A. Phase 1 :

1. présentation de la problématique « Prédiction du diabète »:

Le diabète est une maladie qui empêche le corps d’utiliser correctement l’énergie fournie par Les aliments ingérés.

Cette maladie survient lorsque le pancréas ne s'écrit plus d’insuline ou lorsque le corps devient résistant à la quantité d’insuline produite

Le diabète est un trouble de l’assimilation, de l’utilisation et du stockage des sucres apportés par l’alimentation.

Cela se traduit par un taux de glucose dans le sang élevé : on parle d’hyperglycémie.

Les aliments sont composés de lipides (graisses), protéines (protéines animales ou végétales) et glucides (sucres, féculents). Ce sont eux qui fournissent l’essentiel de l’énergie

dont a besoin le corps pour fonctionner, passent dans l’intestin, puis rejoignent la circulation sanguine.

Le pancréas détecte l’augmentation de la glycémie.

Les cellules bêta du pancréas, regroupées en amas appelés îlots de Langerhans, sécrètent de l’insuline.

Insuline fonctionne comme une clé, elle permet au glucose de pénétrer dans les cellules de l’organisme où il va pouvoir être transformé et stocké le glucose diminue alors dans le sang.

Il ya 463 millions de personnes sont atteintes du diabète dans le monde d’après la Fédération internationale du diabète.



2) les différentes méthodes existantes :

 Les algorithmes d’apprentissage automatique supervisé qui traites la problèmes de la prédiction du diabètes sont  :

 K nearest neighbors

Decision Trees

 Random Forest

Support Vector Machine

Na¨ıves Bayes

 Les plus hauts taux de classification obtenus par l’algorithme de Random Forest , l’arbre de décision  et KNN sont respectivement 39% ,91% et 87%.

Phase 2:

1. Le choix de méthodes :

* K nearest neighbors (KNN)
* Arbre de décision
* Random Forest

2. Analyse détaillée des méthodes choisies:

1. K nearest neighbors (KNN) ou K plus proche voisins :

l’un des méthodes d’apprentissage supervisé le plus simple, utilisé pour résoudre des problèmes de classification et de régression. Son fonctionnement est de classer les nouveaux points de données en fonction de la similarité aux points de données voisins .

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

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

1. Arbre de décision :

C’est un algorithme parmi les algorithmes d’apprentissage supervisé le plus utilisé et le plus pratique, qui est adapté pour résoudre tout type de problèmes (classifications ou régressions) telle-que :

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

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

1. Radom Forest :

* Est Un algorithme d’apprentissage supervisé très populaire est également utilisé pour les problèmes de régression ou de classification.
* Basé sur un ensemble d’algorithmes d'apprentissage, qui est un processus de combinaison de plusieurs algorithmes pour résoudre un problème complexe et améliorer les performances du modèle.
* C’est un algorithme qui crée de nombreux arbres de décision sur divers sous-ensembles de l’ensemble de données.
* Elle prend la prédiction de chaque arbre et sur la base des votes majoritaires des prédictions, et elle prédit le résultat final.

**Algorithme de construction de Random forest :**

**1. Sélectionnez des échantillons aléatoires `a partir d’un ensemble de données d’entrainement.**

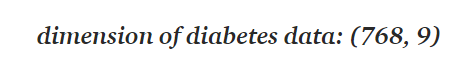
**2. Créer des arbres de décision pour chaque échantillon (sous-ensembles).**

**Ensuite on obtient le résultat de prédiction de chaque arbre de décision**

**3. Pour les nouveaux points le vote sera effectué pour chaque résultat prédit.**

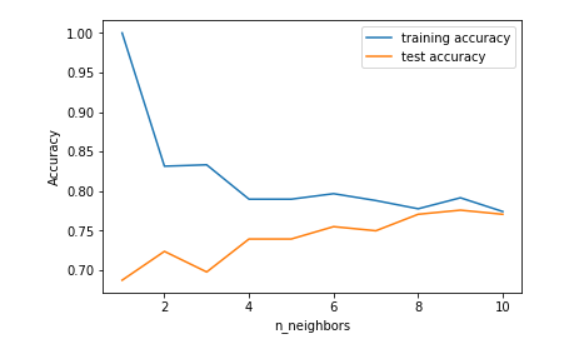
**4. sélectionnez le résultat de prédiction le plus voté comme résultat de prédiction final.**

4. Etude comparatives des méthodes choisies:



"Outcome" est la caractéristique que nous allons prédire, 0 signifie pas de diabète, 1 signifie diabète. Sur ces 768 points de données, 500 sont étiquetés comme 0 et 268 comme 1.

**Pour le KNN:**

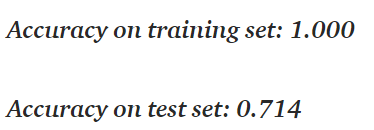


Le graphique ci-dessus montre la précision de l'ensemble de formation et de l'ensemble de test sur l'axe des y en fonction du paramètre n\_neighbors sur l'axe des x.

