TEK-UP
Ecole Supérieure Privée
Technologie & Ingénierie

BI & ML Project Report

EL JED Khalil
ALOUI Arbia
ABIDI Amani

SOMMAIRE

Introduction générale.....2
Context Générale du ProJet.....4
Partie BI : Explication théorique et résultats.....6
Partie machine learning.....10
Conclusion générale.....21

Introduction générale

Le diabète est une maladie chronique caractérisée par la présence d'un excès de sucre dans le sang appelé hyperglycémie. Il est avéré si le taux de glycémie à jeun est égal ou supérieur à 1,26 g/l ou 7 mmol/l de sang lors de deux dosages successifs.

Avant de diagnostiquer un diabète , il existe dans certains cas des indices révélateurs de son développement. Il s'agit du stade prédiabète.

Le prédiabète se caractérise par un taux de sucre élevé dans le sang, à un niveau supérieur à la normale et inférieur à celui pathologique du diabète. Il correspond à des taux qui dépassent la norme pour la glycémie à jeun et pour l'hyperglycémie provoquée. Ce dernier aspect reflète une situation d'intolérance au glucose souvent associée au prédiabète.

Les chercheurs et les médecins s'intéressent au stade de prédiabète car cet état physiopathologique correspond à une phase critique où il est possible d'agir. À ce stade, les anomalies caractéristiques du diabète (insulinorésistance et diminution de la production de l'insuline) commencent à se mettre en place sur une période prolongée et de manière discrète.

En effet, les patients intolérants au glucose ont un risque très élevé de développer un diabète à long terme. Cependant, il est intéressant de noter que ce risque peut être diminué si cette anomalie métabolique est prise en charge suffisamment tôt.

Ainsi, les études cliniques montrent qu'une intervention basée sur la modification du mode de vie des personnes prédiabétiques permet de prévenir l'apparition du diabète.

Il s'agit de modifications concernant le régime alimentaire (baisse de l'apport calorique journalier) et de l'activité physique (modérée et régulière) qui permettent :

- De contrôler la glycémie,
- De stabiliser le poids,
- Et d'améliorer les autres paramètres tels que le taux du cholestérol et la pression artérielle.

L'amélioration de l'ensemble de ces paramètres participe de manière globale à réduire le risque de maladies cardiovasculaires et de prévenir ainsi l'apparition des autres complications liées au diabète.

Le stade de prédiabète constitue donc une étape clé dans l'apparition du diabète. Ceci justifie une prise en charge dès l'observation d'anomalies au niveau de la glycémie notamment chez les sujets qui présentent un haut risque de développement du diabète.

Dans ce contexte, nous proposons un système d'analyse sentimentale qui prédire le risque de diabète à un stade précoce en utilisant technique de Business Intelligence ainsi que les algorithmes de Machine Learning. Notre rapport est composé de trois chapitres. Dont le premier chapitre « Cadre général du

projet », par ailleurs, la problématique afin de proposer une solution par la suite l'adoption de la méthodologie CRISP-DM dans notre travail.
Le deuxième chapitre « Système de prédiction proposé » est consacré à l'étude théorique de nos systèmes proposés et qui présente en première section la partie Business Intelligence données à l'aide d'un outil de tableau de bord. Le troisième chapitre « Réalisation du ssystème de prédiction proposé »

Context Générale du ProJet

I.Introduction:

L'objectif de notre projet est de proposer un système de prédiction du risque de diabète à un stade précoce. Dans ce chapitre, une présentation du cadre générale du projet sera élaborée, ainsi une étude de l'existant, leur critique et la solution proposée et leurs objectifs seront présentés.

II.Présentation du projet

Cette partie est composée de trois sous-section. Dans la première sous-section, nous expliquons la problématique. Ensuite une étude de l'existant et leurs critiques seront élaborés. Dans la dernière sous-section, nous présentons notre solution proposée.

III.Problématique

Le diabète peut entraîner des complications touchant de nombreuses parties du corps comme le cerveau, les yeux, le cœur, les reins et les nerfs. Les complications liées au diabète peuvent être à long terme (chroniques) ou à court terme (aiguës). Le stade de prédiabète constitue donc une étape clé dans apparition du diabète.

IV.Solution proposée

La solution proposée consiste à réaliser un système de Prédiction du risque de diabète à un stade précoce afin de minimiser le risque et éviter le diabète. Notre solution est basée sur :

- Traitement de données avec des outils de BI.
- Les algorithmes de Machine Learning

V.Méthodologie de travail

Dans cette partie nous présentons la méthodologie utilisée durant notre projet, notre choix s'arrête sur la méthodologie Cross-Industrie Standard Process for Data Mining (CRISPDM). Dans les années 1990, l'informatique et les données

sont passées d'un atout marginal à une nécessité pour toutes les entreprises, les organisations ont cherché à trouver un processus efficace et structuré avec lequel elles pouvaient se sentir à l'aise. En réponse, plusieurs leaders de l'industrie ont formé un consortium pour trouver un processus standardisé pour l'exploration de données. Le consortium a donné naissance au processus CRISP-DM, où le processus standard intersectoriel pour l'exploration de données. Cadre général du projet CRISP-DM reste la méthodologie standard pour attaquer aux projets centrés sur les données car elle avère robuste tout en offrant simultanément la flexibilité et la personnalisation. Le modèle CRISP-DM décrit les étapes impliquées dans l'exécution des activités de science des données, du besoin métier au déploiement, mais définit surtout un cadre qui permet des itérations à travers toutes les phases. Dans les applications du monde réel, la nature itérative permet une amélioration constante via le retour en arrière aux tâches précédentes et la répétition de certaines actions.

