle intègre diverses variables cliniques, y compris la fréquence des battements du cœur (HeartRate), la température du corps (BodyTemp), l'âge, la pression artérielle systolique et diastolique, et la glycémie. Ces mesures ont été collectés via des dispositifs connectés à IoT. Après des recherches et après consultation des spécialistes (médecin généraliste et gynécologue), nous avons opté pour la prédiction du taux de probabilité d'accoucher par césarienne. Pour cela, nous avons sélectionné seulement quatre variables explicatives dans notre ensemble de données : L'âge de la patiente, la pression artérielle maximale, La pression artérielle en phase diastolique, la glycémie (BS).

Par la suite, dans la seconde phase, nous procédons à la prédiction des données, un processus que nous allons décrire en détail dans le reste du document.

Pour finir, la troisième phase porte sur l'interface graphique, que nous examinerons dans les sections suivantes du manuscrit.

III.4 Cycle de vie de la prédiction du taux de probabilité d'accoucher par césarienne

Dans cette partie, nous allons expliquer les phases du cycle de vie d'une prédiction des données appliquer à notre contribution.

III.4.1 Collecte des données

Il est essentiel de collecter des données issues de diverses sources cliniques et démographiques pour prédire le risque d'accouchement par césarienne. Notre étude s'appuie sur deux catégories majeures de données :

• **Informations démographiques et médicales :** Ces informations englobent l'âge de la patiente, les valeurs de pression artérielle systolique et diastolique, ainsi que le taux glycémique (BS), qui sont des éléments essentiels pour apprécier le risque.

• **Informations obstétricales :** Cela peut inclure des dossiers médicaux, des états de la grossesse, l'avis d'un spécialiste et d'autres éléments pertinents qui déterminent le type d'accouchement.

Dans le cadre de cette étude, nous supposons que toutes les données ont été mesurées exclusivement sur des **femmes enceintes à terme**, c'est-à-dire ayant une grossesse de 38 semaine complète. Par ailleurs, nous avons considéré aussi que le **fœtus était en bonne santé**.

Dans ce contexte, nous avons exploité un lot de données médicales recueillies avec l'approbation nécessaire, comprenant des indicateurs cliniques significatifs pour anticiper le taux de probabilité d'accouchement par césarienne. Nous avons choisi ce jeu de données car il répond à nos critères de pertinence et de qualité des variables.

Comme déjà mentionner, nous avons choisi seulement quatre variables de la base de données « Maternal Health Risk Data qui sont : l'âge de la patiente, la pression artérielle maximale, la pression artérielle en phase diastolique, la glycémie (BS).

III.4.2 Explication de la définition de la variable cible

Dans le contexte de notre recherche, nous avons introduit une variable cible nommée « taux de probabilité d'accouchement par césarienne », catégorisée en trois niveaux : bas, moyen ou haut. Cette variable a été marquée en fonction des standards médicaux accessibles et constitue le fondement de l'apprentissage supervisé dans notre système prédictif.

Le tableau III.1 catégorise les risques liés à l'accouchement par césarienne en trois niveaux (Haut, Moyen, Bas), en se basant sur quatre critères médicaux : âge de la patiente, Pression artérielle systolique (SystolicBP), Pression artérielle diastolique (DiastolicBP), Niveau de glucose dans le sang (Blood Sugar ou BS). Chaque paramètre est lié à des plages de valeurs qui représentent un degré de risque.

Paramètre	Haut	Bas	Moyen
Age	>40ans	<35ans	35ans à 40ans
SystolicBP	>140mmhg	<130mmhg	130 à 140mmhg
DiastolisBP	>100mmhg	<90mmhg	90 à 100mmhg
BS	>10mmol/L	<7mmol/L	7 à 10mmol/L

Table III.1 : Paramètres médicaux liés à la grossesse avec les valeurs correspondantes

a. **Age :** Le risque d'une césarienne s'accroît avec l'âge. Les femmes âgées de plus de 40 ans ont un risque accru de complications obstétricales (comme la dystocie, l'hypertension gravidique et le diabète gestationnel), ce qui explique la classification du risque comme élevé.
b. **Pression artérielle systolique (systolicBP) :** Une pression systolique supérieure à 140 mmHg indique une hypertension pendant la grossesse ou une pré-éclampsie, des états liés à un risque accru d'accouchement par césarienne.

Ces complications peuvent imposer une intervention rapide pour l'extraction du fœtus afin de protéger la santé de la mère et du bébé.

c. **Pression artérielle diastolique (diastolicBP) :** Tout comme pour la pression systolique, une pression diastolique supérieure à 100 mmHg accentue l'indication d'une hypertension grave. Un risque accru est donc directement lié à une perfusion placentaire insuffisante et à une éventuelle détresse fœtale.
d. **Taux de sucre dans le sang(BS) :** Un niveau élevé de sucre (> 10 mmol/L) indique un diabète gestationnel non maîtrisé. Cela peut entraîner une macrosomie fœtale (un gros bébé), ce qui augmente le risque d'échec de l'accouchement par voie basse.
   Les complications du diabète gestationnel comprennent un travail prolongé, la nécessité d'une césarienne d'urgence ou des dangers pour le nouveau-né (hypoglycémie néonatale, détresse respiratoire).

- Pour conclure une patiente répondant à plusieurs critères dans la colonne « Haut » (par exemple : âge > 40, SystolicBP > 140, BS > 10) présente une forte prédisposition à un accouchement par césarienne. Il s'agit d'une candidate typique pour une surveillance intensive, voire une planification de la césarienne.
  En revanche, une patiente jeune (moins de 35 ans), ayant une pression artérielle normale et un taux de sucre faible, représente un bon candidat pour un accouchement par voie basse (risque « Bas »).
  Les situations « Moyennes » se réfèrent à des zones de transition où le risque peut varier selon d'autres critères cliniques (indice de masse corporelle, antécédents, anomalies obstétricales, etc.). Ils requièrent une surveillance plus stricte.

III.4.3 Prétraitement des données

Le prétraitement des données constitue une phase essentielle dans la prédiction du taux de probabilité d'accouchement par césarienne. Il offre la possibilité de purger et de convertir les données brutes en un format approprié pour l'entraînement des modèles. Voici les étapes de prétraitement principales que nous avons mises en œuvre :

A) Traitement des données manquantes : Il est possible que les données cliniques comportent des valeurs manquantes, comme c'est le cas pour les relevés de pression artérielle ou de concentration en sucre dans le sang. Afin d'éviter que ces données absentes n'affectent négativement la formation du modèle, nous avons recours à des techniques comme l'élimination

des lignes incomplètes ou le remplacement par la médiane.

B) Normalisation des données : Des variables comme l'âge, la pression systolique, la pression diastolique et le taux de glucose (BS) disposent de différentes unités et plages de valeurs. Afin de s'assurer que chaque variable participe de façon équilibrée à la prédiction, nous avons mis en œuvre une standardisation (z-score), qui recentre les données et les ajuste à une échelle comparable.

C) Codage de la variable cible : Le taux de probabilité d'accoucher par césarienne, qui est la variable à prédire, est qualitatif (bas, moyen, haut). Pour l'apprentissage supervisé, cette variable a été convertie en valeurs numériques. Par exemple, pour une classification binaire, la valeur 0 indique un accouchement normal, englobant les niveaux de risque 'bas' et 'moyen', alors que la valeur 1 fait référence à une césarienne, liée à un niveau de risque 'haut'.

D) Contrôle de la répartition des variables : L'examen de la distribution statistique des données nous a aidés à identifier d'éventuelles irrégularités (valeurs aberrantes ou incohérentes), tout en confirmant l'équilibre approprié entre les catégories de sortie.

E) Éventuelle création de nouvelles variables : Nous avons aussi considéré l'élaboration de variables dérivées, à l'instar d'une moyenne de la pression artérielle, pour mieux saisir l'effet de certains paramètres cliniques sur le danger d'accouchement.

F) Analyse de la corrélation entre les variables : Une matrice de corrélation a été créée afin de représenter les liens linéaires entre les variables indépendantes (âge, BS, etc.) et la variable dépendante. Cette étude nous a aidés à valider la pertinence des variables choisies pour la formation du modèle.

