


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en Informatique

Spécialité : *Ingénierie des systèmes d'information et logiciel*

Thème

La prédiction du diabète en utilisant les algorithmes de machine learning

---

Encadré par

— Brahimi Farida

Réalisé par

— Sidahmed Amel

— Rabhi Karima

2019/2020

Remerciements

Nous remercions le DIEU de nous avoir donné la patience, la santé et le courage pour réaliser ce travail.

A travers ce modeste travail , nous tenons à remercier vivement notre encadreuse Farida Brahimi pour ses conseils et ses encouragements qui nous ont permis de réaliser ce travail dans les meilleures conditions.

Nous remercions sincèrement les membres du jury d'avoir accepter d'examiner et d'évaluer notre travail .

Nous exprimons également notre gratitude à tous les professeurs et les enseignants qui ont collaboré à notre formation depuis notre premier cycle d'étude jusqu'à la fin de notre cycle universitaire.

Sans oublier bien sûr de remercier profondément tous ceux qui ont contribué de près ou de loin à la réalisation du présent travail.

MERCI À TOUS

Dédicaces

Je dédie ce travail à la personne qui rêvait de cette journée plus que moi à Aicha maman et mon papa Youcef .

Je dédie ce travail à toute ma famille, mes amis et bien sur ma collègue du travail .

Je dédie ce travail à tous les jours difficiles que j'ai vécus au cours de ces cinq années à l'université et aujourd'hui je peux dire que je dédie ce travail à moi même .

Sidahmed Amel.

Dédicaces

Je dédie ce travail

A mes chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études.

A mes chères sœurs, pour leurs encouragements permanents, et leur soutien moral.

A mes chers frères, pour leur appui et leur encouragement.

A ma collègue du travail pour leur encouragement, et leur aide.

A toute mes amies, pour l'amitié et les moments agréables que nous avons passés ensemble.

A toute ma famille pour leur soutien tout au long de mon parcours universitaire.

Rabhi Karima.

Résumé

Au cour de ce mémoire, nous avons conçu et développé une application web pour la prédiction précoce du diabète de type 2, afin de réduire le risque des complications de cette maladie sur la santé du patient.Pour atteindre cet objectif, nous avons utilisé des algorithmes d'apprentissage automatique supervisé (K nearest neighbors, Decision Trees, Random Forest, Support Vector Machine, Naïves Bayes) et le data set extrait du l'hôpital Frankfurt (Allemagne). Les performances des classifieurs ont été comparées en fonction du taux de précision et la sensibilité de modèle. Les plus hauts taux de classification obtenus par l'application de Random Forest et l'arbre de décision sont respectivement 91% et 87%, en appliquant les deux méthode d'évaluation train/test et validation croisé (10-folds).

Mots clés : IA , ML , Prédiction du diabète ,K nearest neighbors, Decision Trees, Random Forest, Support Vector Machine, Naïves Bayes .

Abstract

In this modest work, we designed and developed a web application for the early prediction of type 2 diabetes, in order to reduce the risk of complications of this disease on the patient's health.To achieve this goal, we used algorithms supervised machine learning (K nearest neighbors, Decision Trees, Random Forest, Support Vector Machine, Naïves Bayes) and the data set extracted from the hospital in Frankfurt (Germany). The performance of classifiers was compared based on accuracy rate and model sensitivity. The highest classification rates obtained by the application of Random Forest and the decision tree are respectively 91% and 87%, by applying the two methods of evaluation train /test and cross validation (10-folds).

Key words : IA , ML , Diabetes prediction,K nearest neighbors, Decision Trees, Random Forest, Support Vector Machine, Naïves Bayes.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

IA	Intelligence Artificielle
ML	Machine Learning
OMS	Organisation mondiale de la santé.
Ceed	Centre européen d'étude du diabète.
ASG	Autosurveillance glycémique
HAS	Haute autorité de santé
DT1	Diabète de type 1
DT2	Diabète de type 2

Introduction générale

L'intelligence artificielle (IA) est devenue le nouveau terme que l'on entend tous les jours ces dernières années, l'IA en général définit la capacité d'une machine capable d'agir par elle-même et qui n'est pas explicitement programmée pour reproduire des actions ou des fonctions qui sont généralement celles des êtres humains . Aujourd'hui, on la retrouve dans nos machines informatiques , les réseaux sociaux, les transports et dans le secteur médical. L'application de l'IA en médecine permettant à la machine d'analyser les données par elle-même et de fournir des estimations, dans le but de prédire de nombreuses maladies afin que les médecins puissent intervenir le plus rapidement possible pour réduire le risque de complications des maladies sur la santé du patient et lutter contre la mort prématurée.

L'apprentissage automatique est une discipline de l'intelligence artificielle qui cherche à trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience.

A travers ce mémoire de Master, nous intéresserons à l'utilisation des algorithmes d'apprentissage automatique pour la prédiction du diabète de type 2 qui est un dysfonctionnement du système de régulation de la glycémie, afin de réduire les risques de complications de cette maladie chronique sur la santé du patient
Notre problématique nous permettent de définir le diagnostic médical comme un processus de classification et l'utilisation de l'informatique devient de plus en plus fréquente pour mettre en œuvre cette classification bien que la décision de médecin soit le facteur le plus important dans le diagnostic. Les systèmes de classification sont d'une grande aide car ils réduisent les erreurs dues à la fatigue et au temps nécessaire au diagnostic.

La méthode utilisée dans ce travail est l'application les différents algorithmes de classification d'apprentissage supervisé (K nearest neighbors, Decision Trees, Random Forest, Support Vector Machine, Naïves Bayes) aux données extrait de l'hopital frankfurt et déduire le meilleur algorithme qui donnera comme résultat une classification des patients en termes de taux du précision et de la sensibilité du modèle . Ce travail est organisé en trois principaux chapitres comme suit :

1. Le 1er chapitre présente un aperçu général sur la maladie du diabète, leur différent types, les symptômes ainsi que le diagnostic et le traitement de la maladie et a la fin quelques préventions pour évité le diabète .
2. Le 2ème chapitre donne un aperçu sur l'apprentissage automatique , les algorithmes d'apprentissage supervisé qui peuvent nous aider a détecter l'apparition précoce du diabète et une étude critique de quelques solutions récemment proposées sur la prédiction du diabète type 2 .
3. Le dernier chapitre présente d'abord une étude technique dans laquelle nous définissons l'environnement logiciel utilisé pour construire notre application, puis une définitions détaille de la base du données utilisée . Ensuite, les résultats sont présentés, comparés et interprétés. Finalement nous finissons par une représentation des interfaces d'application d'apprentissage dans la prédiction du diabète type 2.

A la fin, ce travail est clôturé par une conclusion générale résumant les idées fondamentales que nous avons apportées et les perspectives.

Chapitre 1

Chapitre 1 :Généralités sur le diabète

1.1 Introduction

Le diabète est une maladie qui empêche le corps d'utiliser correctement l'énergie fournie par les aliments ingérés. Par ailleurs, la maladie survient lorsque le pancréas ne sécrète plus d'insuline ou lorsque le corps devient résistant à la quantité d'insuline produite. Il existe principalement deux types de diabète : Le type 1 appelé diabète insulinodépendant ou diabète juvénile se caractérise par une production insuffisante d'insuline dans l'organisme pour lequel la survie du patient nécessite des injections d'insuline. Ces symptômes sont notamment les suivants : émission d'urine, soif excessives, faim constante, perte de poids, altération de la vision et la fatigue. Le type 2 appelé diabète non insulinodépendant ou diabète de l'adulte, il résulte de l'utilisation inefficace de l'insuline par l'organisme. Les symptômes peuvent être similaires à ceux du diabète de type 1, mais ils sont souvent moins marqués ou absents. En outre, il existe un autre type de diabète appelé diabète gestationnel qui se développe pendant la grossesse il est associé à un risque à long terme de diabète de type 2. Le sur-poids, le manque d'exercice, les antécédents familiaux et le stress est augmenté le risque possible de diabète et le mauvais contrôle de dosage de sucre (glucose) dans le sang peut entraîner des complications tré grave (cécité, cataracte, thrombose, néphropathie. . .) .

1.2 Définition de diabète

Le diabète est une maladie chronique connue aussi sous le nom de « une maladie silencieuse ». L'organisation mondiale de la santé (OMS) définit le diabète comme une maladie chronique grave qui se déclare lorsque le pancréas ne produit pas suffisamment d'insuline (hormone qui régule la concentration de sucre dans le sang, ou glycémie) ou lorsque l'organisme n'est pas capable d'utiliser correctement l'insuline qu'il produit.[1]

Plus clairement

Lorsque nous mangeons, les aliments sont dégradés en glucose (sucre). Ce glucose fournit de l'énergie au corps afin qu'il puisse fonctionner correctement en puisant dans ses ressources. Pendant la digestion, le sang transporte le glucose dans tout le corps et vient alimenter les cellules. Cependant, pour que le sucre présent dans le sang puisse ensuite être transmis aux cellules, le corps a besoin d'insuline, une hormone sécrétée par le pancréas. L'insuline agit donc comme une clé permettant au glucose de passer du sang aux cellules de notre corps.[2]

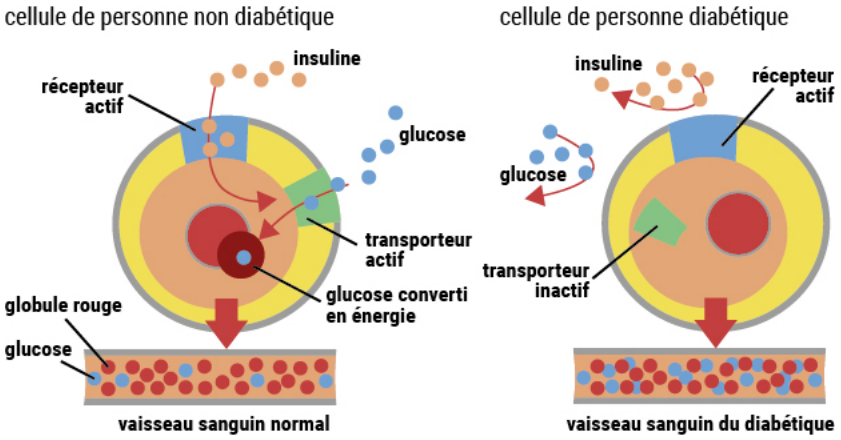


FIGURE 1.1 – Insulinorésistance et insulinopénie [3]

Si le glucose reste dans le sang, la glycémie augmente. À long terme, cela peut entraîner le dysfonctionnement et la détérioration de nombreux organes comme les yeux et les reins.

1.3 Diabètes et complications

Quel qu'en soit le type de diabète, ce dernier peut entraîner des complications à court terme (hypoglycémie, malaise...) ,et des complications à long terme (L'hyperglycémie) en cas de mauvais contrôle de la glycémie , conséquence courante d'un diabète non maîtrisé , peut, au fil du temps, ces complications chroniques provoquer de graves complications touchant de nombreuses parties de l'organisme et accroître le risque général de décès prématuré chez les patients causées par une atteinte des vaisseaux sanguins . Au nombre des complications possibles figurent l'infarctus du myocarde, l'accident vasculaire cérébral, l'insuffisance rénale, l'amputation des jambes, la perte de vision et des lésions nerveuses .

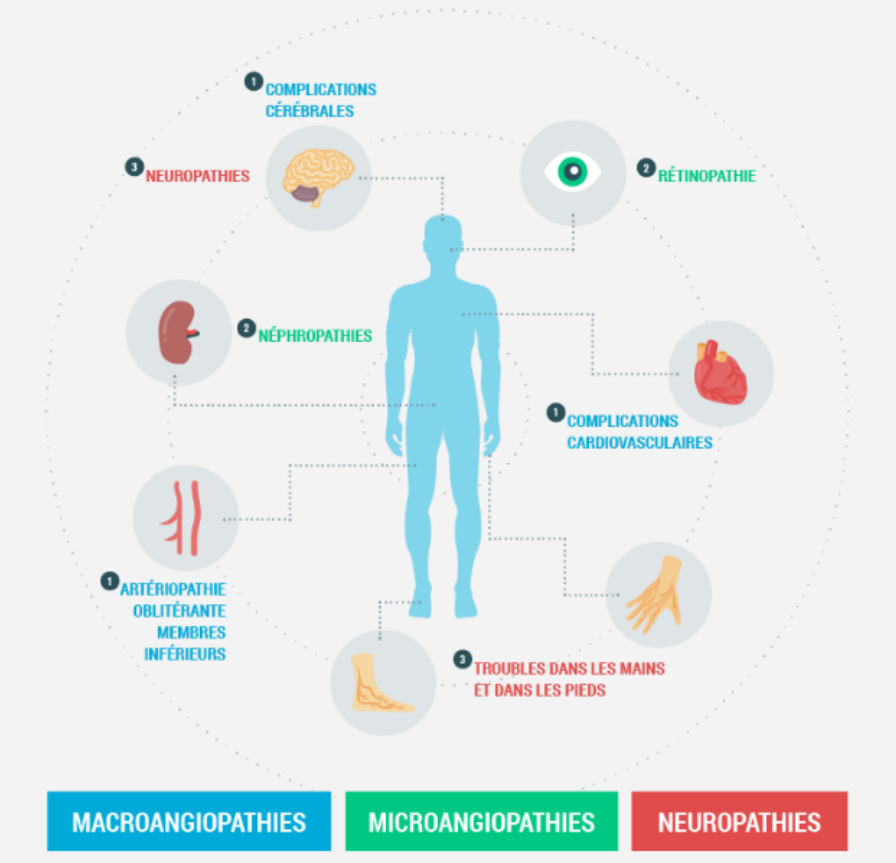


FIGURE 1.2 – Diabètes et complications à long terme [4]

— **Macroangiopathies** : macro-angiopathie est L'atteinte des grosses artères due à la sclérose des vaisseaux (athéro-sclérose) secondaire à un dépôt à l'intérieur des vaisseaux, que l'on appelle "plaque d'athérome" .Les artères les plus touchées sont les artères du cœur, des jambes et du cou.[5]

— **Microangiopathies** : Atteinte de la paroi des petits vaisseaux (artérioles, capillaires et veinules) qui est épaissie. Les lésions de micro angiopathie sont une des complications du diabète sucré, localisées au niveau de la rétine et des reins, elles sont à l'origine, respectivement, de cécité et d'insuffisance rénale. [6]

— **Neuropathie** : atteinte des nerfs, le plus fréquemment au niveau des membres inférieurs c-a-d perte de la sensibilité (chaud, froid, douleur)[4]

1.4 Diagnostic du diabète

Cette maladie silencieuse et indolore est détectée le plus souvent lorsque les complications à long terme s'expriment. Cette découverte peut notamment être brutale dans le cas de diabète de type 1 (pas de sécrétion d'insuline), allant jusqu'au coma diabétique.

1.4.1 Qui est concerné par le dépistage du diabète ?

Toute personne ayant des membres de sa famille atteints de diabète doit se faire dépister régulièrement car un risque héréditaire existe , les personnes en sur-poids ou souffrant de troubles de la glycémie doivent également se plier au dépistage , Il en va de même pour les femmes ayant développé du diabète pendant leur grossesse et le dépistage est également recommandé aux personnes de plus de 65 ans.[7]

1.4.2 Comment savoir si l'on est diabétique ?

La diagnostique du diabète se fait par un test de prise du sang mesurant la glycémie ou le taux de sucre sanguin, qui varie selon les apports alimentaires .il existe deux façons de test :

— **teste en laboratoire d'analyses médicales** : pour mesurer sa glycémie à jeun et tous les 3 mois, son hémoglobine glyquée (HbA1c).

— **auto-teste** :un lecteur de glycémie pour contrôler plusieurs fois par jour sur une goutte de sang à des moments précis. C'est ce qu'on appelle l'autosurveillance glycémique (ASG).[14]

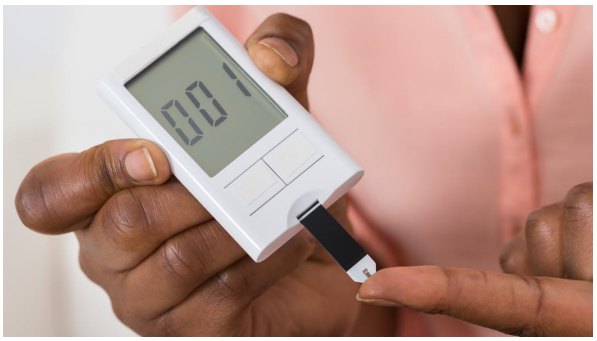


FIGURE 1.3 – Glucomètre [8]

pour mieux savoir si vous êtes diabétique ou non, le mieux est de tester par deux examens sanguins pratiqué en laboratoire d'analyses médicales par une simple prise de sang pour mesurer le dosage de la glycémie, c'est à dire du taux de sucre dans le sang .La prise de sang se pratique le matin, à jeun c-a-d il ne faut rien avoir mangé ou bu, sauf de l'eau, depuis au moins 8 heures pour assurer de la véracité des résultats.[9]

1.4.3 Décoder et comprendre les résultats de la glycémie

le dépistage du diabète peut être fait à tout moment de la journée mais Il est primordial de tester la glycémie après un minimum de 8 heures de jeûne . On considère qu'il y a présence de diabète si la glycémie est supérieure ou égale à 1,26 g/l à jeun et supérieure ou égale à 2 g/l après le repas .

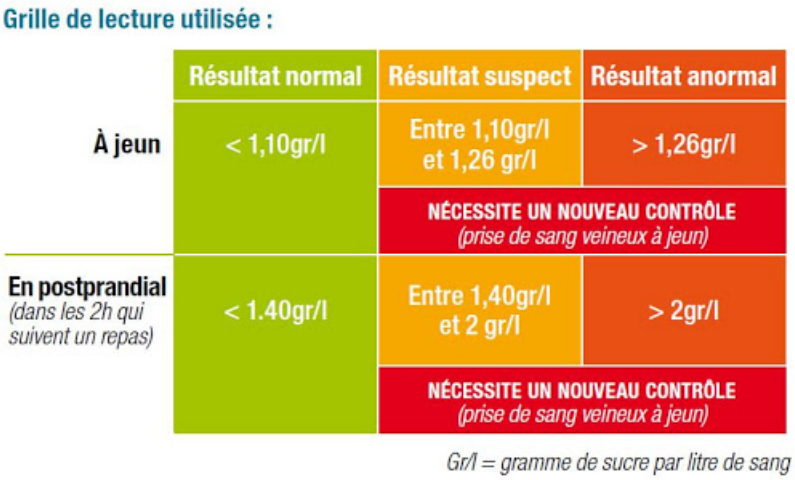


FIGURE 1.4 – Grill de lecteur de la glycémie [10]

1.4.4 à quelle fréquence dois-je contrôler ma glycémie ?

Selon les recommandation de la haute autorité de santé (HAS) [14] :

— **le diabète de type 1** :au moins quatre tests par jour. Les objectifs glycémiques sont fixés entre 70 et 120 mg/dl avant le repas et $<$ 160 mg/dl en post-prandial.

— **le diabète de type 2** : dans tous les cas, les objectifs glycémiques sont fixés entre 70 et 120 mg/dl avant les repas et 180mg/dl en post-prandial. Selon le type de traitement, la fréquence est variable.

— **le diabète gestationnel** :à jeun $<$ 0,95 g/l et $<$ 1,20 g/l en postprandial.