VI.Conclusion

Durant ce premier chapitre, nous avons cerné les problèmes en donnant des solutions adéquates avec une présentation de la méthodologie adoptée. Dans le chapitre suivant, nous détaillerons le système proposé pour la prédiction du risque.

Partie BI : Explication théorique et résultats

I.Introduction

Après avoir présenté le contexte général de notre projet dans le chapitre précédant, nous consacrons ce chapitre à l'étude théorique de nos systèmes proposés pour l'analyse du risque. Tout d'abord nous présentons la partie Business intelligence (BI) qui représente une grande partie de notre projet pour L'extraction, la transformation et le chargement des données. Puis, notre système de l'analyse du risque pour l'amélioration de la prévision de la demande sera détaillé. Ce système est basé sur les algorithmes de l'intelligence artificielle.

La Business Intelligence (BI) est un processus technologique d'analyse des données et de présentation d'informations pour aider les dirigeants, managers et autres utilisateurs finaux de l'entreprise à prendre des décisions commerciales éclairées. La Business Intelligence englobe une grande variété d'outils, d'applications et de méthodologies qui permettent aux organisations de collecter des données à partir de systèmes internes et de sources externes. Ces données sont ensuite préparées pour l'analyse afin de créer des rapports, tableaux de bord et d'autres outils de visualisation de donnée pour rendre les résultats analytiques disponibles aux décideurs et au personnel opérationnel.

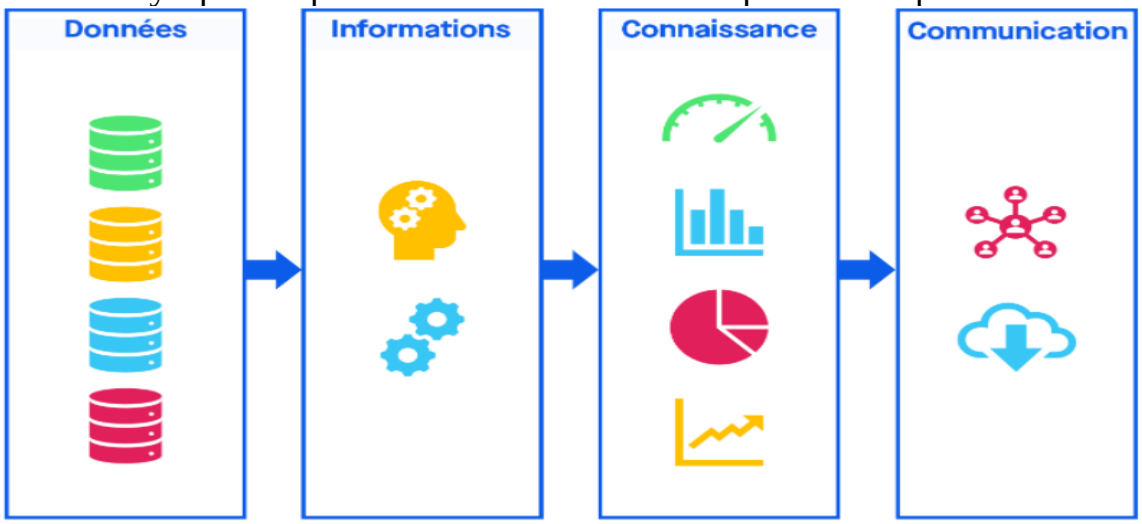


Figure : Les outils d'intelligence d'affaires

II.Les données

Ce sont des ensembles d'éléments de valeur brute qui sont utilisés aux fins de calculs, de mesures ou de raisonnements. Les données n'ont aucune signification si elles ne sont pas analysées dans un contexte déterminé.

III.L'information

C'est le résultat de collecter et d'organiser les données de façon telle qu'il devient possible d'établir des relations entre les éléments de données et sa signification dans un contexte précis.

IV.Les connaissances

C'est le concept de comprendre l'information à partir de modèles établis qui permettent de fournir une conceptualisation de l'information. Les connaissances sont la résultante d'un processus d'intelligence d'affaires et se traduisent par des tableaux de bord, des rapports et des alertes. La force des bons outils BI réside entre autres dans la facilité d'utilisation et de manipulation de ces connaissances. Effectivement, avec ses outils, il est très simple en quelques clics de comparer les performances de l'organisation par rapport à ses objectifs.

Le Data Warehousing Institute définit l'intelligence d'affaires comme suit :

« Les processus, les technologies et les outils nécessaires pour transformer des données en information, l'information en connaissances et les connaissances en plans d'action. L'intelligence d'affaires est une combinaison d'entreposage de données, d'outils d'analyse d'affaires et de gestion de contenu-connaissance ».

