Université Ubn Zohr
Faculté Polydisciplinaire Ouarzazate
Filière Sciences Mathématiques et Informatique

Numéro d'ordre :.../SMI-2024

Mémoire de projet tutoré :

Modélisation prédictive des maladies cardiaques et le diabète à l'aide de l'apprentissage automatique et de science de donnée.

Projet tutoré présenté par :
HANANE IQLI
HAFIDA RAMI
AMINA AZIZE
LATIFA AIT SI ARABI

Sous la direction de :

Prof. **ABDELBASSET BOUKDIR**

Soutenu le :15/05/2024

Devant le jury :
Prof. BOUKDIR ABDELBASSET (Encadrant) : Professeur à la Faculté Polydisciplina
Prof. (Examinateur) : Professeur à la Faculté Polydisciplina

Année Universitaire : 2023-2024

Table des matières

Table des figures

Remerciement

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux qui nous a donné la force et la patience d'accomplir ce modeste travail. En second lieu, nous tenons à remercier notre encadrant professeur Abdelbasset Boukdir, pour ses précieux conseils et son aide durant toute la période d'élaboration de ce mémoire. Nos vifs remerciements sont également adressés aux membres du jury ** ***pour l'intérêt qu'ils ont porté à notre travail ainsi que l'acceptation de le juger et l'enrichir par leurs propositions. Enfin, ça nous fait plaisir de délivrer un bouquet de remerciement pour toutes les personnes qui ont participé de près ou de loin à la réalisation et à la réussite de ce travail.

Introduction

Le diabète touche des millions de personnes à travers le monde,c'est une maladie chronique qui se definie par un taux élvé de sucre dans le sang, Cela peut augmenter le risque d'hypertension artérielle ,entrainent plusieurs maladies ,Dont la maladie cardiaque,Ce dernier se réfère à un ensemble de maladies qui affectent le fonctionnement du coeur,ces maladies pouvent toucher le muscle cardiaque ,ses valves,sa membrane environnante,ainsi queles artères et les veines principales de coeur.

Pour éviter que les patients ne subissent d'autres dommages à cause des maladies cardiaques et du diabète ,il est essentiel de faire La pré-détection précoce qui revêt une importance capitale se manifeste en :

Prévention des complications graves : Initier rapidement des mesures de prévention et de gestion pour réduire le risque de complications graves telles que les crises cardiaques, les accidents vasculaires cérébraux, les lésions rénales et la neuropathie.

Amélioration des résultats de santé : Une détection précoce permet une intervention médicale et des changements de mode de vie appropriés, ce qui peut améliorer considérablement les résultats de santé et la qualité de vie des individus affectés.

Réduction des coûts de santé : Identifier les maladies à un stade précoce peut contribuer à réduire les coûts de traitement à long terme, en évitant les complications graves et en réduisant la nécessité de soins médicaux intensifs.

Promotion du bien-être général : En favorisant la prévention et la gestion précoce des maladies cardiaques et du diabète, on peut encourager une meilleure santé globale et un bien-être accru au sein de la population.

Prédire le diabète cardiaques maladies reste une tache compliquée,voire impossible,pour les médecins et les professionnels de la santé, face à un volume important de données médicales,il est difficile pour les spécialistes de comprendre et de prédire les aspects complexes de ces deux maladies. Dans cette optique, le recours à des techniques avancées offre de nouvelles possibilités pour mieux anticiper et prendre en charge ces deux maladies,et nous ne trouverons rien de mieux que les technologies d'intelligence artificielle, en particulier l'apprentissage automatique,qui est un moyen avancé de traiter facilement les données massives, de les reconnaitre et de prédire rapidement et avec précision les résultats.

Etant donné ce qui précède,notre sujet vise à utiliser l'apprentissage automatique pour détecter le diabète et les maladies cardiaques en utilisant les données des patients à travers la maintenance d'une application qui effectue cette tache,où nous avons base la construction de ce sujet sur trois axes ; le premier chapitre présente une introuduction générale sur les maladies cardiaques et le diabète se qui conserne les types, les facteurs de risque, les symptomes et aussi les analyses nécessaires.

Dans le deuxième chapitre nous avons détaillé l'analyse des données où nous sommes concentrés sur les variables et les caractéristiques avec leurs Description Statique et leurs visualisation pour chaque maladie , plus la construction du modèle précisément la Sélection des algorithmes avec la validation et performance du modèle.Le troisième chapitre tourne autour de l'analyse comparative des différents modèles et l'interprétation des résultats. finallement, une conclusion générale présentant un résumé des résultats et recommandations pour la pratique clinique.

chapitre 1

Maladies cardiaques et diabète

1.1 Aperçu des sur la des maladies cardiaques

Les maladies cardiaques :
une description générale Les maladies cardiaques regroupent un large éventail de troubles affectant le cœur et son système vasculaire. Elles constituent la première cause de mortalité dans le monde, frappant aussi bien les hommes que les femmes.

Types de maladies cardiaques :

- **Maladie coronarienne :** rétrécissement ou blocage des artères coronaires, réduisant le flux sanguin vers le muscle cardiaque. Cela peut entraîner une angine de poitrine, un infarctus du myocarde ou une insuffisance cardiaque.
- **Cardiomyopathie :** affaiblissement ou rigidification du muscle cardiaque, altérant sa capacité à pomper efficacement le sang.
- **Valvulopathie cardiaque :** dysfonctionnement des valves cardiaques, perturbant le flux sanguin à travers le cœur.
- **Arythmies cardiaques :** troubles du rythme cardiaque, tels que la fibrillation auriculaire ou la tachycardie ventriculaire.
- **Congénitales cardiaques :** malformations cardiaques présentes dès la naissance.

Facteurs de risque des maladies cardiaques :

- **Antécédents familiaux :** risque accru si des membres de la famille ont souffert de maladies cardiaques prématurément.
- **Âge :** risque accru avec l'âge.
- **Tabagisme :** endommage les artères et augmente le risque de caillots sanguins.
- **Hypertension artérielle :** force excessive exercée sur les parois artérielles.
- **Hypercholestérolémie :** accumulation de plaque dans les artères.
- **Diabète :** altère la capacité de l'organisme à utiliser l'insuline.
- **Obésité :** augmente le risque de facteurs de risque tels que l'hypertension artérielle et l'hypercholestérolémie.
- **Sédentarité :** manque d'activité physique régulière.
- **Régime alimentaire malsain :** riche en gras saturés, en cholestérol et en sodium, et pauvre en fruits, légumes et fibres.
- **Stress** : facteur de risque potentiel.

Symptômes des maladies cardiaques :

Les symptômes varient en fonction du type de maladie cardiaque, mais peuvent inclure :

FIGURE 1.1 – les symptômes des maladies cardiaques.

Diagnostique des maladies cardiaque :

- **Électrocardiogramme (ECG) :** Un ECG enregistre l'activité électrique de votre cœur.
- **Test d'effort :** Un test d'effort consiste à marcher ou à courir sur un tapis roulant ou à pédaler sur un vélo stationnaire pendant que votre fréquence cardiaque et votre tension artérielle sont surveillées.
- **Échocardiographie :** Une échocardiographie utilise des ultrasons pour créer une image de votre cœur en mouvement.
- **Angiographie coronarienne :** Une angiographie coronarienne est un type de radiographie qui utilise un colorant pour montrer les vaisseaux sanguins de votre cœur.

1.2 Aperçu des sur la des maladies diabètes

Le diabète :

une maladie chronique qui affecte la façon dont votre corps régule le sucre (glucose) dans le sang. Le glucose est la principale source d'énergie de votre corps et provient des aliments que vous mangez. L'insuline, une hormone produite par le pancréas, aide au transport du glucose dans les cellules pour être utilisé comme énergie.

Types de diabète :

Diabète de type 1 : Le corps ne produit pas d'insuline. Il s'agit d'une maladie auto-immune, ce

qui signifie que le système immunitaire attaque les cellules productrices d'insuline du pancréas. Le diabète de type 1 survient généralement chez les enfants et les jeunes adultes.
Diabète de type 2 : Le corps ne produit pas suffisamment d'insuline ou n'utilise pas l'insuline efficacement. C'est le type de diabète le plus courant et survient généralement chez les adultes de plus de 40 ans. Le diabète de type 2 est souvent associé à l'obésité et à un mode de vie sédentaire.