G) Entraînement des modèles prédictifs : Suite au processus de prétraitement des données, divers modèles d'apprentissage automatique ont été formés dans le but d'estimer le danger d'un accouchement par césarienne. Les algorithmes choisis englobent diverses méthodes, facilitant ainsi une évaluation comparative des performances : Gradient Boosting, Forêts aléatoires (Random Forest), Machines de vecteurs-support (SVM), Régression logistique, XGBoost.

Ces modèles ont été sélectionnés en raison de leur capacité à traiter des données cliniques complexes et de leur solidité démontrée dans la classification multi-classes. Chaque modèle a

subi une formation sur un jeu de données standardisé et équilibré, puis a été testé sur un ensemble distinct pour juger sa précision prédictive.

III.5 Estimation du risque de césarienne basée sur des modèles de classification supervisée

III.5.1 Estimation via Régression Logistique

Dans notre étude comparative, nous avons appliqué la régression logistique comme méthode de classification supervisée pour l'estimation du taux de probabilité de Césarienne à partir de données cliniques issues du jeu de données « Maternal Health Risk » [95]. Cette approche s'est révélée particulièrement pertinente en raison de sa simplicité d'interprétation et de sa capacité à modéliser une probabilité binaire à partir de variables continues.

Nous avons retenu les variables explicatives les plus influentes : l'âge de la patiente, la pression artérielle systolique (SystolicBP), **la** pression artérielle diastolique (DiastolicBP) et le taux de glycémie (BS)**.** La variable cible, représentant le risque d'accouchement par césarienne, a été encodée en deux classes : 1 pour un risque haut (probabilité d'accouchement par césarienne) et 0 pour un risque bas ou moyen (accouchement normal).

L'algorithme de régression logistique repose sur une fonction logistique (sigmoïde) qui permet de transformer une combinaison linéaire des variables en une probabilité comprise entre 0 et 1. Le modèle ainsi construit peut s'écrire sous la forme suivante :

$$\mathbf{P(Césarienne)} = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\beta 0 + \beta 1.\text{Age} + \beta 2.\text{SystolicBP} + \beta 3.\text{DiastolicBP} + \beta 4.\text{BS})}}$$

Dans notre cas, les valeurs des coefficients estimés sont les suivantes :

- Intercept (β_0) : -25.9805
- Coefficient âge (β_1) : 0.1990
- Coefficient de la pression systolique (β_2) : 0.0189
- Coefficient de la pression diastolique (β_3) : 0.0448
- Coefficient du taux de glycémie (β_4) : 1.6934

Les coefficients β de cette formule sont estimés automatiquement en Python via la bibliothèque scikit-learn en Python, en utilisant un algorithme d'optimisation numérique (généralement la

descente de gradient). Ce dernier vise **à** minimiser la fonction de coût, qui, dans le cas de la régression logistique, est **la** log-vraisemblance négative :

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m} [y^{(i)} \log\left(\hat{y}^{(i)}\right) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}]$$

Où :

- m est le nombre d'observations dans l'échantillon d'apprentissage,
- $y^{(i)}$est la classe réelle (0 ou 1),
- $\hat{y}^{(i)}$est la probabilité prédite par le modèle pour cette observation.

L'optimisation consiste à ajuster les coefficients β0, β1,…,β4 de manière itérative jusqu'à obtenir le meilleur ajustement possible entre les données et les prédictions.

Une fois le modèle entraîné, les coefficients peuvent être analysés de manière informative :

- **Signe du coefficient** :
  - Un coefficient $\beta i>0$ indique que la variable augmente la probabilité d'accouchement par césarienne.
  - Inversement, $\beta i<0$ signifie que cette variable diminue cette probabilité.
- **Valeur absolue** :
  - Plus le coefficient est éloigné de zéro, plus la variable associée a un impact fort sur la décision du modèle.
- **Odds Ratio** :
  - En calculant $e^{\beta i}$, on obtient **l'**odds ratio, c'est-à-dire l'effet multiplicatif de cette variable sur le rapport de chances (odds) d'un accouchement par césarienne. Cela offre une interprétation clinique concrète, utile pour les professionnels de santé.

Ainsi, le modèle nous a permis de :

- générer une probabilité personnalisée pour chaque patiente en fonction de ses données médicales,

- et intégrer ces résultats dans une interface graphique professionnelle, dans laquelle l'utilisateur peut saisir les données d'une patiente (nom, prénom, âge, pression, glycémie) et obtenir instantanément le taux de probabilité d'un accouchement par césarienne.

Grâce à sa transparence et sa fiabilité, la régression logistique a été identifiée comme l'un des modèles les interprétables de notre étude.

III.5.2 Estimation via Support Vector Machine (SVM)

Dans notre étude, le modèle SVM a été utilisé pour classifier le risque d'accouchement en deux classes : accouchement par voie normale ou par césarienne. Ce modèle est particulièrement adapté aux problèmes de classification binaire, car il cherche à trouver l'hyperplan optimal qui sépare au mieux les deux classes dans l'espace des variables.

Grâce aux données cliniques (âge, pression artérielle systolique, pression artérielle diastolique et glycémie), le SVM projette les observations dans un espace multidimensionnel et trouve une frontière de décision qui maximise la marge entre les deux groupes. Nous avons utilisé la version SVC de scikit-learn, avec un noyau radial (RBF**)** pour gérer les non-linéarités.

Ce modèle est le plus performants comparé aux quarte autres modèles comme montré dans la figure III.2 qui compare les cinq modèles, mais en termes d'interprétabilité la régression logistique reste le meilleur. Néanmoins, SVM est efficace pour capturer des relations complexes entre des variables médicales.

III.5.3 Estimation via Random Forest

Le modèle Random Forest a été mis en œuvre comme algorithme d'ensemble pour prédire le risque d'un accouchement par césarienne à partir des données médicales disponibles. Nous avons utilisé la bibliothèque scikit-learn en Python (RandomForestClassifier) pour entraîner ce modèle à partir des variables suivantes : âge de la patiente, pression systolique (SystolicBP), pression diastolique (DiastolicBP) et taux de glycémie (BS).

Random Forest fonctionne en construisant un grand nombre d'arbres de décision sur différents sous-échantillons des données, puis en combinant leurs prédictions (vote majoritaire pour la classification). Cette méthode permet de réduire le risque de surapprentissage (overfitting), souvent rencontré avec un seul arbre.

Pendant l'entraînement, nous avons évalué la précision globale, mais aussi d'autres métriques comme le rappel et la f-mesure pour mieux comprendre le comportement du modèle dans un contexte médical où les faux négatifs peuvent être critiques.

III.5.4 Estimation via Gradient Boosting

Nous avons également intégré un modèle de Gradient Boosting dans notre étude, en utilisant GradientBoosting de scikit-learn. Ce modèle repose sur la création séquentielle d'arbres de décision faibles, chaque nouvel arbre corrigeant les erreurs de prédiction des arbres précédents.

L'objectif est de minimiser une fonction de perte (log-loss dans notre cas) par descente de gradient. Ce processus permet d'obtenir un modèle plus performant, capable de capturer des relations complexes entre les variables cliniques.

Les données utilisées sont les mêmes que pour les autres modèles (âge, SystolicBP, DiastolicBP, BS), avec une phase de normalisation et d'encodage préalable. Après l'entraînement, nous avons observé que le Gradient Boosting offrait une meilleure précision moyenne et f1 score moyen que Random Forest, régression logistique et XGBoost, notamment dans la différenciation des cas à risque élevé.

En fin de processus, nous avons extrait les coefficients d'importance des variables et intégré les prédictions du modèle dans notre interface, afin d'afficher la probabilité associée à un accouchement par césarienne.

III.5.5 Estimation via XGBoost

Le modèle XGBoost (Extreme Gradient Boosting) a été implémenté via la bibliothèque xgboost en Python (XGBClassifier), qui permet un entraînement rapide, une meilleure gestion de la régularisation et des performances accrues sur les grands jeux de données.