Définition des SQL Server Integration Services

Dans un projet décisionnel, SSIS est le premier service à entrer en action, car il gère l'extraction et l'enregistrement des données en dehors de l'outil de production. Il joue donc le rôle d'un ETL (Extract, Transform and load) et y ajoute toute une série d'outils pour monitorer la performance des processus d'extraction et de transformation de la donnée. Ces données sont alors stockées dans un Datawarehouse fondé sur la technologie SQL Server.

Au passage, un éclaircissement : lorsqu'on commence à vouloir triturer les données, mieux vaut ne pas le faire directement via votre outil métier. Interroger la base de production à des fins de reporting ou Datamining peut en effet ralentir les autres fonctionnalités plus essentielles de votre logiciel professionnel.

Définition des SQL Server Analysis Services

Grâce à SSIS, vos données sont stockées dans une base de données isolée des outils qui les produisent. Mais cela ne suffit pas, car pour un utilisateur lambda la donnée brute reste inexploitable. Il faut la transformer en agrégat, y associer des dimensions pour que l'on arrive à quelque chose d'utile. Un exemple d'agrégat et des dimensions associées : à partir des logs de tracking de votre web analytics, en sortir un nombre de visites par

période, par univers du site, par typologie de clients. Et ne pas passer 3 heures à attendre que les données brutes moulinent. Et bien appliquer des règles de calcul et en restituer rapidement les données sont les fonctionnalités de SSAS.
Pour cela, SSAS génère des cubes, qui précalculent quand vous dormez vos indicateurs et leurs dimensions associées.

Définition des SQL Server Reporting Services

Dernière pierre à l'édifice, SSRS est l'outil de restitution des données. Il vous permet de récupérer vos rapports, fondés sur les données calculées par SSAS, sous forme de fichier Excel, PDF, Word ou HTML. Si vous êtes dans une phase d'exploration des données, vous préférerez Excel, et manipulerez ainsi les indicateurs via des tableaux croisés dynamiques.
Tout ça pour en arriver à l'austérité des tableaux Excel, me direz-vous ? Oui, absolument. Mais des tableaux Excel intelligents, avec des informations que beaucoup d'entreprises vous envieront. Après, vu que vous avez fait le plus compliqué en transformant vos données brutes en indicateurs business, libre à vous d'y ajouter des outils de reporting plus rigolos, qui vous feront des jolis camemberts et histogrammes interactifs.
L'art du tableau de bord se nomme la Data Viualisation (Data Viz pour les intimes).

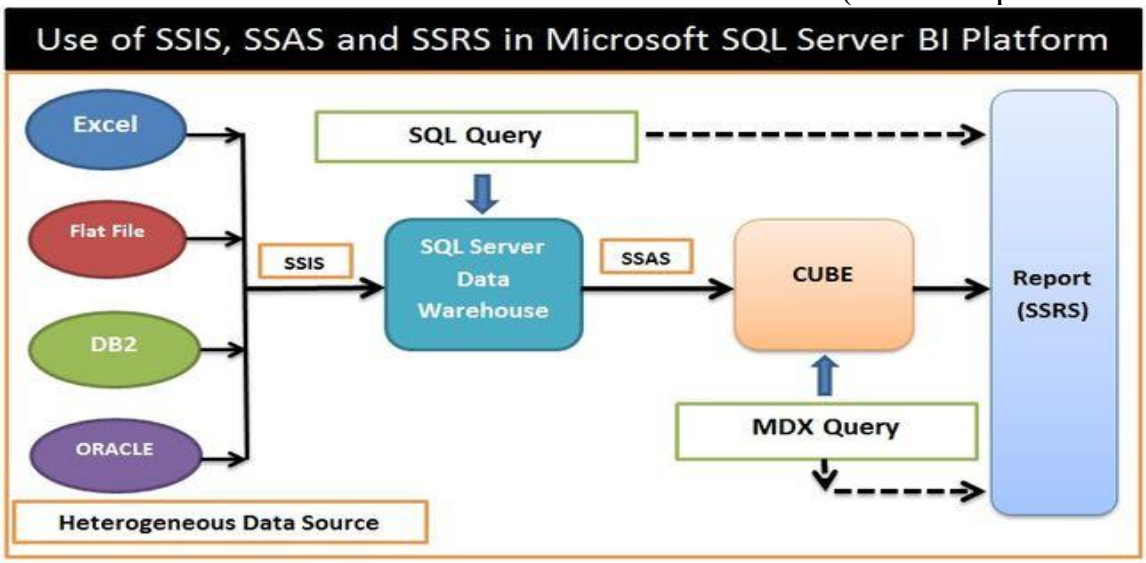


Figure : Processus ETL

ETL, acronyme d'Extraction, *Transformation*, *Loading*, est un système de chargement de données depuis les différentes sources d'information de l'entreprise (hétérogènes) jusqu'à l'entrepôt de données (modèles multidimensionnels). Ce système ne se contente

pas de charger les données, il doit les faire passer par un tas de moulinettes pour les dénormaliser, les nettoyer, les contextualiser, puis de les charger de la façon adéquate.

