

PROJET DE FIN D’ETUDES

En vue de l’obtention du diplôme master spécialisé : Big Data & Cloud Computing

**Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare**

2020/2022

# TABLE DES MATIERES

Contents

[TABLE DES MATIERES 2](#_Toc118649631)

[LISTE DES FIGURES 6](#_Toc118649632)

[LISTE DES TABLEAUX : 8](#_Toc118649633)

[RESUME 10](#_Toc118649634)

[ABSTRACT 10](#_Toc118649635)

[INTRODUCTION GENERALE 11](#_Toc118649636)

[CONTEXTE 11](#_Toc118649637)

[PROBLEMATIQUE 11](#_Toc118649638)

[SOLUTION PROPOSEE 12](#_Toc118649639)

[STRUCTURE DU RAPPORT 13](#_Toc118649640)

[Chapitre 1 : ’’Machine Learning et le domaine médicale’’ 13](#_Toc118649641)

[Chapitre 2 : ‘’LE DIABETE’’ 13](#_Toc118649642)

[Chapitre3 : “Conception et Réalisation” 13](#_Toc118649643)

[Chapitre 4 : “Démonstration de l’application ” 13](#_Toc118649644)

[Chapitre 1 : ’’Machine Learning et le domaine médicale’’ 14](#_Toc118649645)

[3. Introduction 15](#_Toc118649646)

[4. Machine Learning: 15](#_Toc118649647)

[4.1. Définition et Histoire: 15](#_Toc118649648)

[4.2. Méthodes d’apprentissage : 17](#_Toc118649649)

[4.2.1. Apprentissage supervisé : 18](#_Toc118649650)

[4.2.1.1. Classification : 18](#_Toc118649651)

[4.2.1.2. Régression : 20](#_Toc118649652)

[4.2.2. Apprentissage non supervisé : 20](#_Toc118649653)

[4.2.2.1. Clustering 21](#_Toc118649654)

[4.2.3. Différence entre apprentissage supervisé et non supervisé : 22](#_Toc118649655)

[4.2.4. Apprentissage semi-supervisé 23](#_Toc118649656)

[4.2.5. Apprentissage par Renforcement : 24](#_Toc118649657)

[5. Deep Learning : 25](#_Toc118649658)

[5.1. Machine Learning vs Deep Learning: 25](#_Toc118649659)

[5.2. Pourquoi Deep Learning : 27](#_Toc118649660)

[6. Random Forest : 29](#_Toc118649661)

[6.1. Définition : 29](#_Toc118649662)

[6.2. Principe de fonctionnement du random forest : 29](#_Toc118649663)

[6.3. Genèsedel’algorithme 30](#_Toc118649664)

[6.4. Tree bagging: 30](#_Toc118649665)

[6.5. Feature sampling 31](#_Toc118649666)

[6.7. Random Forest : intuition 33](#_Toc118649667)

[6.8. Avantages et inconvénients des random-forests 34](#_Toc118649668)

[7. K-nearset Neighbors (KNN) : 35](#_Toc118649669)

[7.1. Définition 35](#_Toc118649670)

[7.2. Applications de k-NN dans l'apprentissage automatique : 37](#_Toc118649671)

[7.3. Avantages et inconvénients de l'algorithme KNN : 39](#_Toc118649672)

[8. LDA (linear Discriminant Analysis) : 40](#_Toc118649673)

[8.1. Définition : 40](#_Toc118649674)

[8.2. Hypothèses et Formules : 40](#_Toc118649675)

[8.3. La règle bayésienne : 40](#_Toc118649676)

[8.4. L'analyse discriminante paramétrique - L'hypothèse de multi normalité 41](#_Toc118649677)

[8.5. Fonction de classement linéaire 42](#_Toc118649678)

[8.6. Robustesse : 43](#_Toc118649679)

[8.7. Taux d’erreur : 43](#_Toc118649680)

[8.8. Séparabilité - Évaluation globale : 44](#_Toc118649681)

[8.9. Évaluation individuelle des variables prédictives : 45](#_Toc118649682)

[8.10. Lecture des résultats 45](#_Toc118649683)

[8.11. Avantages et inconvénients de LDA : 47](#_Toc118649684)

[**Chapitre 2 : ‘’LE DIABETE ‘’** 49](#_Toc118649685)

[1. Définition : 50](#_Toc118649686)

[51](#_Toc118649687)

[2. Classification : 51](#_Toc118649688)

[2.1. Diabète de type 1 52](#_Toc118649689)

[2.1.1. Les symptômes cliniques du diabète de type 1 52](#_Toc118649690)

[2.1.2. Le traitement du diabète de type 1 52](#_Toc118649691)

[2.2. Diabète de type 2 : 53](#_Toc118649692)

[2.2.1. Les symptômes cliniques du diabète de type 2 54](#_Toc118649693)

[2.2.2. Le traitement du diabète de type 2 54](#_Toc118649694)

[3. Diagnostic du diabète : 54](#_Toc118649695)

[4. Les complications du diabète : 55](#_Toc118649696)

[4.1. Les complications aigües du diabète : 55](#_Toc118649697)

[4.2. Le traitement pharmacologique : 55](#_Toc118649698)

[4.3. Le suivi du diabète : 56](#_Toc118649699)

[5. Le diabète au niveau de Maroc : 58](#_Toc118649700)

[59](#_Toc118649701)

[5.1. Historique diabète : 59](#_Toc118649702)

[6. Problématique et Direction de recherches actuelles : 63](#_Toc118649703)

[7. Conclusion 63](#_Toc118649704)

[8. État de l’art 64](#_Toc118649705)

[8.1. Travaux Rajiv Singla, Ankush Singla, Yashdeep Gupta et Sanjay Kalra 64](#_Toc118649706)

[**Chapitre3 : “Conception et Réalisation”** 66](#_Toc118649707)

[1. Introduction : 67](#_Toc118649708)

[2. Problématique et Contraintes: 67](#_Toc118649709)

[2.1.1. Les travaux de recherche sur l’application des algorithmes de machine Learning pour la prédiction du diabète type 2 : 69](#_Toc118649710)

[2.1.2. Application des méthodes d’apprentissage dans la prédiction du diabète Type 2 : 69](#_Toc118649711)

[2.1.2.1. La précision 69](#_Toc118649712)

[2.1.2.2. L’étude 1 : 70](#_Toc118649713)

[Critique : 70](#_Toc118649714)

[2.1.2.3. Etude 02: Machine Learning Workflow on Diabetes Data 71](#_Toc118649715)

[Méthode 01 :Train/Test Split : 71](#_Toc118649716)

[Méthode 02 :Validation croisé : 71](#_Toc118649717)

[ Logistic Regression : 72](#_Toc118649718)

[ KNN: 73](#_Toc118649719)

[ SVC: 74](#_Toc118649720)

[ DecisionTree: 75](#_Toc118649721)

[ Random Forest: 76](#_Toc118649722)

[ AdaBoost 77](#_Toc118649723)

[ LBGM: 78](#_Toc118649724)

[ Critique 79](#_Toc118649725)

[3. Conclusion 79](#_Toc118649726)

[4. Prédiction du diabète de type 2 par l’apprentissage automatique : 80](#_Toc118649727)

[4.1. Introduction 80](#_Toc118649728)

[4.2. Outils et Librairies utilises 80](#_Toc118649729)

[5. Définition d’ensemble de données utilisé et description des variables 83](#_Toc118649730)

[5.1. Définition l’ensemble de données utilisé 83](#_Toc118649731)

[5.2. Les étapes de pré-traitement de données : 84](#_Toc118649732)

[5.3. Exploration et visualisation de données : 84](#_Toc118649733)

[ Variables : 87](#_Toc118649734)

[ Corrélations 87](#_Toc118649735)

[ Remarque : 89](#_Toc118649736)

[5.4. Nettoyage de données : 90](#_Toc118649737)

[Solutions pour gérer les valeurs aberrantes 90](#_Toc118649738)

[5.5. Sélection de modèle : 93](#_Toc118649739)

[Méthode d’évaluation 93](#_Toc118649740)

[ Train/Test Split : 93](#_Toc118649741)

[5.6. Evaluations des modèles : 94](#_Toc118649742)

[5.7. Réglage de paramètre de modèle : 95](#_Toc118649743)

[5.8. Sauvegarde de modèle : 96](#_Toc118649744)

[Chapitre 4 : “Démonstration de l’application ” 97](#_Toc118649745)

# LISTE DES FIGURES

[Figure 1 : LE ML pour tous : la régresion.les données sont modélisées par y = a\*x+b. 16](#_Toc118649353)

[Figure 2: type d'apprentissage machine learning 18](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649354)

[Figure 3Indices de performance pour les problématiques de classification 19](#_Toc118649355)

[Figure 4l :a proportion de prédictions correctes (Accuracy en anglais) 19](#_Toc118649356)

[Figure 5 : clustering 21](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649357)

[Figure 6: Différence entre apprentissage supervisé et non supervisé 22](#_Toc118649358)

[Figure 7: Semi-supervised Learning 23](#_Toc118649359)

[Figure 8 :Apprentissage par Reinforcement: 24](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649360)

[Figure 9: AI & ML & DL 26](#_Toc118649361)

[Figure 10: ML vs DL 26](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649362)

[Figure 11: Construction de l'arbre N°1 du random forest 30](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649363)

[Figure 12: Prediction du random forest 31](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649364)

[Figure 13 : Mis en œuvre d’une forêt d'arbres 31](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649365)

[Figure 14 : Random forest illustration 33](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649366)

[Figure 15: Euclidian 34](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649367)

[Figure 16: Manhattan 34](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649368)

[Figure 17: Lecture des résultats 45](#_Toc118649369)

[Figure 18 Insulin resistance et insulinopénie 50](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649370)

[Figure 19 – Seringue d’insuline 51](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649371)

[Figure 20 :Resistance a l’ insulin 52](#_Toc118649372)

[Figure 21: Le traitement hygiéno-diététiques 53](#_Toc118649373)

[Figure 22Prévalence (déclaré) du diabète par région ENPSF2018 58](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649374)

[Figure 23 : Statut des mesures antérieures de glycémie par pourcentage par sexe, Steps, Maroc 59](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649375)

[Figure 24: 9 Statut des mesures antérieures de glycémie par pourcentage par Milieu, Steps, Maroc 59](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649376)

[Figure 25: Proportion des personnes prenant des médicaments ou sous 60](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649377)

[Figure 26: Proportion des personnes prenant des médicaments 60](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649378)

[Figure 27 : pourcentage des personnes ayant recours à la médecine 61](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649379)

[Figure 28 : pourcentage des personnes ayant recours à 61](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649380)

[Figure 29 Model performance report (5 folds) logistic Regression 70](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649381)

[Figure 30 Model performance report (5 folds) KNN 71](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649382)

[Figure 31 model performance SVC 72](#_Toc118649383)

[Figure 32 model performance DT 73](#_Toc118649384)

[Figure 33 model performance RF 74](#_Toc118649385)

[Figure 34 model performance Adaboost 75](#_Toc118649386)

[Figure 35 model performance LGBM 76](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649387)

[Figure 36 Aperçu de l’ensemble de données 83](#_Toc118649388)

[Figure 37 Rapport de l’ensemble de données 83](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649389)

[Figure 38 : Analyse des variables de l'ensemble de données 84](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649390)

[Figure 39 table de corrélation 85](#_Toc118649391)

[Figure 40 Matrice de corrélation en couleur 86](#_Toc118649392)

[Figure 41 Répartition des données de train/test 90](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649393)

[Figure 42 Fractionnement de l’ensemble de données 91](#_Toc118649394)

[Figure 43 Les résultats des attributs d ’évaluations pour les diffèrent model 91](#_Toc118649395)

[Figure 44 Meilleur paramètres pour le model Random forest 92](#_Toc118649396)

[Figure 45 Amélioration de précisions du modèle Random forest 92](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649397)

# LISTE DES TABLEAUX :

[Table 1 random forest 29](#_Toc118649398)

[Table 2 Différence entre diabètes de type 1 et 2 50](#_Toc118649399)

[Table 3 Répartition (en %) des individus atteints de diabète, de ceux dont la maladie est confirmée par un médecin et de ceux dont la maladie est confirmée par un médecin et de ceux suivant un traitement régulier selon certaines caractéristiques sociodémographique 57](file:///C:\Users\Bitquark\OneDrive\Bureau\all\PFE\rapport\rapport_pfe.docx#_Toc118649400)

[Table 4 Les résultats de la précision des algorithmes d’étude 01 68](#_Toc118649401)

[Table 5 Description des variables d ’ensemble de données 81](#_Toc118649402)

REMERCIEMENT

Je tiens à remercier dans un premier temps, toute l’équipe pédagogique de *la Faculté des sciences Kenitra* et les intervenants professionnels responsables de la formation *Big Data & Cloud Computing* pour avoir assuré la partie théorique de celle-ci.  
  
Je remercie également Monsieur *Said Tkatek* pour l’aide et les conseils concernant les missions évoquées dans ce rapport, qu’il/elle m’a apporté lors des différents suivis. …..

# RESUME

Au cœur de ce mémoire, nous avons conçu et développe 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….) 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 90% et 87%,

en appliquant les deux méthode d´évaluation train/test et validation croisée.

# 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..) 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 90% and 87%, by applying the two methods of evaluation train /test and cross validation.

# INTRODUCTION GENERALE

# CONTEXTE

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 m´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 m´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 `a trouver un moyen de créer des programmes informatiques qui s’améliorent automatiquement avec l’expérience. A travers ce m´mémoire de Master, nous intéresserons `a 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 d´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 m´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 n´nécessaire au diagnostic.

###### PROBLEMATIQUE

Dans la plupart des cas le diabète est reconnu au stade de complications aiguës qui  
est responsable d'une morbidité et d'une mortalité importante en l’absence de  
traitement approprié et précoce.  
La prise en charge correcte des malades pose d'énormes problèmes s'expliquant par :  
- Le retard de diagnostic ;  
- Le manque de formation du personnel de santé et de spécialistes ;  
- L'insuffisance du plateau technique adapté ;  
- Le coût élevé individuel et collectif de la prise en charge.  
- Le manque d’éducation du diabétique et de son environnement social.

On a estimé le nombre de personnes adultes atteintes de cécité dû à une rétinopathie tardivement diagnostiquée à plus de 93 millions d’où l’importance d’une méthode complète et automatisée, reconnue par la communauté scientifique, pour aider à détecter une cette pathologie de façon précoce.

# SOLUTION PROPOSEE

Notre principal objectif dans ce projet est d'analyser les données que nous avons collectées pour prédire si une personne est diabétique ou non. En outre, nous utiliserons également plusieurs techniques de Machine Learning pour nous aider à atteindre notre objectif. La réalisation de cette analyse est très importante car 9,3 % représente une part significative de la population adulte mondiale. L'observation des tendances et des modèles de données liés au diabète pourrait nous permettre de prédire plus facilement si les gens sont diabétiques ou non.

# STRUCTURE DU RAPPORT

Afin de bien mener ce projet, nous avons structuré le rapport de la manière suivant :

# Chapitre 1 : ’’Machine Learning et le domaine médicale’’

Définit essentiellement le Machine Learning ainsi que le AI dans le domaine médical.

# Chapitre 2 : ‘’LE DIABETE’’

Décrit la diabétique, ainsi que l’organe touché par cette dernière, incluant les études effectuées à but de diagnostiquer le diabète.

# Chapitre3 : “Conception et Réalisation”

Décrit les résultats des tests d’évaluation effectués par rapports aux critères d’évaluation fixés.

# Chapitre 4 : “Démonstration de l’application ”

Présente la solution apportée à la problématique posée.

|  |
| --- |
| Chapitre 1 : ’’Machine Learning et le domaine médicale’’ |

# Introduction

Dans ce chapitre nous allons parler de l'émergence du Machine Learning et du Deep Learning, les raisons de leur apparition ainsi que leur impact, puis on verra un exemple de Random Forest et K-nearset (KNN) et LDA (linear Discriminant Analysis) tout en illustrant son fonctionnement interne dans le domaine médicale .

# Machine Learning:

# Définition et Histoire:

L'apprentissage automatique ou Machine Learning (ML) est une sous-section du domaine de l'intelligence artificielle ou IA en informatique, qui vise à apprendre aux machines à effectuer une tâche sans être explicitement programmé, et cela en utilisant un des algorithmes qu’on appellera modèles et de données.

Le concept d’intelligence artificielle est apparu dans les années 50 dans une assemblée rassemblant toute une flopée de savants célèbres en informatique et en mathématique dont **Alan Turing**. Ce savant de génie a aussi prédit le développement du Machine Learning tel qu’on le connaît.