Symptômes du diabète :

Les symptômes du diabète peuvent inclure :

- Miction fréquente
- Soif excessive
- Faim excessive
- Perte de poids inexpliquée
- Fatigue
- Vision floue
- Lenteur de la cicatrisation des plaies
- Fourmillements ou engourdissements dans les mains ou les pieds

Complications du diabète :

Si le diabète n'est pas bien contrôlé, il peut entraîner de graves complications, notamment :

- Maladies cardiaques
- Accident vasculaire cérébral
- Maladie rénale
- Cécité
- Dommages nerveux
- Amputation des membres

Diagnostique du diabète :

- **Test de glycémie à jeun (glycémie AC) :** Ce test mesure votre glycémie après une période de jeûne d'au moins 8 heures.
- **Test de glycémie postprandiale (glycémie PP) :** Ce test mesure votre glycémie deux heures après avoir mangé un repas.
- **Test de glycémie aléatoire :** Ce test peut être effectué à n'importe quel moment de la journée, avec ou sans jeûne.
- **Test d'hémoglobine A1c (HbA1c) :** Ce test mesure votre glycémie moyenne sur les 2 à 3 mois précédents.

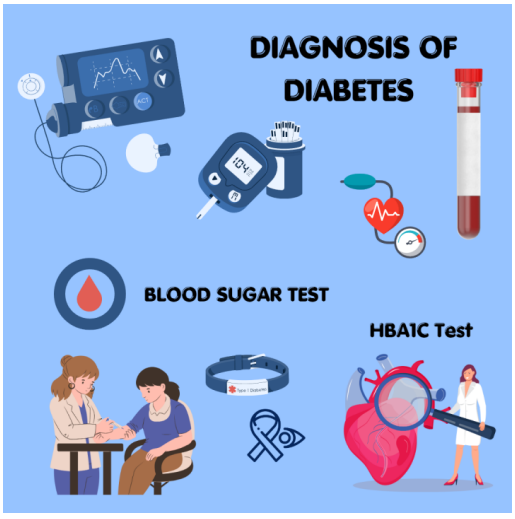


FIGURE 1.2 – Diagnostique du diabète

chapitre 2

Méthodologie

2.1 Architecture du systeme

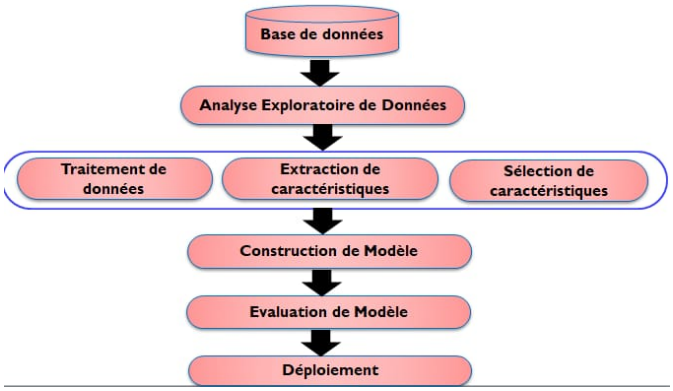


FIGURE 2.1 – processuce du systeme

2.2 Source des données

notre dépenddance à l'apprentissage automatique qui est un sous domaine de l'intélligence artificiel ,se base sur donner au machine la capacité d'apprendre à partir des données et s'améliore par l'expérience, nécessite de choisir une base de donnée avec une extension adaptée et nous n'avons pas trouvé mieux que **CSV** (Comma Separated Values) est un format de fichier courament utilisé pour stocker des dataset dont les données separées par des virgules . D'où on a extrait les données dans la base de données **Heart-Failure-dataframe** pour les maladies cardiaques et la base de données **Dibéte-dataframe** pour le diabéte. les données se divise en 2 parties , parties d'entraînement qui egale dans notre système à 75 % de la base de données et partie de teste qui egale à 25 % .

2.2.1 Analyse exploratoire des données

2.2.1.1 Variables et caractéristiques

Dans le but de faire le diagnostigue des maladies cardiaque et diabète , il est nécessaire de s'appuyer sur des informtions disponibles dans la base de données public de chaque maladie :

Pour les maladies cardiaques la base de donnée**Heart-Failure**-dataframe qui contient 12 attributs et 918 lignes et la figure 2.1 présente les varibles et les caractéristiques et leurs descriptions :

Attribut	Description
Age	Âge d'un patient [années]
Gender	Sexe du patient [M : Homme, F : Femme]
ChestPain	type de douleur thoracique [TA : angine typique, ATA : angine atypique, NAP : douleur non angineuse, ASY : asymptomatique]
RestingBP	Pression artérielle en Hg (Pression artérielle normale - 120/80 Hg)
Cholesterol	Taux de cholestérol sérique dans le sang (taux de cholestérol normal inférieur à 200 mg/dL pour les adultes)
FastingBS	Glycémie à jeun (normale inférieure à 100 mg/dL pour les non-diabétiques pour le diabète 100-125 mg/dL)
RestingECG	résultats de l'électrocardiogramme au repos [Normal : Normal, ST : présentant une anomalie de l'onde ST-T (inversions de l'onde T et/ou élévation ou dépression ST > 0,05 mV), HVG : montrant une hypertrophie ventriculaire gauche probable ou certaine selon les critères d'Estes]
MaxHR	fréquence cardiaque maximale atteinte [Valeur numérique entre 60 et 202]
ExerciseAngina	angine induite par l'exercice [Y : Oui, N : Non]
Oldpeak	oldpeak = ST [Valeur numérique mesurée en dépression]
ST-Slope	la pente du segment ST de l'exercice de pointe [Up : ascendant, Flat : plat, Down : descendant]
HeartDisease	classe de sortie [1 : maladie cardiaque, 0 : Normal]

TABLE 2.1 – les attributs de base données des maladies cardiaques

Et pour le diabète la base de donnée **Diabète**-dataframe contient 9 attributs et 768 lignes et la figure 2.2 présente les informaions correspondantes :

2.2.2 Descriptive Statique

2.2.2.1 Resume statistique de l'ensemble de données

Les deux figure ci-dessous présents un résumé général de toutes les données relatives aux **maladies cardiaques** et leurs description statique, où la figure 2.3 represente des donnée quantitatives expliquées par des valeures numériques telles que l'age ,la pression artérielle au repos (RestingBP) , cholesterol et le sucre dans le sang à jeun (FastingBS) , et la figure 2.4 represente des données qualitatives expliquées par des catégories telles que le genre , le type de douleur thoracique (ChestPainType) et résultats de l'électrocardiogramme au rapos (RestingECG) .

En particulier **pour le diabéte** , il existe des données quantitatives seulement telles que l'insuline , le glycose et l'épaisseur de la peau (SkinThickness) comme le montre la figure ci-dessous .