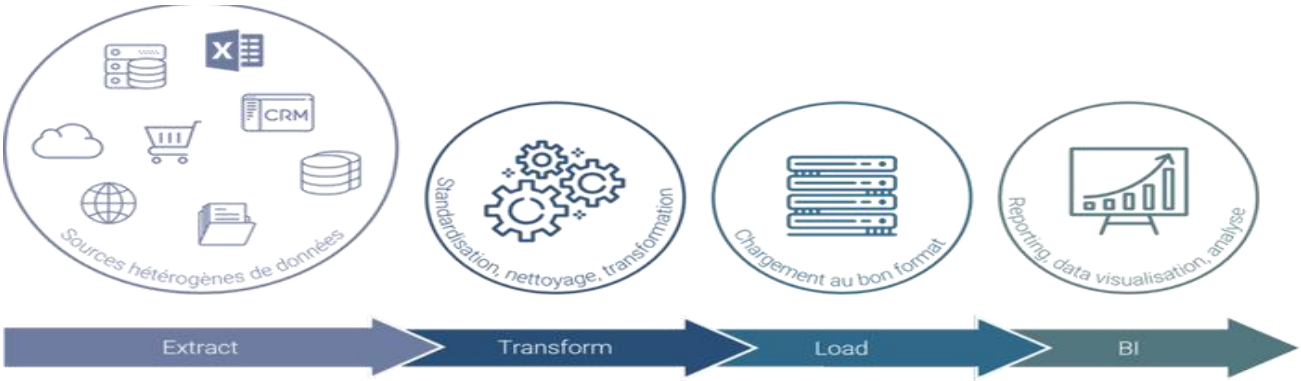
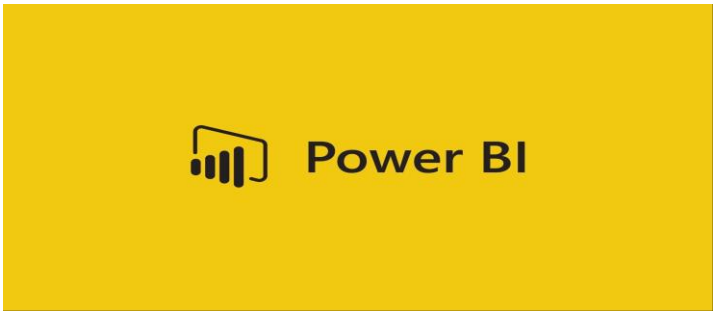


Figure : les étapes de processus etl

V.Les outils d'intelligence d'affaires :

Microsoft Power BI

Depuis sa création en 2014, Microsoft Power BI est reconnue comme une des meilleures solutions d'intelligence d'affaires disponible sur le marché. En proposant différents forfaits (Power BI Free, Power BI Pro, et Power BI Premium), ce logiciel s'adresse réellement aux entreprises de toutes tailles. Ce qui le démarque de ses compétiteurs est qu'il est le seul à pouvoir se connecter à plus d'une centaine de sources de données différentes pour créer les rapports et les tableaux de bord.
De plus, Power BI est offert en version Desktop pour la création et la modification puis en version infonuagique pour la publication et le partage des connaissances. Rassurez-vous, la sécurité des données étant au cœur de la plateforme, vous n'aurez pas à vous inquiéter de la confidentialité des celles-ci puisque vous pourrez la paramétrer grâce à Power BI Services



Partie machine learning

Dans cette partie nous présenterons les techniques utilisées durant notre projet.
Le machine learning est une technique de programmation informatique qui utilise des probabilités statistiques pour donner aux ordinateurs la capacité d'apprendre par eux-mêmes sans programmation explicite. Pour son objectif de base, le machine learning « apprend à apprendre » aux ordinateurs – et par la suite, à agir et réagir – comme le font les humains, en améliorant leur mode d'apprentissage et leurs connaissances de façon autonome sur la durée. L'objectif ultime serait que les ordinateurs agissent et réagissent sans être explicitement programmés pour ces actions et réactions. Le machine learning utilise des programmes de développement qui s'ajustent chaque fois qu'ils sont exposés à différents types de données en entrée.
Il existe plusieurs types d'apprentissage, parmi ces derniers nous citons l'apprentissage supervisé et non-supervisé.

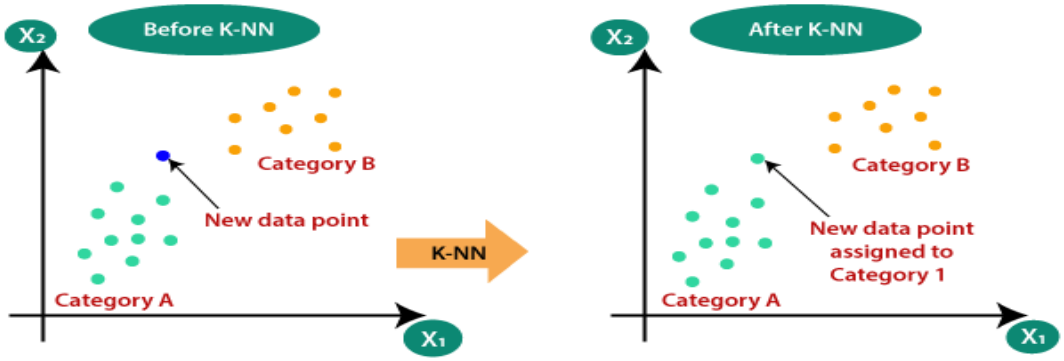
- **Apprentissage Supervisé** : L'algorithme d'apprentissage reçoit des données étiquetées (nombre de classes connu à l'avance) et la sortie souhaitée.