Le ML a refait surface entre les années 70 à 80, l’idée derrière le concept était de créer des algorithmes ayant la capacité d’accumuler de l'expérience, et de la connaissance à partir de données sans être explicitement programmé pour effectuer cette tâche, et c’est vers la fin des années 80 qu’on a le retour des réseaux de neurones (créé plutôt 1943 par **Walter Pitts** et 9 Warren **McCulloch** inventeurs du premier Perceptron qu’on appellera par la suite **Neurone artificiel**. L’idée était de reproduire un neurone biologique d’un cerveau humain de façon artificielle en utilisant d’opérations mathématiques, malheureusement cette approche fut très limitée dans la résolution des problèmes. [1][2]

Pour qu’une machine apprenne et s’adapte automatiquement, la connaissance qu’elle peut extraire de ses expériences doit être stockée sous une certaine forme, de façon à pouvoir être utilisée dans un but précis. En général, le but est de prédire le futur (un comportement, un nombre, etc.), mais une machine peut également être construite pour générer artificiellement de nouvelles données (machines génératives) ressemblant statistiquement aux données originales, ou bien pour détecter des schémas de fonctionnement (causalités, structure d’un réseau, etc.).

Comme nous l’avons mentionné plus haut, le ML a beaucoup puisé dans les mathématiques et les statistiques. Classiquement, les statistiques s’attachent à permettre une compréhension globale des données : quelles sont les caractéristiques d’une variable (minimum, maximum, moyenne) ? De façon plus avancée, quelle est la distribution statistique ? Peut-on modéliser le processus de génération des données en essayant d’extraire le bruit contenu dans les données ? Le modèle le plus simple est la régression, très utilisée dans de nombreux domaines. Entraîner (faire apprendre) un modèle de régression c’est faire du ML. On peut donc faire remonter le machine learning au XVIIème siècle avec Legendre et Gauss et leur méthode des moindres carrés.

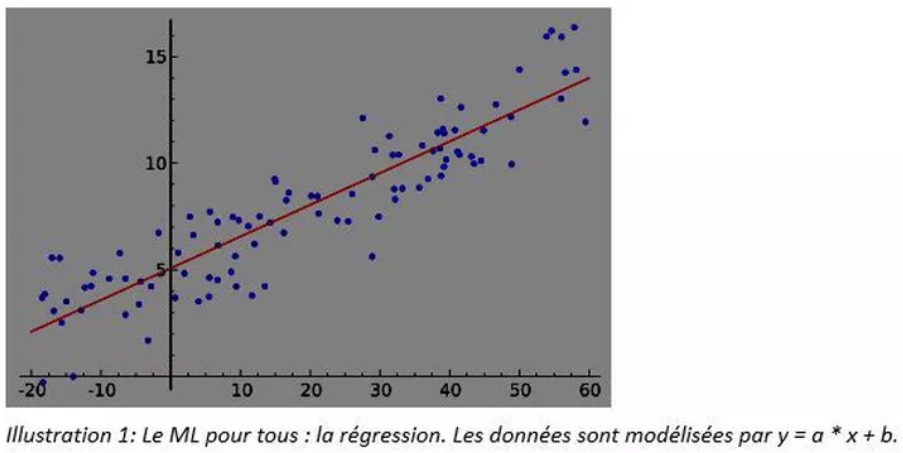


Figure 1 : LE ML pour tous : la régresion.les données sont modélisées par y = a\*x+b.

Vers la fin des années 80, on a vu le retour des réseaux de neurones, avec une réinvention de l’algorithme de **Rétropropagation (Backpropagation),** mais sans succès car le domaine sera finalement laissé à l'abandon faute de capacité de calcul et de stockage suffisantes. Il fallait attendre le milieu des années 2000 pour le grand retour des réseaux de neurones avec le **Deep Learning (3).**

Il faut aussi noter l’importance de l’arrivée du Big Data : la disponibilité d’ensembles de plus variés de données massives et l’augmentation des capacités de calcul et de stockage.

Le développement des technologies Big Data a permis de dépasser des limites importantes du ML : l’absence d’une quantité suffisante de données ne permet pas d’entraîner efficacement un algorithme de prédiction. Par conséquent, en l’absence de telles données, on ne dispose pas d’expérimentations probantes ; il est alors plus difficile de développer les fondements théoriques d’un algorithme de ML.

Le Big Data pose également des challenges d’ordre théorique pour le ML, comme les problèmes de grandes dimensions : le croisement de nombreuses sources de données diluent fortement l’information. On augmente ainsi le nombre de variables à explorer, mais pas le nombre d’exemples disponibles.

# Méthodes d’apprentissage :

Il y a de nombreux sous domaines ou sous sections. On dénote 3 principaux, ses dernières dépendent du type de problèmes que l’on voudrait traiter, on retrouve :

* **Supervised Learning**
* **Unsupervised Learning**
* **Reinforcement Learning.**
* **Semi-supervised Learning**

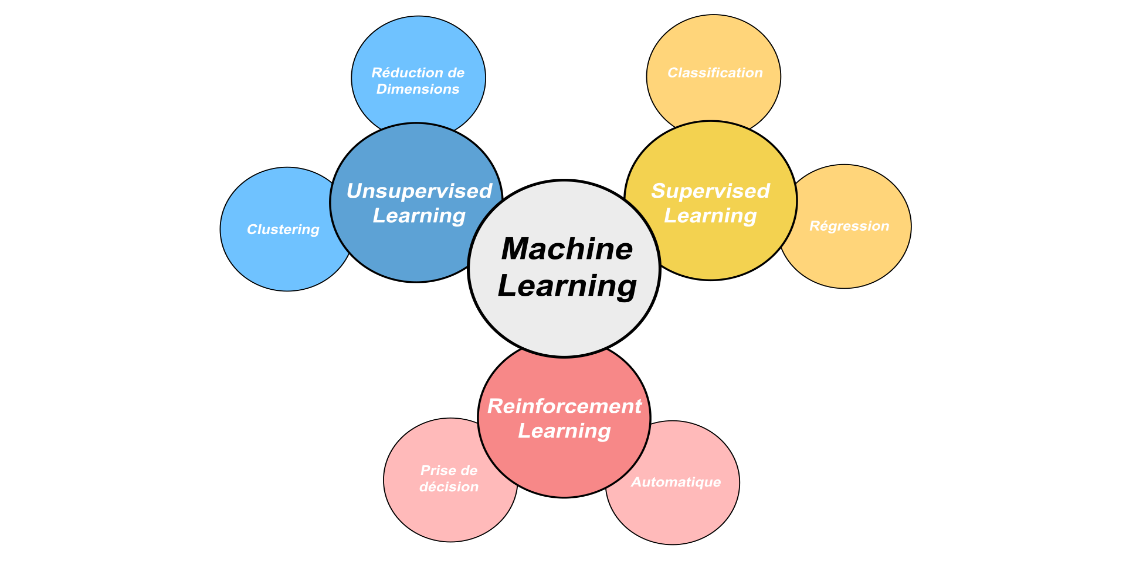
****

Figure 2: type d'apprentissage machine learning

# Apprentissage supervisé :

**L’Apprentissage supervisé** est un ensemble d'algorithmes qui permettent à l'ordinateur d'apprendre à prédire un résultat à partir d'un ensemble de prédicteurs. Le jeu de données doit inclure une variable dépendante aussi appelée variable Y. Il s’agit de la variable que l’ordinateur devra apprendre à prédire. Les autres variables sont les **Prédicteurs**, également appelés **variables X** et utilisées par l'ordinateur pour construire des modèles permettant de prédire Y.

Quand Y est une variable **qualitative** (= catégorielle), nous travaillons sur une problématique de **classification**. Quand Y est quantitative, il s’agit d’une problématique de **régression**.

# Classification :

C’est une tâche d’apprentissage supervisé dont la sortie a des étiquettes définies (valeur discrète) ce qu’on appelle classes ou catégories.

On peut avoir des classifications **Binaire** ou alors **Multi classe**.

La **classification binaire** prédit pour un modèle (0 ou 1) vrai ou faux, dans le cas de **multi classes** le modèle va prédire plus d’une classe.

Le but ici reste le même quel que soit le nombre de classes, **prédire** si **les valeurs discrètes** appartiennent à une classe particulière.

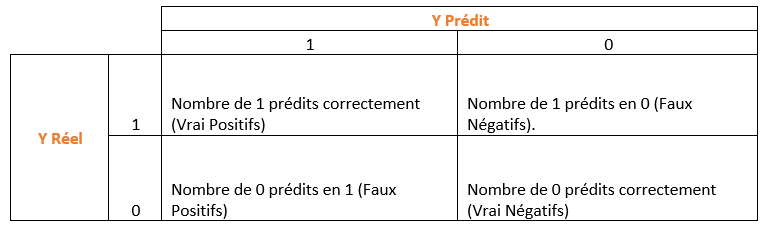


Figure 3Indices de performance pour les problématiques de classification

L’indice le plus intuitif que nous pouvons calculer à partir de cette matrice est la proportion de prédictions correctes (Accuracy en anglais) :

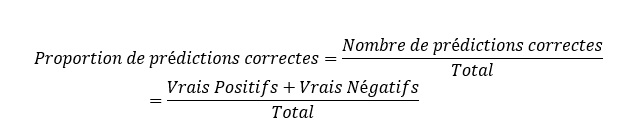


Figure 4l :a proportion de prédictions correctes (Accuracy en anglais)

Cet indice n'est pas fiable en cas de déséquilibre important entre les catégories de Y dans les données. Par exemple, lorsque l'algorithme apprend à détecter une fraude, les données d'entrée contiennent souvent beaucoup plus de cas historiques non frauduleux que de cas frauduleux. D'autres indices peuvent être utilisés en cas de déséquilibre de classe. Par exemple, l'indice d’Aire Sous la Courbe (AUC) calculé à partir d’une [courbe ROC](https://help.xlstat.com/fr/6627-roc-curve-analysis-excel-tutorial).

**Exemple :**

* **E-Commerce**

L'ordinateur apprend à recommander des articles aux clients sur la base de données antérieures, notamment l'historique des achats et les profils d'autres clients.

**Détection de fraude (assurance)**

L'ordinateur apprend à détecter des fraudes à partir d’une base de données de comportement passées, liées à des cas confirmés frauduleux ou non-frauduleux.

#### Médecine

L'ordinateur apprend à prédire la prédisposition de patients à une maladie sur la base de données génétiques associées à des cas de patients confirmés malades ou sains.

#### Web Marketing

L'ordinateur apprend à prédire le risque de désinscription de clients (*churn*) en fonction des données de comportement passées.

#### Industrie

L'ordinateur apprend à prédire une défaillance future sur une chaîne de production en fonction des données de défaillance passées associées à différents signaux. Cela permet aux ingénieurs et techniciens d'intervenir suffisamment tôt sur la chaîne pour éviter la panne. C'est ce qu'on appelle la maintenance prédictive, composante majeure de l'industrie 4.0.

# Régression :

Il s’agit d’une tâche d’apprentissage supervisé dont la sortie a une **valeur continue.**

Le but ici est de prédire une valeur aussi proche de la valeur de sortie réelle que notre modèle le permet, puis l'évaluation est effectuée en calculant la valeur d'erreur. Plus l'erreur est petite, plus la précision de notre modèle de régression est grande.

#### Exemples de problématiques de régression

#### Finance

L'ordinateur apprend à prédire le prix de propriétés immobilières en fonction des caractéristiques et du prix de vente d’autres propriétés vendues récemment dans la région.

# Apprentissage non supervisé :

L'apprentissage non supervisé ou unsupervisé est un type d'algorithme d'apprentissage  
automatique utilisé pour tirer des inférences à partir d'ensembles de données constitués de  
données d'entrée sans réponses étiquetées.

Ce type d’apprentissage est principalement utilisé quand on souhaite étudier un ensemble de données non labellisés

# Clustering

Le clustering permet de séparer les données entrées en un ensemble ou groupe de données  
qui ont des traits similaires et de les affecter à un cluster. Contrairement à la classification  
dans l’apprentissage supervisé ces différents clusters ou groupe ne sont pas connus à l’avance,  
c'est l’algorithme lui-même qui va séparer les données aux nombres de clusters qu’il faut.  
Comme l’on peut le voir pouvez le voir dans l'exemple *Figure* *5*, les points de jeu de  
données ont été divisés en groupes identifiables par les couleurs rouge, vert et bleu.

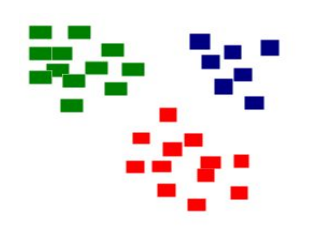


Figure 5 : clustering

* + 1. Différence entre apprentissage supervisé et non supervisé :

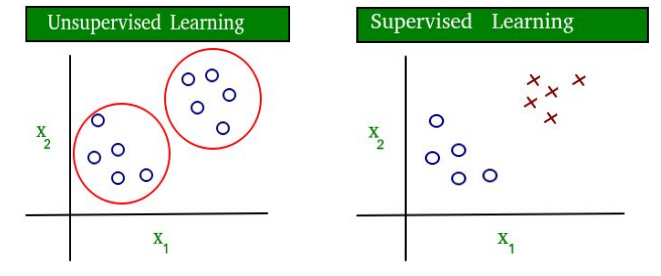
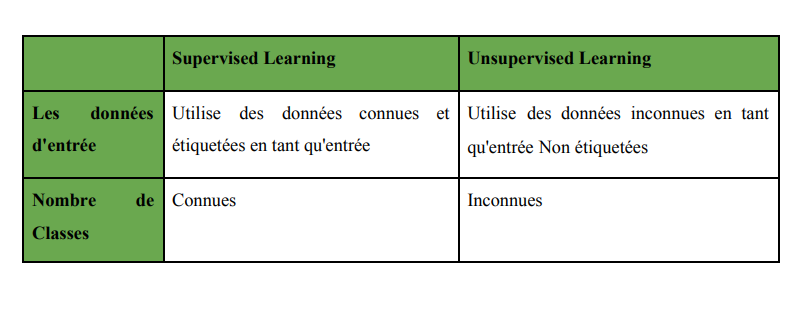
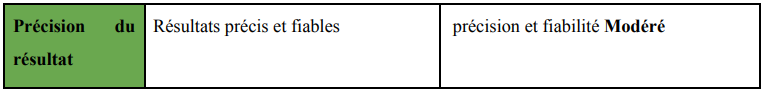
La majeure différence qu’on peut trouver entre ces 2 types d’apprentissages est la disposition des données d’entrée, mais il existe tout de même d’autres différences qu’on peut trouver dans le tableau ci-dessous.[4] Ce tableau résume la différence entre apprentissage supervisé et non supervisé.

Tableau 1 Différence entre apprentissage supervisé et non supervisé

Figure 6: Différence entre apprentissage supervisé et non supervisé

* + 1. Apprentissage semi-supervisé :

Nous avons préalablement vue l’apprentissage supervisé et non supervisé, dont la majeure différence réside dans le fait que les données soient étiquetées ou non, et à cela s’ajoute les méthodes adéquates utilisées pour traiter ses données.

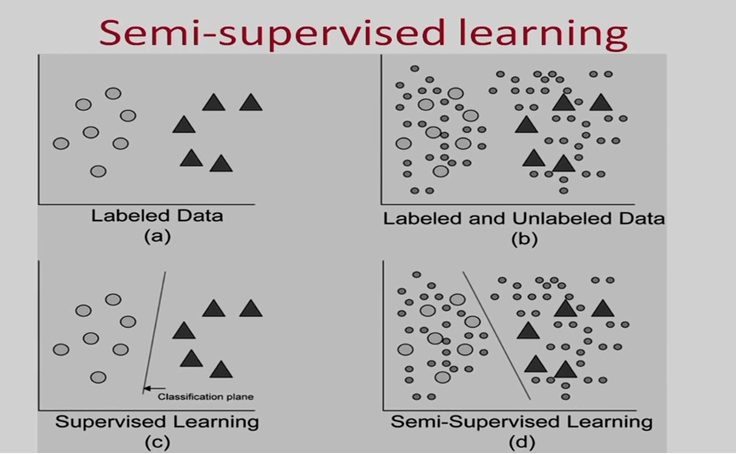


Figure 7: Semi-supervised Learning

L’apprentissage semi-supervisé regroupe ses deux principes, il prend un ensemble réduit de données étiquetées avec un autre ensemble de données non étiquetées du même types. L’avantage de ce type d’apprentissage réside principalement dans le processus d’étiquetage des données prend beaucoup de temps et souvent coûteux. Donc paradoxalement le non étiquetage devient bénéfique pour le processus d’apprentissage, et la construction du modèle et moins coûteuse. [5]

# Apprentissage par Renforcement :

Entraînement par renforcement est une technique dans laquelle on immerge un agent dans un environnement où celui-ci interagit avec son environnement dans le but d’apprendre.

Un agent dans un environnement est muni de capteurs pour capter les informations de son environnement, ainsi que d’actionneurs qui lui permettront d’agir dans son environnement.

Les agents observent l'entrée, puis il effectue une action en prenant des décisions. Une fois l’action réalisées, l'agent reçoit des récompenses en conséquence, ce qui renforce le modèle en stockant ses informations dans une base de données. [6][7].

La récompense peut être positive, négative ou nulle selon les actions effectuées.