Pour notre projet, nous avons utilisé ce modèle pour affiner la prédiction du taux de probabilité de césarienne, toujours à partir des variables cliniques (âge, SystolicBP, DiastolicBP, BS). Le modèle a été entraîné en ajustant des hyperparamètres comme le nombre d'estimators, le taux d'apprentissage (learning_rate) et la profondeur maximale des arbres (max_depth).

III.6 Évaluation des performances des cinq modèles de classification et analyse des résultats

III.6.1 L'évaluation en fonction de l'exactitude, de précision moyenne, du rappel moyen et de F1-score moyenne

Pour évaluer la performance des cinq modèles de classification (GradientBoosting, RandomForest, SVM, LogisticRegression, XGBoost), nous avons tracé un graphique comparatif basé sur quatre métriques essentielles : Accuracy, Précision moyenne, Recall moyen et F1-score moyen comme le montre la figure III.2. Ces valeurs ont été calculées en faisant la moyenne des scores obtenus pour chaque classe ('bas', 'moyen', 'haut'). La figure Figure III.2présente une vue d'ensemble des performances des modèles de classification testés dans le cadre de la prédiction du taux de probabilité d'accoucher par césarienne, à partir des métriques moyennes calculées sur les trois classes.

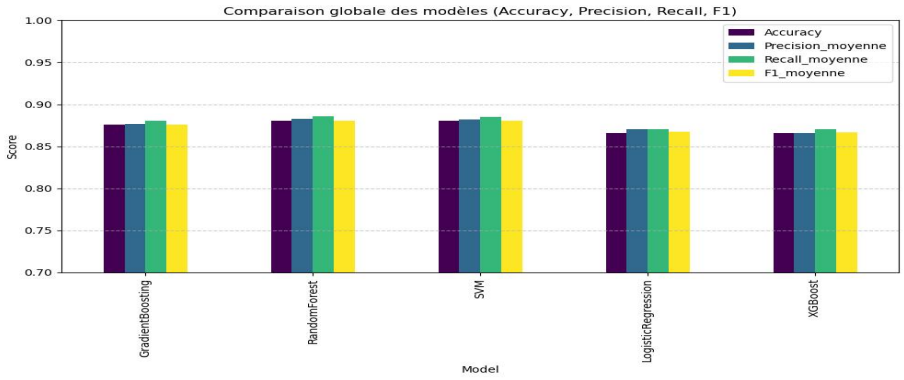

Figure III.2: Comparaison entre les cinq modèles en fonction d'Accuracy, de Precision moyenne, de Recall moyenne et de F1-score moyenne

1. **Exactitude (Accuracy) :**
   - Tous les modèles présentent une accuracy élevée (> 86 %), ce qui reflète une bonne capacité globale à prédire correctement les classes.
   - Le modèle SVM affiche l'accuracy la plus élevée (~88.1 %), suivi de près par GradientBoosting (87.6 %).
   - Les modèles LogisticRegression et XGBoost ferment la marche (86.6 %), mais restent très compétitifs.
2. **Précision moyenne :**
   - SVM et GradientBoosting ont la précision moyenne la plus élevée (~0.88), indiquant qu'ils produisent relativement peu de faux positifs.

- LogisticRegression, RandomForest et XGBoost restent autour de 0.87.

3. **Rappel moyenne :**
   - Le rappel (recall) est également supérieur chez le modèle SVM (0.89), indiquant une meilleure capacité à détecter correctement les classes positives (moins de faux négatifs).
   - Les autres modèles sont assez proches (0.87 à 0.88), avec peu de différence significative.
4. **F1-score moyenne :**
   - Les modèles SVM et GradientBoosting ont encore une fois les meilleurs scores F1 (0.88), ce qui indique un bon équilibre entre précision et rappel.
   - Les autres modèles oscillent entre 0.87 et 0.88.

Interprétation des résultats

- Le **SVM est** globalement le plus performant sur toutes les métriques, ce qui en fait un excellent choix pour cette tâche de classification multiclasse dans un contexte IoT.
- **GradientBoosting est** une alternative très proche, avec des scores légèrement inférieurs mais stables.
- **RandomForest** montre aussi de bonnes performances, mais légèrement en retrait par rapport aux deux premiers.
- **LogisticRegression**, bien que simple, reste compétitif, ce qui peut justifier son usage dans des systèmes où la simplicité et la rapidité sont importantes.
- **XGBoost** offre des résultats comparables à RandomForest et LogisticRegression, bien que son avantage sur des jeux de données plus grands puisse être plus significatif.

III.6.2 Analyse des distributions de probabilité pour une patiente donnée

La figure III.3 présente la distribution des probabilités de classification produites par cinq modèles d'apprentissage automatique (Gradient Boosting, Random Forest, SVM, Logistic Regression, et XGBoost) pour une nouvelle patiente dont les caractéristiques cliniques sont : âge de 32 ans, pression artérielle systolique de 130 mmHg, diastolique de 85 mmHg *et* taux de glycémie de 7 mmol/L.

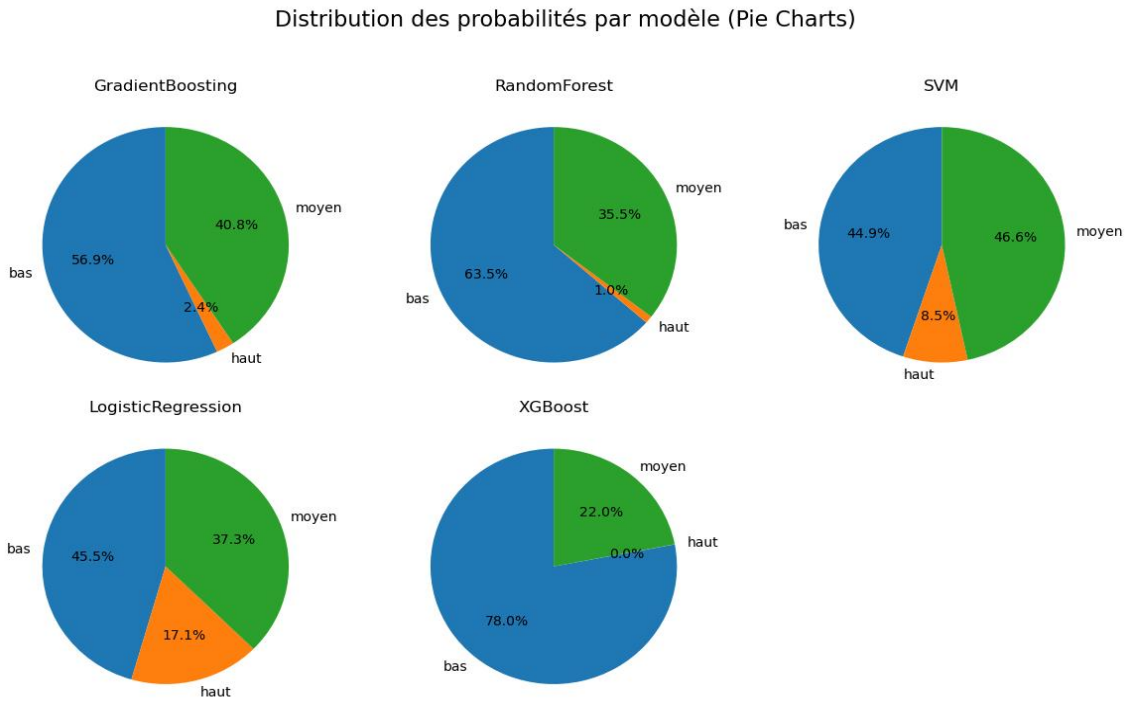

Figure III.3: Distribution des probabilités de classification pour un patient donné selon cinq modèles

Pour chaque modèle, les probabilités associées à chaque classe de risque (*bas*, *moyen*, *haut*) ont été visualisées sous forme des diagrammes circulaires. Cette analyse met en lumière les différences de comportement entre les algorithmes :

- **Gradient Boosting** prédit majoritairement la classe "bas" (56,9 %), avec une probabilité notable de "moyen" (40,8 %), et une faible part de "haut" (2,4 %). Cela indique un certain degré d'incertitude entre les classes "bas" et "moyen", mais un faible risque perçu d'un accouchement à haut risque.
- **Random Forest** montre un profil similaire, avec 63,5 % pour "bas" et 35,5 % pour "moyen", ne laissant qu'1 % pour "haut". Ce modèle est légèrement plus affirmatif dans sa prédiction que Gradient Boosting.
- **SVM (Support Vector Machine)** affiche une répartition plus équilibrée entre "bas" (44,9 %) et "moyen" (46,6 %), avec 8,5 % pour "haut". Cette configuration montre un modèle plus prudent, reconnaissant une possible ambiguïté entre les classes, ce qui peut être utile dans un contexte clinique où les transitions entre niveaux de risque sont parfois subtiles.