- **Apprentissage Non-supervisé** : Les données fournies à l'algorithme d'apprentissage ne sont pas étiquetées (nombre de classes est non connu à l'avance) et l'algorithme est invité à identifier les modèles dans les données d'entrée. Par exemple, le système de recommandation d'un site Web de commerce électronique où l'algorithme d'apprentissage découvre des articles similaires souvent achetés ensemble

I.Classification des données :

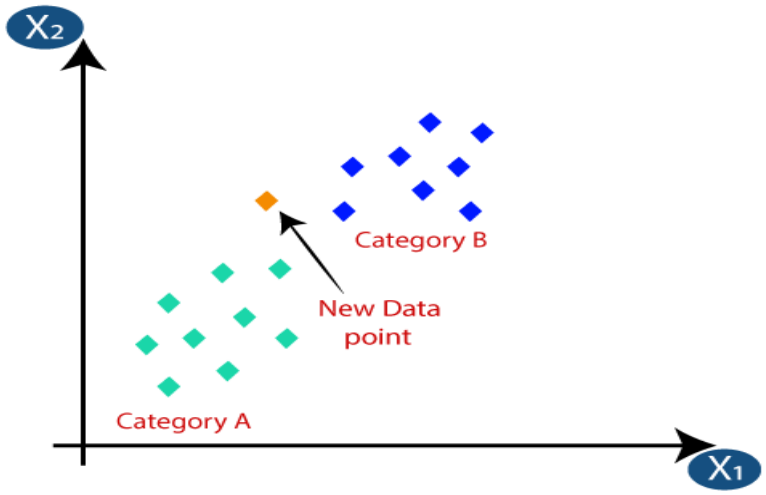
K-Nearest Neighbor est l'un des algorithmes d'apprentissage automatique les plus simples basé sur la technique d'apprentissage supervisé.
L'algorithme K-NN suppose la similitude entre le nouveau cas/données et les cas disponibles et place le nouveau cas dans la catégorie la plus similaire aux catégories disponibles.
L'algorithme K-NN stocke toutes les données disponibles et classe un nouveau point de données en fonction de la similitude. Cela signifie que lorsque de nouvelles données apparaissent, elles peuvent être facilement classées dans une catégorie de suites de puits en utilisant l'algorithme K-NN.
L'algorithme K-NN peut être utilisé pour la régression ainsi que pour la classification, mais il est principalement utilisé pour les problèmes de classification.
K-NN est un algorithme non paramétrique, ce qui signifie qu'il ne fait aucune hypothèse sur les données sous-jacentes.

Il est également appelé algorithme d'apprenant paresseux car il n'apprend pas immédiatement de l'ensemble d'apprentissage, au lieu de cela, il stocke l'ensemble de données et au moment de la classification, il effectue une action sur l'ensemble de données.
L'algorithme KNN lors de la phase d'apprentissage stocke simplement l'ensemble de données et lorsqu'il obtient de nouvelles données, il classe ces données dans une catégorie très similaire aux nouvelles données.

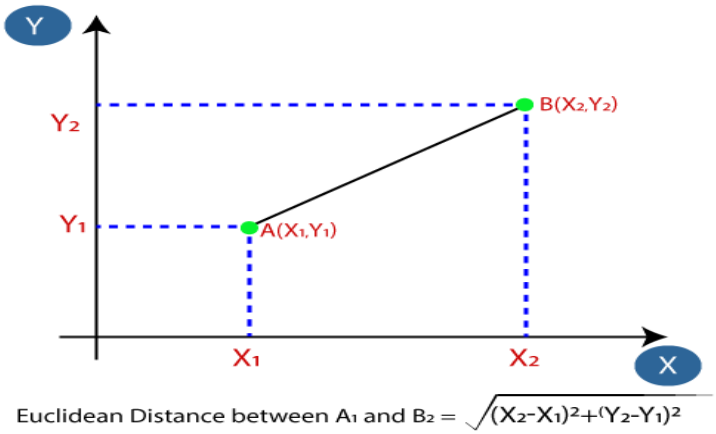


Le fonctionnement de K-NN peut être expliqué sur la base de l'algorithme ci-dessous :

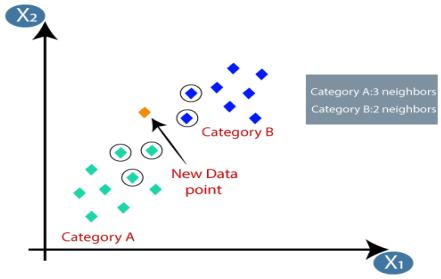
Étape 1 : Sélectionnez le nombre K des voisins
Étape 2 : Calculer la distance euclidienne du nombre K de voisins
Étape 3 : Prenez les K voisins les plus proches selon la distance euclidienne calculée.
Étape 4 : Parmi ces k voisins, comptez le nombre de points de données dans chaque catégorie.
Étape 5 : affectez les nouveaux points de données à cette catégorie pour laquelle le nombre de voisins est maximum.
Étape 6 : Notre modèle est prêt.
Supposons que nous ayons un nouveau point de données et que nous devions le mettre dans la catégorie requise. Considérez l'image ci-dessous :



- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



En calculant la distance euclidienne, nous avons obtenu les voisins les plus proches, en tant que trois voisins les plus proches dans la catégorie A et deux voisins les plus proches dans la catégorie B. Considérez l'image ci-dessous :



Comme nous pouvons le voir, les 3 voisins les plus proches appartiennent à la catégorie A, ce nouveau point de données doit donc appartenir à la catégorie A.