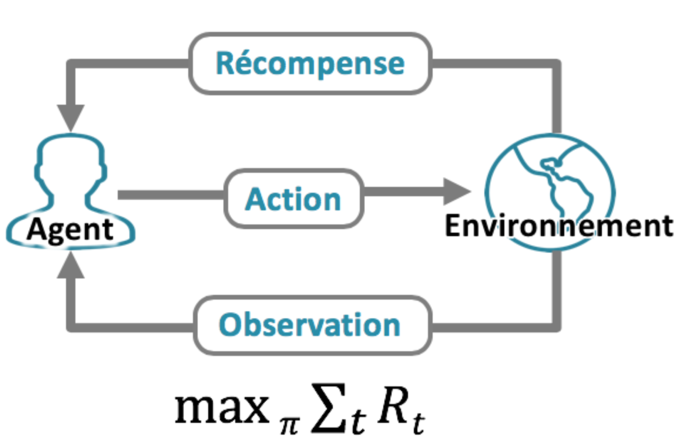


Figure 8 :Apprentissage par Reinforcement:

**Example** : Alpha GO, robot domestique, ou IA dans les jeux vidéo.

# Deep Learning :

Deep Learning ou apprentissage en profondeur ou DL est une branche du Machine Learning entièrement basée sur **des réseaux de neurones artificiels**. [19]

Le concept d'apprentissage en profondeur existe depuis plusieurs années, mais il a été laissé à l'abandon faute de moyens nécessaires.

Dans le milieu des années 2000 le Machine Learning fait rage dans les compétitions de reconnaissance visuelle, en 2012 **Deep Mind** une startup dans le domaine de l’IA arrive dans la compétition avec un algorithme de Deep Learning qui bat largement tous les autres compétiteurs, l’année suivante tous les compétiteurs se sont tournés vers le Deep Learning au vu des résultats obtenus.

L'avancée du DL est dû à l’augmentation en exponentiel qu'ont connu les machines en capacités de **calculs**, et de **stockages**, ainsi que la disponibilité de données de masses (big data), ses 3 ingrédients étaient nécessaires pour exploiter le potentiel du DL qui fût chose impossible dans les années 90.

# Machine Learning vs Deep Learning:

La majeure différence qu’on note entre ses 2 concepts provient de la manière dont les données sont présentées au système (modèle).

* Les algorithmes de ML nécessitent presque toujours des données structurées, alors que les réseaux d’apprentissage approfondis reposent sur des couches de réseaux de neurones artificiels (RNA).
* On voit aussi une différence au sein de l’architecture des modèles qui les composent, on note que les modèles type DL sont plus profond que les modèles type ML.
* Deep Learning n’utilise que les réseaux de neurones, alors que pour le ML les réseaux de neurones sont qu’une approche de conception des modèles parmi tant d’autres.

En considérant le fait que le DL est la prochaine étape de l’évolution du ML inculquant aux machines la manière de prendre leurs décisions de façon précise sans l’intervention de l’expert humain

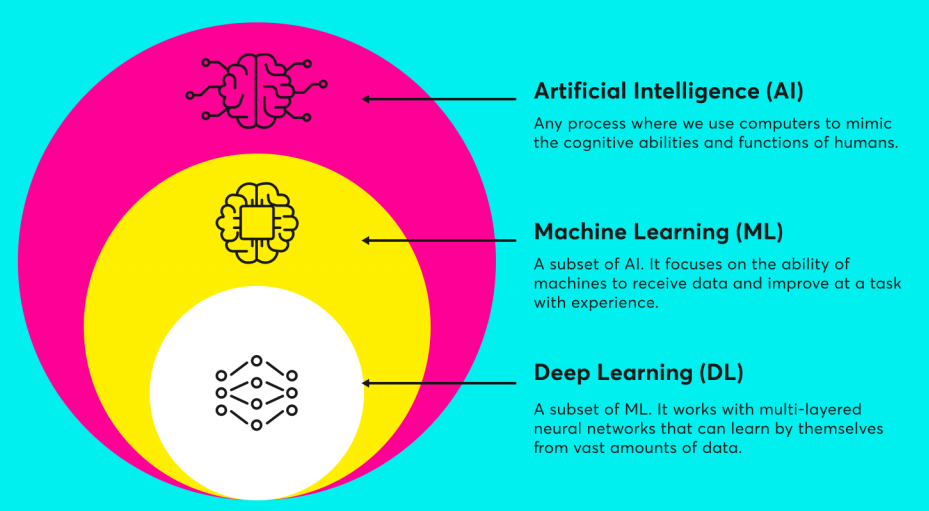


Figure 9: AI & ML & DL

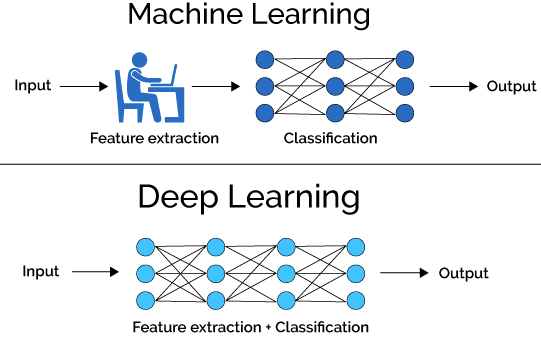


Figure 10: ML vs DL

# Pourquoi Deep Learning :

Il s’agit d’une combinaison de facteurs dont :

* **L’omniprésence des données :** Nous somme dans l’ère de l’informatisation (Internet Of Things), et le propre du Deep-Learning est de tirer parti d’une grande quantité de données pour en estimer une représentation abstraite et en tirer parti.
* **La puissance de calcul** : La théorie des réseaux de neurones existe depuis quelques décennies, mais c’est grâce à la puissance de calcul accessible aujourd’hui qui se démocratise, notamment depuis que les GPUs sont devenus la plateforme de choix pour le Deep-Learning.
* Des **besoins croissants dans le domaine de l’IA** : vision par ordinateur, reconnaissance vocale, traitement du langage, …etc.
* Un **effet de mode**. On a tendance à vouloir appliquer le Deep-Learning partout alors que ça reste un moyen et non une fin. Certains problèmes sont tout à fait solubles par d’autres méthodes d’apprentissage statistique. Cela dit, si beaucoup de gens sont prêts à investir dans le Deep-Learning, il est normal qui deviennent si populaire.
* **Capacité de Stockage :** qui sont devenu beaucoup plus accessible à prix raisonnable.
* Apparition de plateformes et communautés forte encouragent l’évolution de ce domaine, ainsi que sa démocratisation.

# Random Forest :

# Définition :

Le random forest est un algorithme incontournable en Machine Learning. Random forest signifie « forêt aléatoire ». Proposé par Leo Breiman en 2001, c'est un algorithme qui se base sur l’assemblage d’arbres de décision. Il est assez intuitif à comprendre, rapide à entraîner et il produit des résultats généralisables. Seul bémol, le random forest est une boîte noire qui donne des résultats peu lisibles, c’est-à-dire peu explicatifs.

Il est néanmoins possible de limiter cet inconvénient par d’autres techniques de machine Learning ce sera l’objet de notre article N°7 autour de LIME, un algorithme clé pour rendre un modèle explicable.

# Principe de fonctionnement du random forest :

Un random forest est constitué d'un ensemble d'arbres de décision indépendants.

Chaque arbre dispose d'une vision parcellaire du problème du fait d'un double tirage aléatoire :

* Un tirage aléatoire avec remplacement sur les observations (les lignes de votre base de données). Ce processus s'appelle le **tree bagging**.
* Un tirage aléatoire sur les variables (les colonnes de votre base de données). Ce processus s'appelle le **feature sampling**.

A la fin, tous ces arbres de décisions indépendants sont assemblés. La prédiction faite par le random forest pour des données inconnues est alors la moyenne (ou le vote, dans le cas d'un problème de classification) de tous les arbres.

L'idée de base de cet algorithme est assez intuitive. A titre d’exemple, si votre banque vous refuse votre demande de crédit, il y a fort à parier que vous irez consulter une ou plusieurs autres banques. Effectivement, un seul avis ne suffit pas en général pour prendre la meilleure décision.

Le random forest fonctionne sur ce même principe : plutôt que d'avoir un estimateur complexe capable de tout faire, le random forest utilise plusieurs estimateurs simples (de moins bonne qualité individuelle). Chaque estimateur a une vision parcellaire du problème. Ensuite, l'ensemble de ces estimateurs est réuni pour obtenir la vision globale du problème. C'est l'assemblage de tous ces estimateurs qui rend extrêmement performante la prédiction.

# Genèsedel’algorithme

Le défaut majeur de l'arbre de décision est que sa performance est fortement dépendante de l'échantillon de données de départ. Par exemple, l'ajout de quelques nouvelles données dans la base d'apprentissage peut modifier radicalement le modèle et les résultats.

Pour lutter contre ce défaut, on peut utiliser une multitude d’arbres : une forêt d'arbres. Vous comprenez maintenant le terme de forest contenu dans l’anglicisme random forest.

Et vous l’aurez compris, le terme “random” vient du processus de double tirage aléatoire que l’on applique à chaque arbre, à la fois sur les variables et sur les observations.

**\*\*Illustration pratique de l'algorithme**

Table 1 random forest

|  |
| --- |
| Une formula à retenir: **Random Forest = tree bagging + feature sampling.** |

# Tree bagging:

Le Bagging signifie “Bootstrap aggregation”. C'est un processus de tirage aléatoire sur les observations (lignes de données) déterminé par 3 étapes clés :

* Construction de n arbres de décisions en tirant aléatoirement n échantillons d'observations,
* Entraînement de chaque arbre de décision,
* Pour faire une prévision sur de nouvelles données, il faut appliquer chacun de n arbres et prendre la majorité parmi les n prévisions.

# Feature sampling

C'est un processus de tirage aléatoire sur les variables (colonnes de données). Par défaut, on tire Racine n variables pour un problème à n variables au total.

Pour reprendre l'exemple précédent de l’acceptation de crédit, l'idée de base du feature sampling c'est de demander à chaque banque d’étudier votre demande de prêt à partir d'un accès limité aux informations du client. L'une des banques rendra sa décision en ayant, par exemple, uniquement accès aux informations relatifs à l’âge, à la CSP et au revenu annuel du client. Une autre banque quant à elle, aura uniquement pris connaissance des informations relatives à la situation maritale, au sexe et au nombre de crédits en cours du client, ...

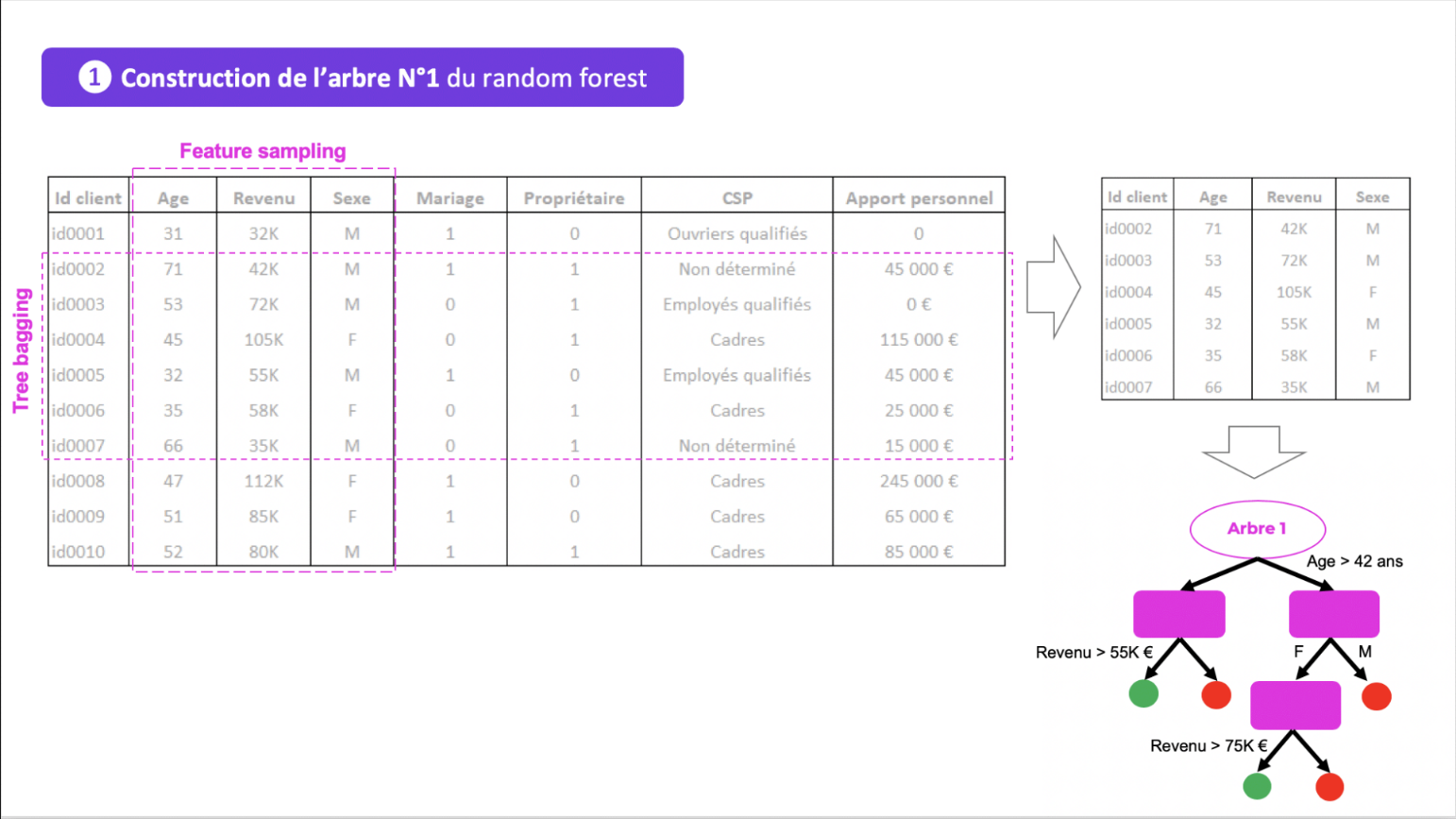
Ce processus permet de baisser la corrélation entre les arbres qui pourrait perturber la qualité des résultats. En statistique, on dit que le feature sampling permet de réduire la variance de l'ensemble créé.