1.5 Classification du diabète

La classification des diabètes proposée par l'OMS se base principalement sur son étiologie et caractéristique physiopathologique en quatre types :

— Diabète de type 1

— Diabète de type 2

— Diabète gestationnel

1.5.1 Diabète de type 1

Le diabète de type 1 ou diabète insulinodépendant survient lorsque le pancréas ne produit plus assez ou, plus du tout, d'insuline. Cette anomalie, se caractérise par une destruction auto-immune de plus de 90% des cellules béta du pancréas productrices de l'insuline, provoquant une carence insulinique totale ou partielle. [10]

A- Les symptômes cliniques du diabète de type 1

Les symptômes cliniques du diabète de type 1 il s'agit des 3 « P » :

— Polydipsie : soif accrue.

— Polyphagie : faim accrue.

— Polyurie : besoin fréquent d'uriner.

Ces symptômes sont souvent associés à une perte de poids importante, un manque d'énergie et des sensations de nausées.[9]

B- Le traitement du diabète de type 1

Pour compenser, celle-ci doit être administrée « artificiellement » au quotidien par une injection sous cutanée d'insuline via une seringue, un stylo ou une pompe. Il s'agit d'un traitement d'insulinothérapie. Le diabète de type 1 touche plus souvent l'enfant, l'adolescent voire le jeune adulte. [11]

FIGURE 1.5 – Seringue d'insuline [12]

1.5.2 Diabète de type 2

Précédemment appelé diabète non insulinodépendant ou diabète de la maturité ou de l'adulte , est une maladie chronique, silencieuse et indolore, qui se caractérise par un taux de sucre (glucose) trop élevé dans le sang (hyperglycémie) . Cette anomalie est causée par un défaut de la sécrétion ou de l'utilisation de l'insuline qu'est la conséquence d'une perte de fonctionnalité des îlots pancréatiques. Cette perte de fonctionnalité est la conséquence de l'interaction de facteurs génétiques, volontiers héréditaires et de facteurs environnementaux liés au mode de vie . Contrairement au diabète de type 1, le diabète de type 2 est le plus souvent asymptomatique. De ce fait, la maladie peut être diagnostiquée plusieurs années après son apparition, une fois les complications déjà présentes [10]

FIGURE 1.6 – Résistance à l'insuline [13]

A- Les symptômes cliniques du diabète de type 2

sont les mêmes que celles du type 1, auxquelles s'ajoutent les risques cardio-vasculaires, mais aussi une incidence sur le développement de certains cancers, troubles du comportement ou maladies mentales.[4]

B- Le traitement du diabète de type 2

Le traitement repose prioritairement sur des alimentation équilibrée et pratique d'une activité physique régulière .si ces deux éléments sont insuffisants, il faudra ajouter un traitement par anti-diabétique oral. Le traitement à l'insuline peut s'avérer nécessaire, si les glycémies restent néanmoins élevées.[4]

FIGURE 1.7 – Le traitement hygiéno-diététiques [13]

1.5.3 Diabète gestationnel

Le diabète gestationnel est un diabète qui survient chez une femme enceinte, du fait des modifications métaboliques provoquées par la grossesse (mais pas toutes les femmes enceintes). Il est appelé aussi « diabète de grossesse ».[4]
Contrairement aux diabètes de type DT1 et DT2 qui sont des pathologies évolutives et à surveiller à vie, le diabète gestationnel disparaît le plus souvent après la naissance du bébé. Lorsqu'une femme souffre de diabète gestationnel au cours de sa grossesse, elle est plus susceptible d'en souffrir à nouveau lors de sa prochaine grossesse et elle est exposée à un risque plus élevé de développer un diabète de type 2 par la suite. Plus une femme est enceinte à un âge avancé, plus le risque de développer un diabète gestationnel au cours de sa grossesse est élevé .

A- Le traitement du diabète gestationnel

selon le Centre européen d'étude du Diabète (Ceed) le traitement par le recours à l'insuline est nécessaire dans 50% des cas et dans quelques cas plus rares un traitement par anti-diabétique oral peut être mis en place. Dans tous les cas, des mesures hygiéno-diététiques doivent rapidement être mises en place, avec la particularité qu'elles doivent prendre en compte à la fois le diabète de la mère et les besoins nutritionnels du foetus.[4]

B- Les conséquences du diabète gestationnel

- **maternelles** : toxémie gravidique .
- **La macrosomie fœtale** : risques pour l'enfant d'être trop gros et d'entraîner des complications à l'accouchement .
- **néonatales** : risques d'hypoglycémie et d'hypocalcémie.

1.6 Prévention du diabète de type 2

Les complications du diabète ne sont pas une fatalité vous pouvez changer l'évolution de votre maladie si vous changez votre style de vie parceque la prévention du diabète de type 2 est étroitement liée à des règles d'hygiène de vie [15] :

- Surveiller votre poids .
- Adopter un régime alimentaire équilibré .
- Pratiquer une activité physique régulière .
- Éviter le stress .
- Arrêter le tabac .
- Faire le point sur ses facteurs de risque c'est le cas si :
  - l'âge $>$ 40 ans .
  - un membre de la famille est atteint de diabète.
  - Des antécédents (diabète gestationnel, accouchement d'un bébé de plus de 4 kilos)...
- Connaître les signes d'alerte et ne pas hésiter à consulter :
  - une soif intense
  - une envie fréquente d'uriner
  - une fatigue et un manque d'énergie
  - une difficulté de cicatrisation
  - une vision floue ...
- Faire attention à certaines pathologies qui conduisent la survenue du diabète de type 2 :
  - l'hypertension artérielle .
  - taux élevé de cholestérol ou de triglycérides dans le sang .

  ...

Même s'il existe des méthodes de prévention qui permettent de réduire le risque d'avoir le diabète, parfois il est impossible de l'éviter comme pour le diabète de type 1. Dans ces cas là, la seule solution est de pouvoir le diagnostiquer très tôt et faire tout son possible pour combattre les complications.

1.7 Conclusion

Dans ce chapitre nous avons présenté la maladie du diabète, leur différent types, les symptômes ainsi que le diagnostic et le traitement de la maladie et a la fin nous avons cité quelques préventions pour évité le diabète . Dans le prochain chapitre, nous présenterons des approches différentes d'aide au diagnostic préventif en utilisant les algorithmes de machine learning dans la prédiction du diabète de type 2 .

Chapitre 2

Chapitre 2 : L'apprentissage automatique et le diabète

2.1 Introduction

Le machine learning ou l'apprentissage automatique permet à une machine d'évoluer par un processus d'apprentissage à effectuer des tâches complexes pour les quelles elle n'est pas explicitement programmée en apprenant avec des données . Selon le patron de l'Intelligence Artificielle (IA) chez Facebook Yann Le Cun « Il n'y a pas d'intelligence sans apprentissage » c'est dans cette logique que s'inscrit le machine learning, une technique d'Intelligence Artificielle qui est aujourd'hui en plein essor avec l'avènement du Big Data . La grande majorité des systèmes d'IA actuels médiatisés utilisent ce processus d'apprentissage qui permet à la machine d'évoluer sans que ses algorithmes ne soient modifiés. Dans ce chapitre, nous définirons l'apprentissage automatique, ses principaux types et les algorithmes utilisés, ainsi nous présenterons quelques travaux de recherche sur l'application d'algorithmes d'apprentissage automatique pour prédire le diabète de type 2, afin de réduire les risques de complications de cette maladie sur la santé d'un patient.

2.2 Apprentissage automatique

« L'apprentissage automatique » ou « Machine Learning » en anglais , permet à une machine d'évoluer par un processus systématique et d'effectuer des tâches pour lesquelles elle n'est pas explicitement programmée en apprenant avec des données.[16]
L'objectif est de rendre la machine capable de traiter une quantité astronomique et inimaginable d'informations , d'effectuer des tâches extrêmement complexes et d'obtenir des résultats en temps réel qu'ils est difficiles à obtenir avec des algorithmes classiques.

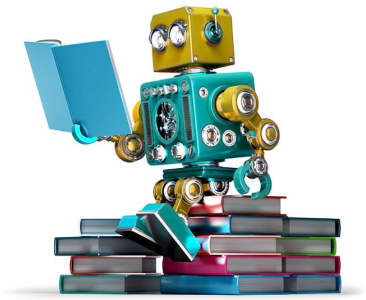


FIGURE 2.1 – Apprentissage automatique [17]

2.3 les types d'apprentissage automatique

L'apprentissage automatique procédé deux principaux types d'apprentissage :

→ Apprentissage supervisé .
→ Apprentissage non supervisé .

2.3.1 Apprentissage Supervisé

L'apprentissage Supervisé ou la méthode statistique d'apprentissage de classes consistant à apprendre une fonction de prédiction de classe des nouvelle éléments à partir d'exemples étiquetés , il s'appelle aussi un « modèle » .

le workflow d'un apprentissage supervisé

l'apprentissage supervisé est le faite de trouver un ensemble de données $D_{données} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,une fonction $f(X)$, pour tout $(x_n, y_n) \in D_{données}$ et $D_{données}$: ensemble fini de données . on ait $f(x_n) = y_n$. [18]

plus clairement :

— **Phase 1** :l'ensemble d'apprentissage ou base d'apprentissage .

$$D_{entrée} = \{(x_1, y_1), \ldots, (x_i, y_i)\}$$

$D_{entrée} \in D_{données}$.

i : indice de donnée

x : donnée et y : classe ou étiquète de donnée .

— **Phase 2** : La création de modèle ou la fonction de prédiction

l'algorithme d'apprentissage reçoit $D_{(x_i,y_i)}$ entrée et construit un modèle Ou bien une fonction de prédiction $f(x_i) = y_i$.

— **Phase 3** : phase de test

on test la qualité de modèle sur un ensemble de variables étiquetées qu'on désigne par :

$$D_{test} = \{(x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$$

$D_{test} \in D_{données}$.

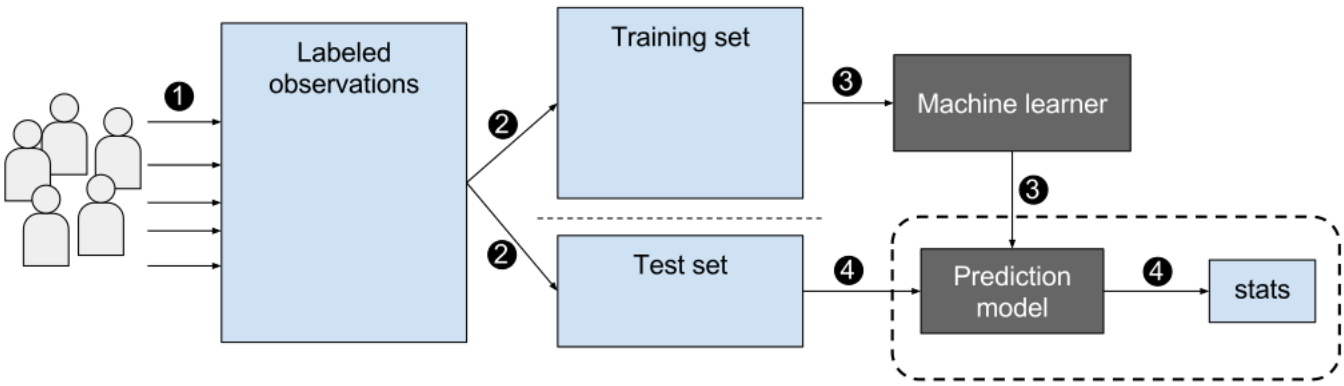


FIGURE 2.2 – workflow d'un apprentissage supervisé [18]

Il existe deux types de modèles d'apprentissages supervisés :

1. le modèle de classification.
2. le modèle de régression .

→ le modèle de classification

Un modèle de classification permet de **prédire une valeur qualitative**. Cela signifie que l'ensemble des valeurs de sortie Y qu'on essai d'estimer avec la fonction f est un ensemble fini : Y=0,1,. . . ,n.

Exemple : On veut créer un modèle $f : X \longrightarrow Y$ qui prédit si un patient P est diabétique ou non.

Dans Cet exemple, X représente l'ensemble des patients à analyser et Y=0,1 ; 1 si le patient est diabétique et 0 sinon. Si on veut analyser un patient P , on calcule f(P)

→ le modèle de régression

Un modèle de régression permet de **prédire une valeur quantitative**. Cela signifie que l'ensemble des valeurs de sortie Y qu'on essai d'estimer avec la fonction f est un ensemble de réels .

Exemple : Prédire l'age d'un humaine en fonction de la taille, du poids, etc.

Dans cet exemple, X représente l'ensemble des humains et Y représente tous les âges. Si nous voulons estimer l'age d'un humaine H de taille T, de poids P, etc., nous calculons f (H).

2.3.2 Apprentissage Non Supervisé

A la différence de l'apprentissage supervisé, le contexte non supervisé est celui où l'algorithme doit opérer à partir d'exemples non étiquetés. il doit extraire automatiquement les catégories à associer aux données qu'on lui soumet , les plus fréquents problèmes connu dans ce type est :

1. Le clustering qui consiste à regrouper un ensemble d'éléments hétérogènes sous forme de sous groupes homogène .
2. La réduction de dimension qui consiste à prendre des données dans un espace de grande dimension, et à les remplacer par des données dans un espace de plus petite dimension sans perdre la variance. [18]

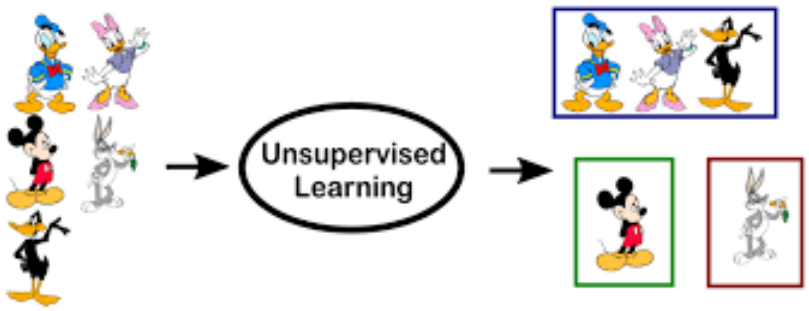


FIGURE 2.3 – Exemple d'apprentissage automatique non supervisé [26]

2.4 Les algorithmes de l'apprentissage automatique utilisés

2.4.1 K nearest neighbors (KNN)

« K nearest neighbors (KNN) » ou « K plus proche voisins » en français est l'un des méthodes d'apprentissage supervisé le plus simple, utilisé pour résoudre des problèmes de classification et du régression.son fonctionnement est de classer les nouveaux points de données en fonction de la similarité aux points de données voisins .

→ KNN est un algorithme qui ne fait aucun hypothèses sur la structure des données et de la distribution, ce qui signifie qu'il s'agit d'un algorithme non paramétrique.

— Il est également appelé algorithme de l'apprenant paresseux, car il n'apprend pas immédiatement de l'ensemble d'apprentissage, mais stocke l'ensemble de données et, au moment de la classification, il exécute une action sur l'ensemble de données.

→ KNN fonctionne par classification ou prédiction sur la base d'un nombre fixe (K) de points de données les plus proches de point d'entrée. Cela signifie que pour une valeur choisie de K, un point d'entrée serait classée ou devrait appartenir à la même classe que la classe la plus proche des nombre des points K voisins.[20]

Exemple

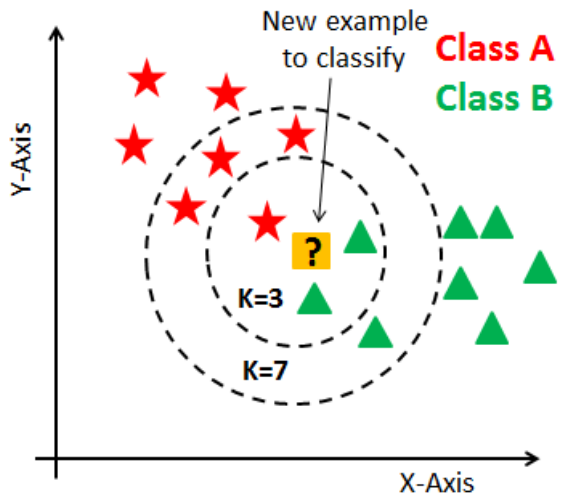


FIGURE 2.4 – Exemple simple sur KNN [21]

L'interprétation de l'exemple

Dans cet exemple nous avons une donnée non classée et tous les autres données sont classée (étoile et triangle) chacun avec leur classe (classe A et B).

- Si k=3 les données les plus proche du nouvelle donnée sont qui ont à l'intérieure de premier cercle, et la classe la plus prédominante c'est triangle (Classe B) car 2 triangles et seulement 1 étoile donc la donnée non classée sera classer un triangle (Classe B).
- Si k=7 les données les plus proches du nouvelle donnée sont qui ont à l'intérieure de deuxième cercle, et la classe la plus prédominante c'est l'étoile (Classe A) car on a 4 étoiles et 3 triangles donc le donnée non classée sera classer un étoile (Classe A).

La distance entre le point non classée et les plus proches voisins

La distance entre le point non classée et les plus proches voisins est mesuré en utilisant différents méthode comme : la distance euclidienne, la distance de Manhattan, la distance de Minkowski, celle de Jaccard, la distance de Hamming. . . etc, le fonction de distance est choisi en fonction de type de données qu'il manipule. Pour les données de même type la

distance euclidienne est le bon candidat, et pour les données qui ne sont pas de même type la distance de Manhattan est le bonne mesure pour l'utiliser.[22]

La représentations mathématiques de quelques distances

. Distance euclidienne

$$D_e(x, y) = \sqrt{\Sigma_{j=1}^{n}(x_j - y_j)^2}$$

. Distance Manhattan

$$D_m(x, y) = \Sigma_{i=1}^{k}|x_i - y_i|$$

Remarque

le choix de la bonne valeur de K est un processus appelé réglage des paramètres est important pour une meilleure précision, la sélection des valeurs plus petites pour k aura une plus grand influence sur le résultat. Et pour la sélection des valeurs plus élevé de k auront des limites de décision plus lisses, ce qui signifie une variance plus faible mais un biais.

Algorithme de construction de KNN

1. Sélectionnez le nombre K des voisins .
2. Pour chaque exemple de l'ensemble de données :
   2.1. Calculez la distance entre l'exemple de requête et l'exemple actuel à partir des données .
   2.2. Ajouter la distance et l'index de l'exemple à une collection ordonnée .
3. Trier cette collection de distances et d'indices du plus petit au plus grand (par ordre croissant) ordonnée par les distances .
4. Choisi les k premiers entrée de collections
5. Attribuer l'exemple de requête à la classe où laquelle le nombre de k voisins est maximal (classe le plus fréquent). [27]

Avantage de KNN

1. Simple à implémenter
2. Gérer naturellement les cas multi classes
3. Peut être utilisé pour la classification et la régression

Inconvénients de KNN

1. le choix de la valeur de k (le nombre de voisins le plus proche)
2. Le cout de calcul est élevé (pour chaque instance de l'ensemble de données on a besoin de calculer la distance)
3. Stockage de données
4. Sensible aux fonctionnalités non pertinentes

2.4.2 Decision Trees (Arbre de décision)

Decision Trees ou L'arbre de décision c'est un algorithme parmi les algorithmes d'apprentissage supervisé le plus utilisé et le plus pratique, qui est adapté pour résoudre tout type de problèmes (classifications ou régressions) telle-que :

→ Un arbre de décision est une structure arborescente semblable à un organigramme où un nœud interne représente une caractéristique (ou un attribut), la branche représente une règle de décision et chaque nœud feuille représente le résultat, cette structure aide pour prendre la décision.