Column Name	Description
Pregnancies	Nombre de fois enceinte
Glucose	Concentration de glucose plasmatique pendant 2 heures lors d'un test oral de tolérance au glucose
BloodPressure	Pression artérielle diastolique (mm Hg)
SkinThickness	Épaisseur du pli cutané du triceps (mm)
Insulin	Insuline sérique sur 2 heures (mu U/ml)
BMI	Indice de masse corporelle (poids en kg/(taille en m)^2)
DiabetesPedigreeFunction	Fonction généalogique du diabète
Age	Années d'âge)
Outcome	Variable de classe (0 ou 1)

TABLE 2.2 – les attributs de base données des maladies diabètes

	count	mean	std	min	25%	50%	75%	max
Age	918	53,51089	9,4326	28	47	54	60	77
RestingBP	918	132,3965	18,5142	0	120	130	140	200
Cholesterol	918	198,7996	109.3841	0	173,25	223	267	603
FastingBS	918	0.2331	0.423	0	0	0	0	1
MaxHR	918	136.8094	25.4603	60	120	138	156	202
Oldpeak	918	0.8874	1.0666	-2.6	0	0.6	1.5	6.2
HeartDisease	918	0,5534	0,4974	0	0	1	1	1

FIGURE 2.2 – données relatives de cardiaque

	count	unique	top	freq
Gender	918	2	M	725
ChestPainType	918	4	ASY	496
RestingECG	918	3	Normal	552
ExerciseAngina	918	2	N	547
ST-Slope	918	3	Flat	460

FIGURE 2.3 – données relatives de cardiaque

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768	3.8451	3.3696	0	1	3	6	17
Glucose	768	120.8945	31.9726	0	99	117	140.25	199
BloodPressure	768	69.1055	19.3558	0	62	72	80	122
SkinThickness	768	20.5365	15.9522	0	0	23	32	99
Insulin	768	79.7995	115.244	0	0	30.5	127.25	846
BMI	768	31.9926	7.8842	0	27.3	32	36.6	67.1
DiabetesPedigreeFunction	768	0.4719	0.3313	0.078	0.2438	0.3725	0.6263	2.42
Age	768	33.2409	11.7602	21	24	29	41	81
Outcome	768	0.349	0.477	0	0	0	1	1

FIGURE 2.4 – les attributs de base de données du diabète

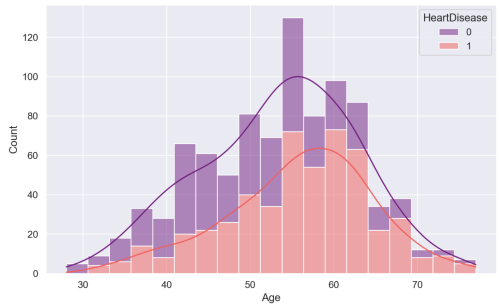


FIGURE 2.5 – Histogramme répartition cardiaque par âge

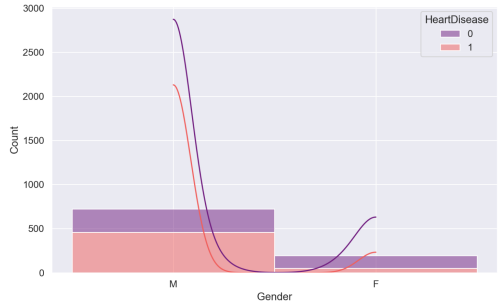


FIGURE 2.6 – Diagramme répartition cardiaque par Gendre

2.2.2.2 visualisation des caracteristiques

Les indicateurs de diagramme(figure 2.6) indique que la répartition des patients par âge , et il est clair que la probabilité d'avoir une maladie cardiaque est élevé chez les patientes âgée de 50 ans à 65 ans .

La diagramme montre que le nombre des Hommes atteintes de maladies cardiaques est plus élevé que celui des femmes .

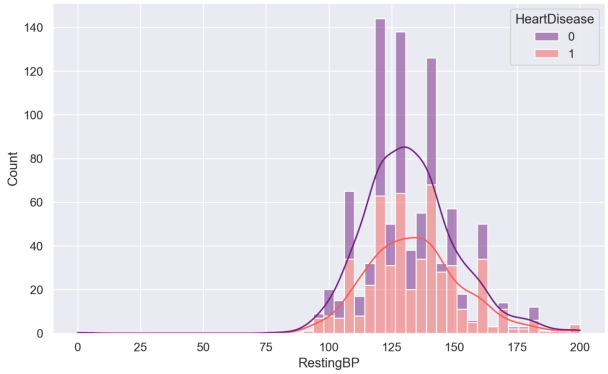


FIGURE 2.7 – diagramme RestingBP en fonction de variable HeartDisease

le graphe montre la brobabilité accrue de maladie cardiaque lorsque la tension artérielle varie de 120 à 150 .

comme indiqué dans la figure ci-dessus,plus que le taux de glucose dans le sang augment (entre 100 et 200) ,plus que le risque de souffre du diabéte augmente .

la figure 2.5 est un diagramme circulaire qui illustre des cas d'insuffisance cardiaque ,dont la partie rouge occupe 55 % du diagramme et représente le pourcentage de patients ne souffrant pas d'insufisance cardiaque, puis la partie bleue occupe 45 % du diagramme et représente le pourcentage de patient souffirant ca rdiaque.

Lorsque le taux d'insuline dans le sang diminue , le risque de développer le diabéte augmente , comme le montrre la fidure 2.10 .

La figure représent un graphe circulaire montrant le pourcentage de personnes atteintes de diabéte , où la partie rouge represente le pourcentage des non atteits , qui 65 % tandis que la partie blue represente le pourcentage des atteits , qui est de 35 % .

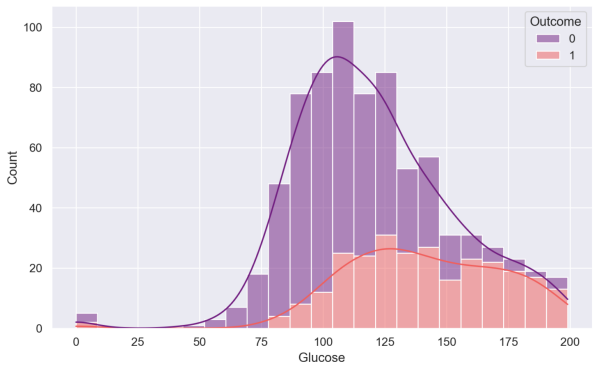


FIGURE 2.8 – taux de glucose en fonction d'Outcome du diabète

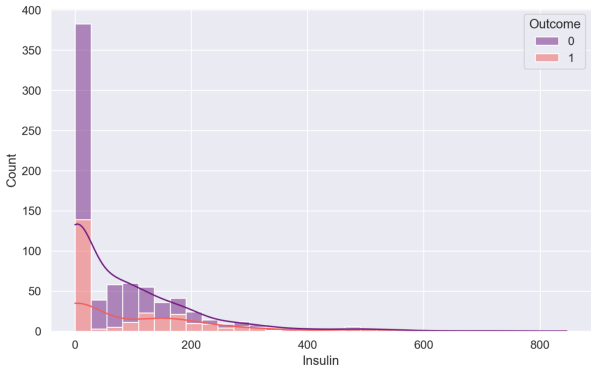


FIGURE 2.9 – taux d'insulin en fonction de variable Outcome

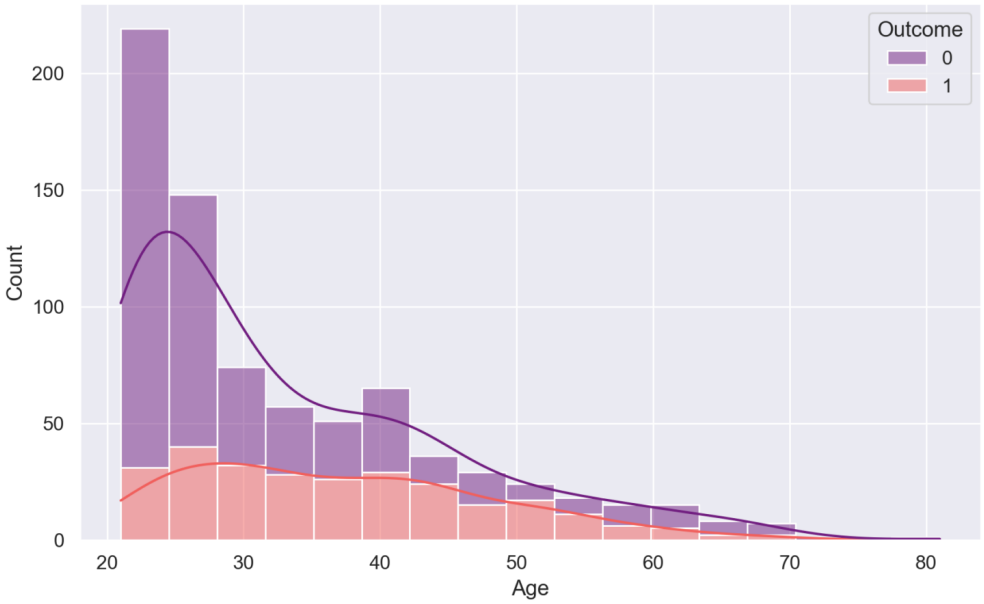


FIGURE 2.10 – Histogramme répartition de diabéte par Age
Il est claire que le nombre de patient souffrant de diabéte est souvent de 21 à 50 ans .