Les avantages	Les désavantages
Il est simple à mettre en œuvre.	Il faut toujours déterminer la valeur de K qui peut être complexe à un moment donné.
Il est robuste aux données d'entraînement bruyantes	Le coût de calcul est élevé en raison du calcul de la distance entre les points de données pour tous les échantillons d'apprentissage.
Cela peut être plus efficace si les données d'entraînement sont volumineuses.	

Étapes pour implémenter l'algorithme K-NN :

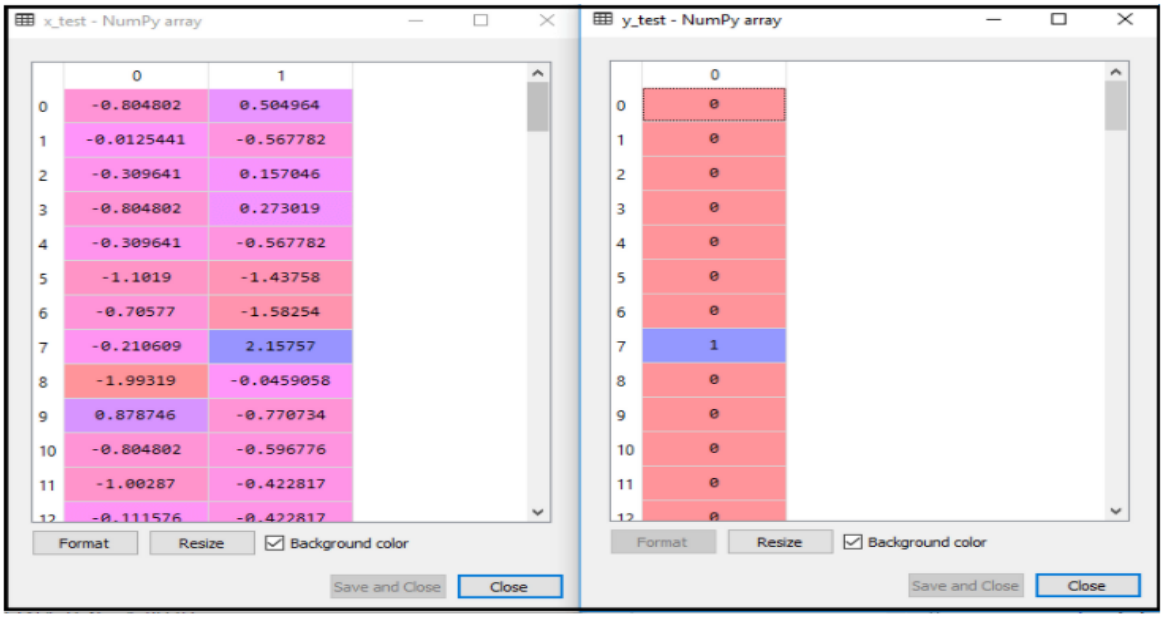
1. Étape de pré-traitement des données
2. Ajustement de l'algorithme K-NN à l'ensemble d'apprentissage
3. Prédire le résultat du test
4. Tester la précision du résultat (Création d'une matrice de confusion)
5. Visualisation du résultat de l'ensemble de test.

Étape de pré-traitement des données :

L'étape de pré-traitement des données restera exactement la même que la régression logistique. Ci-dessous le code pour cela :

```
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd

#importing datasets
data_set= pd.read_csv('user_data.csv')

#Extracting Independent and dependent Variable
x= data_set.iloc[:, [2,3]].values
y= data_set.iloc[:, 4].values

# Splitting the dataset into training and test set.
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_state=0)

#feature Scaling
from sklearn.preprocessing import StandardScaler
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)
```

En exécutant le code ci-dessus, notre ensemble de données est importé dans notre programme et bien prétraité. Après la mise à l'échelle des fonctionnalités, notre ensemble de données de test ressemblera à :



À partir de l'image de sortie ci-dessus, nous pouvons voir que nos données sont mises à l'échelle avec succès.