→ C'est un algorithme non-paramétrique signifie qu'il n'y a pas d'hypothèse sous-jacente sur la distribution des données. [23]

Mesure de sélection d'attribut

Le principal problème qui se pose lorsque la construction d'un arbre de décision si comment choisi ou sélectionné le meilleur attribut pour le nœud racine et qui sépare mieux l'ensemble de données ? Pour résoudre ce problème il existe un technique qui appelé Mesure de sélection d'attribut ou ASM qui contient deux mesures principales et populaires sont :

1. Indice de Gini
2. Gain d'information

Exemple

C'est un petit exemple pour prédire si une personne est diabétique ou non, cette modèle contient trois attributs qui sont « *minimun systolic blood pressure* », « *age* » et « *glucose* » avec deux classes « diabetic » et « non-diabetic » .

Dans cette exemple les attributs représente les nœuds interne, lequel basé pour l'arbre divise en branche, la fin de branche qui ne sépare plus est le feuille (la décision) où on peut prédire si un personne est diabétique ou non.

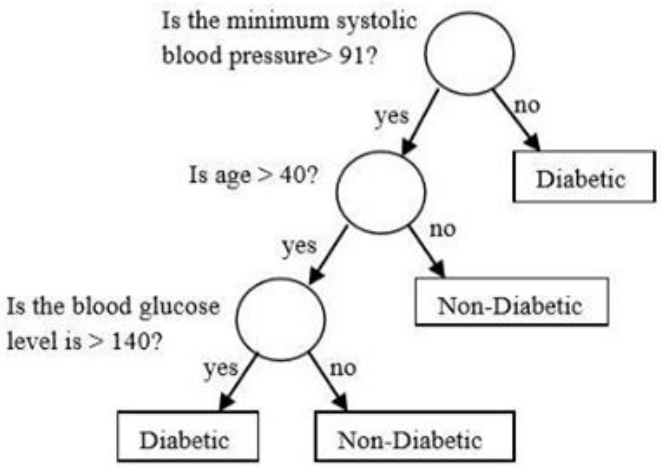


FIGURE 2.5 – Arbre de décision répondre à la question «si un personne diabétique ou non ? »[24]

L'interprétation de l'exemple

Les arbres de décision sont bien adaptés aux problèmes de catégorisation où les attributs sont vérifiés pour déterminer une catégorie finale a cause de sa construction naturelle «si... alors... sinon... » . Par exemple la Lecteur de l'exemple :
Si minimun systolic blood pressure > 91 = no , **alors** : personne = diabetic ;
Si minimun systolic blood pressure > 91 = yes **AND** age > 40 = no , **alors** : personne = non-diabetic ;
Si minimun systolic blood pressure > 91 = yes **AND** age > 40 = yes **AND** glucose =no ,**alors** : personne = non-diabetic ;
Si minimun systolic blood pressure > 91 = yes **AND** age > 40 = yes **AND** glucose =yes , **alors** : personne = diabetic ;

Algorithme de construction d'un arbre de décision

1. **sélectionne le meilleur attribut (nœud racine) :** pour chaque attribut le gain d'information est calculé, et celui qui il est maximal est sélectionner et des branches sont créer pour chaque valeurs de cette attribut .

2. **Continuez la division :** pour chaque branche s'étendant à partir de nœud, en répétant récursivement le processus .
3. **Arrête la division si :**
   3.1 nous obtenons un nœud pur, c'est-à-dire un nœud qui ne contient que des points de données positifs ou négatifs .
   3.2 nous obtenons très peu de points dans un nœud .
   3.3 on atteint une certaine profondeur de l'arbre . [25]

Avantage des arbres de décision

1. faciles à expliquer et comprendre .
2. Fonctionne avec des données catégorielles et numériques .
3. peu coûteux en termes de calcul .

Inconvénient des arbres de décision

1. Il faut souvent plus de temps pour former le modèle .
2. L'arbre devient plus complexe à mesure qu'il s'approfondit .
3. Un petit changement dans les données peut entraîner un changement global de la structure de l'arbre de décision .

2.4.3 Random Forest (forêts aléatoires)

Random Forest ou forêts aléatoires est un algorithme d'apprentissage supervisé très populaire Il est également utilisé pour les problèmes de régression ou de classification. Basé sur un ensemble des algorithmes d'apprentissage, qui est un processus de combinaison de plusieurs algorithmes pour résoudre un problème complexe et améliorer les performances du modèle. C'est un algorithme qui créer de nombreux arbres de décision (c'est la raison pour laquelle il est appelé une forêt) sur divers sous-ensembles de l'ensemble de données. Elle prend la prédiction de chaque arbre et sur la base des votes majoritaires des prédictions, et elle prédit le résultat final. [28]

La figure suivant explique le fonctionnement et la structure d'algorithme

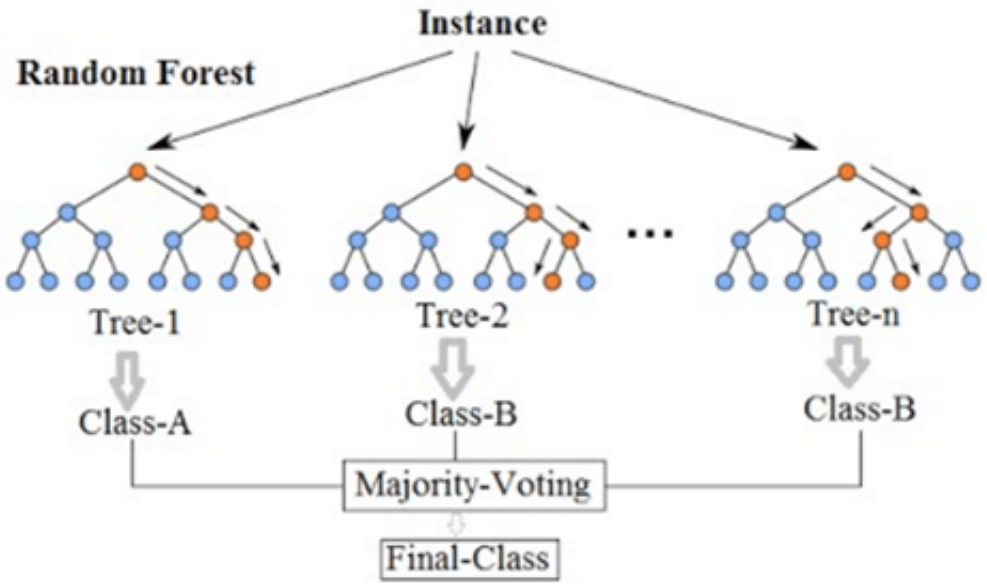


FIGURE 2.6 – Structure de l'algorithme random forest [29]

Exemple

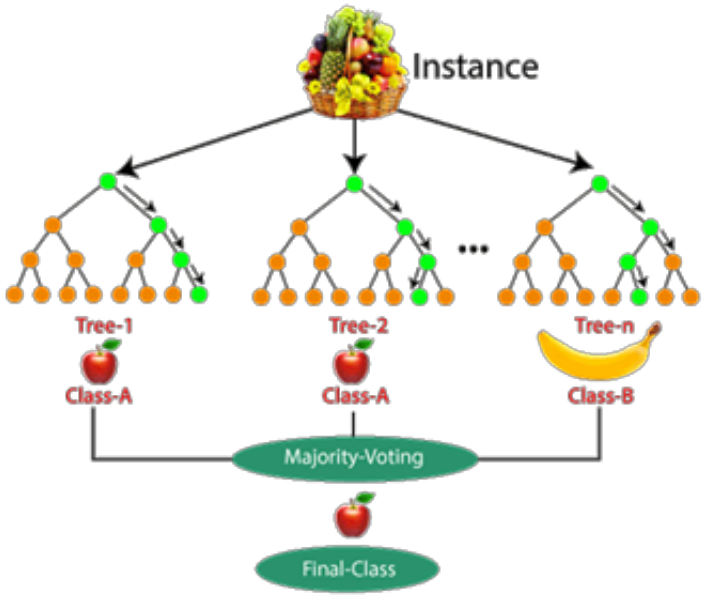


FIGURE 2.7 – Un simple exemple sur l'algorithme random forest [28]

L'interprétation d'exemple

Dans cet exemple l'ensemble de données contenant un ensemble d'images de fruits classifié par l'algorithme random forest, cette ensemble est divisé en sous-ensembles et donné à chaque arbre de décision et dans la phase d'apprentissage chaque arbre produit un résultat de prédiction, et lorsqu'un nouveau point de données se produit, puis sur la base de la majorité des résultats, Random Forest prédit la décision finale (comme l'exemple dans l'image).

Algorithme de construction de Random forest

1. Sélectionnez des échantillons aléatoires à partir d'un ensemble de données d'entrainement.
2. Créer des arbres de décision pour chaque échantillon (sous-ensembles). Ensuite on obtient le résultat de prédiction de chaque arbre de décision
3. Pour les nouveaux points le vote sera effectué pour chaque résultat prédit.
4. sélectionnez le résultat de prédiction le plus voté comme résultat de prédiction final. [30]

Avantage de Random forest

1. Il s'agit de l'un des algorithmes d'apprentissage les plus précis disponibles. Pour de nombreux ensembles de données, il produit un classificateur très précis.
2. Il fonctionne efficacement sur de grandes bases de données.
3. Il dispose d'une méthode efficace pour estimer les données manquantes et maintient la précision lorsqu'une grande partie des données sont manquantes.

Inconvénient de Random forest

Le principal inconvénient de l'algorithme random forest est qu'un grand nombre d'arbres peut rendre l'algorithme trop lent et inefficace pour les prédictions en temps réel. En général, ces algorithmes sont rapides à entraîner, mais assez lents à créer des prédictions une fois qu'ils sont formés. Une prévision plus précise nécessite plus d'arbres, ce qui entraîne un modèle plus lent.

2.4.4 Support Vector Machine (SVM)

Support Vector Machine ou SVM est l'un des algorithmes d'apprentissage supervisé les plus populaires, utilisé pour les problèmes de classification et de régression. Cependant, il est principalement utilisé pour les problèmes de classification dans l'apprentissage automatique. Le but de l'algorithme SVM est de créer la meilleure ligne ou limite de décision qui peut séparer l'espace à n dimensions en classes afin que nous puissions facilement mettre le nouveau point de données dans la bonne classe à l'avenir. Cette meilleure frontière de décision est appelée un **hyperplan**.

SVM choisit les points / vecteurs extrêmes qui aident à créer l'hyperplan. Ces cas extrêmes sont appelés vecteurs de support, et donc l'algorithme est appelé machine de vecteur de support. [31]

Les diagrammes suivant illustre deux classes (classe des points bleus et classe des points roses) différant qui sont classés avec un hyperplan.

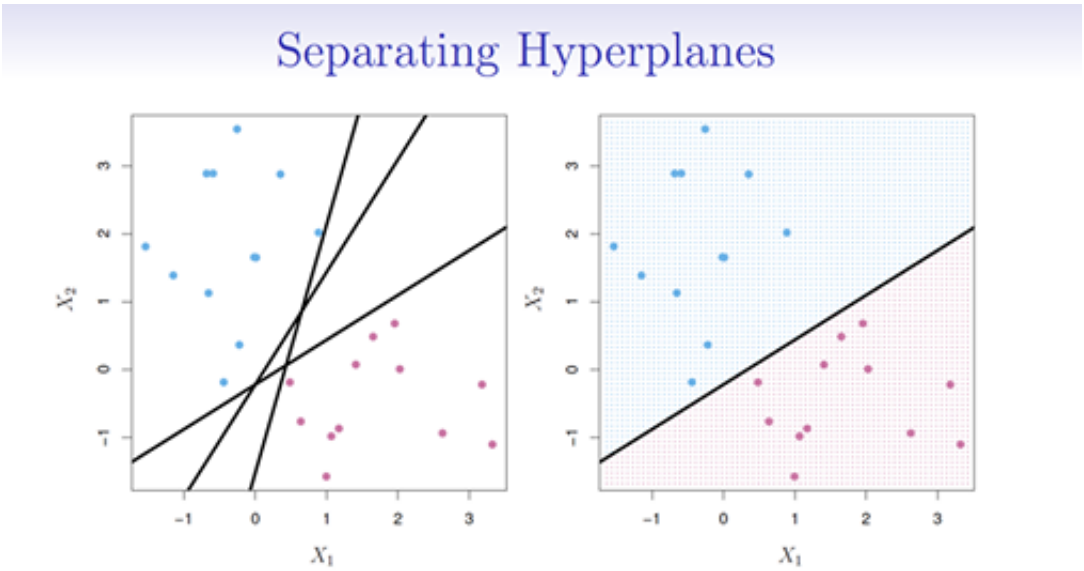


FIGURE 2.8 – Séparation parfait de deux classes avec un hyperplan [32]

Exemple

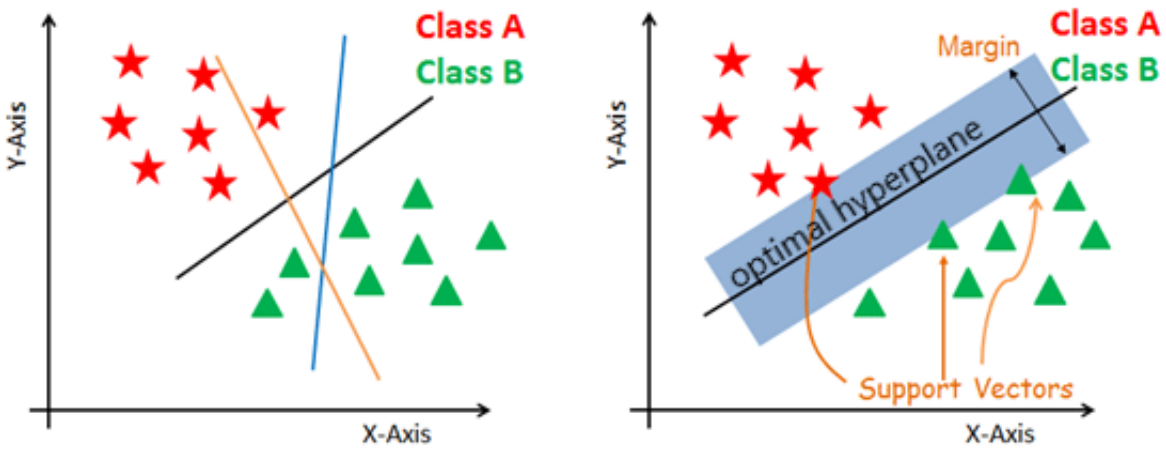


FIGURE 2.9 – Un simple exemple sur le fonctionnement de l'algorithme SVM [33]

→ L'interprétation d'exemple

Dans cette exemple le jeu de données contient des étoiles et des triangles qui sont respectivement classé dans les classe A et B, dans la phase d'apprentissage le classificateur SVM consiste à trouve le meilleur hyperplan qui sépare parfaitement les deux classe, et classe correctement les nouveaux données ainsi comme les vecteurs de support crée une frontière de décision entre les deux classes les nouveaux données sera classé à la base de ces vecteurs .

Hyperplan et vecteur de support et marge dans l'algorithme SVM

→**Hyperplan** : Les hyperplans sont des limites de décision qui aident à classer les points de données dans un espace à n dimensions, ces points de données tombant de chaque côté de l'hyperplan peuvent être attribués à différentes classes.
La dimension de l'hyperplan dépendent au nombre des entités dans le jeu de données, si le nombre d'entité égale à 2 l'hyperplan sera une ligne. Et si le nombre d'entité égal à 3 l'hyperplan devient un plan bidimensionnel.[34]

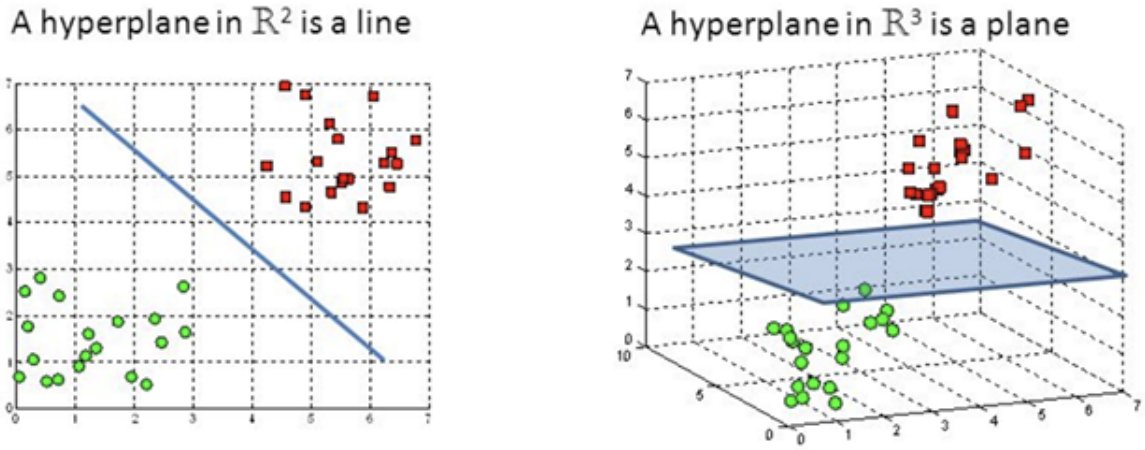


FIGURE 2.10 – Hyperplan dans les entités 2D et 3D [34]

→**Vecteur de support :** Les vecteurs de support sont des points de données plus proches de l'hyperplan, et influencent la position et l'orientation d'hyperplan, la suppression de ces vecteur modifier la position de l'hyperplan. [34]

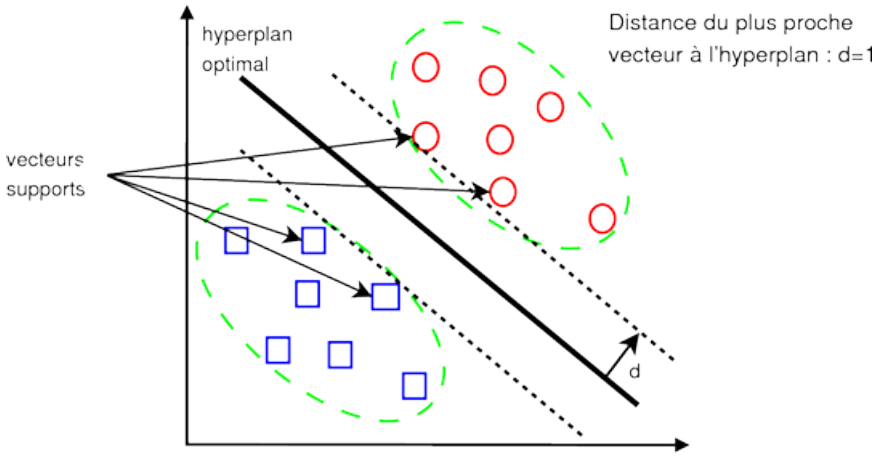


FIGURE 2.11 – Les vecteurs de support [35]

→**Marge :** c'est la distance entre les vecteurs de support et l'hyperplan. l'hyperplan optimal c'est qui a le plus grand marge, car une plus grande marge garantit que de légères déviations dans les points de données ne doivent pas affecter le résultat du modèle. [34]

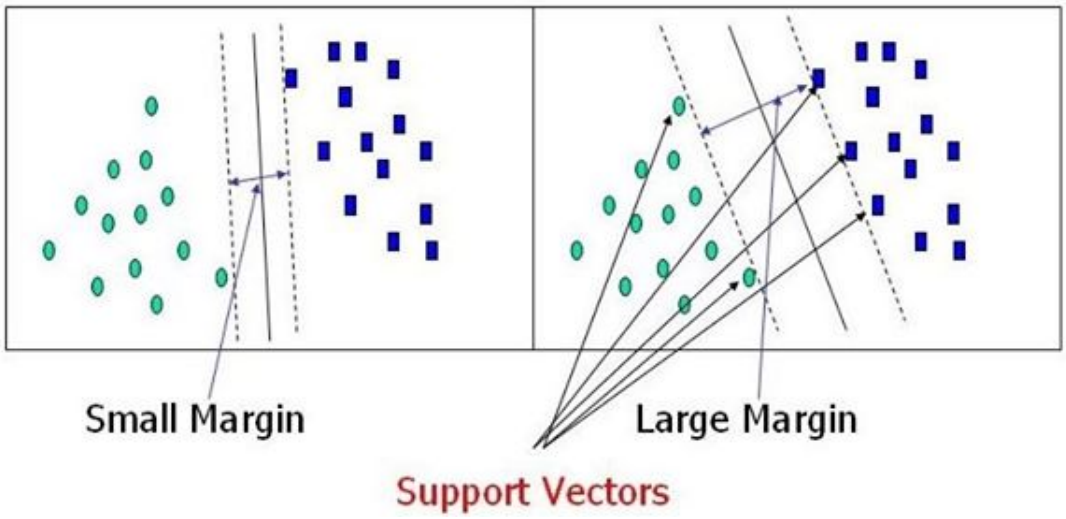


FIGURE 2.12 – Marge dans l'algorithme SVM[34]

Avantage de SVM

1. Il a la capacité de gérer de grands espaces fonctionnels.
2. Fonctionne bien avec même des données non structurées et semi-structurées comme du texte, des images et des arbres.
3. Il s'adapte relativement bien aux données de grande dimension.