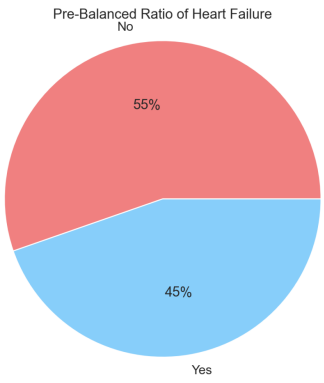


FIGURE 2.11 – pourcentage d'inffluence cardiaque

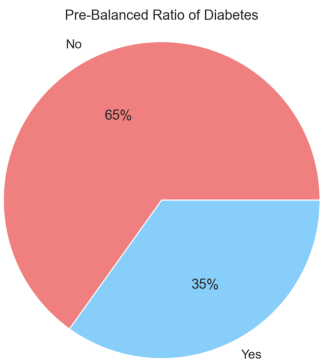


FIGURE 2.12 – pourcentage d'inffluence Diabète

2.2.2.3 Analyse de corrélation

Ce tableau montre la relation entre les variable , deux à deux , permet de quantifier la force et la direction de ces relation , le coeffitient de corrélation est représenté par un nombre compris entre -1 et 1 . En analysant cette matrice on trouve que les variables les plus fortement corrélée avec la variable "Heart disease" sont Age (0.282) RestingBP (0.107) et Oldpeak (0.404) ils sont de coeffitient positifs .

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1	0.2544	-0.0953	0.198	-0.382	0.2586	0.282
RestingBP	0.2544	1	0.1009	0.0702	-0.1121	0.1648	0.1076
Cholesterol	-0.0953	0.1009	1	- 0.261	0.2358	0.0501	-0.2327
FstingBS	0.198	0.0702	-0.261	1	-0.1314	0.0527	0.2673
MaxHR	-0.382	-0.1121	0.2358	- 0.1314	1	-0.1607	-0.4004
Oldpeak	0.2586	0.1648	0.0501	0.0527	-0.1607	1	0.404
HeartDisease	0.282	0.1076	-0.2327	0.2673	-0.4004	0.404	1

FIGURE 2.13 – Matrice de corrélation des variables de cardiaque

En examinant la matrice de corréllation on observe que le Glucose (0.466) , Age (0.238) , BMI (0.292) et l'insulin (0.13) présentent les corrélation les plus élevées avec "Outcome".

	Grossesse	Glucose	Sang pression	Skin thickness	insuline	BMI	Diabètes pedigree fonctionnalité	Age	Résultat
Grossesse	1	0.1295	0.1413	-0.0817	-0.0735	0.0177	-0.0335	0.5443	0.2219
Glucose	0.1295	1	0.1526	0.0573	0.3314	0.2211	0.1373	0.2635	0.4666
Sang pression	0.1413	0.1526	1	0.2074	0.0889	0.2818	0.0413	0.2395	0.0651
Skin thickness	-0.0817	0.0573	0.2074	1	0.4368	0.3926	0.1839	-0.114	0.0748
insuline	-0.0735	0.3314	0.0889	0.4368	1	0.1979	0.1851	-0.0422	0.1305
BMI	0.0177	0.2211	0.2818	0.3926	0.1979	1	0.1406	0.0362	0.2927
Diabètes pedigree fonctionnalité	-0.0335	0.1373	0.0413	0.1839	0.1851	0.1406	1	0.0336	0.1738
Age	0.5443	0.2635	0.2395	-0.114	-0.0422	0.0362	0.0336	1	0.2384
Résultat	0.2219	0.4666	0.0651	0.0748	0.1305	0.2927	0.1738	0.2384	1

FIGURE 2.14 – matrice de corrélation des variables de diabète

2.2.3 Visualisation des données

Le tableau suivant présent un échantillon de la base de données , contient les dix premiers patients dont les informations ont été collectées comme l'age les resultas d'analyses, et le resulta de diagnostic du médecin .

2.3 traitement de données

Une fois les données préparées , le traitement de données peut commencer . il s'agit de manipuler et d'analyser les données pour en extraire des informations exploitables .

	Pregnancies	Glucose	BloodPressure	SKinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31	0.248	26	1
7	10	115	0	0	0	35.5	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0	0.232	54	1

FIGURE 2.15 – Données dix premiers patient diabète

	Age	Genre	Chest Pain Type	Resting BP	Cholestérol	Fasting BP	Resting ecg	Max HR	Exercice Angina	Oldpeak	ST_SLOP	HEART DISEASE
0	40	M	ATA	140	289	0	NORMAL	172	N	0	Up	0
1	49	F	NAP	160	180	0	ST	156	N	1	Flat	1
2	37	M	ATA	130	283	0	NORMAL	98	N	0	Up	0
3	48	F	ASY	138	214	0	NORMAL	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	NORMAL	122	N	0	Up	0
5	39	M	ATA	120	339	0	NORMAL	170	N	0	Up	0
6	45	F	ATA	130	237	0	NORMAL	170	N	0	Up	0
7	54	M	ASY	110	208	0	NORMAL	142	N	0	Up	0
8	37	M	ASY	140	207	0	NORMAL	130	Y	1.5	Flat	1
9	48	F	ATA	120	284	0	NORMAL	120	N	0	Up	0

FIGURE 2.16 – Données dix premiers patient cardiaques

2.4 Extraction des caractéristiques

Permet de crée des caractéristique entiérement nouvelles à partir des caractéristiques existantes dans notre ensemble de données , ces nouvelles caractéristiques sont souvent plus informatisés et peuvent capturer des relations complexes entre les caractéristiques d'origine .

2.5 Selection des caractéristiques

choisit un sous-ensemble des caractéristiques les plus pertinentes de l'ensemble de données originale , elle écarte les caractéristiques non pertinentes ou redondantes qui pourraient ne pas contribuer au peocessus d'apprentissage .

2.6 Construction du modèle

l'apprentisage automatique est une branche de l'intelegence artificielle qui se concentre sur le d evelopement d'agorithmes capable d'apprendre à partire de données et de faire des prédictions . ces algorithmes utilisent des modéles statistiques pour identifier les facteurs de risques , rédire les résultats et faire des recomandation et traitement . c'est l'objectif d'inclure ces techniques dans la détection des maladies cardiaques et du diabète . la figure montre les principaux modéles de l'apprentissage automatique .

L'apprentissage supervisé est une technique d'apprentissage automatique où l'on cherche à produire automatiquement des règles qui ne sont pas définies a priori ,à partir d'une base de données d'apprentissage contenant des exemples .Ces algorithmes d'apprentissage répondent à des problématiques de régression ou de classification . La bibliothèque sklearn nous permet d'entraîner la grande majorité de ces modèles.

2.6.1 Sélection des algorithmes

Il existe de nombreux modèles de machine learning .Presque tous sont basés sur certains algorithmes de machine learning .les algorithmes les plus couramment utilisés pour la classification et la régression appartiennent au machine learning supervisé.

La forêt aléatoire est un algorithme d'apprentissage automatique couramment utilisé, qui permet d'assembler les sorties de plusieures arbres de décision pour attiendre un résultat unique. Les algorithmes de forêt aléatoire disposent de trois hyperparamètres principaux qui n'ont pas besoin d'être définis avant l'entraînement.Ce sont **la taille des noeuds, le nombre d'arbres et le nombre de fonctions échantillonnées.**

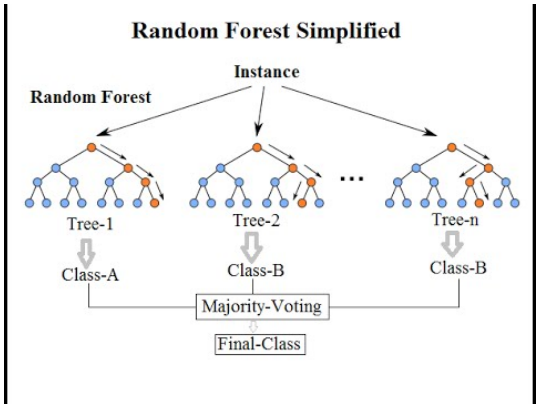


FIGURE 2.17 – Random Forest Simplified

Chaque arbre d'une forêt aléatoire échantillonne de manière aléatoire des sous-ensembles de données d'entraînement dans le cadre d'un processus appelé agrégation bootstrap **(bagging)**. Le modèle est adapté à ces ensembles de données plus petits et les prédictions sont agrégées. Plusieurs instances des mêmes données peuvent être utilisées à plusieurs reprises grâce à un échantillonnage de remplacement, et le résultat est que des arbres qui sont non seulement formés sur différents ensembles de données, mais également sur différentes fonctionnalités, sont utilisés pour prendre des décisions.

