Réaliser par : Amel MOURAH M1 Informatique

Classification de fruits avec KNN, SVM et Decision trees

Présentation du projet :

Dans ce projet je fais une étude sur l'ensemble de données "**Fruit 36O**", cet ensemble de données contient 103 classes de fruits différents et 53177 images totales.

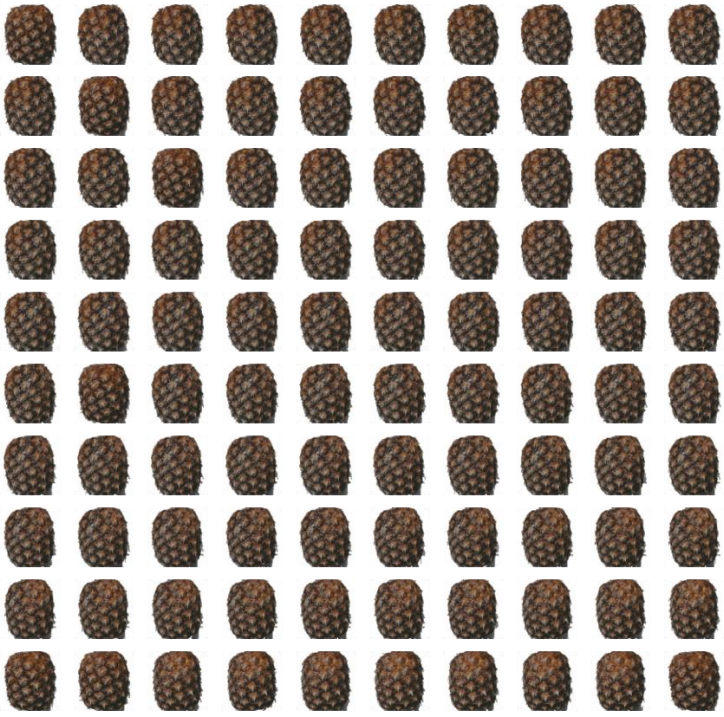
Je vais leurs appliquées algorithmes de classification, notamment SVM, K-NN, Decision Tree, pour une tâche de classification binaire.

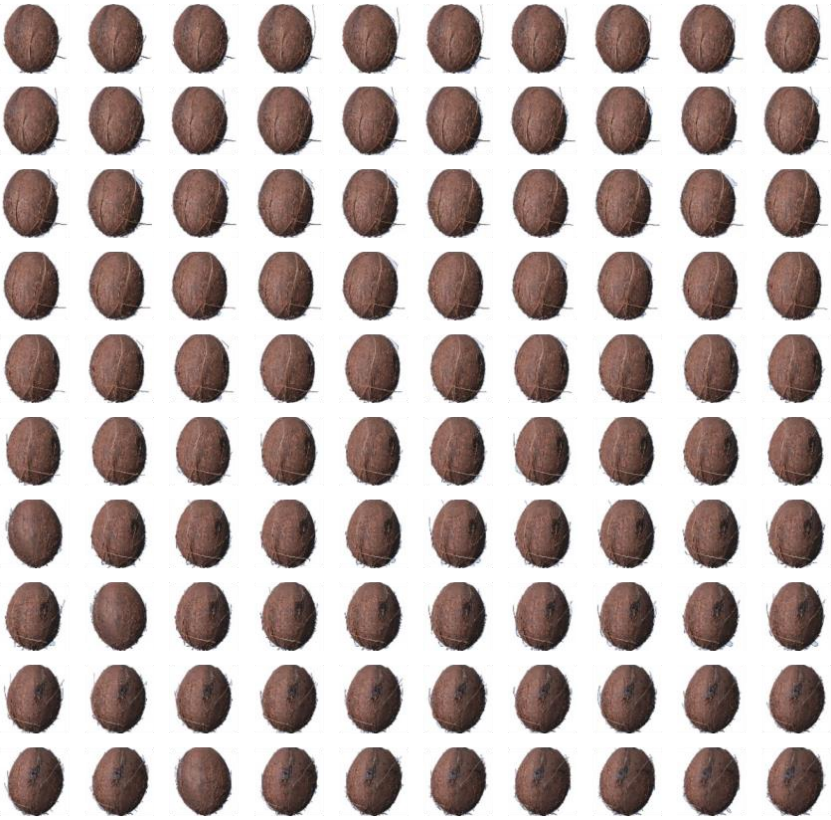
À la fin, je ferai une comparaison entre toutes les méthodes afin de trouver lesquelles d'entre elles fonctionnent mieux sur cet ensemble de données.