Inconvénient de SVM

1. Il est sensible au bruit .
2. Difficile de comprendre et d'interpréter le modèle final, les poids variables et l'impact individuel.
3. L'extension de la classification à plus de deux classes est problématique.

2.4.5 Naïves Bayes

Naïve bayésienne fait partie des algorithmes d'apprentissage automatique supervisé qui sont principalement utilisés pour la classification. c'est un classificateur probabiliste simple basé sur l'application de *théorème de bayes* et qui aide à construire des modèles d'apprentissage automatique rapides qui peuvent faire des prédictions rapides.
«Naive» dans l'algorithme se réfère à l'hypothèse naïve que l'algorithme fait, qui est que chaque fonctionnalité est indépendante des autres fonctionnalités.[36]

→ Théorème de bayes

Le théorème de Bayes (alternativement la loi de Bayes ou la règle de Bayes) décrit la probabilité d'un événement , basée sur la connaissance préalable des conditions qui pourraient être liées à l'événement. La formule est comme suit :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Où :

$P(A|B)$: la probabilité conditionnelle que l'événement A se produise, étant donné que B s'est produit. Ceci est également connu comme la probabilité postérieure.

$P(B|A)$: la probabilité conditionnelle que l'événement B se produise, étant donné que A s'est produit.

$P(A)$ et $P(B)$: probabilité de A et B sans égard l'un à l'autre .[37]

Avantage de Naïves Bayes

1. fonctionne également bien dans la prédiction multi-classes.
2. Lorsque l'hypothèse d'indépendance est vérifiée, un classificateur Naïve Bayésienne fonctionne mieux que d'autres modèles.
3. Fonctionne mieux que les modèles plus compliqués lorsque l'ensemble de données est petit.

Inconvénient de Naïves Bayes

limitation de Naïve Bayésienne est l'hypothèse de fonctionnalités indépendantes. Dans la vraie vie, il est presque impossible d'obtenir un ensemble de fonctionnalités complètement indépendants.

Remarque

Un bon modèle d'apprentissage choisit la fonction de prédiction qui réalise la plus faible erreur de prédiction.

2.5 Les travaux de recherche sur l'application des algorithmes de machine learning pour la prédiction du diabète type 2

Ci-dessous, nous présenterons trois études prédictives conçues pour prédire le diabète et qu'utilisent la précision comme un facteur de comparaison entre les algorithmes d'apprentissage utilisés :

1. Prediction of Diabetes Using Machine Learning Algorithms in Healthcare [38]
2. Machine Learning Workflow on Diabetes Data [39]
3. Application des méthodes d'apprentissage dans la prédiction du diabète Type 2 [40]

2.5.1 Etude 01 : Prediction of Diabetes Using Machine Learning Algorithms in Healthcare

1. La précision :

La précision (Accuracy) est le taux de réussite global de l'algorithme défini par l'équation suivante :

$$Accuracy = (TP + TN)/(P + N) \text{ ....(01)}$$

TP : True Positive . TN : True Negative .

FN : Faux négatif . FP : False Positive .

Dans notre cas d'étude, la signification de TP, TN, FP et FN est comme suit :

— TP : signifier qu'une personne est réellement diabétique et elle a été prédit qu'elle est diabétique.

— TN :signifie qu'une personne est réellement non diabétique et elle a été prédit qu'elle est non diabétique.

— FP :signifie qu'une personne est réellement non diabétique et elle a été prédit qu'elle est diabétique.

— FN : signifie qu'une personne est réellement diabétique et elle a été prédit qu'elle est non diabétique.

2. L'étude :

Cette étude a conclu que SVM et KNN sont appropriés pour prédire l'état du diabète des patients entre six algorithmes d'apprentissage automatique. Ces algorithmes sont : K-Nearest Neighbours (KNN), Support Vector Classifier(SVM), Logistic Regression(LR),Decision Tree Classifier(DT), Gaussian Naive Bayes(NB) et Random Forest (RF). Tous ces algorithmes ont été appliqués à Dataset PIMA Indian comprenant 768 enregistrements et 9 attributes . L'ensemble de données a été divisé en deux parties, les données d'entraînement (training data)et les données de test(test data), ces deux parties constituant respectivement 70% et 30% des données. dans ce travail, le principal paramètre d'évaluation entre les algorithmes est la précision de prédiction définie par l'équation (01) [38].

algorithme	KNN	SVM	NB	LR	DT	RF
Précision	77%	77%	74%	74%	71%	71%

TABLE 2.1 – Les résultats de la précision des algorithmes d'étude 01

$\longrightarrow$ **Critique :**

Pour construire un modèle de prédiction du diabète avec la plus grande précision possible qui nécessite un grand dataset avec des milliers d'enregistrements et avec un minimum ou aucune valeur nulle , révélera plus d'informations et une meilleure précision et les limites de cette étude sont :

1. la taille de l'ensemble de données et les valeurs d'attribut aberrantes inattendues (les valeurs nulle).
2. tous les patients de dataset sont des femmes d'au moins 21 ans d'origine indienne Pima .

2.5.2 Etude 02 : Machine Learning Workflow on Diabetes Data

Cette étude concerne la prédiction de diabète type 2 avec l'utilisation de quelque algorithme d'apprentissage automatique dans un workflow qui contient plusieurs tâches tels que l'exploitation des données, le nettoyage des données jusqu'à la sélection de modèle optimal.Il utilise 7 classificateurs qui sont appliqué à l'ensemble de données « pima indian

» à savoir K-Nearest Neighbours (KNN), Support Vector Classifier(SVM), Logistic Regression(LR),Decision Tree Classifier(DT), Gaussian Naive Bayes(GNB), Random Forest (RF)et Gradient Boost (GB) afin d'évaluer leur précision avec deux méthode d'évaluation [39].

Méthode 01 :Train/Test Split

Fractionner l'ensemble de données en deux parties :

1. Partie d'entrainement : pour former le modèle .
2. Partie de test : pour tester le modèle et évaluer la précision .

Méthode 02 :Validation croisé

Subdiviser l'ensemble de données en k sous-ensembles de même taille (Folds) et on utilise (01) fold comme partie de test et l'union des autres c'est la partie d'entrainement . Les résultat de précision de chaque algorithme selon chaque méthode : A la fin le modèle

	algorithme	Train Test Split	Validation croisé
1	KNN	0.711521	0.711521
2	SVM	0.656075	0.656075
3	LR	0.776440	0.776440
4	DT	0.681327	0.685494
5	GNB	0.755681	0.755681
6	RF	0.739165	0.747519
7	GB	0.765452	0.765442

TABLE 2.2 – Les résultat de précision des algorithmes d'étude 02

régression logistique (LR) est choisie comme le modèle qui fonctionne le mieux pour l'ensemble de données avec une précision égale à 77.64%

⟶ Critique :

1. La suppression des valeurs nulles pour la phase de nettoyage de données qui génère une perte de plus d'une 40 observation qui est important pour la prédiction

2. Manque des paramètres des chaque classificateurs utilisé dans la phase de sélection de modèle qui peut modifier les valeurs de la précision

2.5.3 Etude 03 : Application des Méthodes d'Apprentissage dans la Prédiction du Diabète de Type 2

Cette étude fait une comparaison entre les cinq algorithmes ci-dessous avec la méthode d'évaluation « validation croisé » en 10 itérations. À chaque itération elle est sélectionné la meilleure itération qui a la meilleure précision et la comparer aux meilleurs itérations des autres algorithmes afin de sélectionner l'algorithme optimal ,ces algorithmes sont appliquées sur deux bases de données issue du Centre Hospitalo-Universitaire "CHU" de Bejaia et une d'une cabinet médicale de Dr Djamel MEHIDI privé au centre de la daïra d'Adekar Wilaya de Bejaia. [40]

Algorithme	Meilleure itérations pour la base de données du cabinet médicale	Meilleure itérations pour la base de données du CHU
Random Forest	Itérations 2 : 85%	Itération 9 : 85%
Décision tree	Itérations 1 : 75%	Itération 7 : 80%
K-plus proche voisin	Itérations 2 :83%	Itération 3 :83%
SVM	Itérations 2 :80%	Itération 3 :79%
Gaussian naive	Itérations 2 :83%	Itération 5 : 85%

TABLE 2.3 – Les résultat de précision des algorithmes d'étude 03

En conclusion, cette étude a sélectionné l'algorithme « Random Forest » comme le modèle optimal avec une précision 85% pour les deux meilleures itérations sur les deux bases de données .

⟶ Critique :

Manque des attributs d'évaluation comme la sensitivité pour la comparaison entre les différents algorithmes utilisés .

2.6 Conclusion

Dans ce chapitre, nous avons présenté les algorithmes d'apprentissage automatique qui peuvent nous aider à détecter l'apparition précoce du diabète, ce qui peut aider à réduire les risques des complications de cette maladie sur la santé du patient.
Dans l'étude qui suit, l'objectif principal est d'appliquer ces différents algorithmes (K nearest neighbors, Decision Trees, Random Forest, Support Vector Machine, Naïves Bayes) de classification aux données extrait de l'hôpital frankfurt concernant le risque de développer un diabète de type 2 .

Chapitre 3

Chapitre 3 : Prédiction du diabète de type 2 par l'apprentissage automatique

3.1 Introduction

Dans ce dernier chapitre, nous présentons d'abord une étude technique dans laquelle nous définissons l'environnement logiciel utilisé pour construire notre application, puis nous définirons notre dataset avec une description de ses caractéristiques et les étapes de pré-traitement des données (explorer, nettoyer, sélection de modèle ...) pour corriger les valeurs aberrantes et choisir le meilleur modèle à suivre.
A la fin, c'est la partie application où nous fournissons des interfaces graphiques importantes développées pour clarifier les performances des activités du système et nous terminerons par une conclusion.

3.2 outils et Librairies utilisés

3.2.1 Anaconda

Anaconda est une distribution python pour les applications de data science et d'apprentissage automatique. C'est un logiciel gratuit et open source qui contient plusieur packages. Le principal avantage de l'utilisation d'anaconda est que, anaconda est comme un point central pour les bibliothèques qui auraient besoin pour le traitement de données, l'analyse prédictive et les calculs scientifiques.[20]

→ jupyter notebook

Jupyter Notebook est un environement de programmation qui prend en charge plusieur langages de programmation, dont Python. Jupyter Notebook nous permet de créer des documents contenant du code , des équations, des visualisations et du texte . ses utilisations comprennent : le nettoyage et la transformation des données, la simulation numérique, la modélisation statistique, la visualisation des données, l'apprentissage automatique et bien plus encore.[42]

3.2.2 Python

Python C'est un langage de programmation multi-paradigme et le langage de programmation dominant dans la data science avec de nombreuses implémentations ce qui le rend encore plus intéressant.concernant le domaine de l'apprentissage automatique Python se distingue tout particulièrement en offrant une pléthore de librairies de très grande qualité, couvrant tous les types d'apprentissages disponibles qui combine la facilité d'utilisation et d'apprentissage avec la puissance des librairies qu'elles possèdent. Parmi ces bibliothèques, nous avons utilisé :

→ Matplotlib

Matplotlib est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python.

→ **Seaborn**

Seaborn est une bibliothèque de visualisation de données Python basée sur matplotlib . Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayants et informatifs.

→ **Pandas**

Pandas est une autre bibliothèque Python utilisée pour la manipulation et l'analyse des données, le point fort de cette bibliothèque est qu'elle possède une fonctionnalité importante appelée nettoyage des données qui résout le problème du temps passé à nettoyer les données dans un projet d'apprentissage automatique car de nombreux ensembles de données disponibles contiennent des champs vides ou nuls, ce qui peut avoir un impact négatif énorme sur notre modèle.

→ **NumPy**

NumPy est une extension du langage de programmation Python, destinée à manipuler des tableaux multidimensionnels.

→ **Scikit-learn**

elle est la bibliothèque Python la plus importante pour ce qui concerne l'apprentissage automatique telle que il contient de nombreux algorithmes (forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines à vecteurs de support) .

3.2.3 Flask

Flask est un petit framework web Python léger, qui fournit des outils et des fonctionnalités utiles qui facilitent la création d'applications web en Python. Il offre aux développeurs une certaine flexibilité et constitue un cadre plus accessible pour les nouveaux développeurs puisque vous pouvez construire rapidement une application web en utilisant un seul fichier Python. Flask est également extensible et ne force pas une structure de répertoire particulière ou ne nécessite pas de code standard compliqué [46]

3.3 Définition d'ensemble de données utilisé et description des variables

3.3.1 Définition l'ensemble de données utilisé

C'est un ensemble de données sur le diabète, extrait de l'hôpital de Frankfort, Allemagne, il se composent de plusieurs variables prédictives médicales et d'une variable cible « Outcome ».

Les variables sont les suivants :

1. Glucose : Concentration plasmatique de glucose à 2 heures dans un test oral de tolérance au glucose.
2. Pregnancies : Nombre de fois enceinte.
3. BloodPressure : Pression artérielle diastolique (mm Hg).
4. SkinThickness : Epaisseur de pli cutané du triceps (mm).
5. Insulin : Insuline sérique 2 heures (mu U/ ml).
6. BMI : (ou IMC) Indice de masse corporelle (poids en kg / (taille en m)^2).
7. DiabetesPedigreeFunction : Fonction généalogique du diabète.
8. Age : l'âge en années.
9. Outcome : variable de classe (0 ou 1) où 0 indique que le patient ne souffre pas de diabète et 1 indique que le patient est diabétique.

3.3.2 Description des variables

Variable	Description	Analyse de données
Glucose	Une valeur de 2 heure entre (140 et 200 mg)/dl (7.8 et 11.1 mmol/L) est appelé tolérance au glucose altéré signifie que il y a un risque accru de développe le diabète au fil de temps.Un taux de glucose de 200 mg/dL(11.1 mmol/L) ou plus utilisé pour diagnostiquer le diabète.	Minimum : 0 Maximum : 199

Variable	Description	Analyse de données
Pregnancies	Nombre de fois enceinte	Minimum : 0 Maximum : 17
BloodPressure	Si un TA diastolique > 90 signifie une pression artérielle élevé (probabilité élevé de diabète) Un TA diastolique < 60 signifie une pression artérielle base (mois probabilité de diabète)	Minimum : 0 Maximum : 122
SkinThikness	Valeur estimé pour la graisse corporelle. épissure normal du pli cutané chez les femmes est de 23 mm. Une épissure plus élevée conduit à l'obésité et les chances de diabète augmente.	Minimum : 0 Maximum : 110
Insulin	Insuline sérique 2 heures (mu U/ ml) et niveau d'insuline normal 16-166 mUI/L, les valeurs au-dessus de cette plage peuvent être alarmante.	Minimum : 0 Maximum : 799
BMI	(poids en kg / taille en m^2) IMC de 18.5 à 20 c'est normal IMC entre 25 et 30 situer dans une plage surpoids Et de 30 ou plus situer dans la fourchette d'obésité	Minimum : 0 Maximum : 80.6
DiabetePredigme Function	Fournit des informations sur les antécédentes chez les parents et la relation génétique avec les patients. Une fonction de pedigree plus élevée signifie que le patient plus susceptible de souffrir un diabète	Minimum : 0.078 Maximum : 2.42
Age	Age d'une personne en années	Minimum : 21 Maximum : 81
Outcome	Indique si une personne est diabétique ou non	0(non diabétique) :1316 1(diabétique) : 684

TABLE 3.1 – Description des variables d'ensemble de données

3.4 Les étapes de pré-traitement de données

Créer un modèle de Machine Learning est un processus en plusieurs étapes.Chaque étape présente ses propres défis techniques et conceptuels. Le prétraitement des données pour le Machine Learning implique à la fois la visualisation des données pour définir des informations et faire des analyses sur les caractéristiques d'un ensemble de données, le nettoyage de données qui consiste de faire la suppression ou la correction des enregistrements contenant des valeurs corrompues ou non valides pour un ensemble de données dans le but d'améliorer la qualité de données, et enfin la sélection de modèle qui est capable de faire la prédiction mieux que les autres modèle candidats.

3.4.1 Exploration et visualisation de données

La visualisation des données est définis comme l'exploration visuelle et interactive des données de toutes volumétries. Qui aident à voir des choses n'étaient pas évidentes auparavant. La visualisation facilite la transmission des informations de façon universelle et facilite le partage d'idées avec les autres. L'ensemble de données ressemble à :

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0
...	...	...	...	...	...	...	...	...	...
1995	2	75	64	24	55	29.7	0.370	33	0
1996	8	179	72	42	130	32.7	0.719	36	1
1997	6	85	78	0	0	31.2	0.382	42	0
1998	0	129	110	46	130	67.1	0.319	26	1
1999	2	81	72	15	76	30.1	0.547	25	0

2000 rows × 9 columns

FIGURE 3.1 – Aperçu de l'ensemble de données

Pour visualiser notre ensemble de données on utilise la bibliothèque « pandas profiling » qui génere un rapport de profil à partir d'un ensemble de données et qui aident d'obtenir et connaitre des informations globales et approfondies sur l'ensemble de données et les variables qui contient.