En détail :

Construction de la forêt :

Échantillonnage aléatoire : On commence par créer plusieurs sous-ensembles aléatoires de données (avec remplacement) à partir de l'ensemble de données d'origine. Chaque sous-ensemble représente une partie de la forêt.

Construction d'arbres de décision : Pour chaque sous-ensemble, on construit un arbre de décision distinct. Lors de la construction de chaque arbre, on utilise également un sous-ensemble aléatoire des caractéristiques disponibles.

Limitation de la profondeur : On limite la profondeur maximale des arbres pour éviter le surapprentissage.

Prédiction :

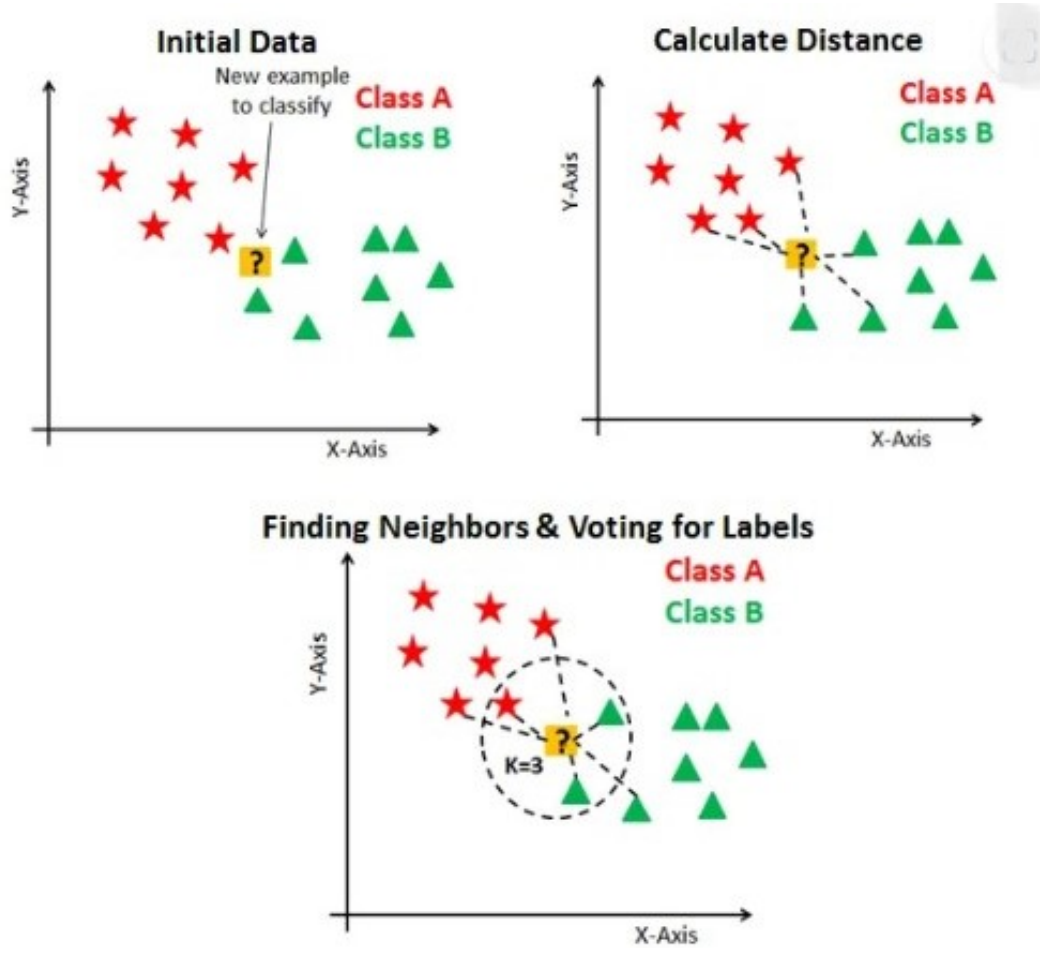
Appliquer chaque arbre : Pour une nouvelle donnée, on la fait passer par chaque arbre de la forêt. Chaque arbre prédit la classe de la nouvelle donnée.

Vote majoritaire : On regroupe les prédictions de tous les arbres. La classe la plus prédite est la prédiction finale de la forêt aléatoire.

Avantages des forêts aléatoires :

- **Robuste au surapprentissage :** Moins sujettes au surapprentissage que les arbres de décision individuels.
- **Précision élevée :** Produisent souvent des prédictions plus précises que d'autres algorithmes.
- **Facilité d'utilisation :** Ne nécessitent pas beaucoup de paramètres à régler.
- **Interprétabilité :** On peut comprendre l'importance relative des caractéristiques en analysant les arbres individuels.

KNN est basé sue le principe que les objets ou les points de données qui sont proche les une des autres dans l'espace des caractéristiques sont susceptibles d'appartenir à la même classe ou d'avoir des sorties similaires .L'algorithme KNN attribue un nouveau point de données à la classe la plus courante parmi ses K plus proches voisins. La valeur de k est un hyperparamètre qui doit être spésifié avant la formation du modèle.



Données initiales et calcul de la distance :

— Cette section représente les données d'apprentissage initiales, probablement un espace de caractéristiques bidimensionnel. Chaque point de données est visualisé par un cercle ou un carré, de couleur selon sa classe (Classe A en bleu et Classe B en vert).

— La distance entre un nouveau point de données non classifié (représenté par un point d'interrogation) et chaque point de données dans l'ensemble d'apprentissage est calculée.
Calculez la distance :

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Recherche des voisins et vote pour les étiquettes :

Cette section illustre la façon dont l'algorithme KNN classifie le nouveau point de données. Voici une description de ce processus :

— Un cercle autour du nouveau point de données englobe ses k plus proches voisins (k=3 dans ce cas), à la fois de la classe A (bleu) et de la classe B (vert).

— Par le principe du vote majoritaire, le nouveau point de données est affecté à l'étiquette de classe de la majorité parmi ses k plus proches voisins. Dans cet exemple, comme deux des trois voisins les plus proches appartiennent à la classe A (bleu), le nouveau point de données est également classé comme classe A.
Le modèle KNN présente plusieurs avantages intéressants pour notre dataset qui le rendent attrayant pour une variété de problèmes d'apprentissage automatique :

— **Simplicité :** L'algorithme KNN est l'un des algorithmes d'apprentissage automatique les plus faciles à comprendre et à implémenter. Le concept de base repose sur l'idée intuitive que les points de données similaires ont tendance à appartenir à la même classe. Cela rend KNN accessible aux utilisateurs sans une formation approfondie en apprentissage automatique.

— **Efficacité pour les petits ensembles de données et les problèmes à faible dimensionnalité :** Lorsque vous travaillez avec des ensembles de données de taille modeste et des problèmes à faible dimensionnalité (nombre limité de caractéristiques), KNN peut être étonnamment efficace. Il peut rapidement identifier les voisins les plus proches et faire des prédictions précises.

2.6.2 Entraînement et validation du modèle

Après le traitement de la base de données , la machine passe à la phase d'entrainement qui implique l'ajustement des paramétres du modèle afin de minimiser une fonction de coût ou d'erreur sur l'ensemble de données d'entraînement , et d'optimiser l'algorithme afin de trouver certains modèles ou certaines données de sortie. La fonction qui en résulte, dotée de règles et de structures de données est appelée modèle entraînement de machine learning. Après l'entraînement, evaluez les performances du modèle sur l'ensemble de validation. Cela vous donne des indications de la capacité du modèle à généraliser à de nouvelles données, qui se présente en mesure de performance .