- ## Choisir les classes

Pour la classification binaire, j'ai décidé de prendre des Cocos et des Ananas car ils se ressemblent et donc la tâche de classification ne sera pas trop facile.

```
There are  490   TRAINING  images of  PINEAPPLE
There are  490   TRAINING  images of  COCOS
There are  166   TEST  images of  PINEAPPLE
There are  166   TEST  images of  COCOS
```

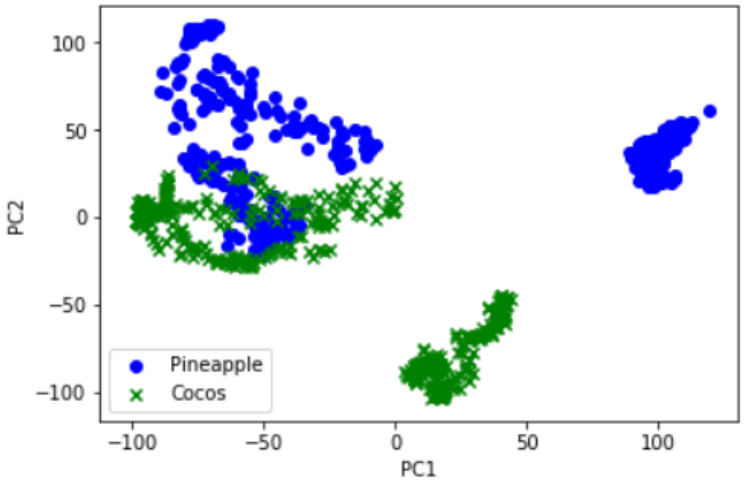


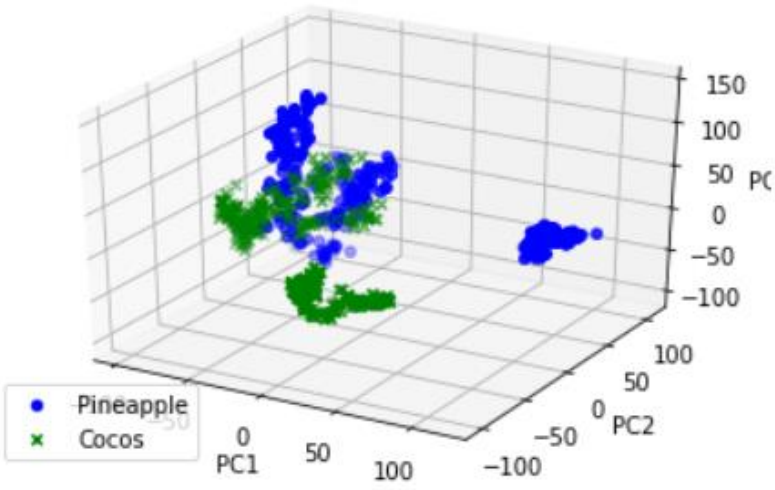


- **Données en dimension 2D et 3D**

Afin de découvrir comment nos données apparaissent dans une dimension inférieure, nous devons réduire la dimensionnalité de l'ensemble de données en 2 ou 3 dimensions afin de pouvoir les tracer et les visualiser. Pour ce faire, j'ai décidé d'utiliser l'analyse en composantes principales.

L'analyse en composantes principales est une technique utilisée pour réduire la dimensionnalité d'un ensemble de données tout en préservant les informations les plus complètes possible. Les données sont reprojetées dans un espace de dimension inférieure, en particulier nous devons trouver une projection qui minimise l'erreur quadratique lors de la reconstruction des données originales.





- ## Binary classification :

Nous allons classer notre ensemble de données, nous allons prendre deux classes afin d'effectuer une classification binaire classique, à la fin l'ensemble de données (ou une sous-partie de celui-ci) sera classé. On va utiliser 3 techniques différentes : SVM, KNN, Arbre de décision.

À la fin de la classification binaire, il y aura une comparaison entre toutes les méthodes.

1. SVM :

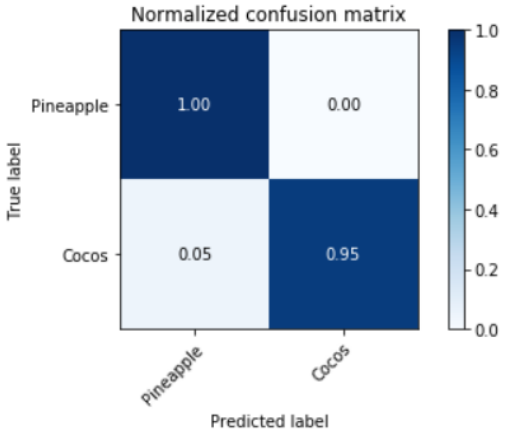
Ce sont des algorithmes d'apprentissage initialement construits pour la classification binaire.

L'idée est de rechercher une règle de décision basée sur une séparation par hyperplan de marge optimale.

Méthode relativement récente qui découle de premiers travaux théoriques de Vapnik et Chervonenkis en 1995, démocratisés à partir de 2000.