La sortie est enregistré sous forme de rapport HTML (voir la Figure suivant)

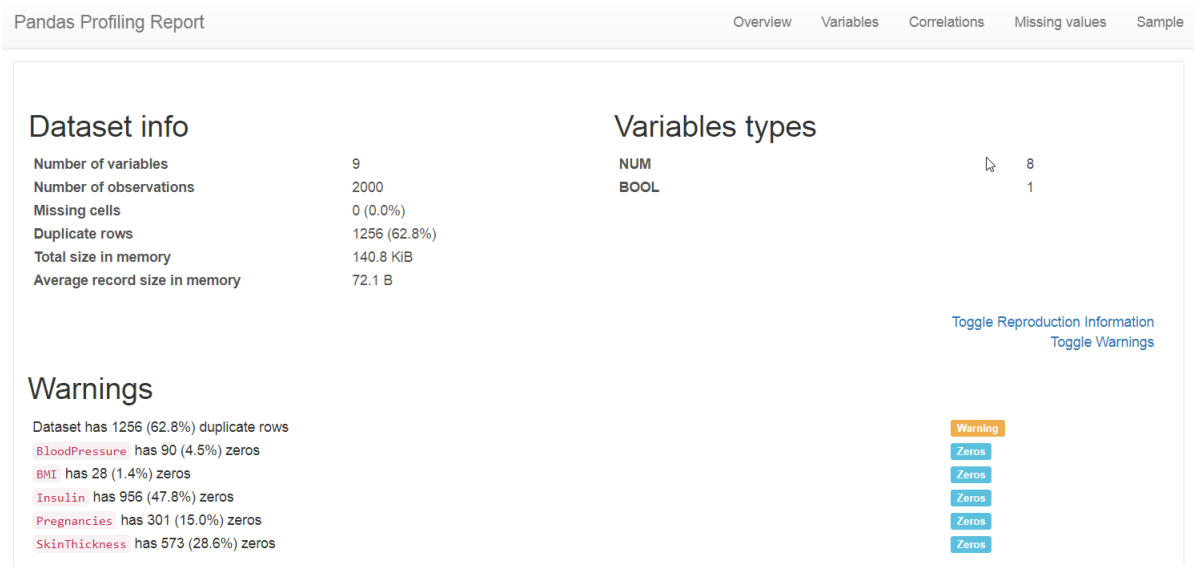


FIGURE 3.2 – Rapport HTML de l'ensemble de données

On extrait de ce rapport les informations de base sur l'ensemble de données tel que :

— Le nombre des observations (2000 patient) .

— Les nombres de variables 8 numérique et 1 variable booléenne (variable résultat) .

— La taille de l'ensemble de données .

— Les valeurs manquantes .

— Le pourcentage (%) des valeurs égale a 0 pour chaque variable .

Dans ce rapport on peut observer même les caractéristiques des variables impliquées dans l'étude et ces informations (voir les figures suivant) :

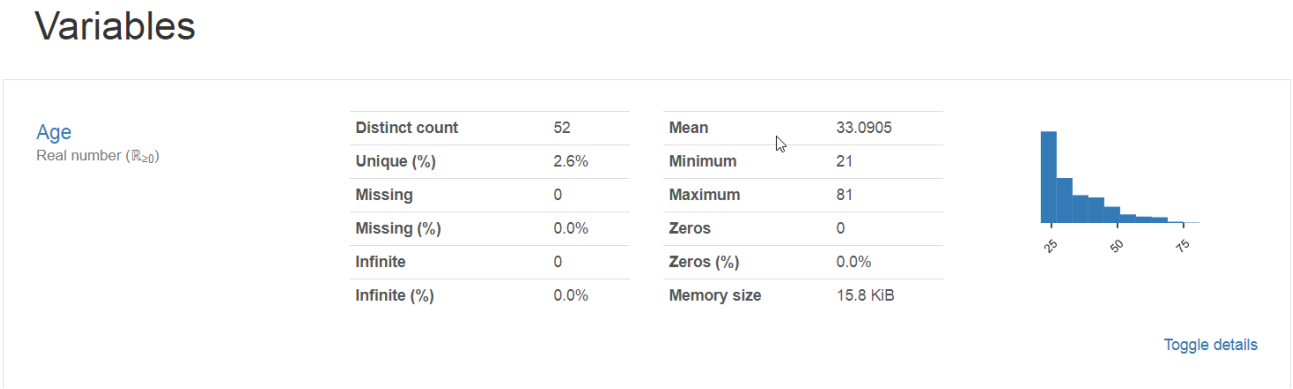
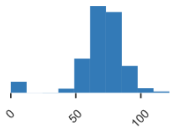


FIGURE 3.3 – La visualisation de variable « AGE »

BloodPressure
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

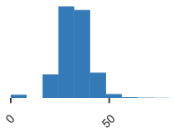
Distinct count	47	Mean	69.1455
Unique (%)	2.4%	Minimum	0
Missing	0	Maximum	122
Missing (%)	0.0%	Zeros	90
Infinite	0	Zeros (%)	4.5%
Infinite (%)	0.0%	Memory size	15.8 KiB



Toggle details

BMI
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

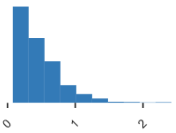
Distinct count	247	Mean	32.193
Unique (%)	12.3%	Minimum	0
Missing	0	Maximum	80.6
Missing (%)	0.0%	Zeros	28
Infinite	0	Zeros (%)	1.4%
Infinite (%)	0.0%	Memory size	15.8 KiB



Toggle details

DiabetesPedigreeFunction
Real number ($\mathbb{R}_{\geq 0}$)

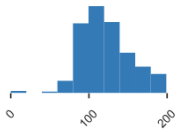
Distinct count	505	Mean	0.47093
Unique (%)	25.2%	Minimum	0.078
Missing	0	Maximum	2.42
Missing (%)	0.0%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	15.8 KiB



Toggle details

Glucose
Real number ($\mathbb{R}_{\geq 0}$)

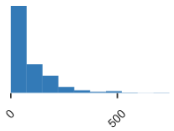
Distinct count	136	Mean	121.1825
Unique (%)	6.8%	Minimum	0
Missing	0	Maximum	199
Missing (%)	0.0%	Zeros	13
Infinite	0	Zeros (%)	0.7%
Infinite (%)	0.0%	Memory size	15.8 KiB



Toggle details

Insulin
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

Distinct count	182	Mean	80.254
Unique (%)	9.1%	Minimum	0
Missing	0	Maximum	744
Missing (%)	0.0%	Zeros	956
Infinite	0	Zeros (%)	47.8%
Infinite (%)	0.0%	Memory size	15.8 KiB



Toggle details

FIGURE 3.4 – La visualisation des variables 01

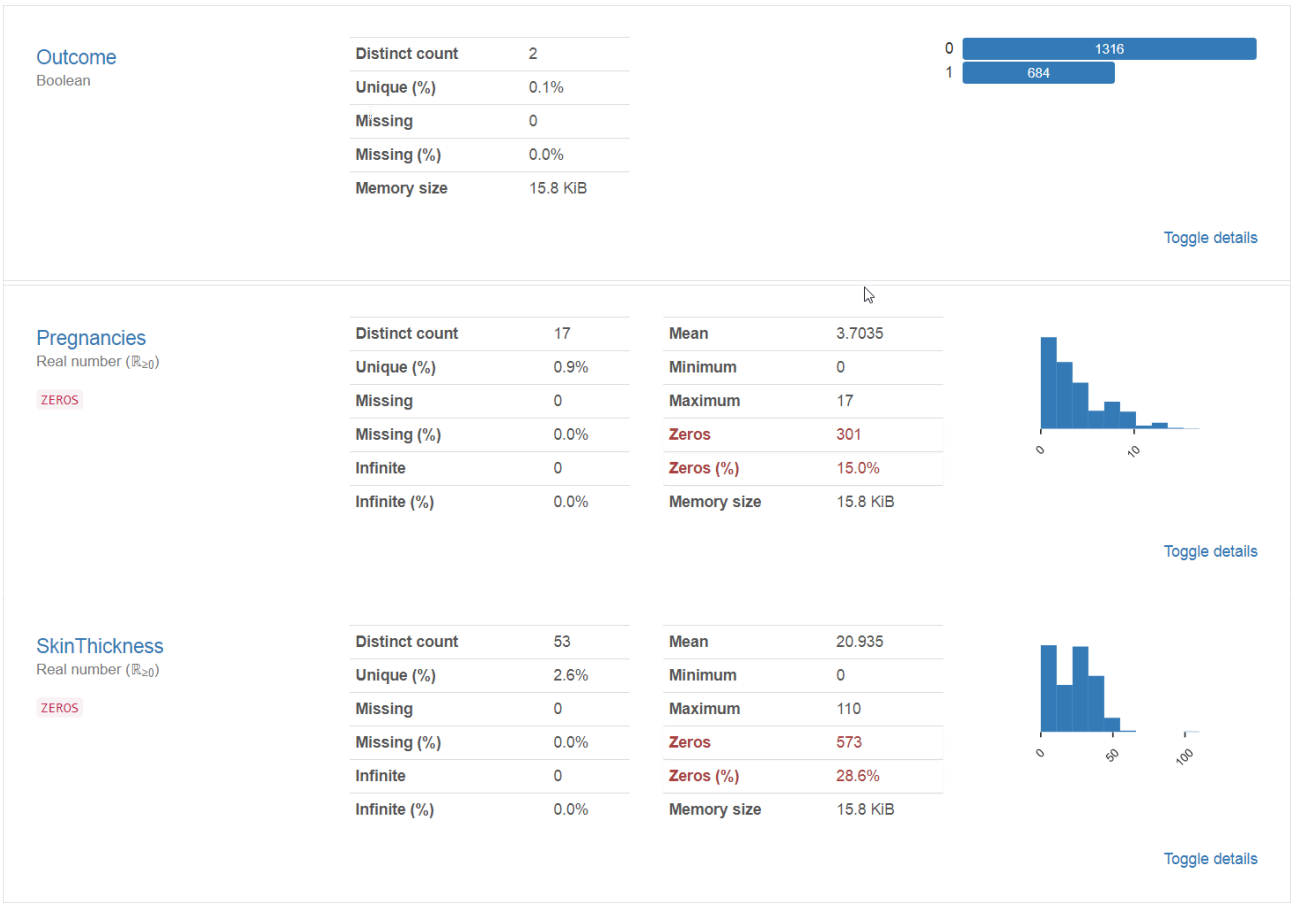


FIGURE 3.5 – La visualisation des variables 02

→ Interprétation des figures

1. Pour chaque variable on constate :
   - Le nombre des valeurs distinctes .
   - Le pourcentage des valeurs unique .
   - Le nombre et le pourcentages (%) des valeurs manquantes.
   - La taille.
   - La moyenne, minimum et maximum .
   - Le nombre et le pourcentages (%) des valeurs nulles (zéro).
   - La distribution de données en graphe.
2. Pour la variable booléenne (Outcome) on constate :
   - Le nombre des valeurs qui égal à 0 (1316 patients non diabétiques) et les valeurs qui égale à 1 (684 patients diabétiques) .

Après avoir explore et visualiser les données récoltées, il nous est apparu qu'une corrélation était nécessaire afin de répertorier les différentes relations entre ces données.

Corrélations

Un bon ensemble de données est un ensemble dans lequel les caractéristiques sont fortement corrélées à la classe cible et sont fortement non corrélées les unes aux autres. Pour trouver les attributs non corrélés, la sélection des caractéristiques se fait via une approche basée sur la corrélation utilisant un coefficient de corrélation.

→ Coefficient de corrélation : est un nombre qui indique la force de la relation entre deux variables. Il existe plusieurs types de coefficients de corrélation, mais le plus commun de tous est le coefficient de Pearson noté r , défini par :

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

ou :

$Cov(X,Y)$ désigne la covariance des variables X et Y ,

σ_X et σ_Y désignent leurs écarts types .

La valeur du coefficient de corrélation comprise entre **-1** et **+1**.

⟶ 1 signifie qu'ils sont fortement corrélés (forte relation positive) .

⟶ 0 signifie aucune corrélation.

⟶ -1 signifie qu'il existe une corrélation négative (forte relation négative).[41]

Le tableau suivant montre la corrélation entre les différentes variables de l'ensemble de données :

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.120405	0.149672	-0.063375	-0.076600	0.019475	-0.025453	0.539457	0.224437
Glucose	0.120405	1.000000	0.138044	0.062368	0.320371	0.226864	0.123243	0.254496	0.458421
BloodPressure	0.149672	0.138044	1.000000	0.198800	0.087384	0.281545	0.051331	0.238375	0.075958
SkinThickness	-0.063375	0.062368	0.198800	1.000000	0.448859	0.393760	0.178299	-0.111034	0.076040
Insulin	-0.076600	0.320371	0.087384	0.448859	1.000000	0.223012	0.192719	-0.085879	0.120924
BMI	0.019475	0.226864	0.281545	0.393760	0.223012	1.000000	0.125719	0.038987	0.276726
DiabetesPedigreeFunction	-0.025453	0.123243	0.051331	0.178299	0.192719	0.125719	1.000000	0.026569	0.155459
Age	0.539457	0.254496	0.238375	-0.111034	-0.085879	0.038987	0.026569	1.000000	0.236509
Outcome	0.224437	0.458421	0.075958	0.076040	0.120924	0.276726	0.155459	0.236509	1.000000

FIGURE 3.6 – Table de corrélation

Un autre outil qui représente la relation entre les variables est la matrice de corrélation où chaque cellule remplit en couleur en fonction du coefficient de corrélation de la paire qu'elle représente (voir la figure suivante) :

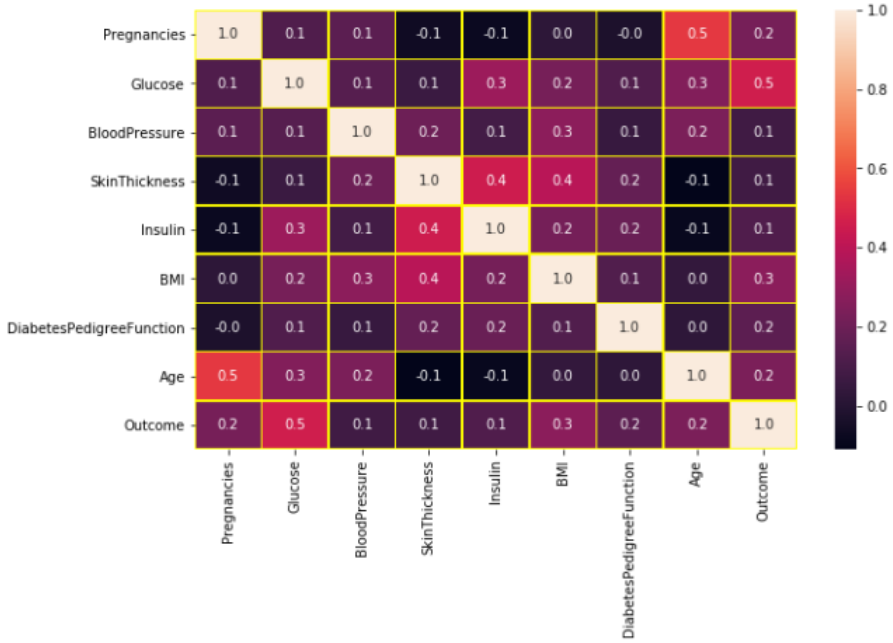


FIGURE 3.7 – Matrice de corrélation en couleur

Les deux figures montre que « Age », « Glucose » et « BMI » sont des caractéristiques important pour le diagnostic du la maladie de diabète et qui sont fortement non corrélés les uns aux autres.

→ Remarque

D'après la visualisation des données on constate qu'il n'y aucun point de données manquant ou nulle dans l'ensemble de données mais il existe des valeurs zéro pour certain colonnes qui rendre la lecture de l'ensemble de données fausse et qu'il avait besoin de nettoyage .

3.4.2 Nettoyage de données

Le nettoyage des données est un processus visant à identifier et à corriger les données altérées, inexactes ou non pertinentes. Cette étape fondamentale de prétraitement des données améliore la cohérence, la fiabilité et la valeur des données, et se traduit par de meilleures données qui fournissent de meilleurs modèles résultants.
Lors de la visualisation de données on a constaté qu'il existe des valeurs aberrantes dans certains colonnes comme :

Blood Pressure : Une personne normale ne peut pas avoir une pression artérielle diastolique 0 mm Hg et cette colonne a 90 observations ayant 0 comme valeur .

BMI : Le poids d'une personne ne sera jamais égal à 0 ou proche de zéro sa mettre sa vie en danger et on constate qu'il y a 28 observations où la valeur égal à 0 .

Glucose : Le taux de glucose ne doit pas être égal à 0 mais d'après l'analyse des données, on trouve cette colonne ayant 13 observations égale à 0.

Insulin : dans une situation rare une personne peut avoir 0 insuline mais cette colonne ayant 956 observations où la valeur égal à 0 .

Pregnancies : c'est normal d'avoir un zéro valeur pour cette colonne donc Il n'est pas besoin de nettoyage .

Skin Thikness : Pour les personnes normales, l'épaisseur du pli cutané ne peut pas être inférieure à 10 mm, mais il y a 573 observations égales à 0 .

Solutions pour gérer les valeurs aberrant :

Option 1 : supprimer toute les observations ayant zéro valeurs mais dans cette option on obtient une perte de données important (plus de 50% de l'ensemble de données)

Option 2 : calcule la valeur médiane d'une colonne spécifique et remplace cette valeur dans cette colonne où nous avons zéro .

Dans notre cas, nous avons choisi d'appliquer l'option 2 en remplaçant chaque colonne a besoin de nettoyage par :

- la valeur médiane(en utilisant la fonction median() pour déterminer le médiane de chaque colonne mentionnée) où les valeurs égal à 0 avec la fonction replace().