Figure 11: Construction de l'arbre N°1 du random forest

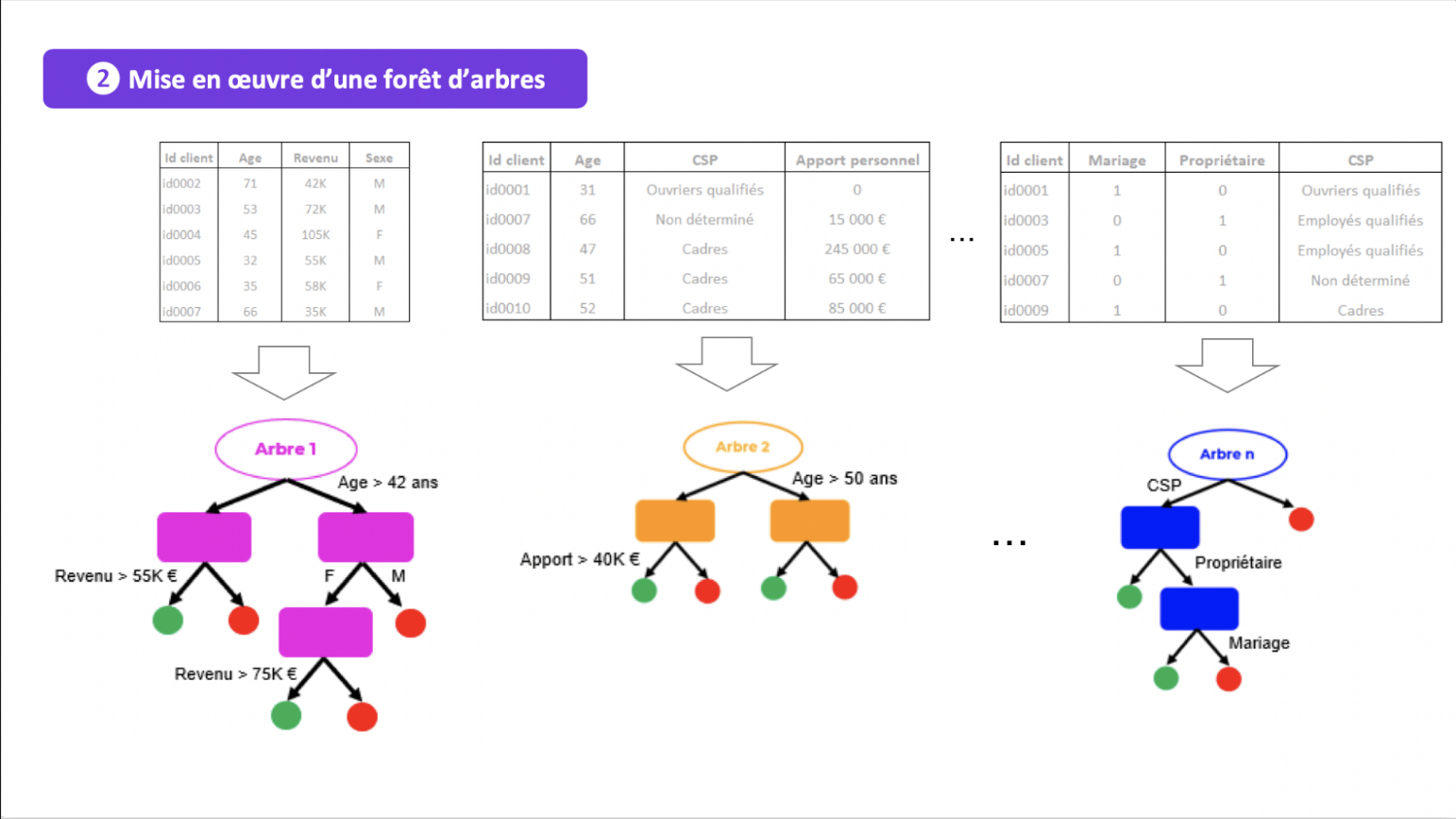
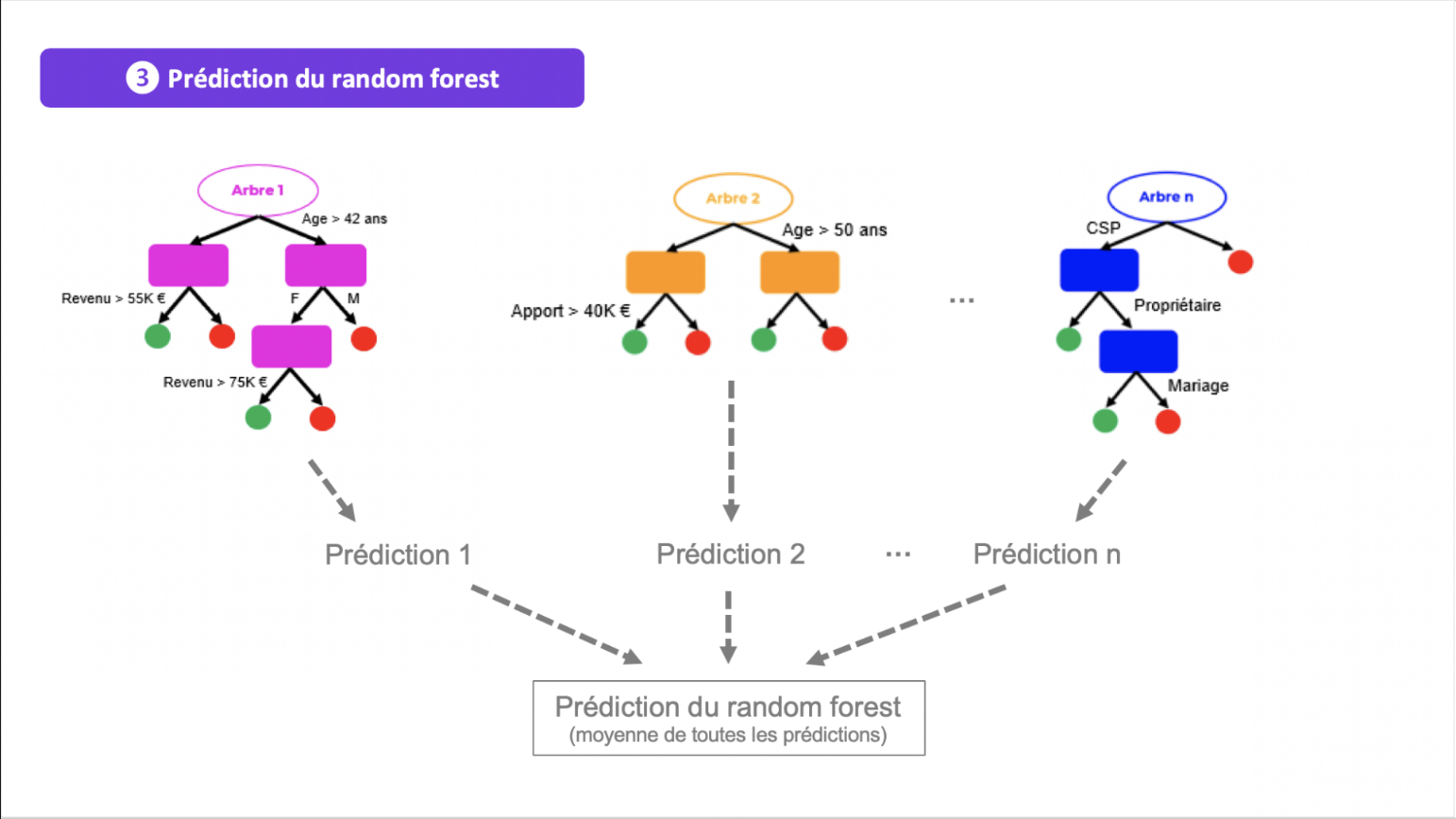


Figure 12: Prediction du random forest

Figure 13 : Mis en œuvre d’une forêt d'arbres

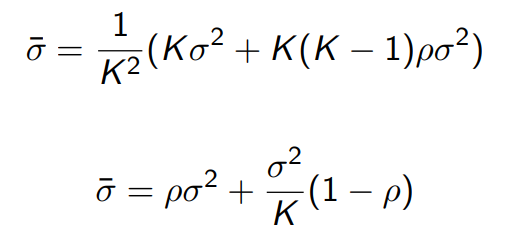
* 1. Critère de division/split

Vous le savez un arbre de décisions construit des sous-populations par séparations successives des feuilles d'un arbre.

Il existe différents critères de séparation pour construire un arbre :

* Le **critère de Gini** organise la séparation des feuilles d'un arbre en se focalisant sur la classe la plus représentée dans le jeu de données : il faut la séparer le plus rapidement possible.
* Le **critère d'entropie** est basé sur la mesure du désordre (comme en thermodynamique) qui règne dans la population étudiée. La construction de l'arbre vise à baisser l'entropie globale des feuilles de l’arbre à chaque étape.

# Random Forest : intuition

Si arbres sont identiquement distribués, de variance , avec un coefficient de corrélation deux à deux ρ la variance de leur moyenne est alors :

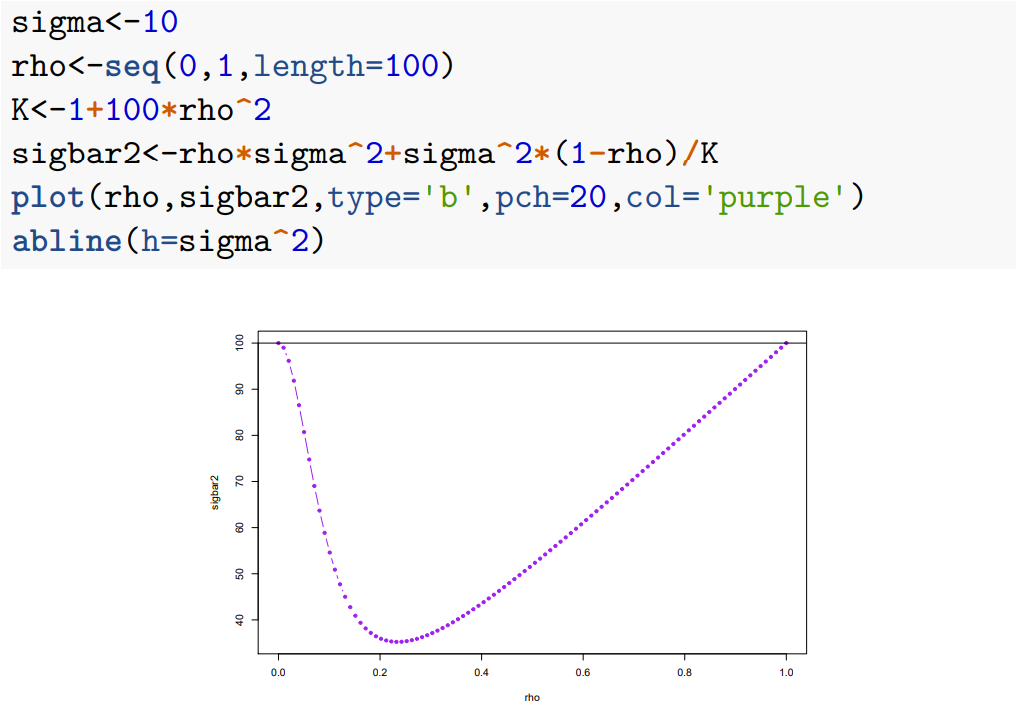


Figure 14 : Random forest illustration

# Avantages et inconvénients des random-forests

#### Avantages

* Pas de sur-apprentissage
* En général : meilleure performance que les arbres de décision, calcul de  
  l’erreur “Out-of-Bag” direct
* Paramètres faciles à calibrer
* Parallélisation possible
* Souvent utilisées comme benchmark dans les compétitions de machine Learning

#### Inconvénients

* Boite noire : difficilement interprétable, difficilement améliorable
* Entrainement plus lent

Les random forest fonctionnent tout le temps bien mais excellent plus rarement même si les forêts sont robustes à un nombre de variable important, en grande dimension (n << p) il est nécessaire d’effectuer de la sélection de variable.

# K-nearset Neighbors (KNN) :

* 1. Définition :

L’algorithme des K plus proches voisins ou K-nearest Neighbors (kNN) est un algorithme de Machine Learning qui appartient à la classe des algorithmes d’apprentissage supervisé simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de classification et de régression. Dans cet article, nous allons revenir sur la définition de cet algorithme, son fonctionnement ainsi qu’une application directe en programmation.

L’intuition derrière **l’algorithme des K** plus proches voisins est l’une des plus simples de tous les algorithmes de Machine Learning supervisé :

* **Etape1 :** Sélectionnez le nombre K de voisins
* **Etape 2 :** Calculez la distance

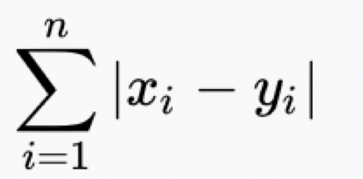


Figure 15: Euclidian

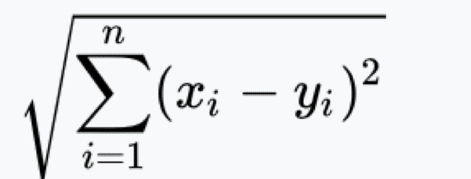
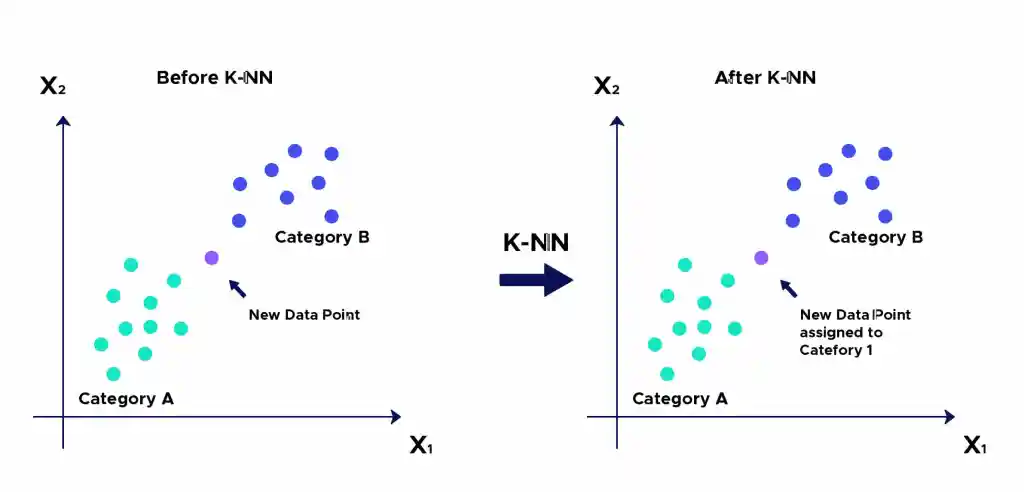
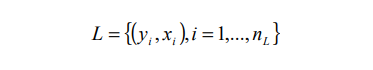


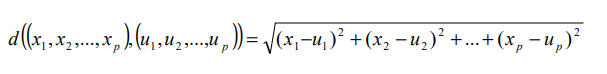
Figure 16: Manhattan

Du point non classifié aux autres points.

* **Étape 3** : Prenez les K voisins les plus proches selon la distance calculée.
* **Étape 4** : Parmi ces K voisins, comptez le nombre de points appartenant à chaque catégorie.
* **Étape 5** : Attribuez le nouveau point à la catégorie la plus présente parmis ces K voisins.
* **Étape 6** : Notre modèle est prêt :

La méthode du plus proche voisin est une méthode non paramétrique où une nouvelle observation est classée dans la classe d’appartenance de l’observation de l’échantillon d’apprentissage qui lui est la plus proche, au regard des Co variables utilisées. La détermination de leur similarité est basée sur des mesures de distance.

Formellement, soit L l’ensemble de données à disposition ou échantillon d’apprentissage :

où yi ∈{1,...,c} dénote la classe de l’individu i et le vecteur ) xi =(,.., ou x représente les variables prédicatrices de l’individu i. La détermination du plus proche voisin est basée sur un fonction distance arbitraire d (.,.). La distance euclidienne ou dissimilarité entre deux individus caractérisés par p Co variables est définie par :

# Applications de k-NN dans l'apprentissage automatique :

L'algorithme k-NN a été utilisé dans une variété d'applications, principalement dans le cadre de la classification. Certains de ces cas d'utilisation incluent :

* **Pré-traitement des données** : les jeux de données ont souvent des valeurs manquantes, mais l'algorithme KNN peut estimer ces valeurs dans un processus connu sous le nom d'imputation des données manquantes.
* **Moteurs de recommandation** : en utilisant les données de parcours de navigation des sites Web, l'algorithme KNN a été utilisé pour fournir des recommandations automatiques aux utilisateurs sur du contenu supplémentaire. Cette [recherche](https://www.researchgate.net/publication/267572060_Automated_Web_Usage_Data_Mining_and_Recommendation_System_using_K-Nearest_Neighbor_KNN_Classification_Method) (lien externe à ibm.com) montre qu'un utilisateur est affecté à un groupe particulier et, en fonction du comportement de l'utilisateur de ce groupe, il reçoit une recommandation. Cependant, étant donné les problèmes de mise à l'échelle avec KNN, cette approche peut ne pas être optimale pour les jeux de données plus volumineux.
* **Finance** : il a également été utilisé dans une variété de cas d'utilisation financière et économique. Par exemple, un [article](https://iopscience.iop.org/article/10.1088/1742-6596/1025/1/012114/pdf) (PDF, 391 Ko)  (lien externe à ibm.com) montre comment l'utilisation de KNN sur les données de crédit peut aider les banques à évaluer le risque d'un prêt à une entreprise ou à un individu. Il est utilisé pour déterminer la solvabilité d'un demandeur de prêt. Un autre [journal](https://www.ijera.com/papers/Vol3_issue5/DI35605610.pdf) (PDF, 447 Ko) (lien externe à ibm.com) met en évidence son utilisation dans les prévisions boursières, le change de devises, les contrats à terme et les analyses de blanchiment d'argent.
* **Soins de santé** : KNN a également eu des applications dans l'industrie des soins de santé, en faisant des prédictions sur le risque de crises cardiaques et de cancer de la prostate. L'algorithme fonctionne en calculant les expressions géniques les plus probables.
* **Reconnaissance de modèle** : KNN a également aidé à identifier des modèles, tels que la [classification de texte et de caractères](https://www.researchgate.net/profile/D-Adu-Gyamfi/publication/332880911_Improved_Handwritten_Digit_Recognition_using_Quantum_K-Nearest_Neighbor_Algorithm/links/5d77dca692851cacdb30c14d/Improved-Handwritten-Digit-Recognition-using-Quantum-K-Nearest-Neighbor-Algorithm.pdf) (lien externe à ibm.com). Cela a été particulièrement utile pour identifier les numéros manuscrits que vous pourriez trouver sur des formulaires ou des enveloppes postales.
  + fonctionnalités supplémentaires augmentent le nombre d'erreurs de classification, en particulier lorsque la taille d'échantillon est plus petite.
  + **Enclin au surajustement** : en raison de la « malédiction de la dimensionnalité », KNN est également plus enclin au surajustement. Bien que les techniques de sélection des fonctions et de réduction de la dimensionnalité soient utilisées pour éviter que cela ne se produise, la valeur de k peut également avoir un impact sur le comportement du modèle. Des valeurs inférieures de k peuvent sur ajuster les données, tandis que des valeurs plus élevées de k ont tendance à « lisser » les valeurs de prédiction, car il calcule la moyenne des valeurs sur une plus grande zone ou sur le voisinage. Cependant, si la valeur de k est trop élevée, il peut sous-ajuster les données.

# Avantages et inconvénients de l'algorithme KNN :

Comme tout algorithme d'apprentissage automatique, k-NN a ses forces et ses faiblesses. Selon le projet et l'application, il peut s'avérer être le bon choix ou le mauvais.