- **Logistic Regression** est le seul modèle à attribuer une part significative à la classe "haut" (17,1 %), bien qu'il penche vers "bas" (45,5 %). Ce résultat pourrait refléter une sensibilité plus élevée à certains indicateurs cliniques de ce cas précis.
- **XGBoost** se distingue nettement par une prédiction très marquée en faveur de "bas" (78 %), avec presque aucune probabilité pour "haut" (0,02 %), ce qui témoigne d'une forte confiance du modèle dans son verdict. Cela peut être perçu comme une force (confiance) ou une faiblesse (manque de prudence).

L'analyse visuelle des diagrammes circulaires confirme que tous les modèles s'accordent sur une prédiction finale de risque "bas", mais avec des degrés de certitude différents. Des modèles comme SVM ou Logistic Regression offrent des répartitions plus nuancées, ce qui peut être particulièrement pertinent pour prévenir les faux négatifs dans les cas cliniques complexes.

III.6.3 Comparaison du F1-Score par modèle et par classe :

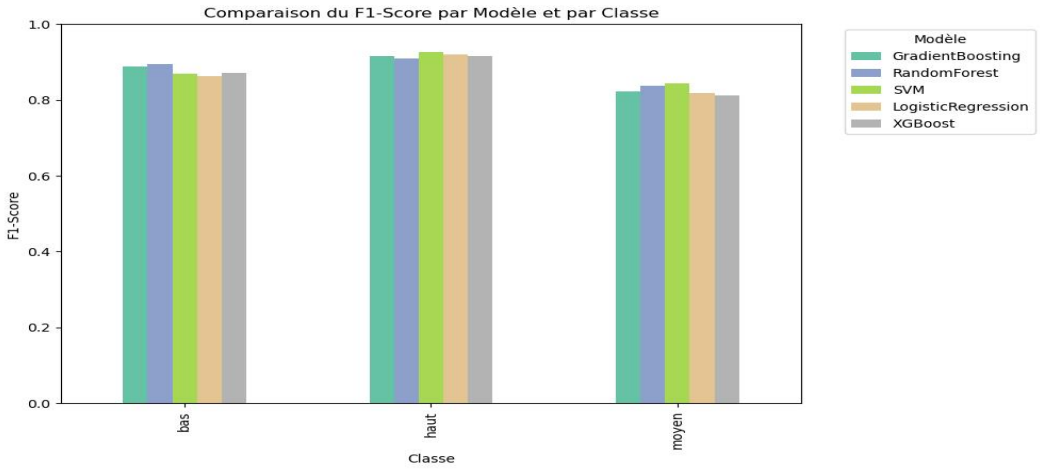

Figure III.4: Comparaison de F1 Score par modèles et par classe

Le graphique intitulé « Comparaison du F1-Score par modèles et par classe » permet d'évaluer la capacité des différents algorithmes à classer correctement les trois catégories de risque d'accouchement (bas, moyen, haut), en s'appuyant sur la métrique du F1-score, qui équilibre à la fois la précision et le rappel.

On observe que :

- Pour la classe haut, correspondant aux cas à fort risque de césarienne, les modèles Gradient Boosting, Random Forest et XGBoost obtiennent des F1-scores

particulièrement élevés, supérieurs à 0.90. Cela montre qu'ils sont très efficaces pour détecter les patientes à risque élevé, un aspect critique en contexte médical.

- Pour la classe bas, tous les modèles atteignent des performances proches, avec des F1-scores avoisinant ou dépassant 0.88. Cela traduit une bonne capacité globale des modèles à identifier les patientes à faible risque, ce qui permet d'éviter les interventions inutiles.
- La classe moyen présente les F1-scores les plus faibles, généralement autour de 0.82–0.85. Cette baisse peut s'expliquer par la nature intermédiaire de cette classe, qui partage des caractéristiques communes avec les deux autres, rendant sa détection plus incertaine.

Ces résultats confirment que, bien que tous les modèles soient relativement performants, les algorithmes d'ensemble comme Gradient Boosting et Random Forest parviennent mieux à modéliser les extrêmes (haut et bas) que les cas intermédiaires (moyen), qui exigeraient peut-être des méthodes de rééchantillonnage ou un affinement des seuils de décision.

III.6.4 Analyse des matrices de Confusion

Nous avons généré une matrice de confusion pour les modèles les plus performants : la régression logistique, Gradient Boosting et SVM. La matrice de confusion illustre la qualité de la classification des observations selon les trois classes de risque (bas, moyen, haut). Ce type de représentation est essentiel pour évaluer la capacité réelle du modèle à identifier correctement les différentes catégories de patientes.

Dans une matrice de confusion, chaque ligne représente la classe réelle, et chaque colonne la classe prédite. Elle permet ainsi de visualiser à la fois les prédictions correctes (valeurs sur la diagonale) et les erreurs de classification (valeurs hors de la diagonale).

Pour obtenir ces résultats, nous avons divisé notre ensemble de données en deux sous-ensembles : 80 % des données ont été utilisées pour entraîner le modèle (X_train, y_train), et 20 % restantes pour l'évaluation (X_test, y_test). C'est uniquement sur ce jeu de test que la matrice de confusion est calculée, afin de mesurer la capacité du modèle à généraliser sur des données jamais vues pendant l'entraînement.

a. Régression Logistique

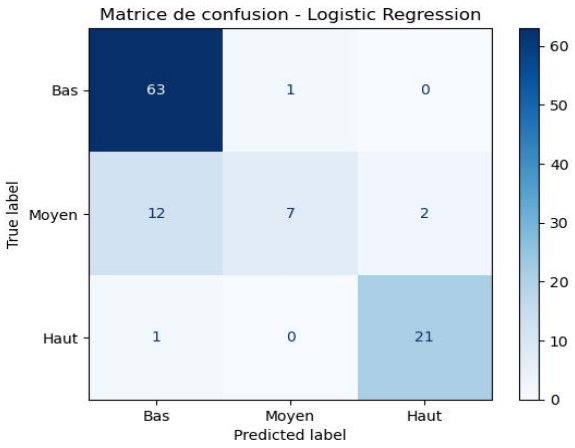

Figure III.5: Matrice de confusion-Logistic Regression

Classe bas :

Sur 64 patientes réellement classées comme bas risque, 63 ont été correctement identifiées. Une seule a été prédite à tort comme "moyen", et aucune comme "haut".

Cela démontre une excellente précision dans la prévision des patientes à faible risque, avec quasiment aucune confusion vers les classes plus graves. Le modèle se montre ici particulièrement fiable et rassurant sur les cas non critiques.

Classe moyen :

Parmi 21 patientes réellement à risque moyen, 7 ont été correctement classées, mais 12 ont été confondues avec la classe "bas", et 2 avec la classe "haut".

Cette distribution montre que la classe moyen reste la plus confuse, ce qui est typique dans les modèles supervisés lorsqu'une classe intermédiaire partage des caractéristiques avec les deux extrêmes. Le modèle a donc tendance à sous-estimer ou surestimer le risque pour ces patientes.

Classe haut :

Sur 22 patientes à haut risque, 21 ont été correctement identifiées, et 1 seule a été classée comme "bas".

Ce résultat est très satisfaisant, car la capacité à détecter les cas critiques est essentielle en contexte médical. Le modèle produit très peu de faux négatifs sur les cas à haut risque, ce qui contribue à renforcer sa fiabilité pour les décisions de suivi renforcé ou d'intervention anticipée.

Le modèle de Régression Logistique offre une très bonne précision pour les classes "bas" et "haut", le rendant fiable pour les cas cliniquement clairs.

Bien que la classe "moyen" reste plus difficile à distinguer, cela reflète la nature floue de cette catégorie.