2.6.3 Métrique de performance :

si le patient est malade ou non (réel)

La prédiction du modéle		**Est malade**	**N'est pas malade**
	Est malade	Vrai positif (VP)	Faux positif (FP)
	N'est pas malade	Faux négatif (FN)	Vrai négatif (VN)

les mesure de performance sont des indicateurs de la correspondance entre Valeurs prédites et valeurs obtenus à partire du modèle obtenu via la matrice de confusion qui vont nous permettre de juger de la qualité de nos prédictions **Precision** , **Rappel** , **F1-score** et **Accuracy** :

Precision : C'est la proportion des individues qui sont correctement identifiées par le modéle . l'équation de précision est représentée comme suit :

$$\text{Précision} = \frac{VP}{VP+FP}$$

Rappel : C'est lapetite proportion des individus par rapport à la quantité globale des individus applicables . l'equoition de rappel est représentée comme suit :

$$\text{Rappel} = \frac{VP}{VP+FN}$$

F1-score : C'est la moyanne entre la précision et le rappele :

$$\text{F1-score} = 2 * \frac{Prcision*Rappel}{Prcision+Rappel}$$

Accuracy : L'exactitude se réfère à la proximité d'une mesure ou d'une valeur à la valeur réelle ou attendue. Il s'agit de la distance entre la valeur mesurée et la valeur réelle. Plus la valeur mesurée est proche de la valeur réelle, plus l'exactitude est élevée.

$$\text{Accuracy} = \frac{VP+VN}{VP+VN+FP+FN}$$

chapitre 3

Implémentation et Résultats

3.1 Implémentation du système

3.1.1 Outils et langages de développment

3.1.2 Python :

Python est un langage de programmation interprété,multi-plateforme et orientée objet.il favorise la programmation impérative structurée, et orientée objet. il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire et d'un système de gestion d'exceptions

3.1.3 jupyter :

Jupyter est une application web qui permet de stocker des lignes de code Python, les résultats de l'exécution de ces dernières (graphiques, tableaux, etc.) et du texte formaté.

3.1.4 Numpy

Numpy est une bibliothèque composée d'objets de tableau multidimensionnel et d'une collection de routines pour traîter ces tableaux.l'utilisation de Numpy permet de faciliter les opérations mathématiques et logiques sur les tableaux.et son rôle est similaire au package de pandas.

3.1.5 seaborn

Seaborn est une bibliothèque python de visualisation de données basées sur matplotlib.il fournit une de haut niveau pour dessiner des graphiques statistiques attrayantes et informatives.

3.1.6 Streamlit

Streamlit est un framework gratuit et open-source permettant de créer et de partager rapidement de belles applications web d'apprentissage automatique et de science des données. Il s'agit d'une bibliothèque basée sur Python spécialement conçue pour les ingénieurs en machine learning. Les data scientists ou les ingénieurs en machine learning ne sont pas des développeurs web et ils ne sont pas intéressés à passer des semaines à apprendre à utiliser ces frameworks pour créer des applications web. Au lieu de cela, ils veulent un outil plus facile à apprendre et à utiliser, tant qu'il peut afficher des données et collecter les paramètres nécessaires.

3.1.7 Pandas

Pandas est une bibliothèque Python sous licence BSD open source fournissant des structures de données et des outils d'analyse de données haute performance et faciles à utiliser pour le langage de programmation Python. Le Python avec Pandas est utilisé dans un large éventail de domaines, y compris académique et commerciale domaines tels que la finance, l'économie, les statistiques, l'analyse,....,etc. C'est pour ces avantages que nous l'avons choisi dans notre projet pour faciliter la manipulation de l'ensemble de donnée que nous avons utilisé sous l'extension (.csv).

3.1.8 Scikit-learn

Scikit-learn est un module Python intégrant un large éventail de rythmes algorithme d'apprentissage automatique de pointe pour les problèmes supervisés et non supervisés à moyenne échelle. Cette trousse met l'accent sur l'apprentissage machine pour les

nonspécialistes qui utilisent un langage général de haut niveau pour faciliter d'utilisation, les performances, la documentation et la cohérence des API. Il a des dépendances minimales et est distribué sous la licence BSD simplifiée, encourageant son utilisation dans les deux milieux académiques et commerciaux .

3.1.9 Matplotlib

Matplotlib est probablement le paquet Python le plus utilisé pour les graphiques 2D.Il fournit à la fois un moyen très rapide de visualiser les données à partir Python et des chiffres de qualité de publication dans de nombreux formats.Nous allons pour explorer matplotlib en mode interactif couvrant les plus courants affaires .

3.2 Selection de caractéristiques

La sélection de caractéristiques est une étape cruciale dans le développement de modèles d'apprentissage automatique pour la prédiction des maladies cardiaques et du diabète. Elle consiste à identifier les sous-ensembles de caractéristiques les plus pertinents et informatifs à partir d'un ensemble de données initialement volumineux.

Pour cardiaque :

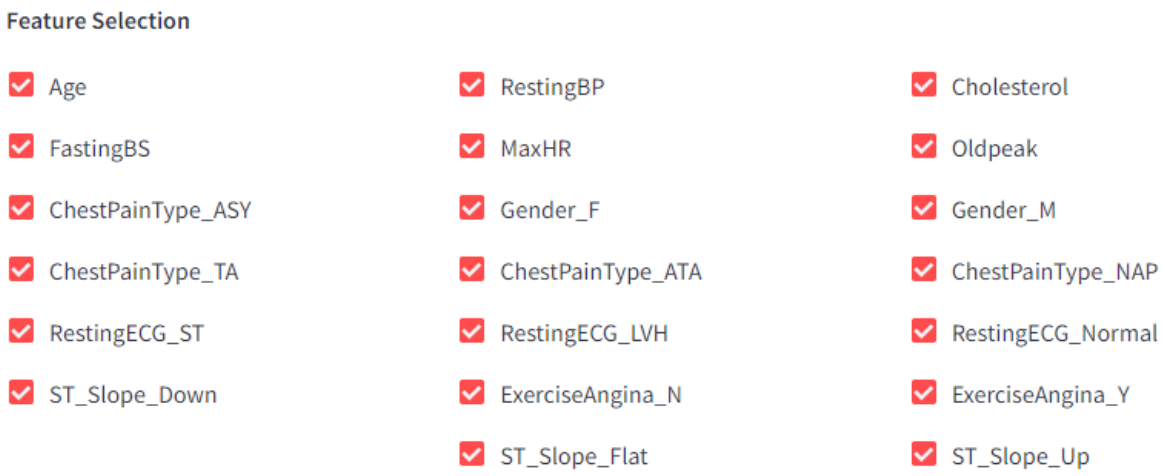


FIGURE 3.1 – Les caractéristiques de cardiaque

Pour Diabète :

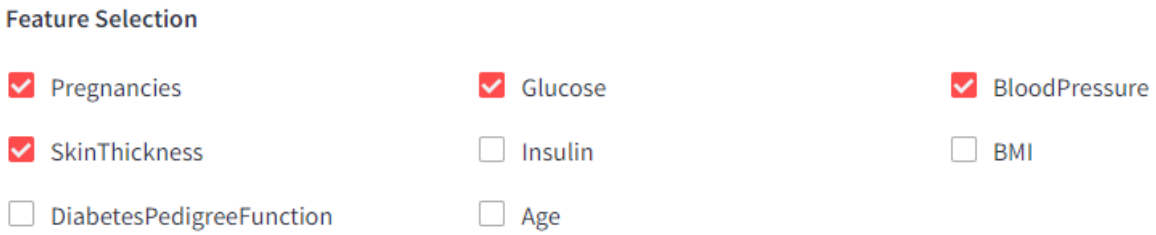


FIGURE 3.2 – Les caractéristiques de diabète

3.3 Analyse comparative des différents modèles

Modèles utilisés :

Nous avons réalisé plusieurs modèles pour ces différentes approches à savoir :
• **KNN (K-Nearest Neighbors) :** Un modèle de classification basé sur la proximité, où les prédictions sont faites en fonction des classes des k voisins les plus proches.
• **Random Forest (RandomForestClassifier) :** Un ensemble d'arbres de décision qui agrège leurs prédictions pour améliorer la robustesse et la généralisation du modèle.