→ Le résultat obtenu dans les graphes :

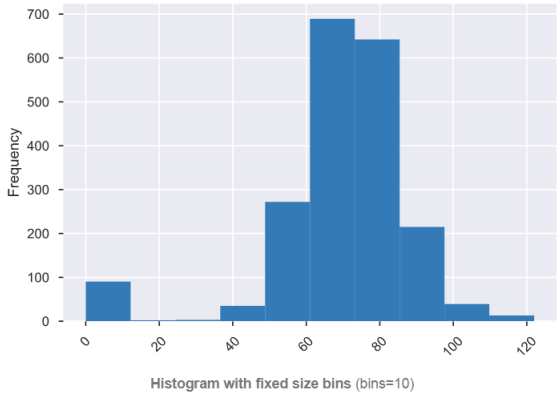
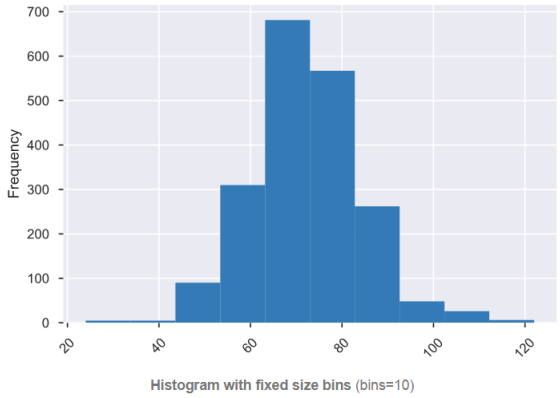
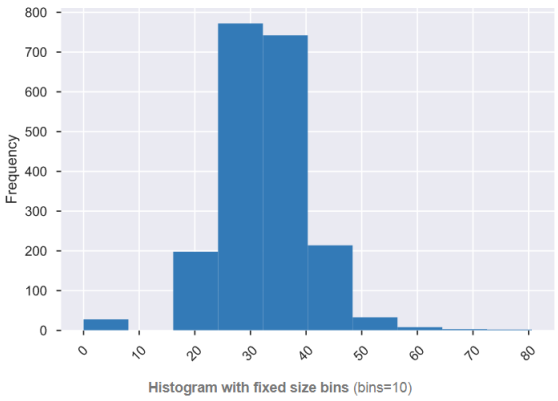
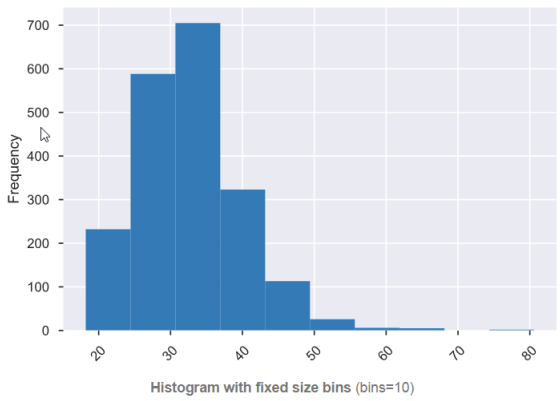
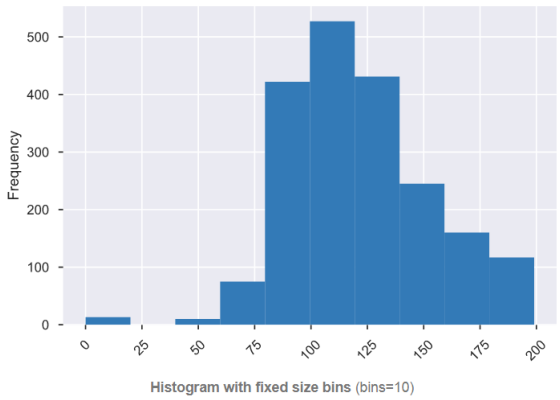
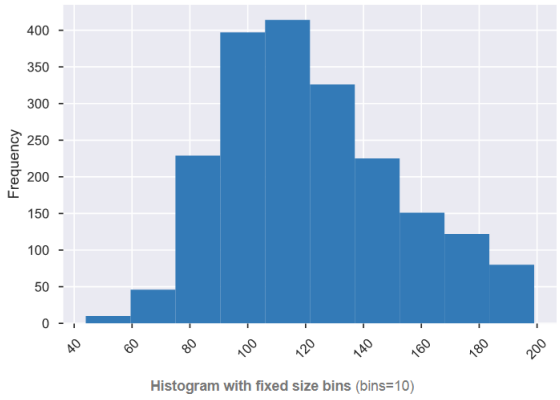
Variable	Avant	Après
Blood Pressure		
BMI		
Glucose		

TABLE 3.2 – Distribution des variables avant et après nettoyage 01

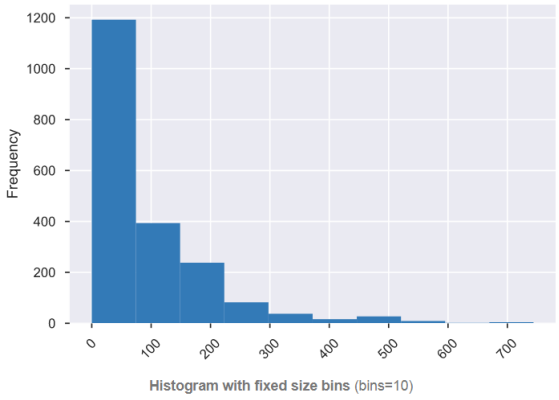
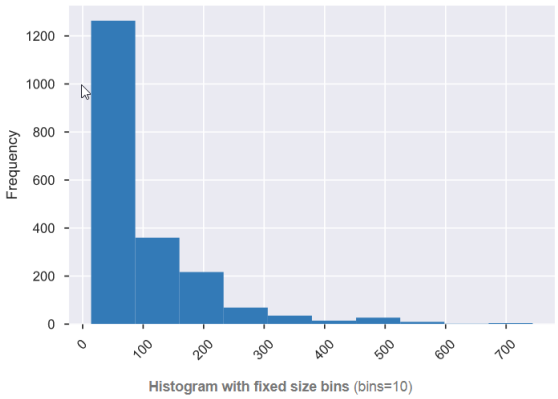
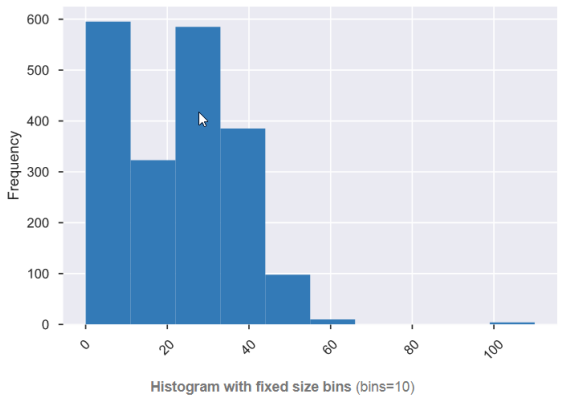
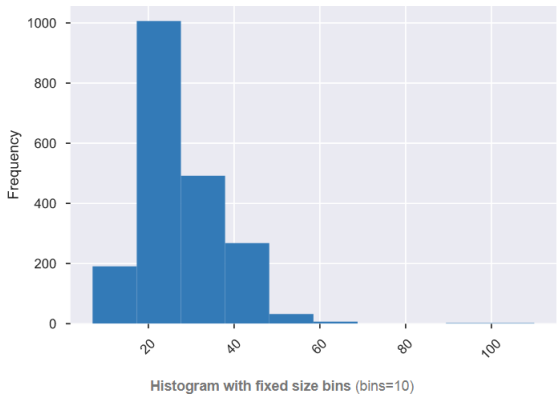
Variable	Avant	Après
Insulin		
Skin Thikness		

TABLE 3.3 – Distribution des variables avant et après nettoyage 02

3.4.3 Sélection de modèle

La sélection de modèle c'est une phase très importante et le cœur de l'apprentissage automatique où on sélectionne le modèle qui fonctionne mieux pour l'ensemble de données parmi une collection de modèles d'apprentissage automatique candidats.

Les modèles utilisés pour la prédiction de diabète sont :

1. KNN (K-Nearest Neighbors)
2. L'arbre de décision (Decision tree)
3. SVM (support vector machin)
4. Random Forest (forêt aléatoire)
5. Naïve bayésienne (Gaussian Naive Bayes)

Méthode d'évaluation

Utilise pour donne la capacité au modèle de prédire les données hors échantillons et évite le problème de sur-ajustement (overfiting) qui correspond à l'incapacité de modèle du généraliser sur des données de test car il est appris par cœur sur les données d'entrainement. Les deux méthodes sont :

1. Train/Test Split

Cette méthode consiste à divisé l'ensemble de données en deux partie : partie d'entrainement sur lequel le modèle fait son apprentissage et partie de test sur lequel on a testé le modèle et évaluer sa performance.

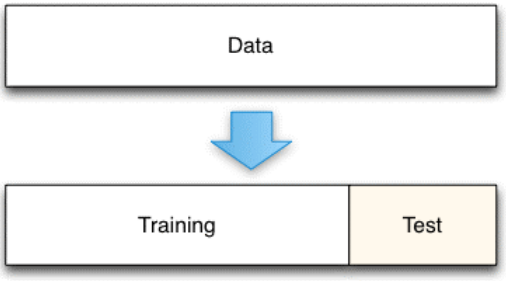


FIGURE 3.8 – Répartition des données de train/test [43]

Nous utilisons la méthode « train_test_ split » importé de la bibliothèque sklearn pour effectuer le fractionnement train/test. « test_ size=0.2 » à l'intérieur de la fonction indique le pourcentage des données qui doivent être conservées pour le test. C'est généralement autour 20% pour le test et le reste de 80% pour l'entrainement ce qui signifie 1600 observations partie d'entrainement et 400 observations partie test.

Le code standard pour fractionner les données dans la figure suivant :

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.2, random_state = 2 )
```

FIGURE 3.9 – Fractionnement de l'ensemble de données

Retourne 04 variables X_train et y_train pour l'entrainement et X_test et y_test pour le test .

Pour évaluer la précision des modèles dans la méthode train/test on a importé la métrique « accuracy_score » de la bibliothèque « sklearn ». Le résultat dans la figure suivant

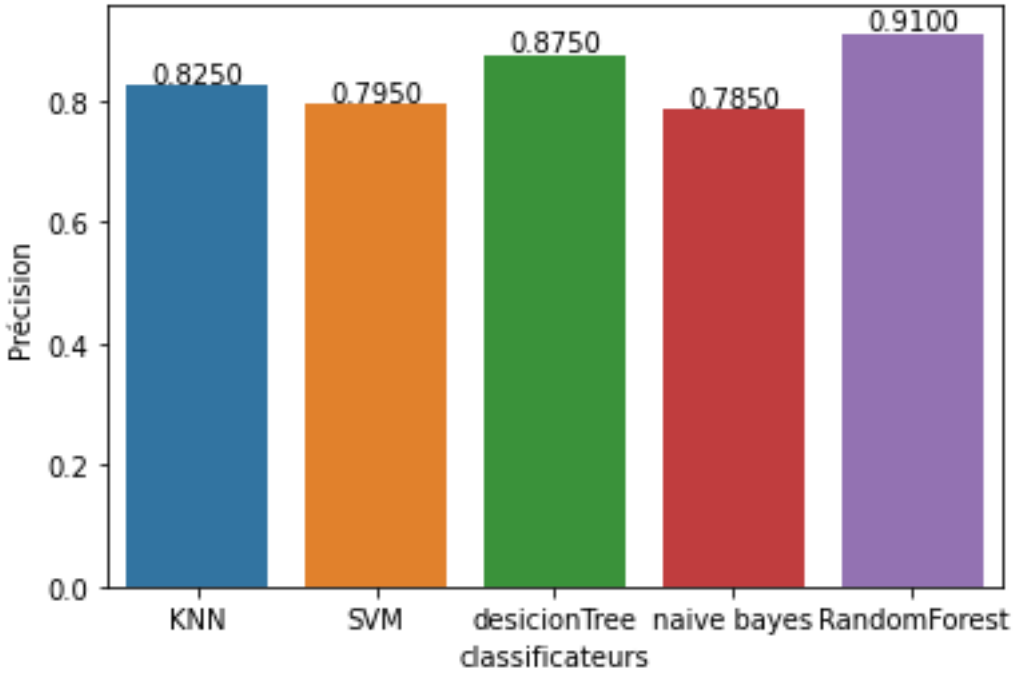


FIGURE 3.10 – La précision des modèles

D'après la figure on constate que les modèles Random Forest et l'arbre de décision ont obtenu les meilleures résultats que les autres.

2. Validation croisé

Validation croisé ou Cross validation en anglais cette méthode consiste à diviser l'ensemble de données en k sous-ensembles (ou plis) différents puis il l'utilise l'union de k-1 sous ensemble pour l'entrainement et le dernier sous-ensemble pour le test. Le processus est répété pour chaque sous-ensemble et la précision moyenne des tests est la précision de test.

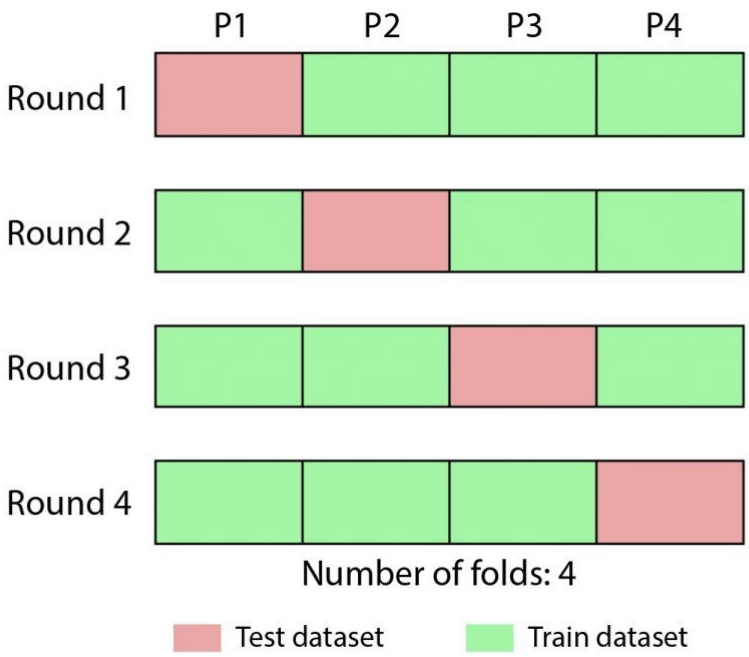


FIGURE 3.11 – Processus de validation croisé en 4 itérations [44]

La méthode validation croisé est importé d'après la bibliothèque « sklearn » avec « n_split=10 » qui indique le nombre de fois l'ensemble de données est subdivisé en sous-ensembles. La précision est calculée avec « cross_val_score » pour chaque itération.
Le code standard pour fractionner les données par la méthode validation croisé est dans la figure suivant :

```
from sklearn.model_selection import KFold
kfold = KFold(n_splits=10, random_state=10)
```

FIGURE 3.12 – Subdivision des données en k-Folds

Avec la bibliothèque « matplotlib » on peut tracer des graphes qui représentent le taux de précisions des modèles en fonction de nombres des itérations de la méthode « validation croisé » (voir les graphes suivant) :

1.Arbre de décision (Decision Tree)

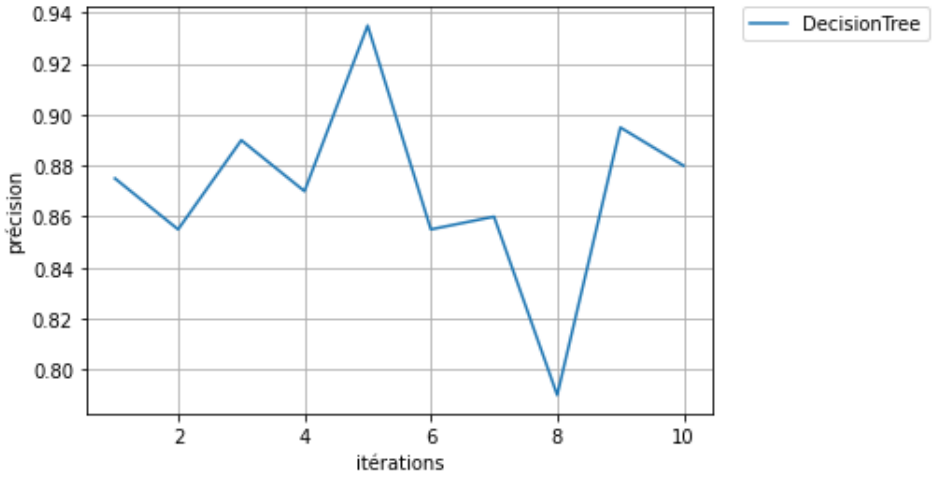


FIGURE 3.13 – Représentation graphique de taux du précision de « Arbre de décision »

Le graphe représente le taux de précision pour le modèle « Arbre de décision » en fonction de nombres d'itérations. On illustre que la meilleure itération est l'itération 5 avec un taux de précision égale à 93%.

2.Random Forest (forêt aléatoire)

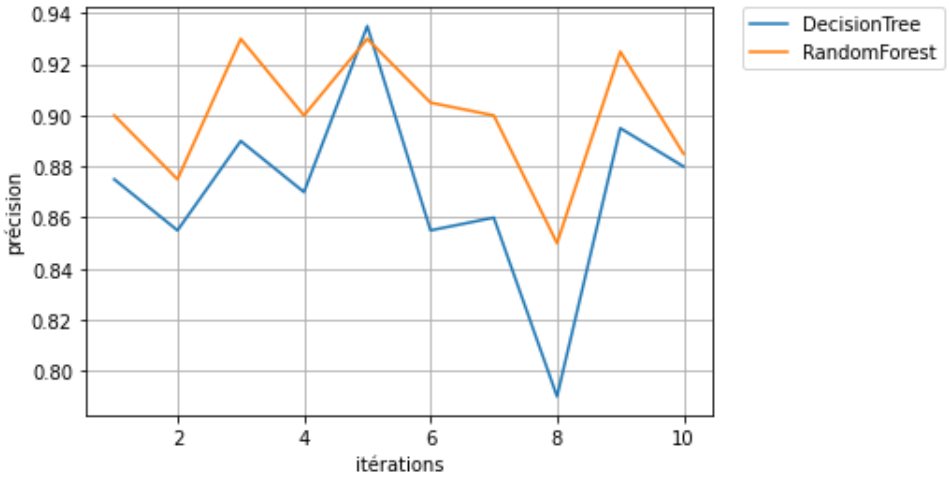


FIGURE 3.14 – Représentation graphique de taux du précision « Random forest »

Le graphe ci-dessus représente une comparaison entre le modèle « Arbre de décision » et le modèle « Random Forest » par le taux de précision en fonction de nombres des itérations. On illustre que la meilleure itération pour le modèle Random forest est l'itération 3 et 5 avec un taux de précision de 93%.
Le modèle Random forest est meilleur que le modèle Arbre de décision dans toutes les itérations sauf l'itération 5 avec une différence de taux de précision égal à 0.5%.

3.Naïve bayésienne (Gaussian Naïve Bayes)

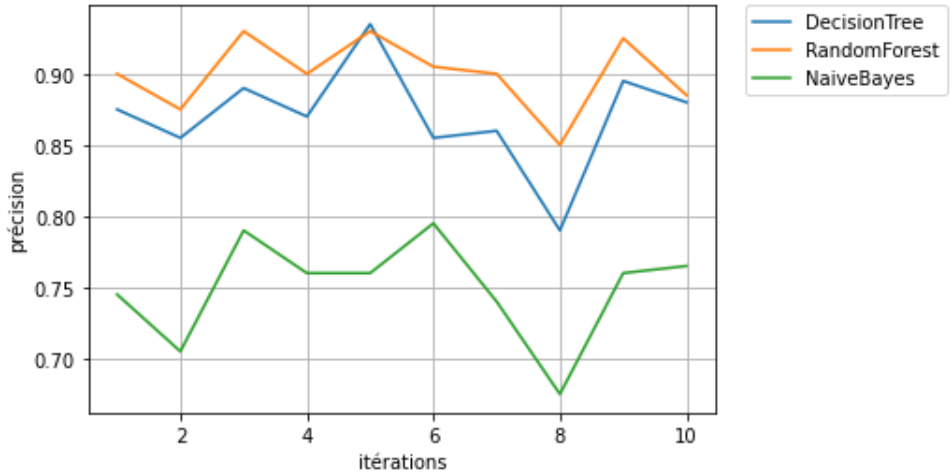


FIGURE 3.15 – Représentation graphique de taux du précision «Naïve bayésienne»

Le graphe représente une comparaison entre trois modèles qui sont «Arbre de décision», « Random forest », « Naïve Bayes » par le taux de précision en fonction de nombres des itérations.
On illustre que la meilleure itération de modèle Naïve Bayes est l'itération 6 avec un taux de précision égale à 79.5%, et le taux de précision pour les modèles Arbre de décision et Random forest dans toute les itérations est mieux que Naïve Bayes.

4. KNN (K-Nearest Neighbors)

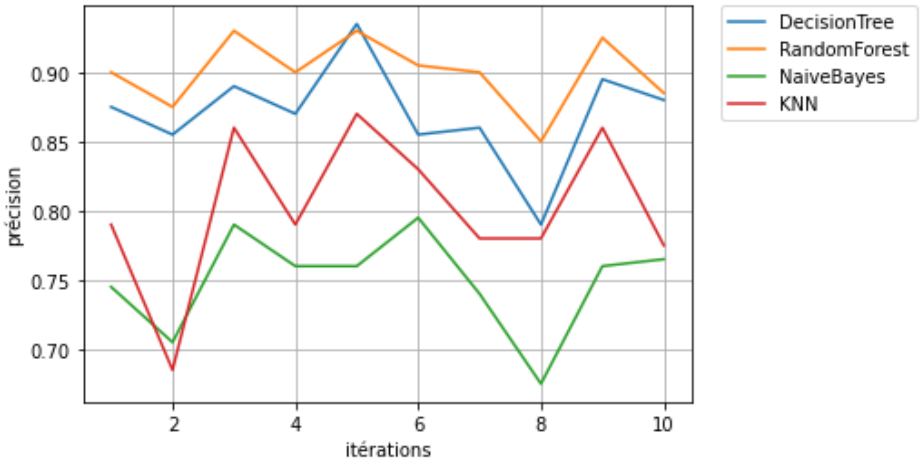


FIGURE 3.16 – Représentation graphique de taux du précision «KNN»

Le graphe représente une comparaison entre quatre modèles qui sont « Arbre de décision », « Random Forest », « Naïve Bayes » et « KNN » par le taux de précision en fonction de nombres d'itérations. On illustre que le modèle KNN est mieux que le modèle Naïve Bayes sauf dans l'itération 2, et son meilleure itération est l'itération 5 avec une taux de précision égal à 87%. Les modèles Arbre de décision et Random forest sont mieux que les modèle KNN et Naïve Bayes dans toutes les itérations.

5.SVM (Support vector machine)

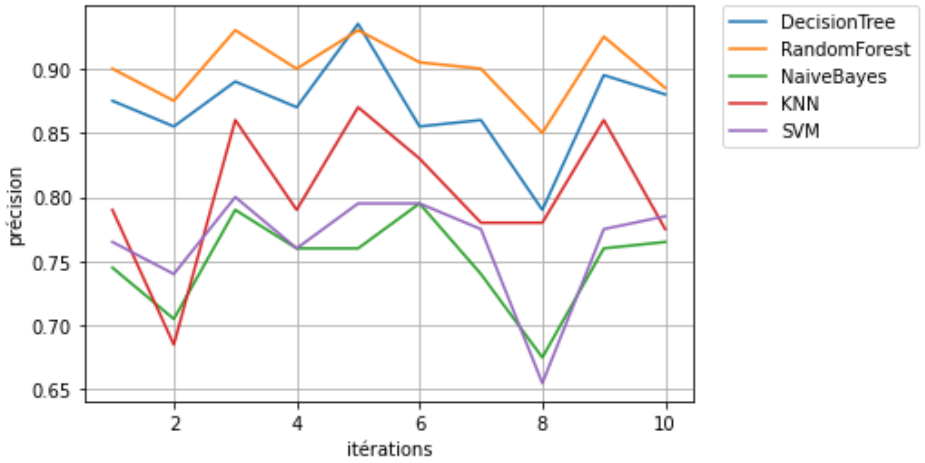


FIGURE 3.17 – Représentation graphique de taux du précision «SVM»

Le graphe ci-dessus représente un comparaison entre les différente modèles qui sont « Arbre de décision », « Random forest », « Naïve Bayes », « KNN » et « SVM » par le taux de précision en fonction de nombres des itérations. On illustre que pour le modèle SVM sa meilleure itération est l'itération 3 avec un taux de précision égal à 80%, et c'est mieux que les modèles KNN et Naive bayes dans les itérations 2 et 10.
Le modèle Arbre de décision et Random forest reste les deux mieux que les trois modèles qui sont : naive bayes, KNN et SVM.

Évaluations des modèles

Pour comparer les différente modèles et évaluer ces performances on utilise 03 mesures qui sont : La précision, recall_score, F1_score

1. Précision : capacité d'un modèle de classification à ne renvoyer que les instances pertinentes, défini comme le nombre de vrais positifs divisés par le nombre de vrais positifs plus le nombre de faux positives.

$$Precision = \frac{true\_positives}{true\_positives + false\_positives}$$

   true positives (TP), false positives (FP)
   true negatives (TN), false negatives (FN)

2. recall_score : (rappel ou sensibilité) c'est la capacité d'un modèle de classification à identifier toutes les instances pertinentes, défini comme le nombre de vrais positifs divisé par le nombre de vrais positifs plus le nombre de faux négatifs.

$$recall = \frac{true\_positives}{true\_positives + false\_negatives}$$

3. F1_score : métrique unique qui combine le rappel et la précision en utilisant la moyenne harmonique, en tenant compte des deux métriques dans l'équation suivante. [45]

$$F1 = 2 * \frac{Precision * recall}{Precision + recall}$$

	précision	recall_score	F1_score
Decision Tree	0.87	0.77	0.80
RandomForest	0.91	0.83	0.86
Naïve Bayes	0.78	0.64	0.67
KNN	0.82	0.75	0.74
SVM	0.80	0.63	0.68

TABLE 3.4 – Les résultats des attributs d'évaluations pour les différents modèles

D'après le tableau ci-dessus le modèle Random forest obtenu la meilleure précision qui égal à 91% et le meilleure score de rappel égal à 0.83 c'est-à-dire que sur toutes les patients diabétiques 83% d'entre eux sont correctement classé à l'aide de mesure de diagnostiques médicales.
Nous sélectionnons le modèle Random forest comme le modèle le plus optimale et qui fonctionne mieux pour notre ensemble de données en raison de sa grande précision et score de rappel.

3.5 Réglage de paramètre de modèle

La bibliothèque « sklearn » offre au modèle d'apprentissage automatique des paramètres par défaut, sensible qui donne parfois des scores de précision décents avec la possibilité de modification du ces paramètres pour évaluer la précision de ces modèles.
D'après la phase de sélection du modèle nous avons sélectionné le modèle Random Forest comme le modèle optimal pour notre ensemble de données et afin d'augmenter sa précision par le réglage de ces paramètres. Au lieu de rechercher manuellement les paramètres optimaux pour notre modèle nous avons opté d'utilisé la fonction « GridSearchcv » de la bibliothèque « sklearn » qui permet un recherche exhaustive sur les valeurs de paramètres spécifique pour un modèle.

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=8, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=200,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

FIGURE 3.18 – Meilleure paramètres pour le modèle Random forest

D'après le réglage de paramètres nous avons observé que la précision de modèle est augmente(voir figure suivant)

