*Université Abdelmalek Essaadi Faculté des Sciences et techniques de Tanger Département Génie Informatique*



Rapport sur Atelier 2 :

« Classification »

Réalisé par :

***LASIRI Nouhaila***

***BENSLIMAN Hamza***

Encadré par :

***Pr . ELAACHAk LOTFI***

*Cycle Ingénieur: LSI S4*

*Machine Learning*

*Année : 2019/2020*

Objectif :

L’objective principal de cet atelier est de pratiquer les concepts de la classification, en traitant les données d’une Data Sets, ainsi d’évaluer les algorithmes pour construire le modèle adéquat à notre problématique.

Outils:

Visual Studio Code, Anaconda, Python, Pandas, Sklearn, matplotlib.

Data Sets:

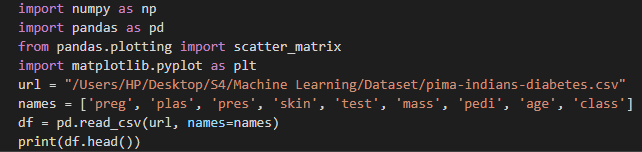
Diabète Data Set :

[https://www.kaggle.com/kumargh/pimaindiansdiabetescsv](%20https:/www.kaggle.com/kumargh/pimaindiansdiabetescsv)

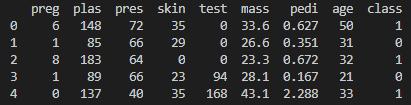
Partie 1 (Data Visualisation et Feature Selection et Normalisation):

1. Exploration des données :

*Code Source :*

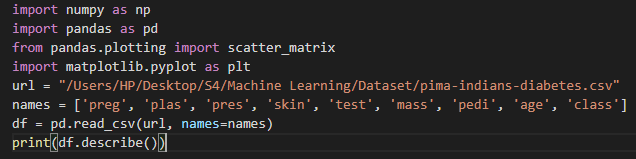


*Résultats :*

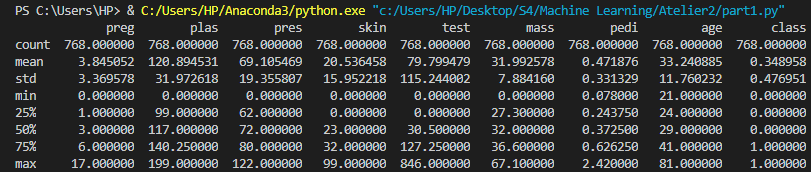


1. Le résumé statistique :

*Code Source :* Avec la fonction describe()