Pour les maladies cardiaques :

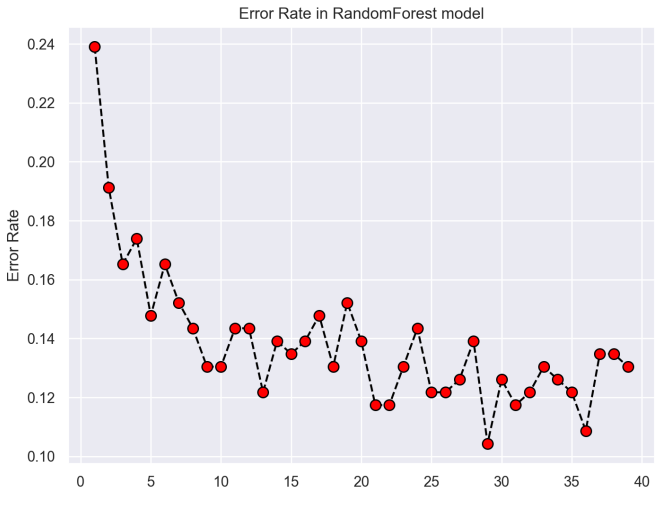


FIGURE 3.3 – diagramme d'error du lmodèle Random Forest

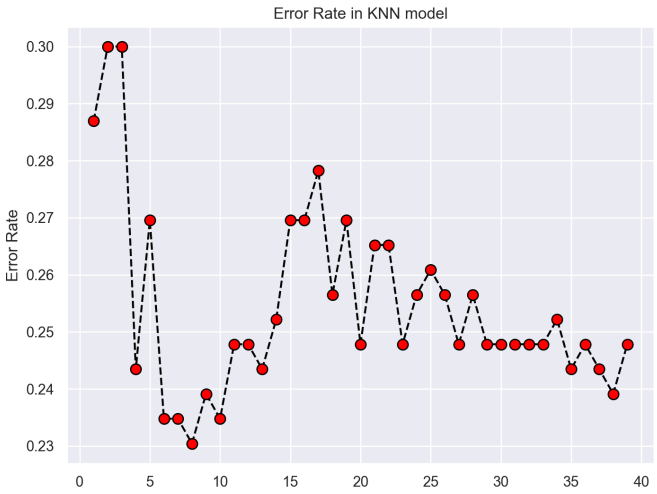


FIGURE 3.4 – diagramme d'error du modèle KNN

D'après cette image, on peut conclure que le modèle **de forêt aléatoire** obtient sa meilleure performance avec environ 20 arbres de décision. En dessous de ce nombre, le modèle ne prédit pas correctement les classes. Au-dessus de ce nombre, le modèle commence à sur-apprendre et sa performance se dégrade.
Il est important de noter que ce graphique ne présente qu'un seul exemple de performance d'un modèle **de forêt aléatoire**. La performance réelle d'un modèle peut varier en fonction des données spécifiques utilisées et des paramètres choisis.

D'après cette image, on peut conclure que le modèle **KNN** obtient sa meilleure performance avec environ 15 voisins. En dessous de ce nombre, le modèle ne prédit pas correctement les classes. Au-dessus de ce nombre, le modèle commence à sur-apprendre et sa performance se dégrade.
Il est important de noter que ce graphique ne présente qu'un seul exemple de performance d'un modèle **KNN**. La performance réelle d'un modèle peut varier en fonction des données spécifiques utilisées et des paramètres choisis.

Nous constaton que **la forêt aleatoire** à une meilleur modèle pour analyser la dataset du

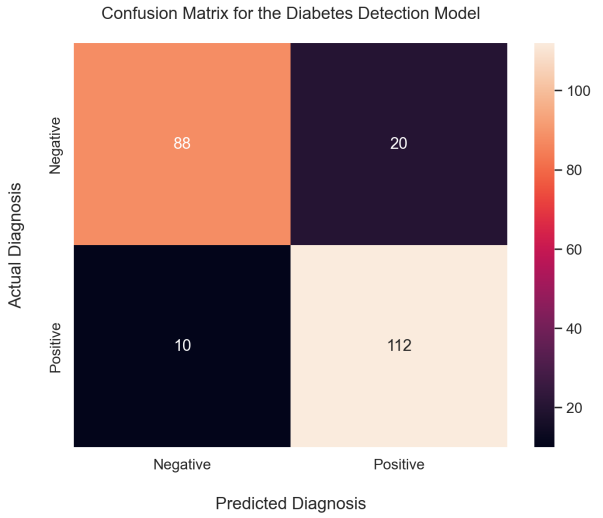


FIGURE 3.5 – matrice de Confusion du modèle RF

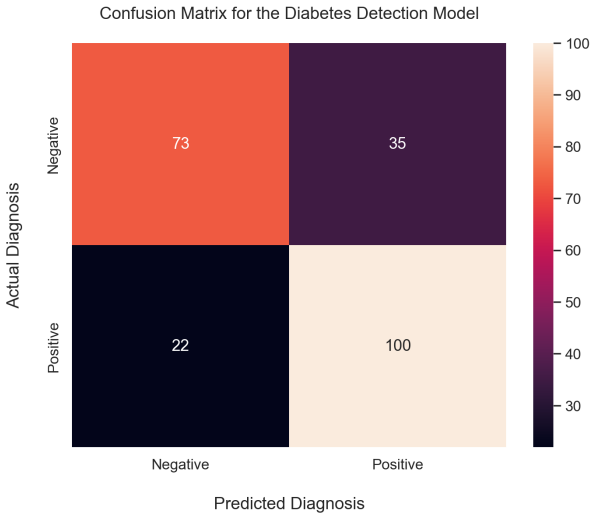


FIGURE 3.6 – matrice Confusion du modèle KNN

modèles	Précision0	Précision1	RAPPELER 0	RAPPELER 1	Exactitude	F1_score 0	F1_score 1
KNN	0.77	0.74	O.68	0.82	0.75	0.72	0.78
RF	0 .9	0.85	0 .82	0.92	0.87	0.86	0.89

FIGURE 3.7 – la classification des modèles

cardiaque et la performance résultat.car les valeurs des indicateurs de métriques ou performance qui déterminent la qualité de modèle sont très grand que l'autre .

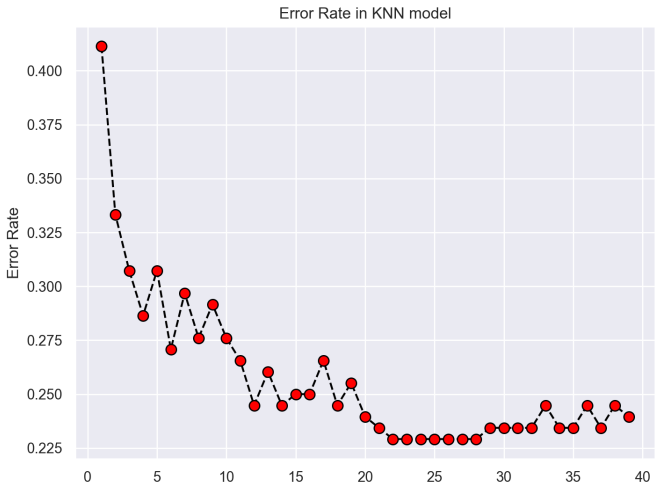


FIGURE 3.8 – diagramme d'error du modèle KNN

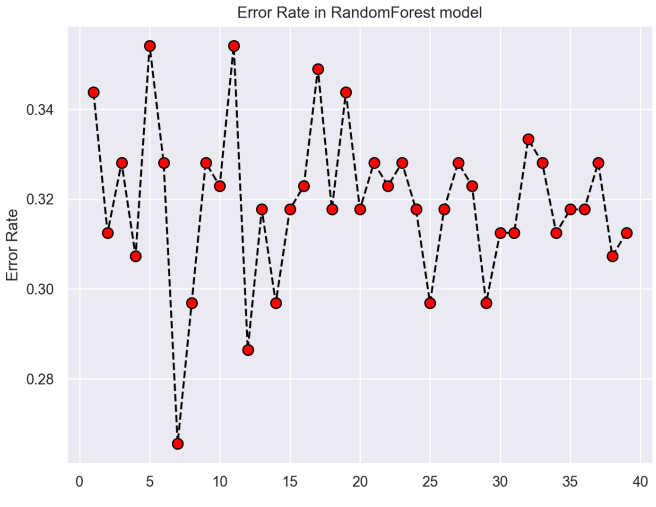


FIGURE 3.9 – diagramme d'error du modèle Random Forest

Pour les maladies diabètes :

D'après cette image, on peut conclure que le modèle **KNN** obtient sa meilleure performance avec environ 15 voisins. En dessous de ce nombre, le modèle ne prédit pas correctement les classes. Au-dessus de ce nombre, le modèle commence à sur-apprendre et sa performance se dégrade.