Sa robustesse, simplicité et transparence font de la régression logistique un outil pertinent pour une première évaluation du risque d'accouchement.

b. Gradient Boosting

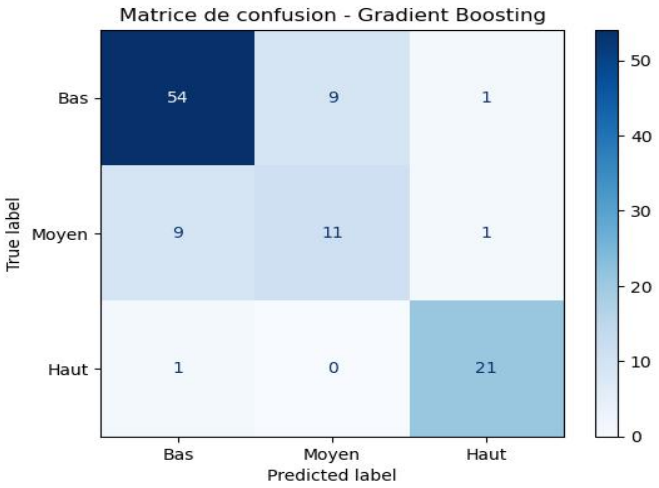

Figure III.6: Matrice de confusion Gradient Boosting

Classe bas :

Sur 64 patientes réellement classées comme bas risque, 54 ont été correctement identifiées. En revanche, 9 ont été classées à tort comme "moyen" et 1 comme "haut".

Cette performance montre que le modèle est globalement fiable pour reconnaître les patientes à faible risque, mais avec une tendance à surestimer légèrement le risque en les plaçant dans la classe "moyen". Cette erreur est cliniquement tolérable, car elle reste prudente dans la gestion du risque.

Classe moyen :

Parmi 21 patientes réellement à risque moyen, 11 ont été correctement prédites, 9 ont été confondues avec la classe "bas", et 1 avec la classe "haut".

Cela met en évidence que la classe moyen reste délicate à modéliser, avec une confusion importante vers la classe bas, ce qui pourrait entraîner une sous-estimation du risque réel chez

certaines patientes. Néanmoins, ce modèle fait mieux que SVM, en reconnaissant une proportion correcte de cas intermédiaires.

Classe haut :

Sur 22 patientes à haut risque, 21 ont été bien identifiées, et 1 seule a été confondue avec la classe "bas".

Ce résultat est excellente nouvelle du point de vue médical, car le modèle parvient à capturer presque tous les cas critiques, avec une faible probabilité de faux négatif, ce qui est essentiel pour éviter des complications lors de l'accouchement.

c. SVM

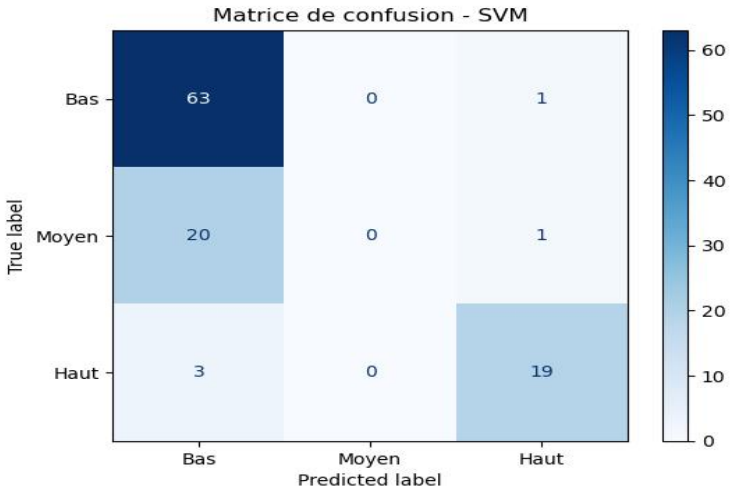

Figure III.7: Matrice de confusion-SVM

Classe bas :

Sur 64 patientes réellement classées comme bas risque, 63 ont été correctement identifiées. Aucune n'a été confondue avec "moyen", et 1 seule a été classée à tort comme "haut".

Ce résultat traduit une très bonne capacité du modèle à reconnaître les patientes à faible risque, avec quasiment aucune confusion vers les autres classes, ce qui est particulièrement rassurant dans un contexte de tri clinique.

Classe haut :

Parmi 22 patientes réellement à haut risque, 19 ont été correctement classées, 3 ont été confondues avec la classe "bas", et aucune n'a été classée comme "moyen".

Le modèle affiche ici une bonne précision, même si l'on observe quelques erreurs graves, car 3

patientes à haut risque ont été considérées à tort comme bas risque. Cela reste acceptable, mais suggère un besoin d'amélioration sur les cas critiques.

Classe moyen :

Sur 21 patientes à risque moyen, aucune n'a été correctement identifiée comme "moyen". 20 ont été classées comme "bas", et 1 comme "haut".

Cette performance révèle que la classe moyen est totalement ignorée par le modèle SVM, ce qui indique une incapacité à détecter les profils intermédiaires. Cette situation est problématique, car elle brouille la frontière entre bas et haut risque, et peut compromettre une prise en charge graduée.

Alors le modèle SVM démontre une excellente précision sur la classe "bas", et une bonne performance sur la classe "haut", ce qui en fait un modèle pertinent pour les prédictions extrêmes (risque faible ou élevé).

Cependant, son incapacité totale à reconnaître la classe "moyen" limite fortement son usage dans une application médicale qui doit proposer des décisions nuancées.

Malgré tout, son comportement binaire fort (bas ou haut) pourrait être exploité dans un système d'alerte simplifié ou de pré-tri médical, à condition de compléter avec un modèle plus fin pour les cas moyens.

III.7 Interface graphique

L'interface graphique a été faite par Spyder et python. La figure III.8 présentent l'interface graphique de notre contribution. Cette figure est sous forme de formulaire et elle contient 6 cases (nom, prénom, âge, tension systolique, tension diastolique, glycémie) à remplir, un ascenseur qui contient les cinq modèles de classification utilisés (régression logistique, SVM, Random Forest, Gradient Boosting et XGBoost) et un bouton de soumission. Après soumission, une nouvelle fenêtre s'affiche comme montrer dans les figures III.9, III.10, III,11, III.12 et III.13. Ces figures permettent de prédire le taux de probabilité d'accoucher par césarienne.

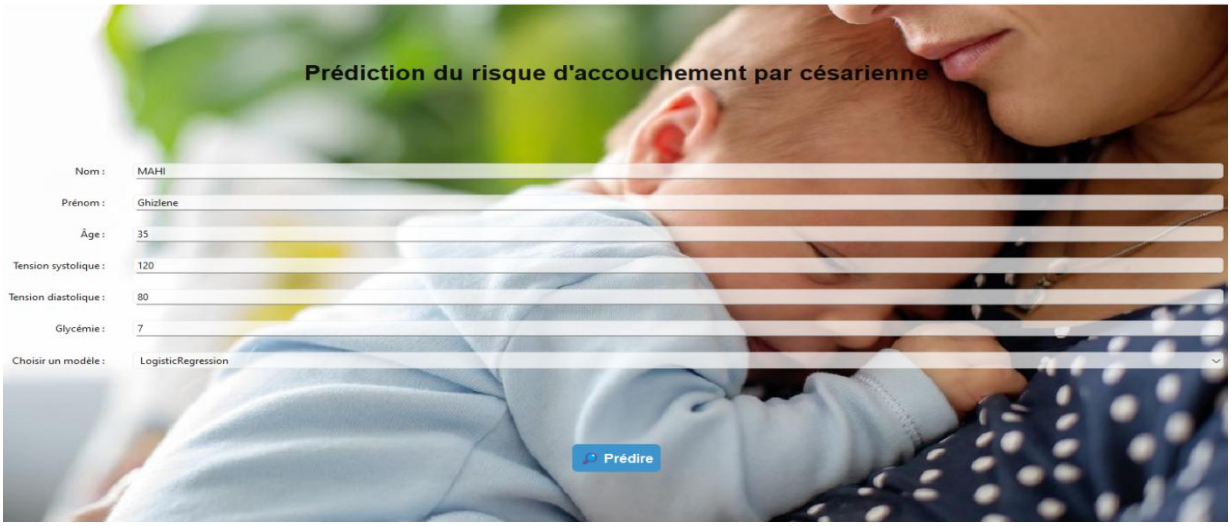

Figure III.8: l'interface d'accueil pour saisir les données

La figure suivante montre la prédiction du taux de probabilité d'accoucher par césarienne avec le modèle de classification régression logistique qui est de 42,59% pour cette patiente et à partir de cette probabilité, nous déduisons que le risque est moyen d'accoucher par césarienne.