* **Avantages**
  + **Facile à implémenter :** compte tenu de la simplicité et de la précision de l'algorithme, c'est l'un des premiers discriminants qu'un nouveau spécialiste des données apprendra.
  + **Il s'adapte facilement** : à mesure que de nouveaux échantillons d'apprentissage sont ajoutés, l'algorithme s'ajuste pour tenir compte de toute nouvelle donnée, toutes les données d'apprentissage étant stockées en mémoire.
  + **Peu d'hyperparamètres :** KNN ne nécessite qu'une valeur k et une mesure de distance, ce qui est peu par rapport aux autres algorithmes d'apprentissage automatique.
* **Inconvénients**
  + **Mise à l'échelle difficile :** puisque KNN est un algorithme paresseux, il utilise plus de mémoire et de stockage de données par rapport aux autres discriminants. Cela peut être coûteux en termes de temps et d'argent. Davantage de mémoire et de stockage augmenteront les dépenses de l'entreprise et plus de données peuvent prendre plus de temps à calculer. Alors que différentes structures de données, telles que Ball-Tree, ont été créées pour remédier aux inefficacités de calcul, un discriminant différent peut s'avérer idéal en fonction de la problématique métier.
  + **Malédiction de la dimensionnalité** : l'algorithme KNN a tendance à être victime de la malédiction de la dimensionnalité, ce qui signifie qu'il fonctionne mal avec des entrées de données de grande dimension. Cette dernière est aussi parfois appelée le [phénomène de peaking](https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.418.6517&rep=rep1&type=pdf) (PDF, 340 Mo) (lien externe à ibm.com), où une fois que l'algorithme a atteint le nombre optimal de fonctions, des

# LDA (linear Discriminant Analysis) :

# Définition :

En anticipant sur la présentation des méthodes de régularisation, nous avons choisi de présenter ici l’Analyse Discriminante Linéaire (LDA en anglais) car elle est généralement considérée comme une méthode d’analyse discriminante à part entière plutôt que comme une méthode de régularisation de QDA. Pourtant, LDA est une régularisation de QDA car elle fait, par rapport à QDA, l’hypothèse supplémentaire d’égalité des matrices de variances . L’analyse discriminante linéaire doit également son nom au fait qu’elle réalise des séparations linéaires entre les classes

# Hypothèses et Formules :

Nous disposons d’un échantillon de observations réparties dans groupes d’effectifs .

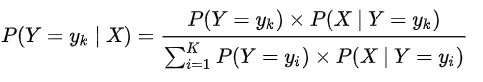
Notons la variable à prédire, elle prend ses valeurs dans l’ensemble .Des classes. Nous disposons de variables prédictives .

Nous notons  {\displaystyle \mu \_{k}} les centres de gravité des nuages de points conditionnels et  {\displaystyle W\_{k}} leurs [matrice de variance-covariance](https://fr.wikipedia.org/wiki/Matrice_de_variance-covariance).

# La règle bayésienne :

L’objectif est de produire une règle d’affectation {\displaystyle X(\omega )\mapsto Y(\omega )} qui permet de prédire, pour une observation {\displaystyle \omega } donnée, sa valeur associée de Y à partir des valeurs prises par X.

La règle bayésienne consiste à produire une estimation de la probabilité *a posteriori* d’affectation

) est la probabilité *a priori* d’appartenance à une classe.) représente la fonction de densité des conditionnellement à la classe .

La règle d’affectation pour un individu  à classer devient alors)

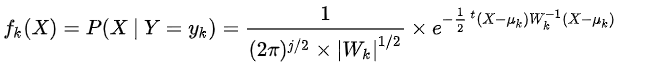
 Toute la problématique de l’analyse discriminante revient alors à proposer une estimation de la quantité ).{\displaystyle P(X~|~Y=y\_{k})}

# L'analyse discriminante paramétrique - L'hypothèse de multi normalité

On distingue principalement deux approches pour estimer correctement la distribution ).

* L’approche non-paramétrique n’effectue aucune hypothèse sur cette distribution mais propose une procédure d’estimation locale des probabilités, au voisinage de l’observation à classer. Les procédures les plus connues sont la méthode d'estimation par noyau et la méthode des plus proches voisins. La principale difficulté est de définir de manière adéquate le voisinage.
* La seconde approche effectue une hypothèse sur la distribution des nuages de points conditionnels, on parle dans ce cas **d’analyse discriminante paramétrique.** L’hypothèse la plus communément utilisée est sans aucun doute l’hypothèse de multi normalité.

Dans le cas de la loi normale multidimensionnelle, la distribution des nuages de points conditionnels s’écrit :



Où représente le déterminant de la matrice de variance covariance conditionnellement à .

L’objectif étant de déterminer le maximum de la probabilité a posteriori d’affectation, nous pouvons négliger tout ce qui ne dépend pas de . En appliquant le logarithme à la relation de Bayes, nous obtenons le score discriminant proportionnel à .



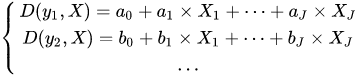
La règle d’affectation devient donc .

Si l’on développe complètement le score discriminant, nous constatons qu’il s’exprime en fonction du carré et du produit croisé entre les variables prédictives. On parle alors d’analyse discriminante quadratique. Très utilisée en recherche car elle se comporte très bien, en matière de performances, par rapport aux autres méthodes, elle est moins répandue auprès des praticiens. En effet, l’expression du score discriminant étant assez complexe, il est difficile de discerner clairement le sens de la causalité entre les variables prédictives et la classe d’appartenance. Il est notamment malaisé de distinguer les variables réellement déterminantes dans le classement, l’interprétation des résultats est assez périlleuse.

# Fonction de classement linéaire

En développant l’expression du score discriminant après introduction de l’hypothèse d’homoscédasticité, on constate qu’elle s’exprime linéairement par rapport aux variables prédictives.

Nous disposons donc d’autant de fonctions de classement que de modalités de la variable à prédire, ce sont des combinaisons linéaires de la forme suivante :



Cette présentation est séduisante à plus d’un titre. Il est possible, en étudiant la valeur et le signe des coefficients, de déterminer le sens des causalités dans le classement. De même, il devient possible, comme nous le verrons plus loin, d’évaluer le rôle significatif des variables dans la prédiction.

# Robustesse :

Les hypothèses de multi normalité et d’homoscédasticité peuvent sembler trop contraignantes, restreignant la portée de l’analyse discriminante linéaire dans la pratique.

La notion clé qu’il faut retenir en statistique est la notion de robustesse. Même si les hypothèses de départ ne sont pas trop respectées, une méthode peut quand même s’appliquer. C’est le cas de l’analyse discriminante linéaire. Le plus important est de le considérer comme un séparateur linéaire. Dans ce cas, si les nuages de points sont séparables linéairement dans l’espace de représentation, elle peut fonctionner correctement.

Par rapport aux autres techniques linéaires telles que la régression logistique, l’analyse discriminante présente des performances comparables. Elle peut être lésée néanmoins lorsque l’hypothèse d’homoscédasticité est très fortement violée.

# Taux d’erreur :

De manière classique en apprentissage supervisé, pour évaluer les performances d'une fonction de classement, nous confrontons ses prédictions avec les vraies valeurs de la variable à prédire sur un fichier de données. Le tableau croisé qui en résulte s’appelle une matrice de confusion avec : en ligne les vraies classes d’appartenance, en colonne les classes d’appartenance prédites. Le taux d’erreur ou taux de mauvais classement est tout simplement le nombre de mauvais classement, lorsque la prédiction ne coïncide pas avec la vraie valeur, rapporté à l’effectif du fichier de données.

Le taux d’erreur a de séduisant qu’il est d’interprétation aisée, il s’agit d’un estimateur de la probabilité de se tromper si l’on applique la fonction de classement dans la population.

Attention cependant, on parle de taux biaisé ou taux d'erreur en resubstituions, le taux d’erreur mesuré sur les données qui ont servi à construire la fonction de classement. Tout simplement parce que les données sont juges et parties dans ce schéma. La bonne procédure serait de construire la fonction de classement sur une fraction des données, dites d'apprentissage ; puis de l’évaluer sur une autre fraction de données, dite de test. Le taux d’erreur en test ainsi mesuré est un indicateur digne de foi.

La pratique veut que la répartition des données en apprentissage et test soit de 2/3 – 1/3. Mais en réalité, il n’y a pas de règle véritable. Le plus important est de concilier deux exigences contradictoires : en avoir suffisamment en test pour obtenir une estimation stable de l’erreur, tout en réservant suffisamment en apprentissage pour ne pas pénaliser la méthode d’apprentissage.

Lorsque les effectifs sont faibles, et que le partage apprentissage-test des données n’est pas possible, il existe des méthodes de rééchantillonnage telles que la validation croisée ou le Bootstrap pour évaluer l’erreur de classement.

# Séparabilité - Évaluation globale :

Le taux d’erreur permet d’évaluer et de comparer des méthodes, quelles que soient leurs hypothèses sous-jacentes. Dans le cas de l’analyse discriminante linéaire, nous pouvons exploiter le modèle probabiliste pour réaliser des tests d’hypothèses.

Un premier test permet de répondre à la question suivante : est-il possible de discerner les nuages de points dans l’espace de représentation. Rapporté dans le cadre multi normal, cela revient à vérifier si les centres de gravité conditionnels sont confondus (hypothèse nulle) ou si un au moins de ces centres de gravité s’écarte significativement des autres (hypothèse alternative).

La statistique du test est le de Wilks, son expression est la suivante :



où représente le déterminant de la matrice de variance covariance infra-classes, le déterminant de la matrice de variance covariance globale.

La table des valeurs critiques de la loi de Wilks étant rarement disponible dans les logiciels, on utilise couramment les transformations de Bartlett et de Rao qui suivent respectivement une loi du KHI-2 et de Fisher.

Avec un prisme différent, nous constatons que ce test peut s’exprimer comme une généralisation multidimensionnelle de l’analyse de variance à un facteur (**ANOVA**), on parle dans ce cas de **MANOVA** (Multidimensionnel Analysis of Variance).

# Évaluation individuelle des variables prédictives :

Comme dans toutes les méthodes linéaires, il est possible d’évaluer individuellement chaque variable prédictive, et éventuellement d’éliminer celles qui ne sont pas significatives dans la discrimination.

La statistique du test s’appuie sur la variation du Lambda de Wilks lors de l'adjonction de la (J+1)-ième variable dans le modèle de prédiction. Sa formule est la suivante :



Elle suit une loi de Fisher à degrés de liberté.

# Lecture des résultats

Une analyse discriminante linéaire a été lancée sur les Flea Beetles décrites dans l'article analyse discriminante. Les résultats sont les suivants.

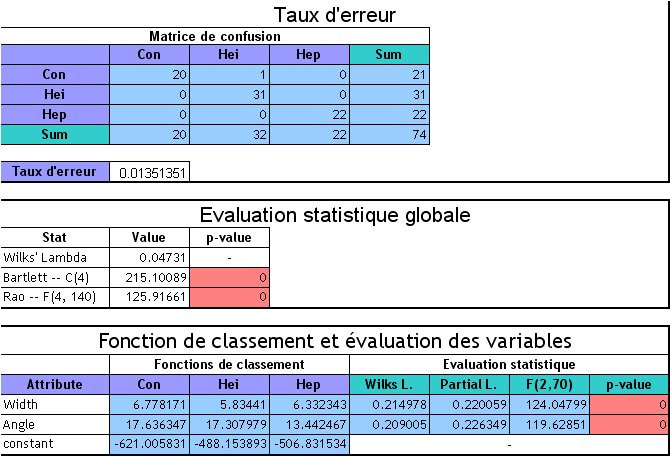


Figure 17: Lecture des résultats

* La matrice de confusion indique qu'une seule erreur a été commise, un « Concinna » a été classé en « Heikertingeri ». Le taux d'erreur associé est de 1,35 %. Ce résultat est à relativiser, il a été établi sur les données ayant servi à l'apprentissage.
* Les centres de gravité des trois nuages de points s'écartent significativement. C'est ce que nous indique la statistique de Wilks dans la section MANOVA. Les probabilités critiques associées, transformation de Bartlett et de Rao, sont proches de 0. Ce résultat numérique confirme l'impression visuelle laissée par la projection des nuages de points dans l'espace de représentation (voir Analyse discriminante).
* La variable à prédire comportant 3 modalités, nous obtenons 3 fonctions de classement linéaires. L'évaluation individuelle des variables dans la discrimination indique qu'elles sont toutes les deux très significatives (p-value proches de 0).

# Avantages et inconvénients de LDA :

Parmi les avantages engendrés par une utilisation de la méthode LDA, nous citons les points suivants :

* LDA maximise l'éparpillement interclasses.
* Réduit l'éparpillement infra-classes.
* La méthode de Fisher faces résout le problème de la robustesse face aux variations de pose, et d'expressions faciales.

Malgré ces avantages, dans la littérature un ensemble de pointes négatives existe encore comme :

* Couteuse en temps de calcul.
* Couteuse en espace mémoire.
* Rend de mauvais résultats quand le nombre d'images d'apprentissage est grand.

Nous avons vu que la LDA est une méthode de classification qui s'intéresse à séparer les classes, en observant les variations intra classes et les variations inter classes. Dans ce chapitre on a expliqué tous les points qui seront utile pour la compréhension de l'algorithme, pour une meilleure implémentation, en fait nous avons expliquées le fondement mathématique de la LDA ; et son utilisation pour la reconnaissance de visage avec exposition de quelques avantages et inconvénients de cette méthode.

# **Chapitre 2 : ‘’LE DIABETE ‘’**

# Définition :

Le diabète, également connu sous le nom de tueur silencieux, est l'une des maladies les plus mortelles du 21e siècle. Des maladies les plus mortelles du 21ème siècle qui provoque une élévation du taux de sucre dans le sang. Il joue un rôle majeur dans le développement de maladies telles que l'insuffisance rénale, la cécité, l'amputation d'un membre, et des maladies cardiaques. Sur une période de 16 ans, c'est-à-dire de 2000 à 2016, on a constaté une augmentation de 5 % du nombre de cas de diabète de type 2. 2016, il y a eu une augmentation de 5% des décès prématurés qui sont directement attribuées au diabète. En 2019, pas moins de 1,5 million de de décès ont été directement causés par le diabète, ce qui lui a valu l'étiquette de l'étiquette de tueur silencieux. Dans des circonstances normales, un patient souhaiterait se rendre dans un centre de soins de santé et qu'un médecin décide de la ligne de traitement à suivre. Cependant, l'analyse expérimentale que j'ai mise en place cherche à aider les patients qui se trouvent dans des endroits éloignés, sans aucune de ressources médicales immédiates à leur disposition. Cette application web a pour but d'être une interface conviviale pour tout individu ayant des données médicales valides, pour vérifier la présence de diabète. Des données médicales valides, pour vérifier la présence de diabète. J'ai implémenté 3 classificateurs différents avec un réglage hyperparamètres appropriés pour améliorer les niveaux de précision niveaux de précision précédemment atteints.

#### 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 secrétée par le pancréas. L’insuline agit donc comme une clé permettant au glucose de passer du sang aux cellules de notre corps.

# 

Figure 18 Insulin resistance et insulinopénie

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

Table 2 Différence entre diabètes de type 1 et 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Diabète de type 1 |  | Diabète de type 2 |
| Fréquence | 10-15% | | 85-90% |
| Age de début | Avant 30 ans | | Après 40 ans |
| Surpoids | Absent | | Présent |
| Physiopathologie | Insulinopénie totale | | Insulinorésistance |
| Antécédents familiaux | Rares | | Fréquents |
| Mode de début | Brutal | | Progressif |
| Cétose | Fréquente | | Très rare |

# Classification :

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

# 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.

# 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.

# 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.

Figure 19 – Seringue d’insuline

# 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 caractérisé par un

taux de sucre (glucose) trop élève 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 ilots 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.



Figure 20 :Resistance a l’ insulin

# 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.

# 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.



Figure 21: Le traitement hygiéno-diététiques

# Diagnostic du diabète :

Le diagnostic du diabète est basé sur :

* La glycémie à jeun (au moins 8heures de jeûne) supérieure ou égale à 1,26g/l (soit 7mmol/l) à 2 reprises.
* Ou une glycémie supérieure ou égale à 2g/l (soit 11mmol/l) à n’importe quel moment de la journée.
* L’hyperglycémie provoquée par voie orale n’est plus jugée utile dans le diagnostic du diabète et reste réservée au diagnostic du diabète gestationnel et de l’intolérance aux hydrates de carbone.
* Le dosage de l’HbA1c aussi ne trouve pas d’indication dans le diagnostic du diabète.

# Les complications du diabète :

# Les complications aigües du diabète :

Elles sont essentiellement des complications métaboliques ; on distingue donc :

* L’hypoglycémie : très fréquente dans le diabétique de type 1, et chez le diabétique de type 2 sous insuline ou sous sulfamides hypoglycémiants. C’est une complication réversible sans séquelle et non mortelle sauf en cas d’intoxication alcoolique chez le diabétique.
* La cétoacidose est une complication relativement fréquente du diabète de type 1, les enfants et les adolescents sont les plus concernés.
* Le coma hyperosmolaire est une forme grave de la décompensation du diabète sucré, son pronostic reste toujours grave en tenant compte de l’âge, du terrain fragilisé et surtout de pathologies associées.
* L’acidose lactique est une complication plus rare mais plus grave que le coma hyperosmolaire ; le traitement par les biguanides constitue sa principale étiologie.