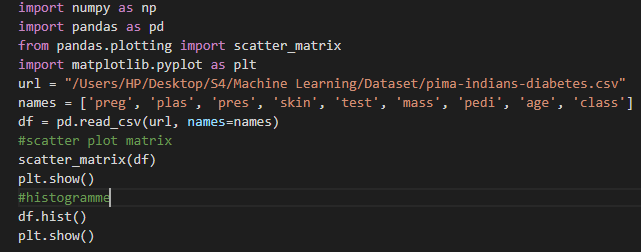


*Résultats :*

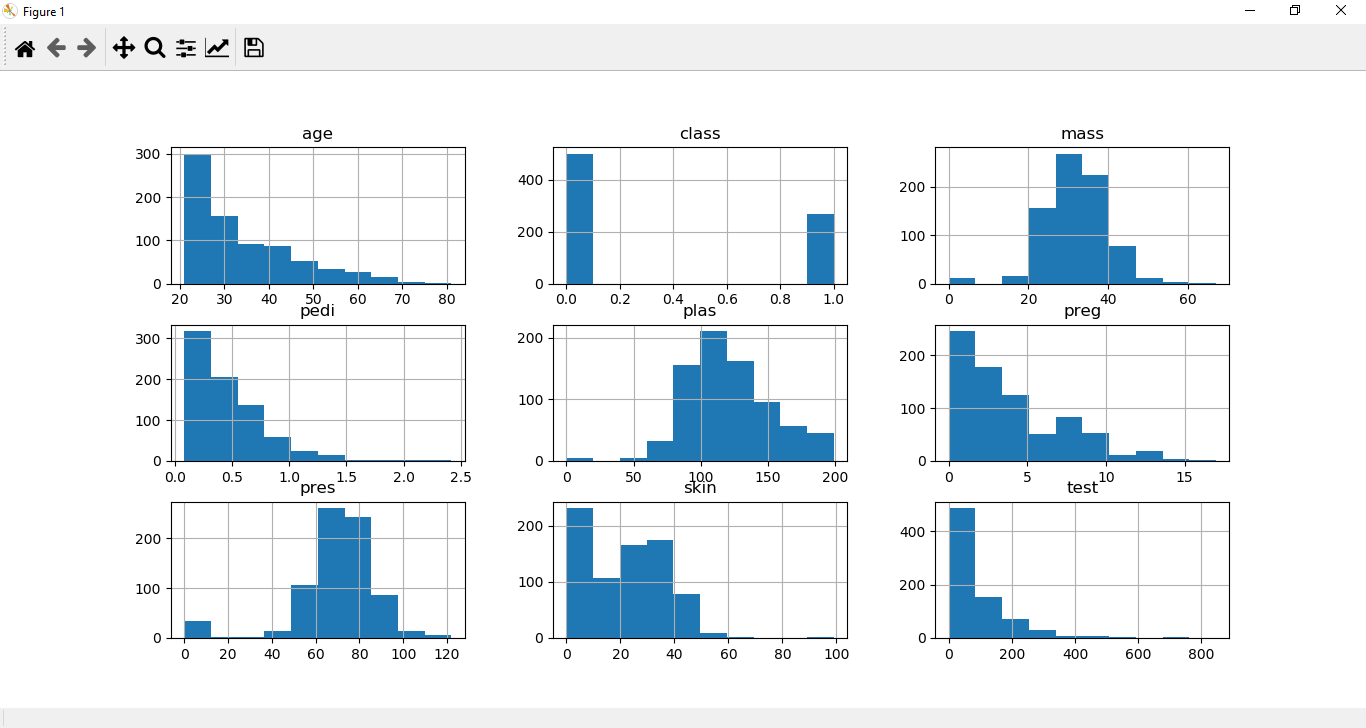


1. Les nuages des points du dataset :

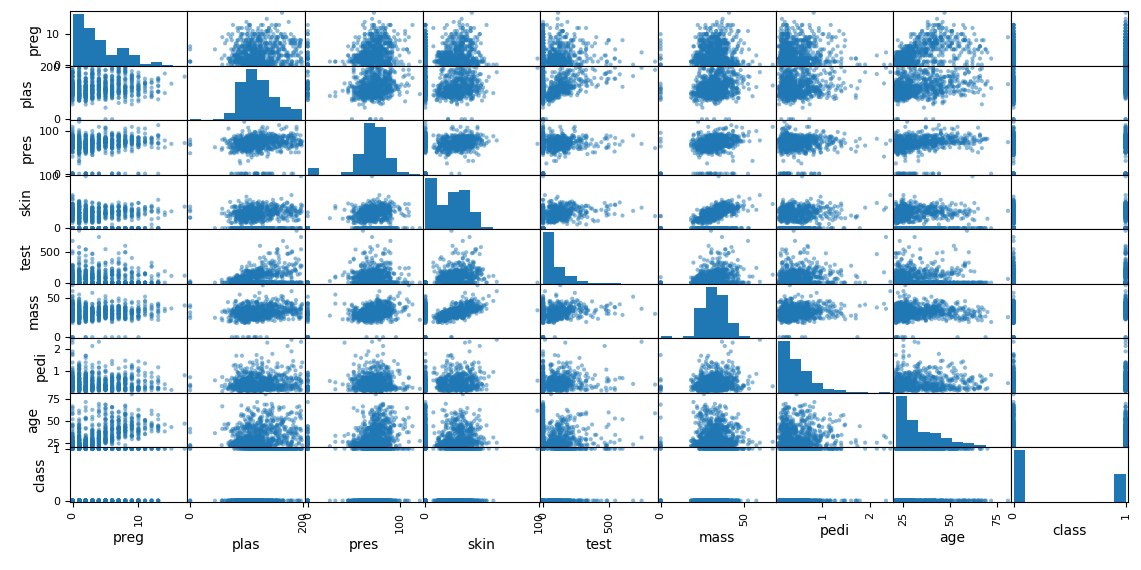
*Code Source :*



* Histogramme :