Figure III.9: Rapport de prédiction de la patiente avec le modèle LogistiqueRegression

La figure suivante montre la prédiction du taux de probabilité d'accoucher par césarienne avec le modèle de classification Random forest qui est de 51,46% pour cette patiente et à partir de cette probabilité, nous déduisons que le risque est bas d'accoucher par césarienne

Figure III.10: Rapport de prédiction de la patiente avec le modèle RandomForest

La figure suivante montre la prédiction du taux de probabilité d'accoucher par césarienne avec le modèle de classification gradient boosting qui est de 58,65% pour cette patiente et à partir de cette probabilité, nous déduisons que le risque est moyen d'accoucher par césarienne

Figure III.11: Rapport de prédiction de la patiente avec le modèle GradientBoosting

La figure suivante montre la prédiction du taux de probabilité d'accoucher par césarienne avec le modèle de classification SVM qui est de 59,88% pour cette patiente et à partir de cette probabilité, nous déduisons que le risque est bas d'accoucher par césarienne.

Figure III.12: Rapport de prédiction de la patiente avec le modèle SVM

La figure suivante montre la prédiction du taux de probabilité d'accoucher par césarienne avec le modèle de classification XGBoost qui est de 54,97% pour cette patiente et à partir de cette probabilité, nous déduisons que le risque est moyen d'accoucher par césarienne

Figure III.13: Rapport de prédiction de la patiente avec le modèle XGBoost

Dans la figure suivante, nous allons essayer avec des données différentes des données précédentes

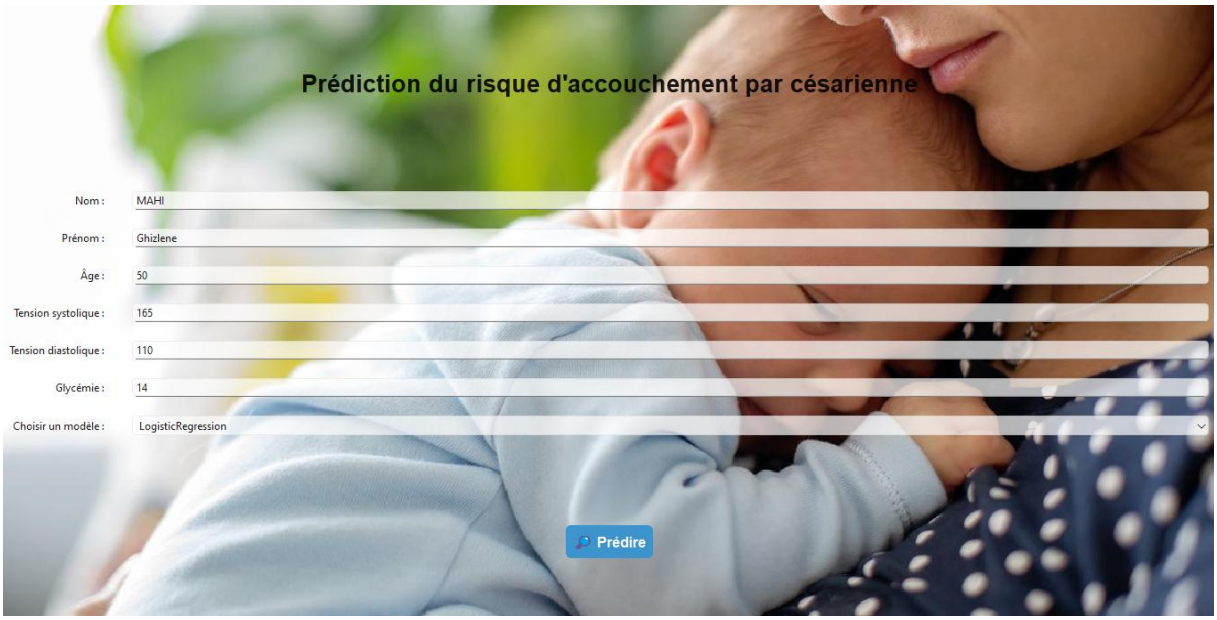

Figure III.14: Exemple de taux haut

La figure suivante montre la prédiction du taux de probabilité d'accoucher par césarienne avec le modèle de classification logistic Regression qui est de 99,95% pour cette patiente et à partir de cette probabilité, nous déduisons qu'il y'a de forte probabilité que cette patiente accouche par césarienne.

Figure III.15: Rapport de prédiction de la patiente de taux haut

III.8 Conclusion

Ce chapitre a exposé le processus de préparation des données et d'entraînement de cinq modèles de classification pour prédire le risque d'accouchement. Après la normalisation des variables cliniques, les modèles ont été évalués selon leur capacité à différencier les niveaux de risque (bas, moyen, haut). Les résultats obtenus montrent que le SVM est le modèle le plus performant selon toutes les métriques, suivi de près par Gradient Boosting et Random Forest, qui constituent également des alternatives fiables pour cette tâche de classification multiclasse dans un contexte IoT.

Conclusion générale

Conclusion Générale

Cette étude a examiné l'impact des technologies IoT et de l'intelligence artificielle dans le secteur du suivi de la grossesse, notamment pour prévoir le risque d'accouchement par césarienne. Nous avons conçu et examiné un système de prévision basé sur divers algorithmes d'apprentissage automatique supervisé, en utilisant des données cliniques élémentaires recueillies grâce à des appareils IoT connectés, comme l'âge, la pression artérielle et le taux de sucre dans le sang.

Dans la première section de notre manuscrit, nous avons présenté les principes de base de l'Internet des Objets et son importance grandissante dans notre quotidien vu son implication dans plusieurs domaines. Par la suite, nous avons examiné les techniques de prédiction des données dans IoT. Finalement, nous avons implémenté et effectué une comparaison de cinq modèles (Régression Logistique, SVM, Forêt Aléatoire, Gradient Boosting et XGBoost) dans l'objectif d'évaluer leur aptitude à juger les niveaux de risque (bas, moyen ou haut) d'un accouchement par césarienne.

Les résultats obtenus montrent que certaines approches offrent une meilleure précision et une meilleure discrimination entre les différentes classes de risque, ce qui confirme l'utilité de ces outils dans un contexte d'aide à la décision médicale. Toutefois, leur utilisation doit être envisagée avec prudence et nécessite une validation clinique stricte avant tout mise en œuvre dans la pratique.

Ce travail ouvre la voie à de futures améliorations : Nous proposons l'intégration d'autre équipements IoT pour avoir des données plus variées et plus riche (les mouvements, le poids et les battements du fœtus, le niveau du liquide amniotique, historique médical de la maman, imagerie, etc.) ; et pour avoir une estimation d'accouchement par césarienne plus optimale. Nous proposons aussi la réalisation d'une application pour le suivi de grossesses. Cette application va permettre non seulement de détecter et faire face aux fausses couches des grossesses mais aussi d'estimer le risque d'accoucher par césarienne. Et pour cela, il faut la combinaison de plusieurs modèles intelligents pour une prédiction des données plus performante. Dans ce contexte, l'union des objets connectés et de l'IA constitue une piste prometteuse pour optimiser le suivi des grossesses et renforcer la prévention des dangers.

Références Bibliographiques

Bibliographie

[1] Dia, I., & Sidoummou, I. (2023). Conception et implémentation d'une application d'internet des objets médicaux pour le suivi des patients diabétiques. Université Saad Dahlab Blida 1.

[2] Belhadj, N., & Abbad, A. (2022). La sécurité de l'Internet des Objets (IoT) (Thèse de doctorat, Université Ibn Khaldoun Tiaret).