Ces 2 dernières complications concernent essentiellement le diabétique de type2.

# Le traitement pharmacologique :

Ils sont utilisés dans le diabète de type 2 lorsque les mesures hygiéno-diététiques se sont révélées insuffisantes. On distingue : les insulines o sensibilisateurs (biguanides et glitazones), les insulines o sécréteurs (sulfamides hypoglycémiants, glinides), les incrétines, et les inhibiteurs de l’absorption digestive du glucose (acarbose).

* Les biguanides représentés essentiellement par la Metformine peuvent être utilisés en première intention chez le diabétique de type 2 obèse ; son effet secondaire le plus redoutable est l’acidose lactique, ce qui impose le respect des contre-indications liées à sa prescription, essentiellement l’insuffisance rénale, l’insuffisance hépatique, l’injection de produits de contraste iodés et toute situation de décompensation aigüe.
* Les glitazones représentées par les thiazolinediones (rosiglitazone et pioglitazone) ont une prescription limitée (échec de la metformine, intolérance ou contre-indication de la metformine), leurs principaux effets secondaires sont : l’hépatotoxicité, la rétention hydrosodée avec prise de poids.
* Les sulfamides sont très nombreux : glimépiride, gliclazide, glibenclamide, gliquidone. Leur utilisation nécessite que le pancréas endocrine soit encore fonctionnel, leur utilisation doit être prudente chez le sujet âgé ; leur effet secondaire principal est l’hypoglycémie.
* Les glinides sont représentés par le répaglinide ; ils agissent surtout sur la glycémie post-prandiale ; leur durée d’action est plus courte et les effets secondaires sont plus rares.
* Les inhibiteurs de l’alpha glucosidase sont représentés par l’acarbose, leur utilisation reste assez limitée du fait de leur puissance inférieure par rapport aux autres classes et de la fréquence d’apparition de troubles digestifs

# Le suivi du diabète :

Le bilan recommandé au moment de la découverte du diabète (pour le diabète de type 2) et chaque année est le suivant :

* Prise du poids, de la pression artérielle tous les 3-4 mois.
* HbA1c tous les 3-4mois.
* Incitation à l’arrêt du tabac.
* Interrogatoire à la recherche de symptômes en faveur d’une atteinte cardio-vasculaire ou neurologique.
* Examen des pieds : état cutané, utilisation du mon filament en nylon et/ou diapason, réflexes ostéo-tendineu.
* Palpation des pouls, recherche de souffles abdominaux, fémoraux et carotidiens.
* Recherche d’une hypotension orthostatique.
* Examen de la bouche, de la sphère ORL, de la peau.
* Examen par un ophtalmologiste (acuité visuelle, tension oculaire, fond d’œil).
* ECG.
* Bilan lipidique à jeun : CT, TG, HDL et LDL-CT.
* Créatininémie et calcul de la clairance par la formule de Cockcroft.
* Protéinurie, microalbuminurie, hématurie, recherche d’infection urinaire.

# Le diabète au niveau de Maroc :

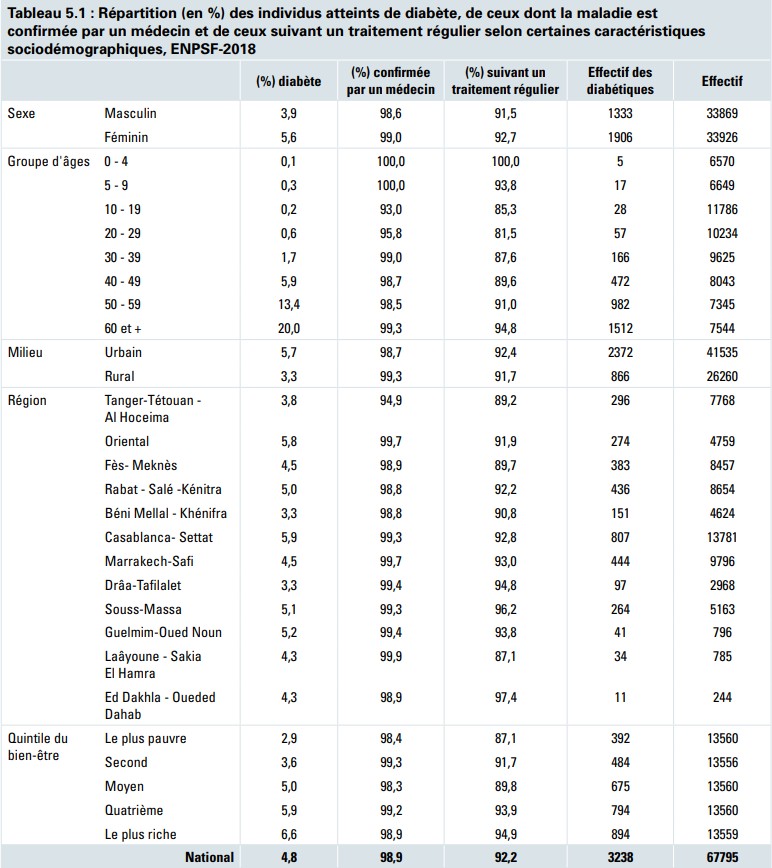
À l’échelle mondiale, on estime à 422 millions le nombre d’adultes qui vivaient avec le diabète en 2014, contre 108 millions en 1980. La prévalence mondiale du diabète (normalisée selon l’âge) a presque doublé depuis 1980, passant de 4,7 à 8,5 % de la population adulte. Ces chiffres reflètent l’augmentation des facteurs de risque associés comme le surpoids et l’obésité. Cette dernière décennie, la prévalence du diabète a progressé plus rapidement dans les pays à revenu faible ou intermédiaire que dans les pays à revenu élevé. Au Maroc, selon les résultats de l’ENPSF-2018, la prévalence du diabète, estimée sur la base des déclarations recueillies au moment de l’enquête ménage, est de 4,8%, montrant une augmentation par rapport aux résultats de ENPSF – 2011 (3,3%). Cette prévalence est 1,7 fois plus élevé en urbain (5,7%) qu’en rural (3,3%) et 1,4 fois plus élevé chez les femmes (5,6%) que chez les hommes (3,9%). Sur les 3238 personnes déclarées atteintes de diabète, 98,9% (98,7% en urbain et 99,3% en rural) ont mentionné avoir consulté un médecin qui a diagnostiqué la maladie et 92,2% parmi elles (92,4% en urbain et 91,7% en rural) affirment suivre régulièrement un traitement médical (tableau 3)

Table 3 Répartition (en %) des individus atteints de diabète, de ceux dont la maladie est confirmée par un médecin et de ceux dont la maladie est confirmée par un médecin et de ceux suivant un traitement régulier selon certaines caractéristiques sociodémographique

Les tranches d’âge de 50-59 ans et de 60 ans et plus enregistrent des prévalences respectivement de 13,4% et de 20%. Chez les moins de 30 ans la prévalence est inférieure à 1% et pour les 30 à 49 ans, elle est inférieure de 5,9%. La prévalence du diabète augmente selon le niveau socio-économique du ménage. En effet, cette prévalence varie de 2,9% chez les plus pauvres à 6,6% chez les plus aisés. Mais il faut s’abstenir à ce stade de l’analyse de toute conclusion hâtive vue le non contrôle de certaines variables clé, en particulier de la structure par âge dans les deux groupes. La répartition par région de la prévalence du diabète montre que sept parmi elles ont des prévalences inférieures au niveau national (4,8%). Le maximum est enregistré à la région du Grand Casablanca (5,9%).

# 

Figure 22Prévalence (déclaré) du diabète par région ENPSF2018

# Historique diabète :

L’historique du diabète des répondants a été pris en compte dans notre étude, l’objectif était de déterminer les mesures antérieures de glycémie, le diagnostic antérieur de cette maladie et de pouvoir détecter les répondants ayant déjà été diagnostiqués pour compléter l’information lors du Step 3 des mesures biochimiques par les nouveaux diagnostiqués suite à notre étude. Deux personnes sur trois, soit 63.2% (61.8-64.6) n’avaient jamais mesuré leur glycémie. Cette donnée était significativement plus importante chez les hommes, 71.5% (69.3-73.7) que chez les femmes 55.2% (53.5-56.9). Aussi, la différence est significative entre les milieux, les résidents du milieu rural dans 72.4%.

n’avaient jamais mesuré leur glycémie, comparés à 58.1% (56.2-59.9) des résidents du milieu urbain. Sur l’ensemble des répondants, 5.8% (5.2-6.4) ont confirmé être diagnostiqués diabétiques durant les 12 derniers mois précédent l’enquête et 1.2% (0.9-1.5) avant cette date. Ce pourcentage est significativement plus important en milieu urbain qu’en milieu rural, il a été conclu que 7.1% (6.2-7.9) et 1.6% (1.2-2.0) des résidents du milieu urbain étaient respectivement diagnostiqués comme diabétiques durant l’année de l’enquête, et auparavant, alors qu’en milieu rural ces pourcentages étaient de 3.7% (2.9-4.4) et 0.5% (0.2-0.9).(figure 23, figure 24).

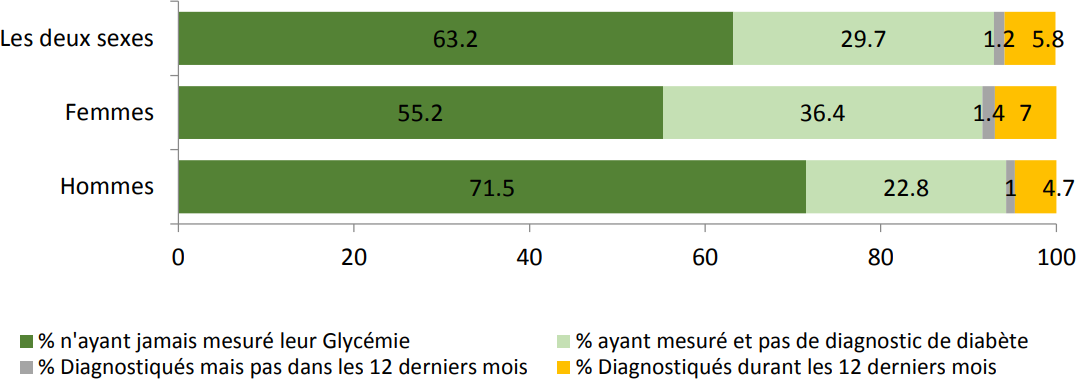


Figure 23 : Statut des mesures antérieures de glycémie par pourcentage par sexe, Steps, Maroc

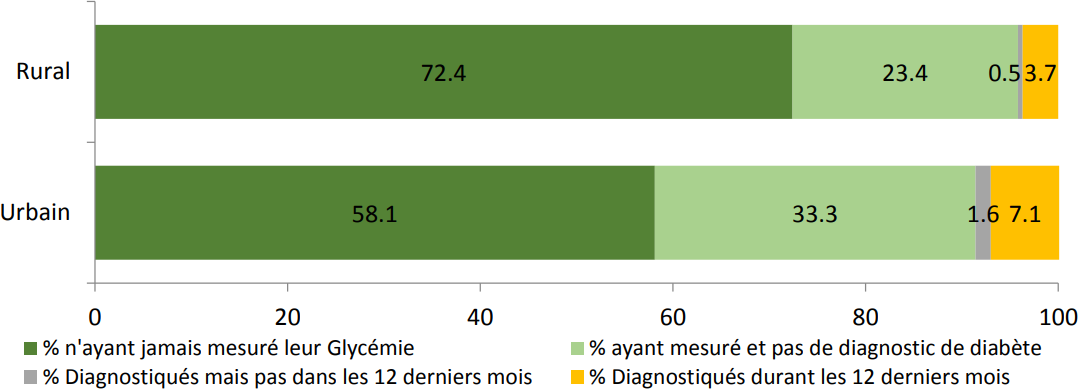


Figure 24: 9 Statut des mesures antérieures de glycémie par pourcentage par Milieu, Steps, Maroc

Sur l’ensemble des personnes ayant déclaré être diagnostiquées diabétiques, 71.6% (67.2-75.9) ont déclaré avoir pris des médicaments pour traitement de diabète prescrits par un médecin ou un professionnel de santé les deux dernières semaines. 76.3% (68.4-84.2) des hommes et 68.5 % (63.3- 73.6) sont des femmes.

Aussi 22% (18.0-25.9) des personnes déclarées diabétiques ont rapportés être sous insuline pour traitement du diabète. Il n’existait pas de différence significative entre les deux sexes alors que cette différence était significative entre les milieux de résidence, 24.2% (19.5-28.9) du milieu urbain et 13.8% (7.5-20.2) du milieu rural.

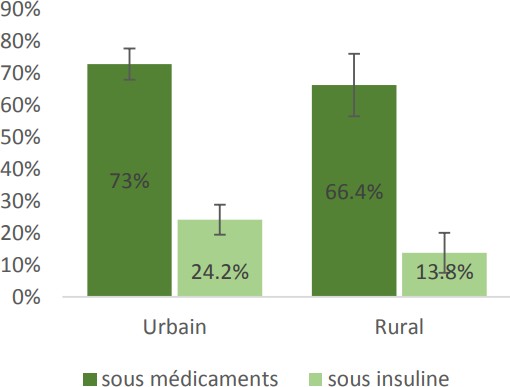
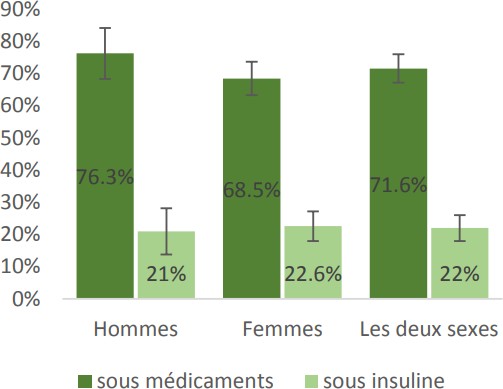


Figure 25: Proportion des personnes prenant des médicaments ou sous

Figure 26: Proportion des personnes prenant des médicaments

Toujours parmi les personnes connues diabétiques, 4.9% (3.0-6.8) étaient sous remèdes traditionnels de traitement de diabète. Une différence significative a été enregistrée entre les deux sexes, les femmes avec 7.1% (4.2-9.9) utilisaient plus ces remèdes que les hommes avec 1.6% (0.0-3.3).

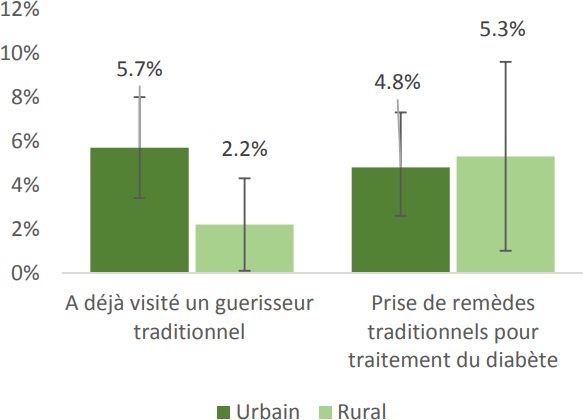


Figure 27 : pourcentage des personnes ayant recours à la médecine

Figure 28 : pourcentage des personnes ayant recours à

# Problématique et Direction de recherches actuelles :

* **Détection de la diabétique à un stade précoce ?**

Dans cette optique nous avons pensé à une solution automatique pouvant détecter cette pathologie à un stade précoce, en utilisant des algorithmes à base de Maching Learning architecture, dans le but d’extraire, traiter et enfin classification des clichés. Dans ce qui suit, nous présenterons un état de l’art sur les travaux réalisés dans le cadre de cette problématique afin d’y voir plus clair.

# Conclusion

Dans ce chapitre nous avons présente la maladie du diabète, leur différent type, les symptômes ainsi que le diagnostic et le traitement de la maladie et à la fin nous avons cité quelques préventions pour évite 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 .

# État de l’art

# Travaux [Rajiv Singla](https://pubmed.ncbi.nlm.nih.gov/?term=Singla%20R%5BAuthor%5D)[, Ankush Singla](https://pubmed.ncbi.nlm.nih.gov/?term=Singla%20A%5BAuthor%5D), [Yashdeep Gupta](https://pubmed.ncbi.nlm.nih.gov/?term=Gupta%20Y%5BAuthor%5D) et [Sanjay Kalra](https://pubmed.ncbi.nlm.nih.gov/?term=Kalra%20S%5BAuthor%5D)

L'intelligence artificielle (IA) est un terme général défini comme la théorie et le développement de systèmes virtuels capables d'effectuer normalement des tâches en utilisant l'intelligence humaine, telles que la perception visuelle, la reconnaissance vocale, la prise de décision et la traduction entre les langues.[ [1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/#ref1) ] peut être aussi simple que basé sur des règles ou piloté par des méthodes statistiques complexes. L'apprentissage automatique est un sous-ensemble de l'intelligence artificielle (IA) qui fournit aux systèmes la capacité d'apprendre et de s'améliorer automatiquement à partir de l'expérience sans être explicitement programmé.[ [1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/#ref1) , [2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/#ref2) ] L'apprentissage automatique peut être supervisé, non supervisé, semi-supervisé ou basé sur le renforcement. Grâce à l'apprentissage en profondeur, la machine tente d'émuler l'intelligence humaine en simulant la structure du cerveau humain à l'aide de réseaux de neurones récurrents.