Le principe de l'algorithme est d'intégrer lors de la phase ´ d'apprentissage une estimation de sa complexité´ pour limiter le phénomène d'over-fitting.

```
Accuracy with SVM: 97.59%
```

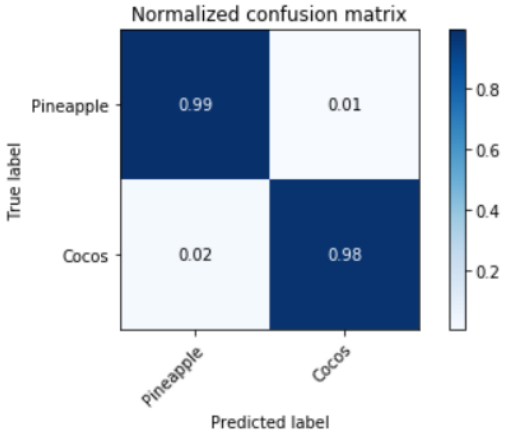


2. KNN :

L'algorithme K-NN (K-nearest neighbors) est une méthode d'apprentissage supervisé. Il peut être utilisé aussi bien pour la régression que pour la classification. Son fonctionnement peut être assimilé à l'analogie suivante "dis-moi qui sont tes voisins, je te dirais qui tu es…".

Pour effectuer une prédiction, l'algorithme K-NN va se baser sur le jeu de données en entier. En effet, pour une observation, qui ne fait pas parti du jeu de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches de notre observation. Ensuite pour ces 'k' voisins, l'algorithme se basera sur leurs variables de sortie (output variable) 'y' pour calculer la valeur de la variable 'y' de l'observation qu'on souhaite prédire.

```
Accuracy with K-NN: 98.80%
```



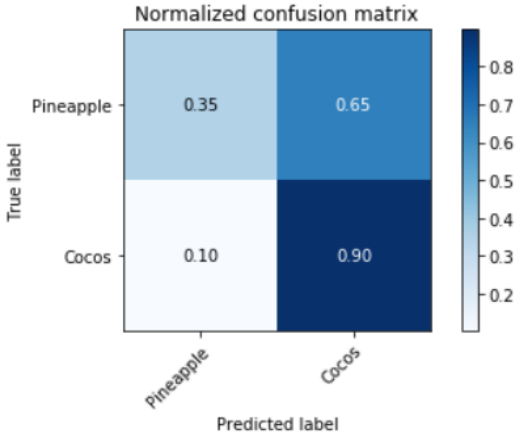
3. Arbre de décision

Ensemble de règles de classification basant leur décision sur des tests associes aux attributs, organises de manière arborescente.

Son principe est de Prédire la valeur d'un attribut (variable cible ou variable exogène) à partir d'un ensemble de valeurs d'attributs (variables prédictives ou variables endogènes). Une méthode simple, supervisée, et très connue de classification et de prédiction. Un arbre est équivalent `a un ensemble de règles de d´décision : un modèle facile à comprendre. Un arbre est composé de :

- Nœuds : classes d'individus de plus en plus fines depuis la racine.
- D'arcs : prédicats de partitionnement de la classe source.

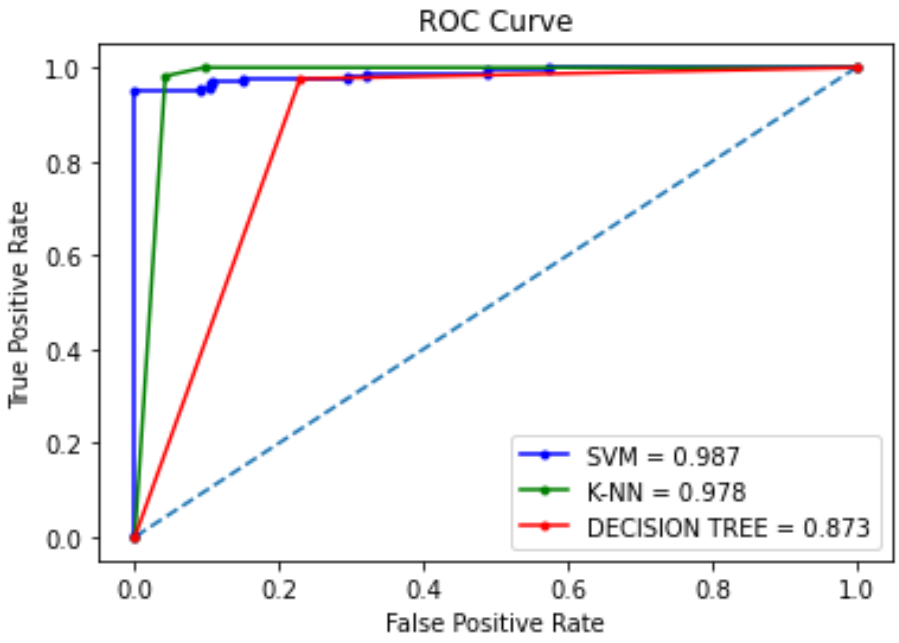
```
Accuracy with Decision Tree: 62.35%
```



- **Définir l'algorithme le plus adapté**

Afin de trouver l'algorithme le plus adapté à cet ensemble de données, on va utiliser la méthode d'évaluation : courbe ROC.

Afin de tracer la courbe ROC (Receiver Operating Characteristic), nous devons calculer le TPR et le FPR et choisir un certain nombre de seuils pour la classification (AUG). L'aire sous la courbe ROC, le traçage effectué TPR et FPR est utilisé comme matrice d'évaluation pour les différents classificateurs.



- ## Conclusion

SVM est l'algorithme de classification le plus optimal pour la classification de ce groupe de données choisi.