[3] Evans, D. (2011). The Internet of Things: How the Next Evolution of the Internet Is Changing Everything. Cisco Internet Business Solutions Group.

[4] Lehsaini, M., & Benmahdi, M. B. (2018). An improved k-means cluster-based routing scheme for wireless sensor networks. 2018 International Symposium on Programming and Systems (ISPS), 1–6. IEEE.

[5] Casablanca, P. J. T. P. C. (s.d.). Suivi des palettes RFID, en temps réel, en utilisant la méthode LANDMARC.

[6] Ashton, K. (2009). That 'Internet of Things' Thing. RFID Journal.

[7] Benghozi, P.-J., Bureau, S., & Massit-Folea, F. (2008). L'Internet des objets : Quels enjeux pour les Européens ?

[8] Vermesan, O., Friess, P., Guillemin, P., Gusmeroli, S., Sundmaeker, H., Bassi, A., ... et al. (2011). Internet of things strategic research roadmap. In Internet of things-global technological and societal trends (Vol. 1, pp. 9–52).

[9] Guillemin, P., & Friess, P. (2009). Internet of things strategic research roadmap. The Cluster of European Research Projects.

[10] Ray, P. P. (2018). A survey on Internet of Things architectures. Journal of King Saud University-Computer and Information Sciences, 30(3), 291–319.

[11] Zikria, Y. B., Yu, H., Afzal, M. K., Rehmani, M. H., & Hahm, O. (2018). Internet of Things (IoT): Operating system, applications and protocols design, and validation techniques. Future Generation Computer Systems, 88, 699–706.

[12] Alvi, S. A., Afzal, B., Shah, G. A., Atzori, L., & Mahmood, W. (2015). Internet of multimedia things: Vision and challenges. Ad Hoc Networks, 33, 87–111.

[13] Ghrib, T., Larouci, A., & Ben Slimane, A. (s.d.). Simulation et comparaison de protocoles de communication pour l'internet des objets (Thèse de doctorat, Université KASDI Merbah-Ouargla).

[14] Benchoula, R., & Hanachi, N. E. (s.d.). Etude comparative de protocoles de communication dans l'IOT. Thèse de doctorat, Université de Kasdi Merbah Ouargla.

[15] Nisrine, M. E., Imane, M. E. K., & Kabira, M. E. S. (s.d.). Conception et réalisation d'un système domotique en utilisant des technologies IoT.

[16] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, 17(4), 2347–2376.

[17] Leila, D. (s.d.). Gestion dynamique du spectre pour l'Internet des objets (IoT).

[18] Djaad, M., & Dahmani, C. (2022). Mémoire de fin d'étude. Université de Tiaret.

[19] Ghrib, T., Larouci, A. Simulation et comparaison de protocoles de communication pour l'internet des objets (Thèse de doctorat, Université KASDI Merbah-Ouargla).

[20] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 29(7), 1645–1660.

[21] M2M France. (2025). Les villes intelligentes et l'IoT. Consulté sur http://www.m2m.fr.

[22] Yick, J., Mukherjee, B., & Ghosal, D. (2008). Wireless sensor network survey. Computer Networks, 52(12), 2292–2330.

[23] International Journal of Communication Systems. (s.d.). Model of agriculture IoT network with sensors.

[24] Challal, Y. (2012). Sécurité de l'Internet des Objets : vers une approche cognitive et systémique. Thèse de doctorat, Université de Technologie de Compiègne.

[25] Biz4Group. (2023). Internet of Things: A Revolution for Public Transportation.

[26] Sultana, S. (2019). Internet of Things (IoT) for Video Surveillance Applications. University of Saskatchewan.

[27] Hafdi, K. (2020). Proposition et validation formelle d'une architecture ReDy fiable et dynamique destinée aux systèmes IoT-Application aux Smart Grids (Thèse de doctorat).

[28] Benkhira, A. (2023). Conception d'une antenne UHF ISM pour l'IOT (Thèse de doctorat).

[29] Atoumi, M. Y., & Bensadi, S. (2018). Approche évolutionnaire pour la composition de services sensible à la QoS dans l'Internet des Objets à large échelle. Mémoire de master, Université de Bejaia.

[30] Halli, T. (2020). Transformation d'équipements classiques télécommandés en objets connectés et intelligents (Thèse de doctorat, Université Mouloud Mammeri).

[31] Messaoud, H., & Sidi Bachir, S. C. (2021). Développement et simulation d'une maison intelligente dédiée pour des personnes handicapées basée sur l'IoT.

[32] Zhang, Y., & Li, X. (2024). Understanding the Trend of Internet of Things Data Prediction. ResearchGate.

[33] WebbyLab. (2025). IoT Data Management: Challenges and Best Practices. WebbyLab Blog.

[34] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[35] Delassus, R. (2018). Apprentissage automatique pour la détection d'anomalies dans les données ouvertes : application à la cartographie (Doctoral dissertation, Université de Bordeaux).

[36] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[37] Chloé-Agathe, A. (2018). Introduction au machine learning. Dunod, Paris.

[38] Purushotham, S., Meng, C., Che, Z., & Liu, Y. (2018). Benchmarking deep learning models on large healthcare datasets. Journal of biomedical informatics, 83, 112-134.

[39] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988)."Using the ADAP learning algorithm to forecast the onset of diabetes mellitus."Proceedings of the Annual Symposium on Computer Application in Medical Care.pp. 261–265.

[40] Jaotombo, F. (2022). Apports des méthodes de Machine Learning et de Deep Learning dans la prédiction des durées de séjours hospitalières et des ré-hospitalisations (Doctoral dissertation, Aix Marseille Université (AMU)).

[41] Sethy, A., Rout, A. K., Uriti, A., & Yalla, S. P. (2023). A comprehensive machine learning framework for automated book genre classifier. Revue d'Intelligence Artificielle, 37(3), 745.

[42] S. Dey, A. Saha, D. Roy, et al. (2020)."Heart Disease Prediction Model using KNN and Gradient Boosting Classifier."International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), Vol. 6, Issue 1, 2020.

[43] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989).International application of a new probability algorithm for the diagnosis of coronary artery disease.American Journal of Cardiology, 64(5), 304–310.

[44] Bénard, C. (2021, April). Random forests: A sensitivity analysis perspective. In Proceedings of the MascotNum Annual Conference, Aussois, France (pp. 28-30).

[45] Bellahmer, H. (2020). Implémentation et évaluation d'un modèle d'apprentissage automatique pour l'estimation de la valeur marchande de propriétés immobilières (Doctoral dissertation, Université Mouloud Mammeri).

[46] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[47] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

[48] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988)."Using the ADAP learning algorithm to forecast the onset of diabetes mellitus."Proceedings of the Annual Symposium on Computer Application in Medical Care.pp. 261–265.

[49] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

[50] Géron, A. (2019). Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts. Aurélien Géron-Google Kitaplar, yy https://books. google. com. tr/books.

[51] Polat, K., & Güneş, S. (2007).Breast cancer diagnosis using least square support vector machine.Digital Signal Processing, 17(4), 694–701.

[52] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995).Breast cancer diagnosis and prognosis via linear programming.Operations Research, 43(4), 570–577.

[53] Jaotombo, F. (2022). Apports des méthodes de Machine Learning et de Deep Learning dans la prédiction des durées de séjours hospitalières et des ré-hospitalisations (Doctoral dissertation, Aix Marseille Université (AMU)).

[54] Shankar, K., & Perumal, E. (2020).Diabetes mellitus prediction using XGBoost classifier.In: Smart Intelligent Computing and Applications. Springer, Singapore.DOI : 10.1007/978-981-15-3284-9_78

[55] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[56] Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12-18.

[57] Kayaer, K., & Yıldırım, T. (2003).Medical diagnosis on Pima Indian diabetes using general regression neural networks.Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP), Istanbul, Turkey.

[58] Barnard, G. A., & Bayes, T. (1958). Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. Biometrika, 45(3/4), 293-315.

[59] Hamidache, N. (2020). Déanonymisation de clients dans le réseau Bitcoin à l'aide de l'apprentissage automatique (Doctoral dissertation, Université Mouloud Mammeri).