Les outils d'IA/ML sont largement utilisés dans tous les domaines scientifiques et sont responsables de la révolution des entreprises à travers le monde. Les systèmes de santé, en revanche, ont été très lents à adopter ces avancées et accusent un retard considérable dans ce domaine.

L'IA/ML peut être utile dans la gestion des maladies chroniques, à savoir le diabète. En fait, ML/AI est déjà utilisé pour prédire le risque de diabète sur la base des données génomiques, le diagnostic du diabète sur la base des données du DSE, pour prédire le risque de complications telles que la néphropathie et la rétinopathie, et également dans le diagnostic de la rétinopathie diabétique [[Tableau 1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/table/T1/)].[ [3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/#ref3) ] Il y a peu de données spécifiques à l'Inde sur tous ces aspects de l'IA dans la littérature publiée. L'unité de recherche Google AI, en collaboration avec quelques centres d'ophtalmologie indiens, a déjà fait de grands progrès dans le domaine du diagnostic et du classement automatisés de la rétinopathie diabétique sur la base de photographies du fond d'œil.[ [4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/#ref4) ] L'adoption de ces technologies peut considérablement augmenter la détection et le traitement précoce des complications diabétiques. [ [4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6844177/#ref4) ]

# **Chapitre3 : “Conception et Réalisation”**

# Introduction :

Nous allons présenter dans ce chapitre la solution que nous proposons pour répondre à la

problématique qu’est la détection précoce de la rétinopathie diabétique.

Enfin, on terminera avec la partie théorique, pour aborder la partie technique dans laquelle

on introduira l’environnement, langage ainsi que la plateforme utilisée, et on terminera par la

présentation de l’application du système de détection.

# Problématique et Contraintes:

L’ensemble du travail qui sera présenté consiste en la mise en œuvre d’un système d’aide au diagnostic médical dont le rôle est de détecter la présence de symptômes liés à la présence ou non de la pathologie diabétique de type 1 ou 2.

Pour le bon fonctionnement du système, la prise en compte de certains critères de performances sont nécessaires:

* un système fiable.
* un système performant à moindre coût (Rapidité du temps d’exécution).
* un système en mesure de continuer à apprendre au cours de sa vie.

Le but ici est de faire en sorte d’avoir un système qui a les meilleurs taux de performance selon les contraintes fixées (capacité de calcule et stockage limité).

* 1. Méthode:

L’ensemble des modèles proposés dans notre solution reposent sur une architecture MVC. Notre choix s’est orienté vers MVC en raison de la nature du problème (apprentissage supervisé type classification), ainsi que du type de données à traiter (CSV). On dispose d’un jeu de données de 769 échantillons (dataset) fournis exclusivement par la plateforme kaggle, contenant 9 paramètres .

Cet ensemble de données provient à l'origine de l'Institut national du diabète et des maladies digestives et rénales. L'objectif est de prédire, sur la base de mesures diagnostiques, si un patient est atteint de diabète.

# 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 utilises :

* Prediction of Diabetes Using Machine Learning Algorithms in Healthcare.
* Machine Learning Workflow on Diabetes Data.
* Application des méthodes d’apprentissage dans la prédiction du diabète Type 2 .

# Application des méthodes d’apprentissage dans la prédiction du diabète Type 2 :

# La précision

La précisions (Accuracy ) est le taux de réussites global de l’algorithme d´défini par l’équation suivante :

|  |
| --- |
| *Accuracy = (T P + T N)/(P + N) ....(01)* |

**TP**: True Positive.

**TN** : True Négative .

**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.

# L’étude 1 :

Cette étude a conclu que GBC et lightGBM 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(SVC), Logistic Regression(LR),Décision Tree Classifier(DT),et Random Forest (RF). Tous ces algorithmes ont été appliques a Dataset PIMA Indian comprenant 768 enregistrements et 9 attributs . L’ensemble de données a été divisé en deux parties, les données d’entrainement (training data )et les données de test(test data ), ces deux parties constituantes respectivement 80% et 20% des données. dans ce travail, le principal paramètre d’évaluation entre les algorithmes est la précision de prédiction définie par l’équation.

Table 4 Les résultats de la précision des algorithmes d’étude 01

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Algorithme* | GBC | lightGBM | LogisticRegression | AdaBoost | KNN | Random Forest | DecisionTree |
| *Precision* | 88% | 88% | 82% | 87% | 80% | 86% | 83% |

# 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 :

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

# 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 taches tels que l’exploitation des données, le nettoyage des données jusqu’ à la s´élection 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),Décision Tree Classifier(DT), Random Forest (RF)et Gradient Boost (GB) afin d’évaluer leurs précisions avec deux méthodes d’évaluation.

# Méthode 01 :Train/Test Split :

Fractionner l’ensemble de données en deux parties :

* Partie d’entrainement : pour former le modèle.
* 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 résultats de précision de chaque algorithme selon chaque méthode : A la fin le modèle

**NB** : Dans notre cas on a fait 5 folds .

# Logistic Regression :

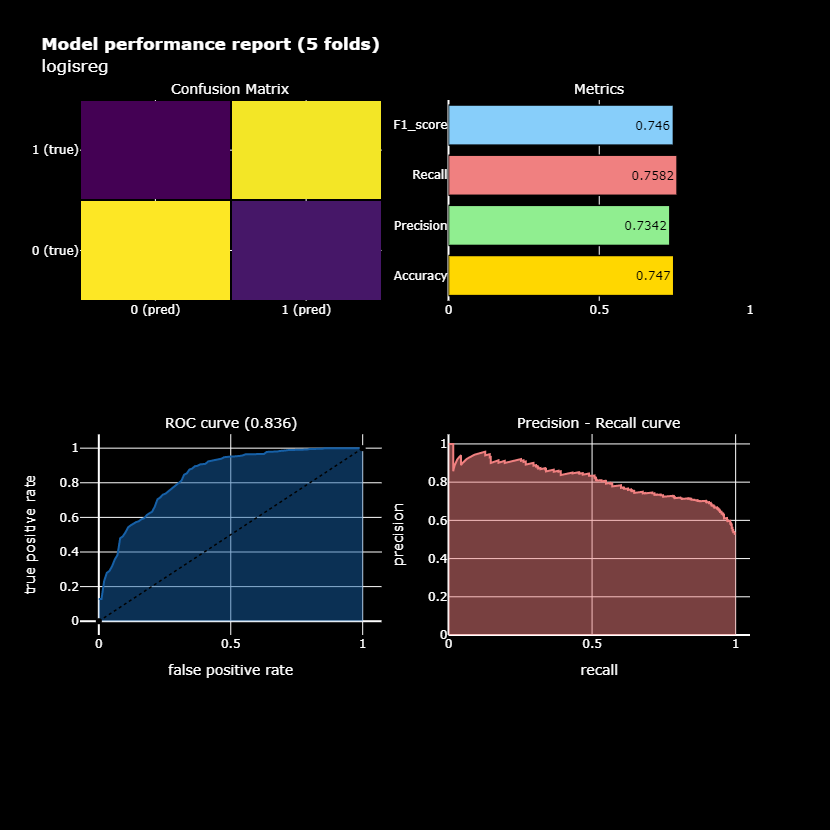
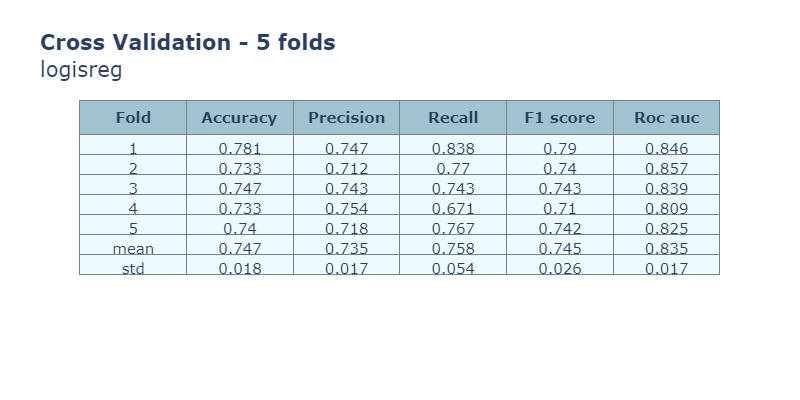


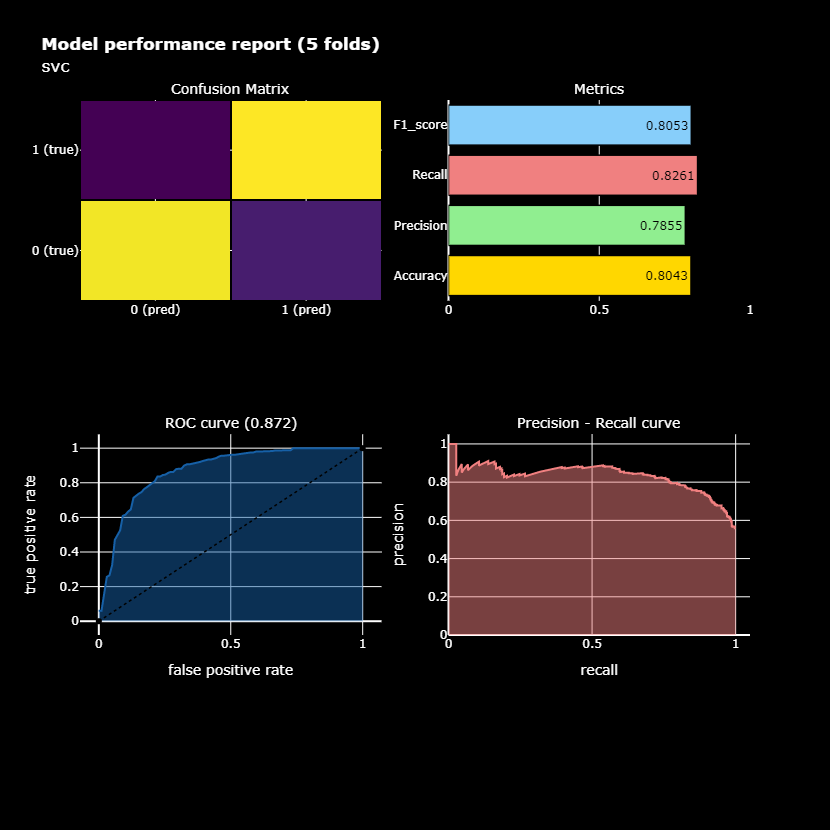
Figure 29 Model performance report (5 folds) logistic Regression

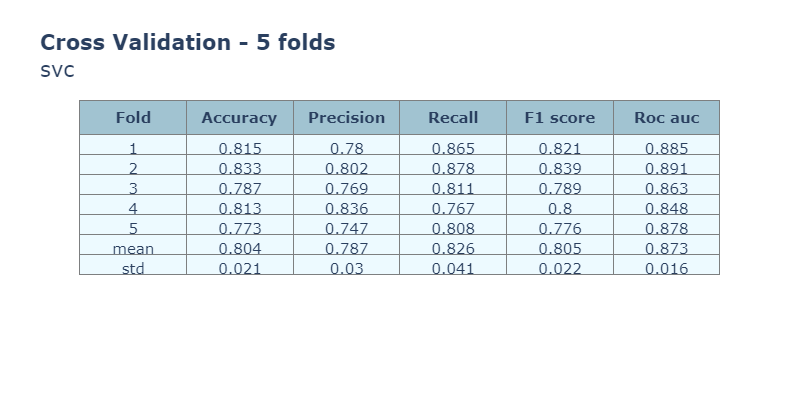


# KNN:

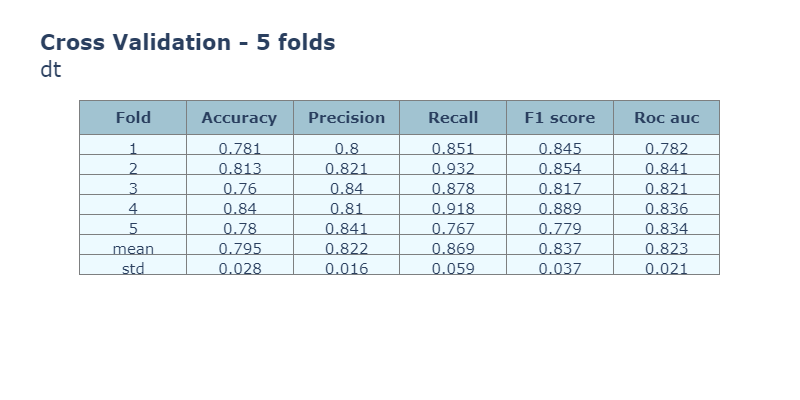
Figure 30 Model performance report (5 folds) KNN

# SVC:

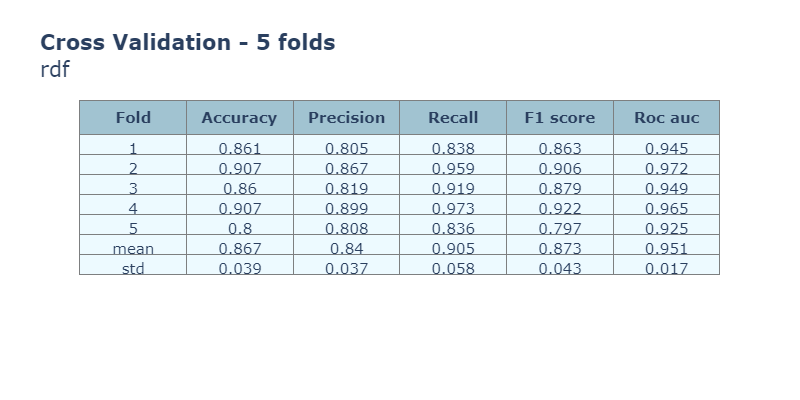


Figure 31 model performance SVC

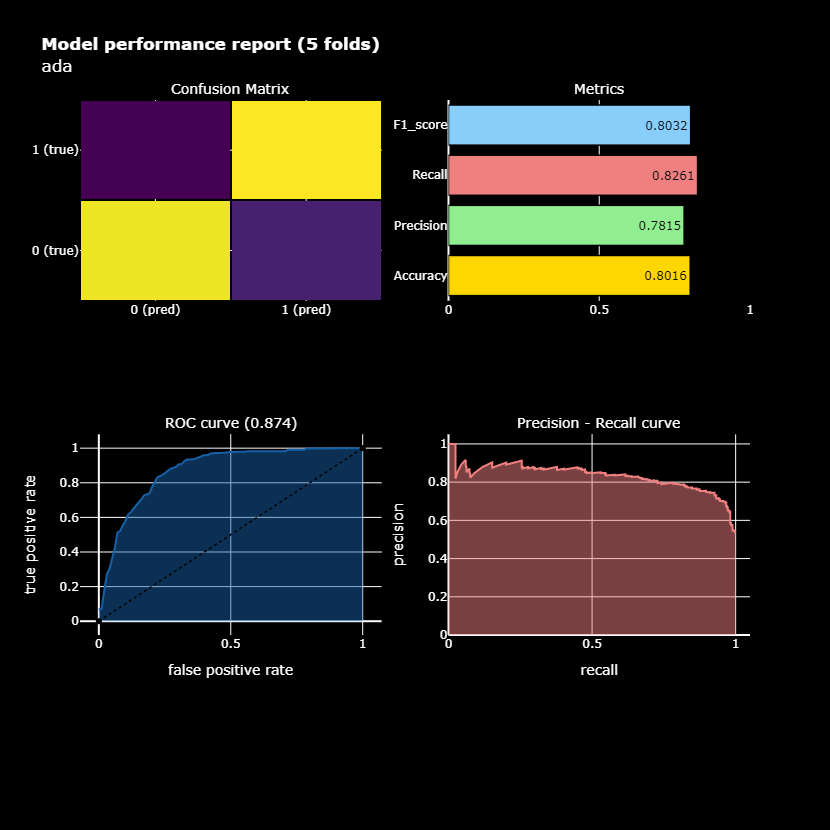
# DecisionTree:

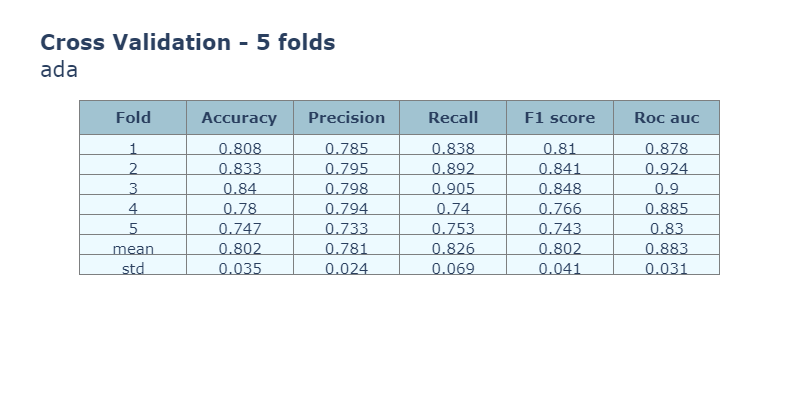
Figure 32 model performance DT

# Random Forest:

Figure 33 model performance RF

# AdaBoost



Figure 34 model performance Adaboost

# LBGM:

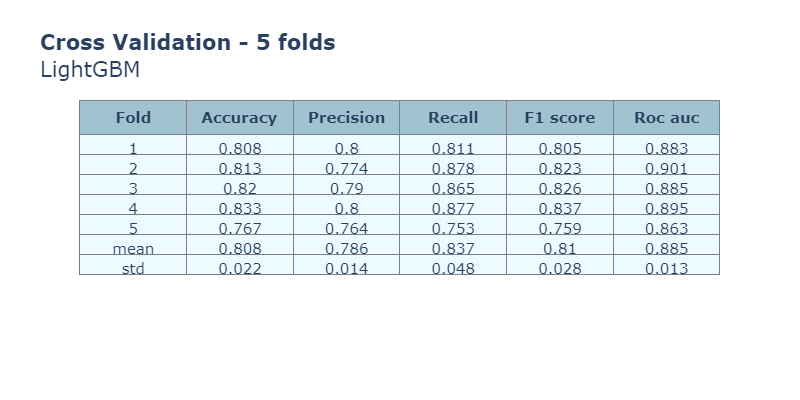


Figure 35 model performance LGBM

Tableau 5 Les résultat de précision des algorithmes d’étude 02

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithme | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 |
| LR | 0,781 | 0,733 | 0,747 | 0,733 | 0,74 |
| KNN | 0,834 | 0,86 | 0,84 | 0,847 | 0,76 |
| SVC | 0,815 | 0,833 | 0,787 | 0,813 | 0,77 |
| DT | 0,781 | 0,813 | 0,76 | 0,84 | 0,70 |
| RF | 0,86 | 0,907 | 0,86 | 0,907 | 0,80 |
| AdaBoost | 0,808 | 0,833 | 0,84 | 0,78 | 0,74 |
| LGBM | 0,808 | 0,813 | 0,82 | 0,833 | 0,767 |

# Critique

Random forest est choisie comme le modèle qui fonctionne le mieux pour l’ensemble de données avec une précision égale à 90 %. (Fold 2)

# Conclusion

Dans ce chapitre, nous avons présente 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 Tree, Random Forest, Support Vector Machine …) de classification aux données extrait de l’hôpital Frankfurt concernant le risque de développer un diabète de type 2 .