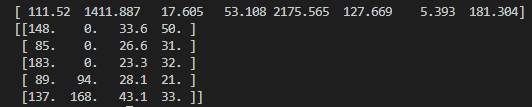
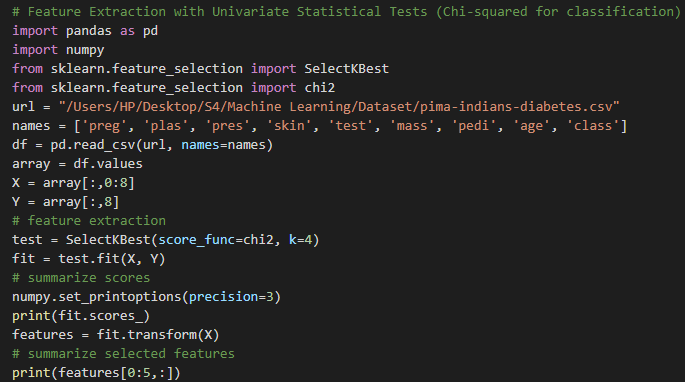


* Scatter plot matrix :

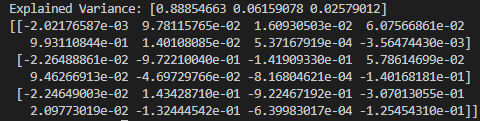
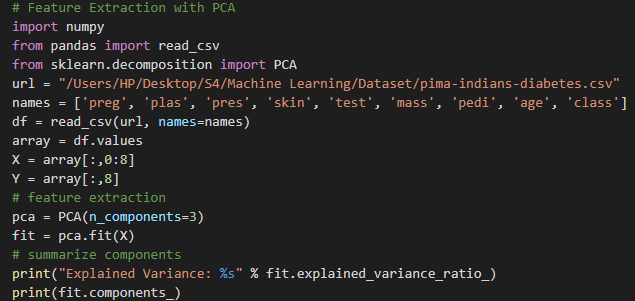


1. Application des 4 méthodes de Features selection :

* *Univariate Selection:*



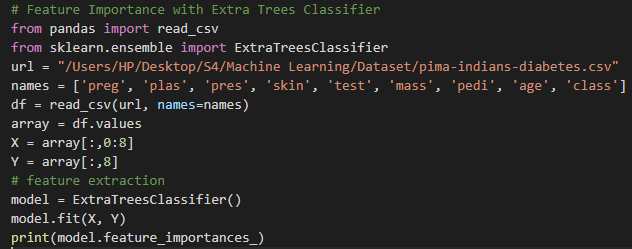
* *PCA :*



* *Recursive Feature Elimination:*

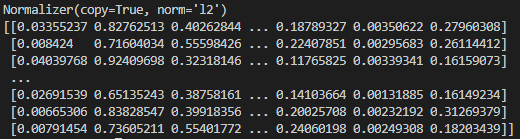
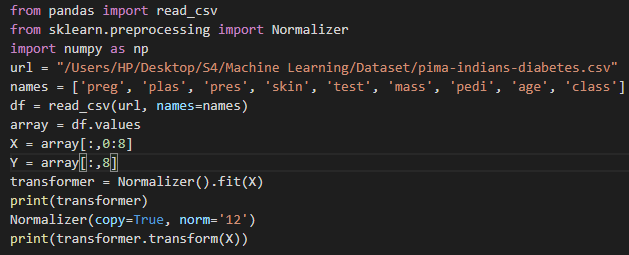


* *Feature Importance:*



1. Normalisation des données :

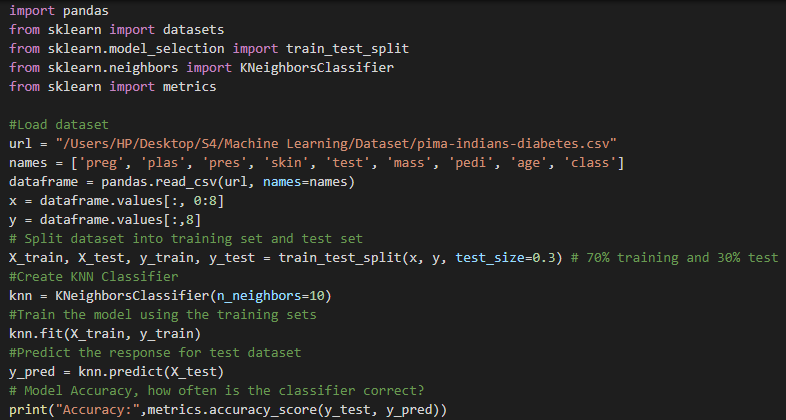
La normalisation s’applique juste sur les lignes et non pas les colonnes comme on a fait dans le code suivant en prenant les lignes du dataset.