[60] Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010).Hybrid prediction model for Type-2 diabetic patients.Expert Systems with Applications, 37(12), 8102–8108.

[61] Elman, J. L. (1990). Finding structure in time. Cognitive science, 14(2), 179-211.

[62] M. A. Alsoufi, S. Razak, M. M. Siraj, B. A. Al-rimy, A. Ali, M. Nasser, and S. Abdo, "A review of anomaly intrusion detection systems in iot using deep learning techniques," Advances in Data Science and Adaptive Analysis, vol. 13, no. 03n04, p. 2143001, 2021.

[63] [63] MEZAACHE, A. (2024). Classification Des Données Basée Sur Les Réseaux De Neurones Récurrents (RNN).

[64] [64] Raza, S., Wallgren, L., & Voigt, T. (2019).Predicting IoT Data Using Recurrent Neural Networks (RNNs).In: Proceedings of the 2019 IEEE International Conference on Communications (ICC)

[65] Ouyang, Z., Jabloun, M., & Ravier, P. (2023, August). Rankformer: un Nouveau Transformer avec un Mécanisme d'Attention Ordinal pour la Prédiction des Séries Temporelles. In GRETSI.

[66] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 11106-11115).

[67] Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., & Eickhoff, C. (2021).A Transformer-based Framework for Multivariate Time Series Representation Learning.Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD 2021).

[68] Charpentier, A. (2012). Modèles de prévision-Séries temporelles. Paris, FrancParis, France.

[69] Fan, C., et al. (2019). "Data-Driven Load Forecasting for Building Energy Consumption Using SARIMA Model."Energies, 12(1), 150.

[70] Durbin, J., & Koopman, S. J. (2012). Time series analysis by state space methods. Oxford University Press (UK).

[71] Jiang, Y., & Fei, Y. (2016). "Smart Vehicle Tracking System Based on Kalman Filter and Internet of Things", International Journal of Distributed Sensor Networks,

[72] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[73] BOUSMAT, Y. (2022). K-means & K-mers pour le regroupement et la comparaison de grands ensembles de séquences biologiques.

[74] Khan, I., Capozzoli, A., Corgnati, S. P., & Cerquitelli, T. (2013). Fault detection analysis of building energy consumption using data mining techniques. Energy Procedia, 42, 557-566.

[75] Benmahdi, M. B., & Lehsaini, M. (2020). Performance evaluation of main approaches for determining optimal number of clusters in wireless sensor networks. International Journal of Ad Hoc and Ubiquitous Computing, 33(3), 184-195.

[76] G. C. Guillen et al., "Smart Water Consumption Measurement System for IoT Home Automation,"Sensors, 2020.

[77] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[78] Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014, February). DBSCAN: Past, present and future. In The fifth international conference on the

applications of digital information and web technologies (ICADIWT 2014) (pp. 232-238). IEEE.

[79] Souri, A., et al. (2021)."IoT-Based Wearable Sensors and DBSCAN for Real-Time Health Monitoring and Anomaly Detection", IEEE Access,

[80] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.

[81] L. Zhang, et al. (2020)."Anomaly detection based on Isolation Forest algorithm for streaming data in IoT-enabled manufacturing", Journal of Intelligent Manufacturing.

[82] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

[83] Chen, K., et al. (2019)."Hybrid ARIMA and LSTM for air quality forecasting", Environmental Science and Pollution Research, 26(22), 22480–22493.

[84] Liu, P., & Zeng, Z. (2021). Theory and application of logistic-SVM two-stage hybrid model. In Journal of Physics: Conference Series (Vol. 1746, No. 1, p. 012017). IOP Publishing.

[85] Naseem, A., Habib, R., Naz, T., Atif, M., Arif, M., & Allaoua Chelloug, S. (2022). Novel Internet of Things based approach toward diabetes prediction using deep learning models. Frontiers in Public Health, 10, 914106

[86] Moursi, A. S., El-Fishawy, N., Djahel, S., & Shouman, M. A. (2021). An IoT enabled system for enhanced air quality monitoring and prediction on the edge. Complex & intelligent systems, 7(6), 2923-2947.

[87] Ahamed, J., Manan Koli, A., Ahmad, K., Alam Jamal, M., & Gupta, B. B. (2022). CDPS-IoT: cardiovascular disease prediction system based on IoT using machine learning.

[88] Wistuba, M., Duong-Trung, N., Schilling, N., & Schmidt-Thieme, L. (2016). Bank card usage prediction exploiting geolocation information. arXiv preprint arXiv:1610.03996.

[89] Ahmed, M., Kashem, M. A., Rahman, M., & Khatun, S. (2020). Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT). In InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019 (pp. 357-365). Springer Singapore.

[90] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., ... & Willing, C. (2016). Jupyter Notebooks–a publishing format for reproducible computational workflows. In Positioning and power in academic publishing: Players, agents and agendas (pp. 87-90). IOS press.

[91] Van Rossum, G., & Drake, F. L. (2009). PYTHON 2.6 reference manual.

[92] Raybaut, P. (2009). Spyder-documentation. Available online at: pythonhosted. org, 769.

[93] Ahmed, M., Kashem, M. A., Rahman, M., & Khatun, S. (2020). Review and analysis of risk factor of maternal health in remote area using the Internet of Things (IoT). In InECCE2019: Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, Kuantan, Pahang, Malaysia, 29th July 2019 (pp. 357-365). Springer Singapore.

[94] https://www.kaggle.com/

[95] https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data

Résumé

Cette recherche s'inscrit dans l'intégration des technologies IoT dans le domaine médical, notamment pour la surveillance prénatale. Il propose une méthode de prévision du taux de probabilité d'accouchement par césarienne, à partir de données cliniques collectées grâce à des capteurs connectés. Les paramètres considérés incluent l'âge, la pression artérielle et le taux de glycémie. Cinq modèles de machine learning ont été appliqués (Régression Logistique, SVM, Random Forest, Gradient Boosting, XGBoost) pour estimer le risque selon trois niveaux : bas, moyen et haut. Les résultats obtenus soulignent le potentiel de l'intelligence artificielle en tant qu'outil d'assistance à la prise de décision médicale, notamment pour prévoir les complications associées à l'accouchement.

Mots clés : IoT, Régression Logistique, SVM, Random Forest, Gradient Boosting, XGBoost, apprentissage supervisè.

Abstract

This research falls within explores the integration of Internet of Things (IoT) technologies in the medical field, with a particular focus on prenatal monitoring. It presents a method for predicting the probability of delivering by cesarean section based on clinical data collected through connected sensors. The considered parameters include age, blood pressure, and blood glucose level. Five machine learning models were applied (Logistic Regression, SVM, Random Forest, Gradient Boosting, and XGBoost) to estimate the risk at three levels: low, medium, and high. The results highlight the potential of artificial intelligence as a decision-support tool in medical practice, particularly in anticipating childbirth-related complications.

Keywords: IoT, Logistic Regression, SVM, Random Forest, Gradient Boosting, XGBoost, supervised learning.

الملخص

يندرج هذا البحث ضمن سياق دمج تقنيات إنترنت الأشياء في المجال الطبي، لاسيما في مجال مراقبة فترة ما قبل الولادة. يقترح منهجية للتنبؤ بمعدل احتمالية الولادة القيصرية، اعتمادًا على بيانات سريرية بواسطة أجهزة استشعار متصلة. تشمل المعايير المعتمدة: العمر، ضغط الدم، ومستويات السكر في الدم. تم تطبيق خمسة نماذج من التعلم الآلي) الانحدار اللوجستي، وآلة المتجهات الداعمة SVM ، والغابة العشوائية، والتعزيز التدريجي، و (XGBoostلتقدير مستوى الخطر وفقًا لثلاث درجات: منخفض، متوسط، وعالٍ. تُبرز النتائج المحصّلة قدرة الذكاء الاصطناعي على دعم اتخاذ القرار الطبي، لاسيما في التنبؤ بالمضاعفات المرتبطة بالولادة.

الكلمات المفتاحية: IoT، الانحدار اللوجستي، آلة الدعم المتجه، الغابة العشوائية، التعزيز المتدرج، XGBoost، التعلم الخاضع للإشراف.