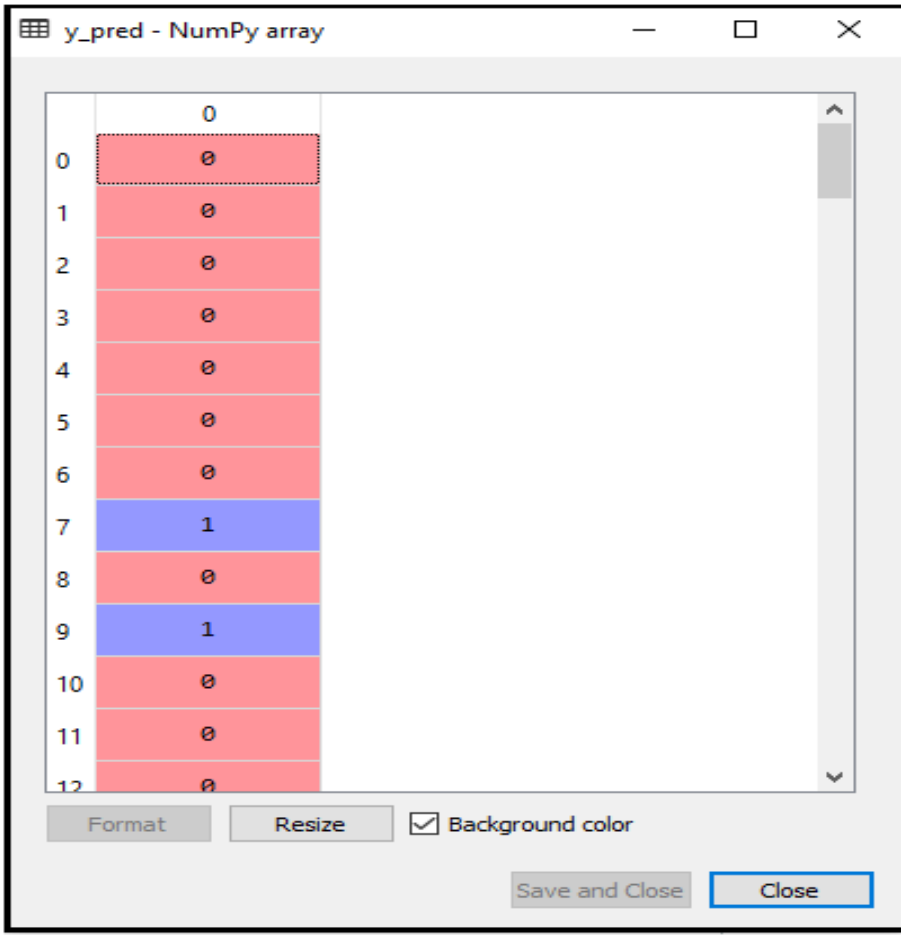
Ajustement du classificateur K-NN aux données d'entraînement :
Nous allons maintenant adapter le classificateur K-NN aux données d'apprentissage. Pour ce faire, nous allons importer la classe KNeighborsClassifier de la bibliothèque Sklearn Neighbors. Après avoir importé la classe, nous allons créer l'objet Classifier de la classe. Le paramètre de cette classe sera
n_neighbors : Pour définir les voisins requis de l'algorithme. En général, il en faut 5.
metric='minkowski' : C'est le paramètre par défaut et il décide de la distance entre les points.
p=2 : C'est équivalent à la métrique euclidienne standard.
Ensuite, nous adapterons le classificateur aux données d'apprentissage. Ci-dessous le code pour cela :

```
#Fitting K-NN classifier to the training set
from sklearn.neighbors import KNeighborsClassifier
classifier= KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2 )
classifier.fit(x_train, y_train)
```

Prédire le résultat du test : pour prédire le résultat de l'ensemble de test, nous allons créer un vecteur y_pred comme nous l'avons fait dans la régression logistique. Ci-dessous le code pour cela:

```
#Predicting the test set result
    y_pred= classifier.predict(x_test)
```

La sortie du code ci-dessus sera :



II.Environnement de travail

Dans cette section, nous présentons l'environnement matériel ainsi que l'environnement Logiciel.

Environnement logiciel

Au niveau de cette partie, nous mentionnons les différents logiciels utilisés au cours de notre projet :
Pour la partie de classification nous avons utilisé le langage de programmation Python avec Spyder comme un environnement de développement ainsi que leurs librairies.

Python est un langage de programmation interprété, orienté objet et de haut niveau avec une sémantique dynamique. Sa structure de données intégrée est de haut niveau. C'est le langage le plus utilisé dans les domaines de Machine Learning, Big Data et Data Science
Python est un langage de programmation interprété, orienté objet et de haut niveau avec

Anaconda Enterprise est une plate-forme de science des données prête pour l'entreprise sécurisée et évolutive qui permet aux équipes de gérer les actifs de science des données, de collaborer et de déployer des projets de science des données

Spyder, l'environnement de développement scientifique Python, est un environnement de développement intégré gratuit (IDE) inclus avec Anaconda. Il comprend des fonctionnalités d'édition, de test interactif, de débogage et d’introspection.

III.Les protocoles expérimentaux

Dans cette partie nous mesurons les performances des algorithmes, cette étape est indispensable et extrêmement importante pour produire un algorithme de qualité et qui réponde aux attentes métiers, donc nous faisons une comparaison entre le vecteur étiquettes réelles des classes et le vecteur d’étiquettes prédites pour chaque ensemble de données testé.