Il est important de noter que ce graphique ne présente qu'un seul exemple de performance d'un modèle **KNN**. La performance réelle d'un modèle peut varier en fonction des données spécifiques utilisées et des paramètres choisis.

Sur la base de cette image, nous pouvons conclure que le modèle **de forêt aléatoire** atteint ses meilleures performances avec une vingtaine d'arbres de décision. En dessous de ce nombre, le modèle ne prédit pas correctement les classes. Au-dessus de ce nombre, le modèle commence à être surajusté et ses performances se dégradent.

Il est important de noter que ce graphique ne présente qu'un exemple des performances d'un modèle **de forêt aléatoire**. Les performances réelles d'un modèle peuvent varier en fonction des données spécifiques utilisées et des paramètres choisis.

modèles	Précision0	Précision1	rappeler0	rappeler1	Exactitude	F1_score 0	F1_score 1
KNN	0.78	0.7	0.91	0.45	0.76	0.84	0.55
RF	0.77	0.52	0.78	0.52	0.69	0.77	0.52

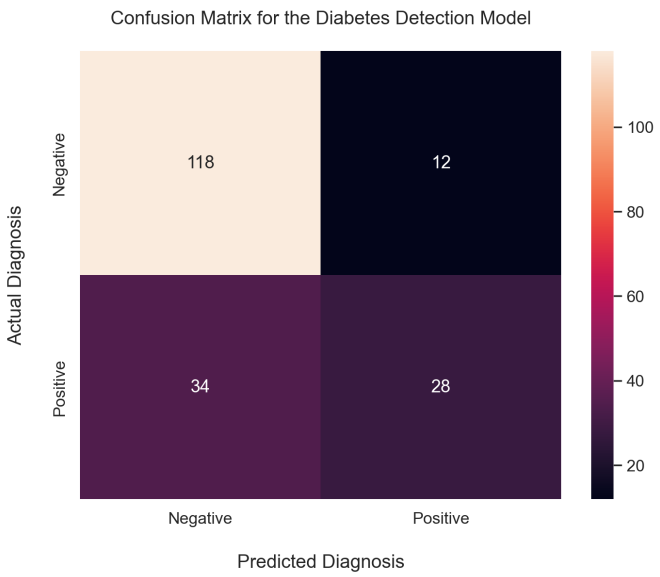


FIGURE 3.10 – matrice de Confusion du modèle KNN

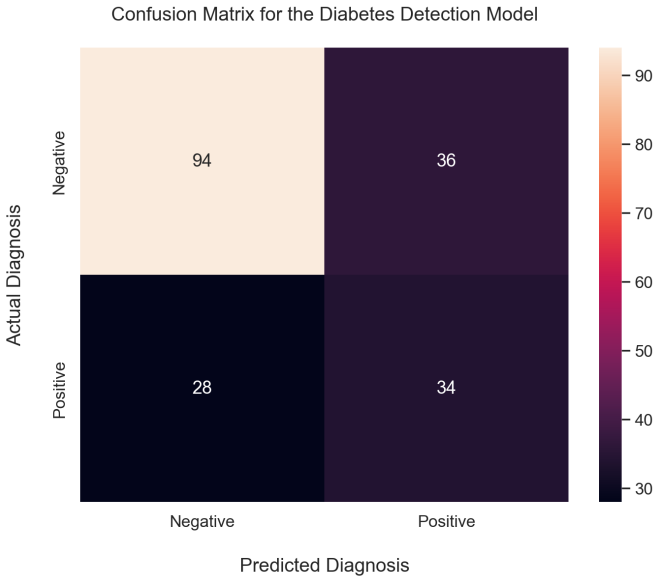


FIGURE 3.11 – matrice de Confusion du modèle RF

Nous constaton que **KNN** à une meilleur modèle pour analyser la dataset du Diabète et la performance résultat.par ce que les valeurs des indicateurs de métriques ou performance qui déterminent la qualité de modèle sont très grand que l'autre.

Note :

Il est important de mentionner que l'algorithme **KNN** (K-Nearest Neighbors) est un algorithme de classification supervisée, tandis que **la forêt aléatoire** est un algorithme d'ensemble. Cela signifie qu'ils fonctionnent de manière différente et peuvent être mieux adaptés à différents types de problèmes. Le choix entre les deux algorithmes dépend de plusieurs facteurs, tels que la nature des données, la taille de l'ensemble de données, la complexité du problème et les performances souhaitées.

Conclusion

L'objectif de notre étude était d'effectuer des modèles capables de prédire si une personne était exposée à un risque de maladie cardiaque ou non et de savoir quelles caractéristiques sont importantes. La question que vous vous posez est " sommes-nous de bons médecins sachant détecter les maladies diabètes et cardiaques ? " . Avant de répondre, nous avons exploré la base de données et effectué des modifications afin de pouvoir utiliser les données : détection et suppression des valeurs atypiques. Ensuite, nous nous sommes intéressés à la relation entre les variables et nous en sommes venus à la conclusion qu'il était nécessaire d'effectuer une sélection des caractéristiques en raison du nombre élevé de liens entre les variables. Nous en avons réalisé deux bases de données : base de diabète et autre cardiaque. Ces deux sélection des caractéristiques nous ont conduit à réaliser 2analyses différentes afin d'effectuer le meilleur modèle possible.

Ensuite, après cette étape préliminaire mais nécessaire, nous avons pu répondre à la question. Pour ce faire, nous avons réalisé 2 modèles différents pour chacune de nos analyses. Pour data set diabète les modèles de KNN semble être le meilleur en termes d'accouracy et le modèle qui est le plus performant pour nos données. A contre le modèle foret aléatoire qui moins performance car leur terme accuracy est faible .pour la deuxième cas de base données cardiaques le modèle foret aléatoire est le meilleure accuracy sur le jeu d'entraînement et le plus performant pour nos données. Alors que le KNN mois performance.

Ce mémoire présente un système de perdition du cardiaque et diabète utilisant les modèles KNN et La forêt aleatoire. Dans notre projet, nous avons étudié le diabète et cardiaque en détail, Pour obtenir les meilleurs résultats et atteindre l'objectif, ce projet a été implémenté en utilisant anaconda et jupyter notebook en plus des bibliothèques telles que seaborn, pandas. Le but de ce système était d'obtenir le meilleur taux de prédiction pour assurer son efficacité.

Bibliographie

— kagglekernelsoutputsanchitakarmakar/heart-failure-prediction-visualization-p/path/to/dest

— https://www.databricks.com/fr/glossary/machine-learning

— https://www.kaggle.com/code/prashant111/svm-classifier-tutorial

— kagglekernelsoutputdejoune/chps0942-td-1-pr-diction-du-diab-te-p/path/to/dest

— https://ultramedica.care

— https://www.msdmanuals.com/ar

— https://ledatascientist.com/support-vector-machines-svm-en-python/

— https://dspace.univ-bba.dz/bitstream/handle/123456789/1393/m%c3%a9moire%20fin%20%c3%a9tude%202021.pdf?sequence=1&isAllowed=y

— https://fr.wikipedia.org/wiki/Scikit-learn

— :https://fr.wikipedia.org/wiki/Pandas

— :https://fr.wikipedia.org/wiki/Matplotlib

— :https://fr.wikipedia.org/wiki/Python_(langage)

— https://jupyter.org/