```
la précision actuel de modèle est: 93.75 %
la précision précédente de modèle est : 91 %
```

FIGURE 3.19 – Amélioration de précisions du modèle Random forest

3.6 Sauvegarde de modèle

L'enregistrement du modèle finalisé fait gagne beaucoup de temps car on a pas besoin d'entrainer le modèle à chaque exécution de l'application. La bibliothèque « pickle » permet de sérialisé les modèles d'apprentissage automatique et enregistrer le format sérialisé dans un fichier pour effectuer avec la prédiction, avec la fonction dump() qui permet de stocker les données de modèle en format binaire dans un fichier et la fonction load() qui permet de récupérer le modèle qui été enregistré dans le fichier binaire pour l'utilisé dans l'application.

3.7 Application

Ci-dessous, nous fournissons nos interfaces d'application "Daisha prediction" dans le but de permettre aux personnes des savoir s'ils ont le risque de développer un diabète avec un taux de prédiction bien défini.L'application est disponible en ligne sur :

https ://diasha-prediction.herokuapp.com/

FIGURE 3.20 – Le logo d'application Web

Explication du logo

1. Feuille : une forme approximative du pancréas.
2. 03 couleurs :

   Jaune pour les enfants

   Rose pour les Femelles

   Bleu pour les Mâles
3. Le mot prédiction se trouve en dessous du mot daisha signifie que moins de Mâles sont diabétiques que de Femelles.

La page d'accueil

La page d'accueil fournit des informations explicatives générales pour les diabétiques, où nous expliquons leur hypersensibilité à l'épidémie que nous connaissons actuellement. Cette page fournit des hyperliens vers les autres interfaces que constitue notre application Web.Voici quelques interfaces que constitue notre application Web :

1. Prediction
2. Doctor Map
3. About diabetes
4. Help

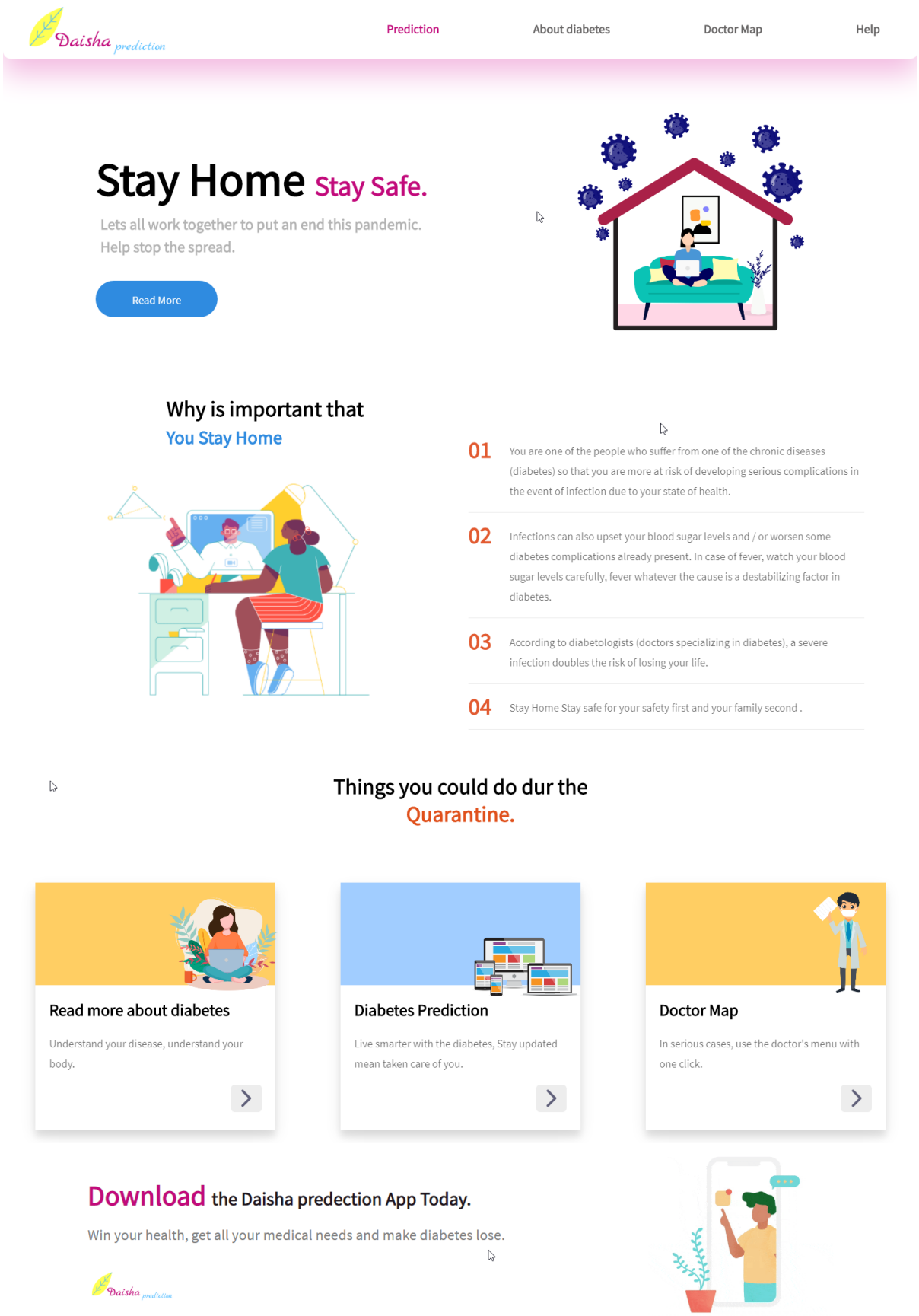


FIGURE 3.21 – La page d'accueil de l'application

L'interface de prédiction

Le but de cette interface est de prédire si une personne est diabétique ou non avec un taux de prédiction, pour cela il doit remplir le formulaire ci-dessous qui contient les informations suivantes : Pregnancies, Glucose, BloodPressure, SkinThikness, Insulin, BMI, DiabetePedigreeFunction et Age.

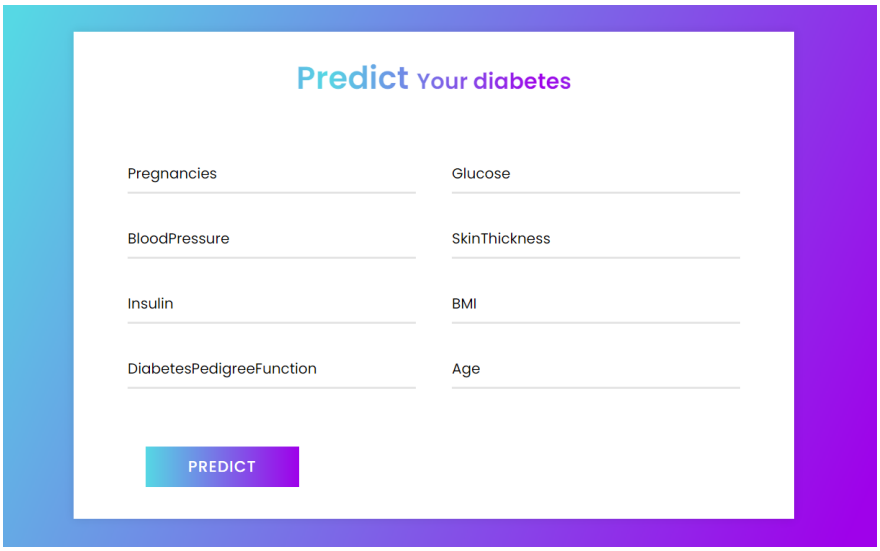


FIGURE 3.22 – L'interface « Prediction »

Le remplissage de tout les champs du formulaire ci-dessus est obligatoire, sinon un message d'erreur sera affiché (voir figure suivant)

FIGURE 3.23 – Message d'erreur

Une fois toute les informations sont remplit il doit cliquer sur le bouton « Predict » pour que les entrées seront récupérées et entraînées par le modèle Random Forest et le résultat de prédiction sera affichée :

1. Cas d'une prédiction non diabétique : Il affiche que l'utilisateur n'est pas diabétique

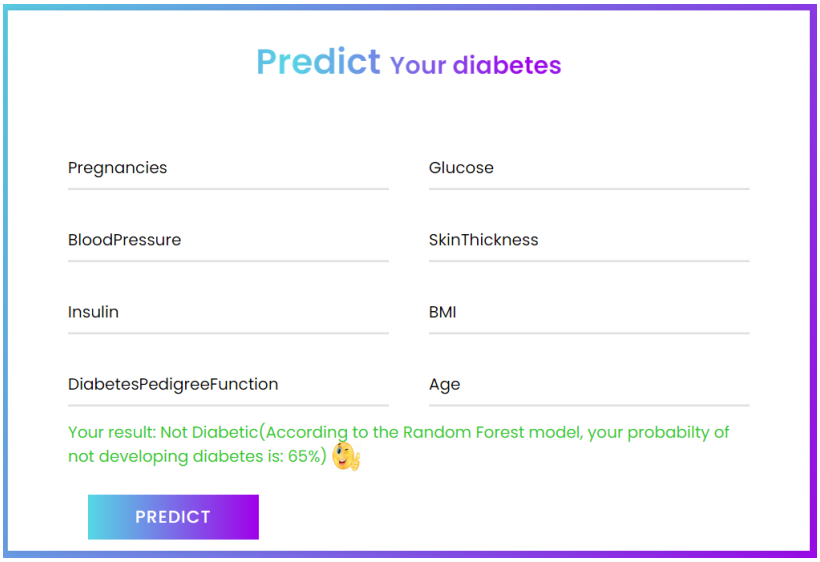


FIGURE 3.24 – Résultat de Prédiction non diabétique

2. Cas d'une prédiction diabétique : Il affiche que l'utilisateur est diabétique.

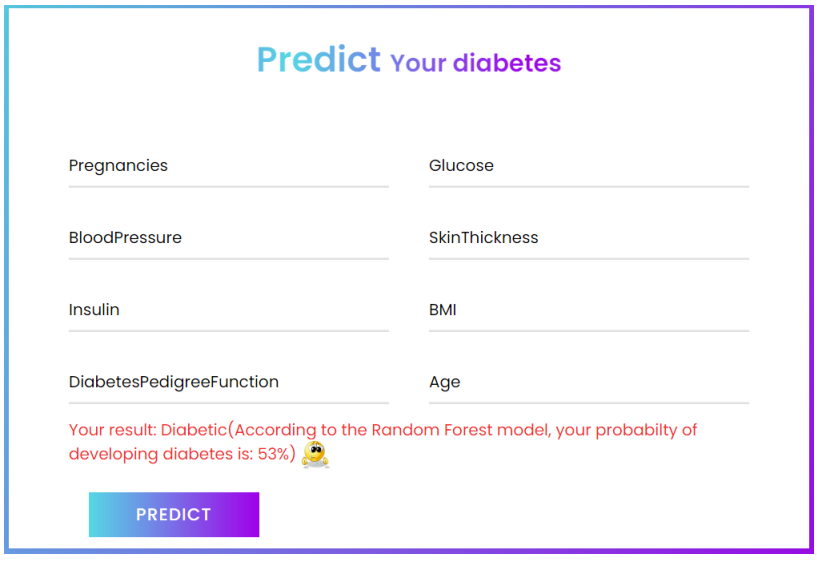


FIGURE 3.25 – Résultat de Prédiction diabétique

L'interface Doctor Map

Cette interface permet d'afficher des informations qui concernent des diabétologues comme : le numéro de téléphone, site web, temps d'entrée et sortie, localisation, sur Google Map pour que l'utilisateur peut l'appelé et prendre des rendez-vous.

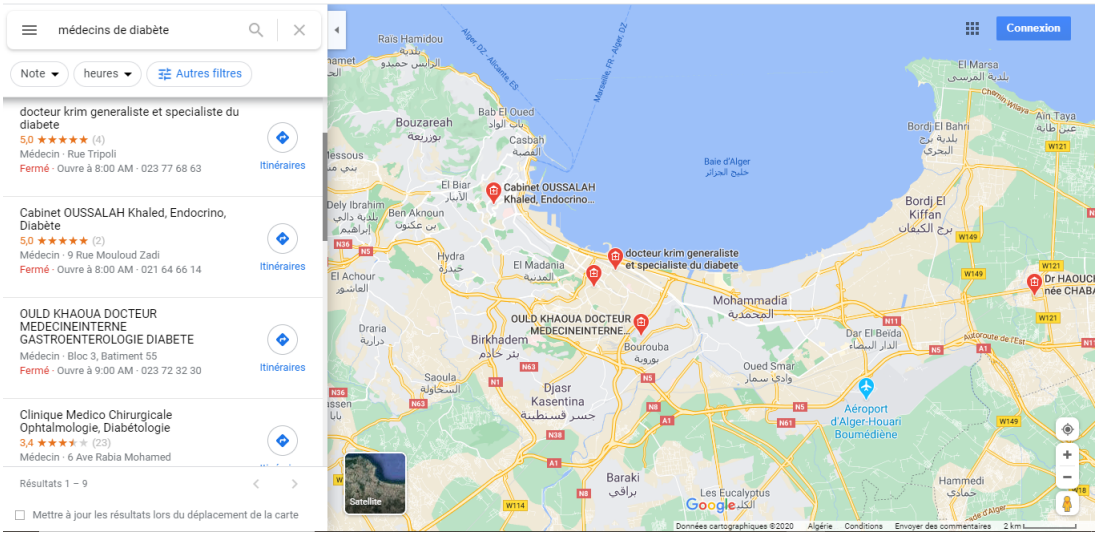


FIGURE 3.26 – L'interface « Doctor Map »

About diabetes

Cette interface permet d'accéder au "chapitre 01" de notre mémoire qui fournit des informations général sur le diabète.

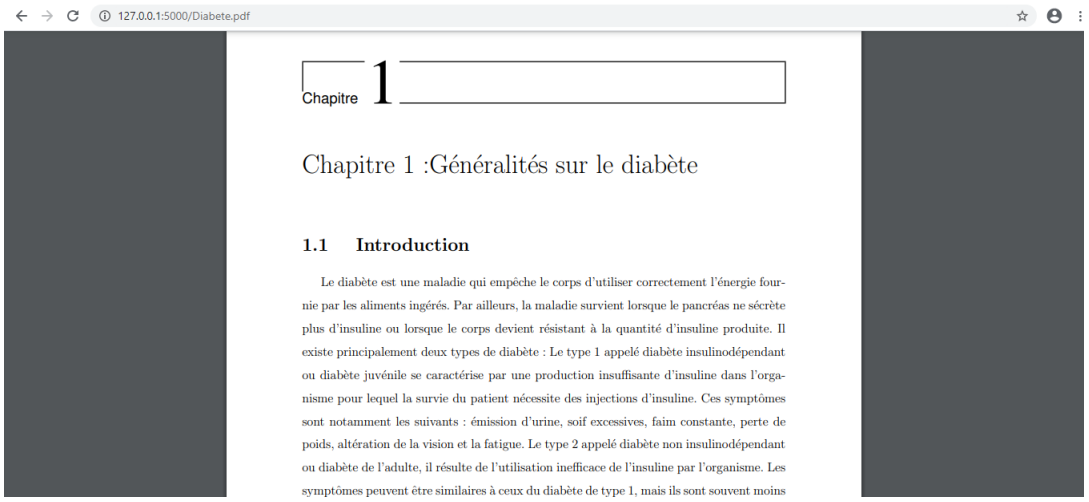


FIGURE 3.27 – L'interface «Généralités sur le diabète »

Help

Cette interface aide les utilisateurs à utiliser notre application web "Daisha Prediction" qui fournit des informations sur les différents hyperliens vers les autres interfaces ainsi que notre contact.

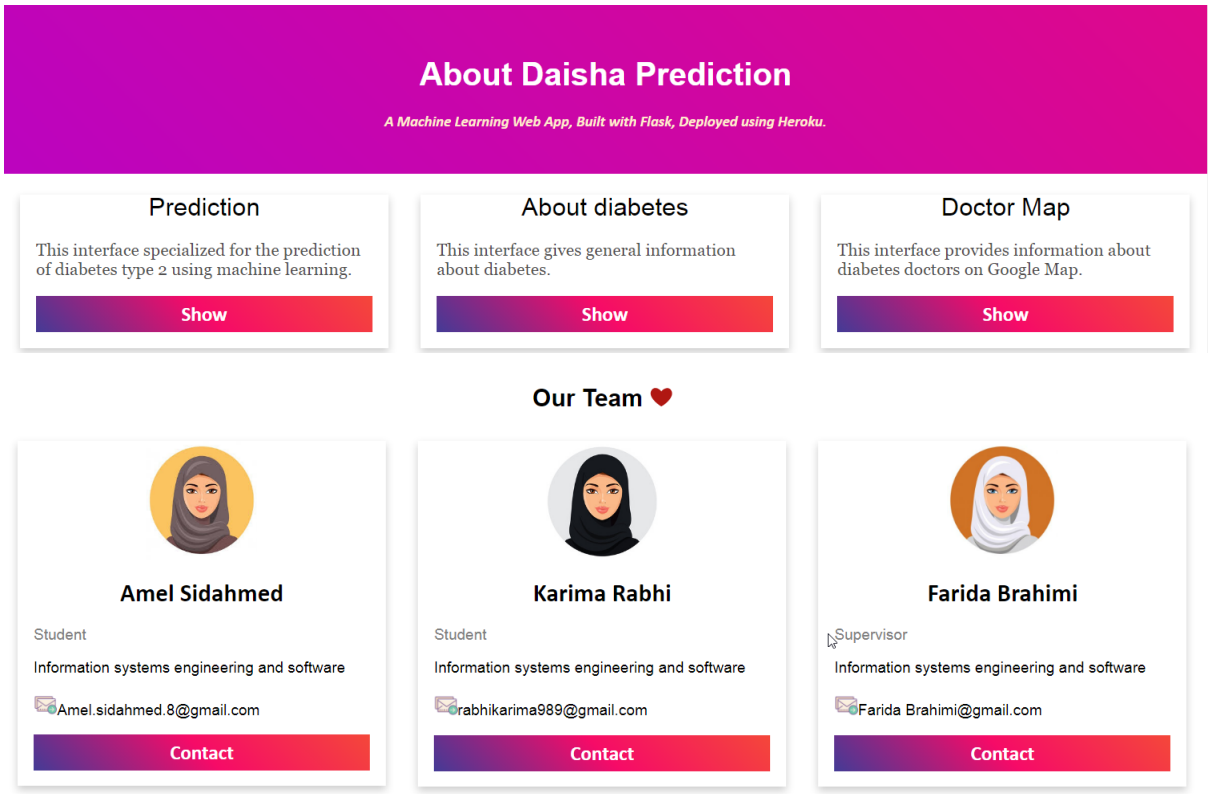


FIGURE 3.28 – L'interface « Help »

3.8 Conclusion

Dans ce chapitre, nous avons présenté les différents étapes de prétraitement des données tels que l'exploration et la visualisation des données ainsi le nettoyage des valeurs aberrantes. L'application des méthodes d'évaluation nous permet de sélectionner le modèle Random forest comme le meilleur modèle qui a un taux de précision élevé. Pour augmenter la performance de notre modèle on a réglé ses hyper paramètres comme « criterion » et « n_estimators »... .etc.

A la fin, on a développé une application web qui nous permet de prédire si une personne donnée est diabétique ou pas à partir de ces informations médicales.

Conclusion générale et perspectives

Le diabète est considéré comme l'une des maladies les plus mortelles et chronique qui provoque une augmentation de la glycémie, Les gens ont commencé à rendre compte que cette maladie chronique a profondément affecté chaque famille et la vie quotidienne de chacun. Aujourd'hui le diabète est une pandémie mondiale qui touche plusieurs millions de personnes et ce augmente chaque année, ici en Algérie la prévalence estimée à 14.4% d'après SANOFI qui une partenaire de santé des patients algérienne dans sa compagne « Diabetes Your Type*». DE nombreuses complications surviennent si le diabète demeure non traité et non identifié, et le processus d'indentification fastidieux entraine la visite d'un patient à un centre de diagnostic mais l'approche d'apprentissage automatique résoudre ce problème critique dans le but de cette étude pour construire un modèle capable de prédire si les personnes sont diabétiques ou non avec l'utilisation des algorithmes d'apprentissage automatique.

Dans cette Mémoire nous avons mené à faire une comparaison entre cinq algorithmes d'apprentissage automatique à savoir : l'arbre de décision, Random forest, naive bayes, K nearest neighbors et support vector machine, les résultats expérimentaux obtenu pour l'ensemble de données de l'hôpital de Frankfurt Allemagne montre que Random forest est meilleure que les autres algorithmes en terme de sa grande précision dans les deux méthodes d'évaluation et sa grande score pour les attributs d'évaluation. Sur la base d'un l'algorithme Random forest et que nous avons besoin d'un moyen pour rendre le modèle applicatif pour tout le monde nous avons développé une solution basé sur une application web dans le but d'aide les personnes de prédire s'il souffre de diabète Type 2.

En termes de perspective :

1. la construction d'une application Android parallèle avec notre application web permet d'aide les personnes qui sont diabétiques de suivre son situation médical, les médicaments, les rendez-vous médical ainsi leur état physique comme le sport et le régime alimentaire équilibré pour leur cas.
2. La prédiction de diabète avec l'approche de deep learning.

Bibliographie

[1] Organisation mondiale de la santé.(2016).Rapport mondial sur le diabète.88p

[2] Medtronic.Le Diabète En Quelque Mots.[en ligne].Disponible sur : https ://www.parlonsdiabete.com/parlons-diabete/le-diabete-en-quelques-mots

[3] fédération français de Cardiologie.Réduire le risque cardio-vasculaire LE DIABÈTE.[en ligne]. Disponible sur : https ://www.fedecardio.org/Je-m-informe/Reduire-le-risque-cardio-vasculaire/le-diabete

[4] CEED :Centre européen d'étude du Diabète.Diabètes et complications.[en ligne].Disponible sur : http ://ceed-diabete.org/fr/le-diabete/diabete-et-complications/

[5] La macroangiopathie diabétique.Complications.[en ligne].(Mis à jour en juin 2015).Disponible sur :http ://www.hegp.fr/diabeto/complicationmacro.html

[6] Doctissimo.Micro-angiopathie.[en ligne].Disponible sur : https ://www.doctissimo.fr/sante/dictionnaire-medical/micro-angiopathie

[7] TopSanté.Maladies chroniques.[en ligne].Disponible sur : https ://www.topsante.com/medecine/maladies-chroniques/diabete/comment-savoir-si-je-suis-diabetique-609768 (consulté 14/03/2020)

[8] santé magazine.Maladies. [en ligne].Disponible sur : https ://www.santemagazine.fr/sante/maladies/maladies-endocriniennes-et-metaboliques/diabete/comment-depiste-t-on-un-diabete-334903(consulter 13/05/2020)

[9] Medtronic.Diagnostic Du Diabète .[en ligne].Disponible sur : https ://www.parlonsdiabete.com/parlons-diabete/le-diagnostic#les-differents-tests-a-effectuer-en-cas-de-doute (consulté 14/03/2020)

[10] CEED :Centre européen d'étude du Diabète.le dépistage au cœur des actions de prévention.[en ligne].Disponible sur :http ://ceed-diabete.org/blog/diabete-le-depistage-au-coeur-des-actions-de-prevention/

[11] Orkyn.Qu'est-ce que le diabète.[en ligne].Disponible sur : https ://www.orkyn.fr/mon-traitement-suivi-domicile-diabete/quest-ce-que-diabete (consulté 14/03/2020)