Dans notre cas nous considérons ces métriques :
• Accuracy : mesure les nombres des instances classées correctement par rapport aux échantillons totales, dont son équation est :

$$Accuracy = \frac{True_{positive} + True_{negative}}{True_{positive} + True_{negative} + False_{positive} + False_{negative}}$$

1. Matrice de confusion

La technique de la matrice de confusion aide à mesurer les performances de la classification par apprentissage automatique. Avec ce type de modèle, vous pouvez distinguer et classer le modèle avec les vraies valeurs connues sur l'ensemble des données de test. Le terme "matrice de confusion" est simple mais déroutant. Cet article va simplifier le concept afin que vous puissiez facilement comprendre et créer une matrice de confusion par vous-même.

Calcul de la matrice de confusion

Suivez ces étapes simples pour calculer la matrice de confusion pour l'exploration de données :

Étape 1

Estimez les valeurs des résultats de l'ensemble de données.

Étape 2

Testez l'ensemble de données à l'aide des résultats attendus.

Étape 3

Prédisez les lignes de votre jeu de données de test.

Étape 4

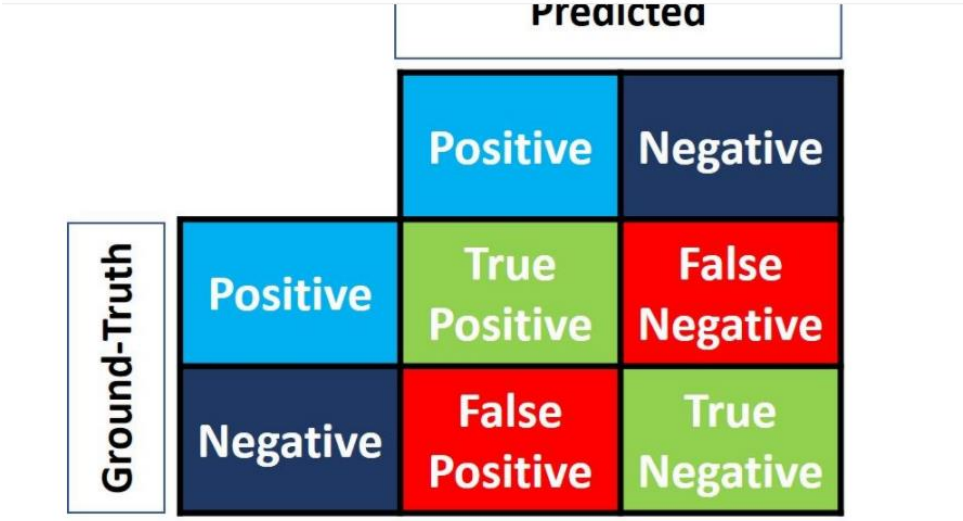
Calculez les résultats attendus et les prédictions. Vous devez prendre en compte les éléments suivants

– Le total des prédictions correctes de la classe
– Le total des prédictions incorrectes de la classe

Après avoir effectué ces étapes, vous devez organiser les chiffres selon les méthodes ci-dessous :

– Faire correspondre chaque ligne de la matrice avec la classe prédite.
– Faire correspondre chaque colonne de la matrice avec la classe réelle
– Inscrivez la classification correcte et incorrecte du modèle dans le tableau
– Incluez le total des prédictions correctes dans la colonne prédite.

Ajoutez également la valeur de la classe dans la ligne prévue.
– Incluez le total des prédictions incorrectes dans la ligne prévue et la valeur de classe dans la colonne prévue.



La courbe ROC

La caractéristique de fonctionnement du récepteur ou la courbe ROC est souvent utilisé comme graphique de visualisation pour mesurer les performances d'un classificateur binaire. Ce n'est pas une métrique du modèle, en soi, mais plutôt la représentation graphique du taux de vrais positifs (TPR) par rapport au taux de faux positifs (FPR) à divers seuils de classification de 0 à 1. Il peut également être considéré comme la puissance du type I Erreur (faux positifs) fonction de règle de décision
Ce graphique de visualisation ROC devrait aider à comprendre le compromis entre les taux. Nous pouvons également quantifier l'aire sous la courbe, également appelée AUC, à l'aide de la métrique roc_auc_score de scikit-learn, afin d'évaluer les performances du modèle. Plus le score est proche de 1, mieux le modèle distingue la classe, s'il est plus proche de 0,5, alors votre modèle fonctionne aussi mal que le tirage au sort.

Conclusion générale

L'intelligence artificielle est devenue un élément important de la société d'aujourd'hui pour que les entreprises peuvent exister dans un environnement très concurrentiel. Ce rapport et les travaux réalisés dans le cadre de la matière du projet BI ont été présentés. Le projet comprend la mise en place d'un système d'analyse du comportement humain. En fait, ce travail aborde d'abord le problème de la prédiction du diabète. Notre système est principalement composé de trois processus, le premier est le traitement Données basées sur la technologie d'intelligence d'affaires (BI), conçues pour cartographier Les données nécessaires. La deuxième partie est la partie analyse numérique, à travers l'application Les algorithmes d'apprentissage automatique sont basés sur la classification et la prédiction, dans cette dernière, nous nettoyons les données nécessaires de la modélisation. Afin d’obtenir le modèle le plus performant, nous avons utilisés comme protocoles expérimentaux :
Accuracy, précision, recall, f-mesure et la complexité de calcul.
Malgré sa grande complexité de calcul KNN reste le meilleur modèle adapté au cas étudié.