# Prédiction du diabète de type 2 par l’apprentissage automatique :

# Introduction

Dans ce 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.

# Outils et Librairies utilises

* 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 plusieurs 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 prédictives et les calculs scientifiques.

* Jupyter notebook

Jupyter Notebook est un environnement de programmation qui prend en charge plusieurs 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.

* Python

Python C’est un langage de programmation multiparadigme 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, animes 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 biblioth`eque Python utilisée pour la manipulation et l’analyse des données, le point fort de cette biblioth`eque est qu’elle possède une fonctionnalité importante appelée nettoyage des donnes 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 biblioth`eque Python la plus importante pour ce qui concerne l’apprentissage automatique telle qu’il contient de nombreux algorithmes ( forêts aléatoires, des régressions logistiques, des algorithmes de classification, et les machines a vecteurs de support ) .

* + **Streamlit**:

Streamlit est un framework open-source Python spécialement conçu pour les ingénieurs en machine learning et les Data scientists. Ce framework permet de créer des applications web qui pourront intégrer aisément des modèles de machine learning et des outils de visualisation de données.

* + **Google Collab** :

Colab permet à n'importe qui d'écrire et d'exécuter le code Python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté au machine learning, à l'analyse de données et à l'éducation.

* + **Pycaret**:

PyCaret est une bibliothèque d'apprentissage automatique simple, facile à apprendre et à faible code en Python. Avec PyCaret, vous passez moins de temps à coder et plus de temps à analyser.

# Définition d’ensemble de données utilisé et description des variables

# Définition l’ensemble de données utilisé

Le Dataset (jeux de données) utilisé pour ce projet a été fourni par Kaggle. Kaggle est une communauté regroupant des data scientiste tout niveau confondu, qu’il leurs permet de publier leurs travaux. Kaggle organise aussi des compétitions proposées par des entreprises ou particuliers dont les vainqueurs seront payés.

Notre dataset comprend 769 linges avec 8 columens .

Table 5 Description des variables d ’ensemble de données

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Description | | Analyse donees |
| Pregnancies | | Nombre de fois enceinte. | Minimum : 0 Maximum : 17 |
| Glucose | | Concentration plasmatique de glucose à 2 heures dans un test oral de tolérance au glucose. | Minimum : 0 Maximum : 199 |
| Blood Pressure | | Pression artérielle diastolique (mm Hg). | Minimum : 0 Maximum : 122 |
| Skin Thickness | | Epaisseur de pli cutané du triceps (mm). | Minimum : 0 Maximum : 110 |
| Insulin | | Insuline sérique 2 heures (mu U/ ml). | Minimum : 0 Maximum : 799 |
| BMI | | (ou IMC) Indice de masse corporelle (poids en kg / (taille en m)2 ). | Minimum : 0 Maximum : 80.6 |
| DiabetesPedigreeFunction | | Fonction généalogique du diabète. | Minimum : 0.078 Maximum : 2.42 |
| Age | | l’age en années. | Minimum : 21 Maximum : 81 |
| 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. | 0(non diabétique) :1316 1(diabétique) : 684 |

# Les étapes de pré-traitement de données :

Créer un modèle de Machine Learning est un processus en plusieurs etapes.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.

# Exploration et visualisation de données :

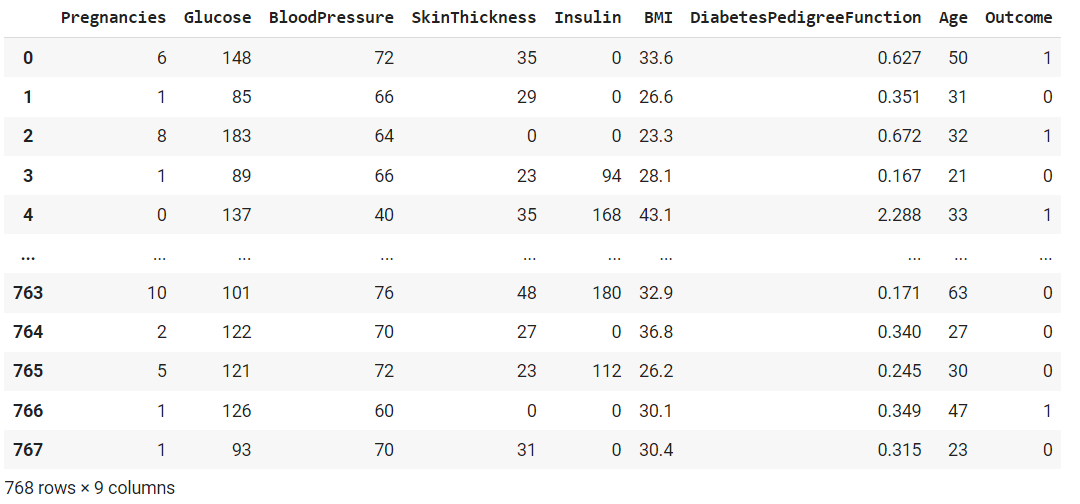
La visualisation des données est définie comme l’exploration visuelle et interactive des données de toutes volumétries. Qui aident à voir des choses n’étaient pas évidentes auparavant. La visualisation facilite la transmission des informations de façon universelle et facilite le partage d’idées avec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

Figure 36 Aperçu de l’ensemble de données

Pour visualiser notre ensemble de données on utilise la bibliothèque, pandas profiling qui générer 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.

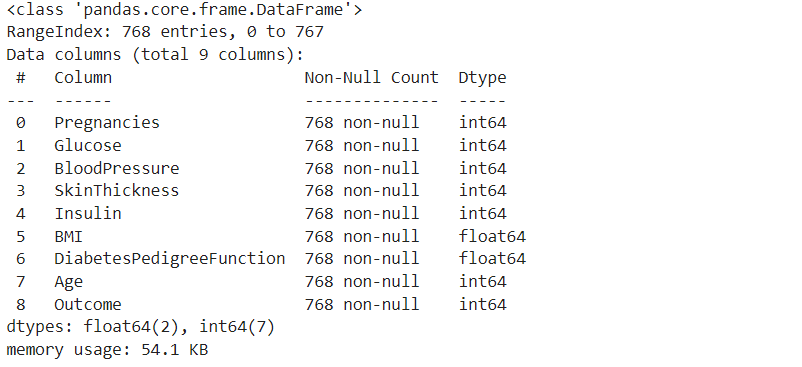


Figure 37 Rapport de l’ensemble de données

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

* Le nombre des observations (768 patients) .
* 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 à 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) :

# Variables :

Figure 38 : Analyse des variables de l'ensemble de données

* Corrélations**:**

Un bon ensemble de données est un ensemble dans lequel les caractéristiques sont

fortement corrélées `a 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 sont le coefficient de Pearson noté, défini par :

Ou :

* : désigne la covariance des variables X et Y , σX et σY 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

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

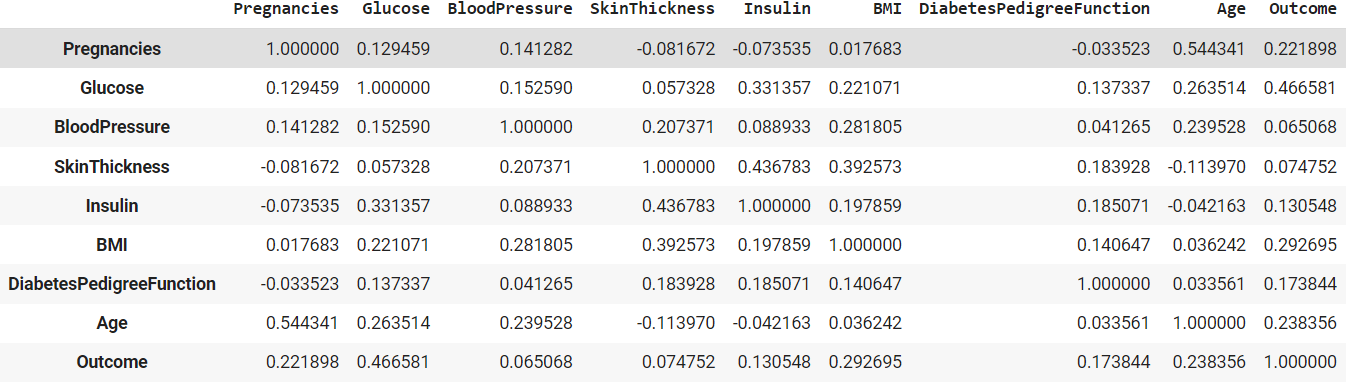


Figure 39 table de corrélation

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 40 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 certaines colonnes qui rendre la lecture de l’ensemble de données fausse et qu’il avait besoin de nettoyage .

# 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 .
* **BMI** : Le poids d’une personne ne sera jamais égal à 0 ou proche de zéro sa mettre sa vie en danger.
* **Glucose** : Le taux de glucose ne doit pas être égal à 0 mais d’après l’analyse des données.
* **Insuline** : dans une situation rare une personne peut avoir 0.
* **Pregnancies** : c’est normal d’avoir une 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.

Solutions pour gérer les valeurs aberrantes :

* **Option 1 :** supprimer toutes les observations ayant zéro valeur 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 la médiane de chaque colonne mentionnée ) où la valeur égale à 0 avec la fonction replace().

Tableau 2 Distribution des variables avant et après nettoyage

|  |  |  |
| --- | --- | --- |
| Variable | Avant | Après |
| Blood Pressure |  |  |
| BMI |  |  |
| Glucose |  |  |
| Insulin |  |  |
| Skin Thikness |  |  |

# 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 :

* + KNN (K-Nearest Neighbors)
  + L’arbre de décision (Decision tree)
  + SVM (support vector machine)
  + Random Forest (forêt aléatoire)
  + LGBM light gradient boosting maching
  + Adaboosting

# Méthode d’évaluation

Utilise pour donne la capacite au modèle de prédire les données hors échantillons et évite le problème de surajustement (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 :

# Train/Test Split :

Cette méthode consiste à divise l’ensemble de données en deux parties : 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.

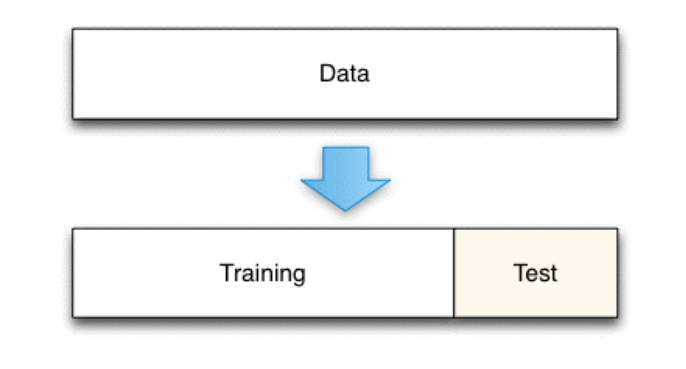


Figure 41 Répartition des données de train/test

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 :

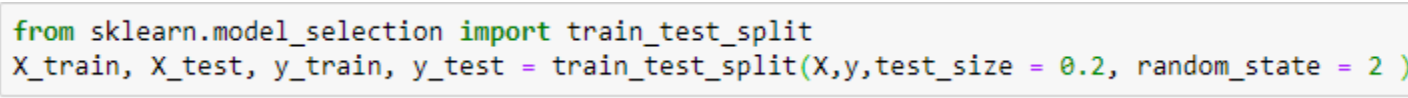


Figure 42 Fractionnement de l’ensemble de données

# Evaluations des modèles :

Pour comparer les différentes modelées et évaluer ces performances on utilise 03 mesures qui sont : La précision, recall score, F1 score.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algo | Accuracy | Precision | recall score | F1 score |
| Decision Tree | 0.848 | 0.744 | 0.878 | 0.833 |
| Random Forest | 0.901 | 0.932 | 0.905 | 0.873 |
| KNN | 0.868 | 0.8 | 0.973 | 0.878 |
| SVM | 0.821 | 0.841 | 0.784 | 0.811 |
| LR | 0.828 | 0.843 | 0.797 | 0.819 |
| Adaboosting | 0.820 | 0.822 | 0.811 | 0.816 |

Figure 43 Les résultats des attributs d ’évaluations pour les diffèrent model

D’après le tableau ci-dessus le modèle Random forest obtenu la meilleure précision qui égal à 90% et le meilleur score de rappel ´égal `a 0.83 c’est-a-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 précisions et score de rappel.

# Réglage de paramètre de modèle :

La bibliothèque **sklearn** offre au modèle d’apprentissage automatique des paramétrés par défaut, sensible qui donne parfois des scores de précisions décents avec la possibilité de modification du ces paramétrés pour évaluer la précision de ces modelées. 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 paramétrés. Au lieu de rechercher manuellement les paramétrés optimaux pour notre modèle nous avons opté d’utilisé la fonction **GridSearchcv** de la bibliothèque **sklearn** qui permet une recherche exhaustive sur les valeurs de paramétrés spécifique pour un modèle.

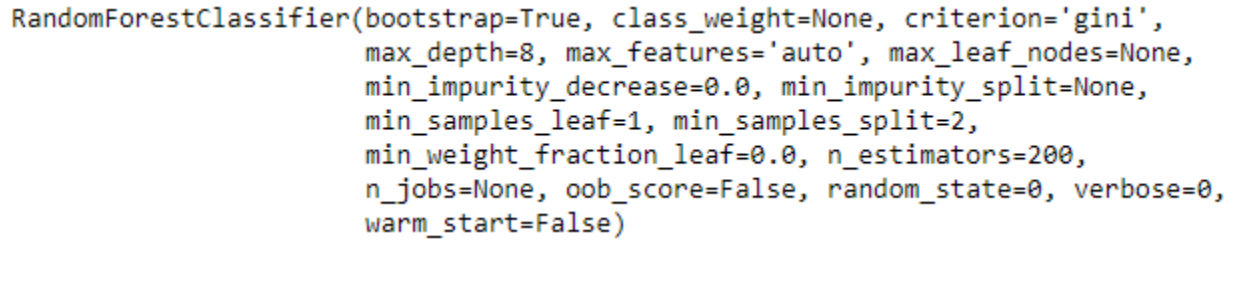


Figure 44 Meilleur paramètres pour le model Random forest

D’après le réglage de paramétrés nous avons observé que la précisions de modèle est augmenté(voir figure suivant) :



Figure 45 Amélioration de précisions du modèle Random forest

# Sauvegarde de modèle :

L’enregistrement du modèle finalisé fait gagne beaucoup de temps car on n’a pas besoin d’entrainer le modèle à chaque exécution de l’application. La bibliothèque pickle permet de sérialise les modelées 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.

# Chapitre 4 : “Démonstration de l’application ”

# 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 prédictions bien défini. L’application est disponible en ligne sur :