Si on choisit un seul voisin le plus proche, la prédiction sur l'ensemble de formation est parfaite. Mais lorsque plusieurs voisins sont pris en compte, la précision de l'entraînement chute, ce qui indique que l'utilisation d'un seul voisin le plus proche conduit à un modèle trop complexe.

La meilleure performance se situe quelque part autour de 9 voisins.

**Pour l’arbre de décision :**

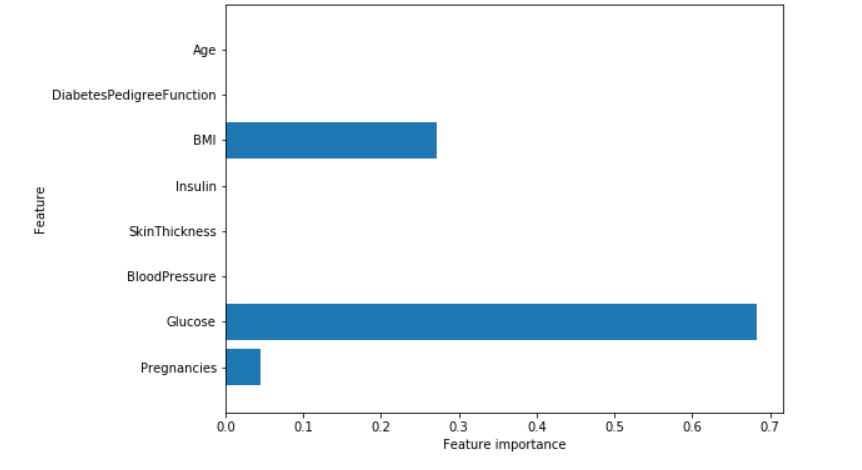
****

La précision sur l'ensemble d'apprentissage est de 100 %, alors que la précision sur l'ensemble de test est bien pire.

Cela indique que l'arbre est sur ajusté et ne se généralise pas bien aux nouvelles données.

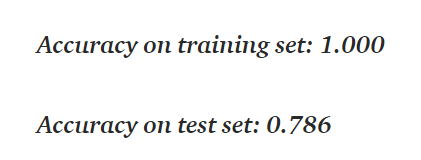
L'importance des caractéristiques évalue l'importance de chaque caractéristique dans la décision prise par l'arbre.

Il s'agit d'un nombre compris entre 0 et 1 pour chaque caractéristique, où 0 signifie "pas du tout utilisé" et 1 signifie "prédit parfaitement la cible". La somme des importances des caractéristiques est toujours égale à 1 :

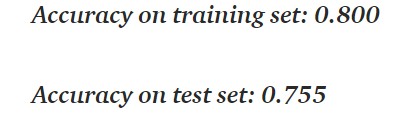


La fonction "Glucose" est de loin la plus importante.

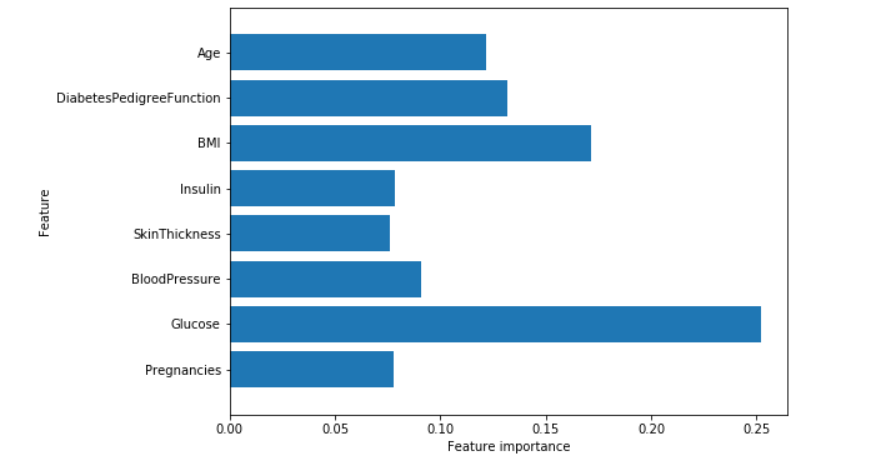
**Pour Random Forest :**

****

La forêt aléatoire donne une précision de 78,6 %, meilleure que le modèle de l’arbre de décision simple, sans réglage des paramètres.



ce qui indique que les paramètres par défaut de la forêt aléatoire fonctionnent bien.



Comme pour l'arbre de décision unique, la forêt aléatoire accorde également beaucoup d'importance à la caractéristique "Glucose", mais elle choisit également "BMI" comme deuxième caractéristique la plus informative dans l'ensemble. Le caractère aléatoire de la construction de la forêt aléatoire oblige l'algorithme à envisager de nombreuses explications possibles, ce qui fait que la forêt aléatoire saisit une image beaucoup plus large des données qu'un arbre de décision.