[12] 123rf.Banque d'images - 100 unités Seringue d'insuline pour le diabète sur fond blanc .[en ligne].Disponible sur : https ://fr.123rf.com/photo_76260135_100-unit%C3%A9s-seringue-d-insuline-pour-le-diab%C3%A8te-sur-fond-blanc.html

[13] Xpnworld.La sensibilité à l'insuline c'est quoi ?.[en ligne].Disponible sur : https ://xpnworld.com/sensibilie-insuline/

[14] fédération Français des diabetiques.MA GLYCÉMIE. [en ligne].Disponible sur : https ://www.federationdesdiabetiques.org/diabete/glycemie (consulté 18/05/2020)

[15] Doctissimo.Facteurs de risque du diabète.[en ligne].Disponible sur : https ://www.doctissimo.fr/html/dossiers/diabete/articles/3617-diabete-vie-saine.htm

[16] https ://www.inprincipio.xyz/machine-learning/ (consulter 03/06/2020)

[17] Lev Kiwi.(2018).Apprentissage et Machine Learning.[en ligne].Disponible sur : https ://levkiwi.ch/apprentissage-et-machine-learning/ (consulter 04/06/2020)

[18] Pensée Artificielle.Machine Learning pour débutant : Introduction au Machine Learning.[en ligne].Disponible sur :http ://penseeartificielle.fr/introduction-au-machine-learning/ (consulter 04/06/2020)

[19] Gaël, P.Makina Corpus.(2017).Initiation au Machine Learning avec Python.[en ligne].Disponible sur : https ://makina-corpus.com/blog/metier/2017/initiation-au-machine-learning-avec-python-pratique

[20] Ilemona S.Atawodi.(2019).A Machine Learning Approach to Network Intrusion Detection System Using K Nearest Neighbor and Random Forest.Thèse

de master : université de Southern Mississippi .52p.[en ligne].Disponible sur :https ://aquila.usm.edu/cgi/viewcontent.cgi ?article=1707&context=masters_theses

[21] Navlani, A.DataCamp.(2018).KNN Classification using Scikit-learn.[en ligne].Disponible sur :https ://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

[22] Benzaki, Y.Mr Mint .(2018).Introduction à l'algorithme K Nearest Neighbors (K-NN).[en ligne].Disponible sur : https ://mrmint.fr/introduction-k-nearest-neighbors

[23] Gupta, P.to wards datascience.(2017).Decision Trees in Machine Learning.[en ligne].Disponible sur :https ://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

[24] Choudhury, A.Analytics India Magazine.(2019).Beginner's Guide To Decision Trees : Why Are They Crucial For Data Science Applications.[en ligne].Disponible sur : https ://analyticsindiamag.com/beginners-guide-to-decision-trees-why-are-they-crucial-for-data-science-applications/

[25] Shrivastav, A.to wards datascience.Almost Everything You Need To Know About Decision Trees (With Code). [en ligne].Disponible sur : https ://towardsdatascience.com/almost-everything-you-need-to-know-about-decision-trees-with-code-dc026172a284

[26] Ismaili, Z.(2019).Le Data Scientist . Apprentissage Supervisé Vs. Non Supervisé.[en ligne].Disponible sur : https ://le-datascientist.fr/apprentissage-supervise-vs-non-supervise

[27] Harrison, O.to wards datascience.(2018).Machine Learning Basics with the K-Nearest Neighbors Algorithm.[en ligne].Disponible sur : https ://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[28] java T point.Random Forest Algorithm.[en ligne].Disponible sur : https ://www.javatpoint.com/machine-learning-random-forest-algorithm

[29] Chakure, A.to wards datascience.(2019).Random Forest Classification.[en ligne].Disponible sur : https ://towardsdatascience.com/random-forest-classification-and-its-implementation-d5d840dbead0

[30] tutorials point.Classification Algorithms - Random Forest.[en ligne].Disponible sur : https ://www.tutorialspoint.com/machine_learning_with_python /machine_learning_with_python_classification_algorithms_random_forest.htm

[31] java T point.Support Vector Machine Algorithm.[en ligne].Disponible sur : https ://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[32] Tandel, A.to wards data science.(2017).Support Vector Machines — A Brief Overview.[en ligne].Disponible sur :https ://towardsdatascience.com/support-vector-machines-a-brief-overview-37e018ae310f

[33] Sharma, N.heartbeat.Understanding the Mathematics behind Support Vector Machines.[en ligne].Disponible sur : https ://heartbeat.fritz.ai/understanding-the-mathematics-behind-support-vector-machines-5e20243d64d5

[34] Gandhi, R.to wards data science.(2018).Support Vector Machine — Introduction to Machine Learning Algorithms.[en ligne].Disponible sur : https ://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca 47

[35] Raymond, C.(2005).Décodage conceptuel : co-articulation des processus de transcription et compréhension dans les systèmes de dialogue.Thèse doctorat :SPÉCIALITÉ : Informatique.Université d'Avignon et des Pays de Vaucluse.144p [en ligne].Disponible sur : https ://www.irisa.fr/texmex/people/raymond/These/these_v1.0ch3.html

[36] laptrinhx.(2019).Naive Bayes Unlocked.[en ligne].Disponible sur : https ://laptrinhx.com/naive-bayes-unlocked-1301819179/

[37] Gupta, P.to wards data science.(2017).Naive Bayes in Machine Learning.[en ligne].Disponible sur : https ://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4

[38] Sarwar, MA.Kamal N.Hamid W .Ali Shah M.(2018). Prediction of Diabetes Using Machine Learning Algorithms in Healthcare.*Keywords-Big data analytics ; Predictive Analytics ; Machine Learning ; Healthcare.*6p

[39] Liyanapathirana, L.to wards data science.(2018).Machine Learning Workflow on Diabetes Data : Part 01.[en ligne].Disponible sur :

https ://towardsdatascience.com/machine-learning-workflow-on-diabetes-data-part-01-573864fcc6b8

[40] MEHIDI, D.MEDJOUDJ, S. (2018).Application des Méthodes d'Apprentissage dans la Prédiction du Diabète de Type 2. mémoire master :Intelligence Artificielle.Département Informatique. Faculté des Sciences Exactes.Université A.MIRA de Bejaia.79p

[41] Ebrahim, M.like geeks.(2020). Python Correlation Matrix Tutoriel.(mise a jour 29-07-2020).[en ligne].Disponible sur : https ://likegeeks.com/python-correlation-matrix/

[42] Jupyter.(2017).Project Jupyter .[en ligne].Disponible sur :https ://jupyter.org/

[43] ResearchGate.Train-Test Data Split .[en ligne].Disponible sur :https ://www.researchgate.net/figure/Train-Test-Data-Split_fig6_325870973

[44] Shah, I.QuantInsti.(2019).Cross Validation In Machine Learning Trading Models.[en ligne].Disponible sur : https ://blog.quantinsti.com/cross-validation-machine-learning-trading-models/

[45] Koehrsen, W.to wards data science.(2018).Beyond Accuracy : Precision and Recall.[en ligne].Disponible sur : https ://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

[46] Dyouri, A.Community.(2020).Comment créer une application web en utilisant Flask en Python 3.[en ligne]. Disponible sur : https ://www.digitalocean.com/community/tutorials/how-to-make-a-web-application-using-flask-in-python-3-fr

Annexe A

Annexe

hémoglobine glyquée (HbA1c)

Ce test représente une "mémoire" globale de la glycémie sur les trois derniers mois, il prend donc en compte tous les états, y compris la glycémie postprandiale.

autosurveillance glycémique (ASG)

autosurveillance glycémique est prescrite par le médecin en fonction du votre type de diabète et de votre type de traitement.elle est indispensable dans le diabète de type 1, nécessaire dans le diabète de type 2 insulinotraité et variable pour les diabétiques de type 2 non insulinotraités.cette autosurveillance sert principalement à contrôler et prévenir les déséquilibres (hypo/hyperglycémies) et à adapter votre traitement. Elle permet aussi de mesurer l'effet d'un aliment, d'une pratique sportive ou d'une activité physique sur sa glycémie.

Gain d'information

C'est une mesure de sélection utilisé pour mesurer la pureté d'un attribut, et définir le degré de désorganisation dans un système appelé l'entropie. Si l'échantillon est complètement homogène, l'entropie est nulle et si l'échantillon est également divisé (50% - 50%), il a une entropie de un. L'entropie peut être calculée à l'aide de la formule suivant :

$$Entropie(S) = -p_+ log_2(p_+) - q_- log_2(q_-)$$

Où S est l'ensemble d'échantillon et p et q respectivement la probabilité de succès et d'échec dans l'ensemble d'échantillon.

Indice de Gini

L'indice de Gini est une mesure d'impureté ou de pureté utilisée lors de la création d'un arbre de décision dans l'algorithme CART (arbre de classification et régression). L'indice de Gini peut être calculé à l'aide de la formule suivant :

$$Gini(S) = 1 - \Sigma_{j=1}^{n} p_j^2$$

Où S est l'ensemble d'échantillon et n le nombre de classe à prédire et Pj est la fréquence de la classe j dans l'ensemble d'échantillon.