Partie 2 (Classification choix de algorithme adéquat):

1. Entraînement du modèle :

* KNN :
* *Division du dataset en 70% pour l’entrainement et 30% pour le test avec train\_test\_split,*
* *Utilisation de la fonction KNeighborsClassifier de la libraire sklearn.neighbors,*
* *Entrainement du modèle.*
* *Comparaison des résultats prédits avec la partie du test.*

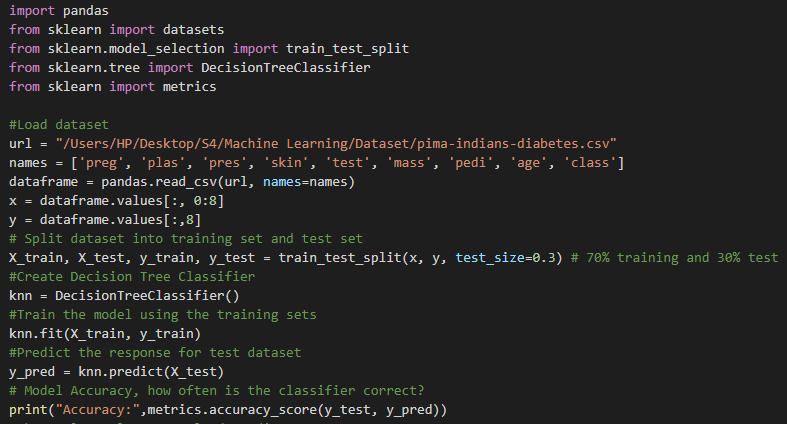


*\*Précision de l’entrainement* :



Presque les mêmes étapes pour les autres algorithmes.

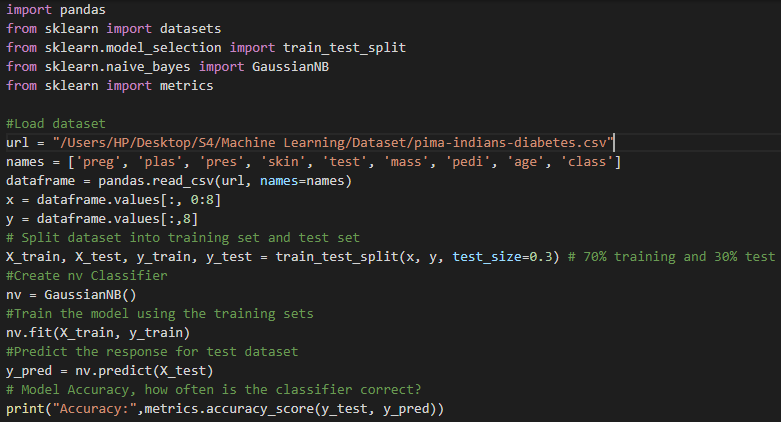
* Decision Tree :



*\*Précision de l’entrainement* :



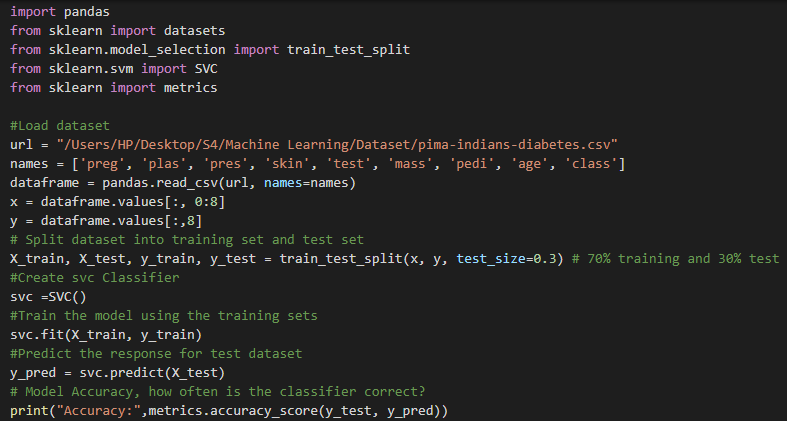
* Naive Bayes :



*\*Précision de l’entrainement* :



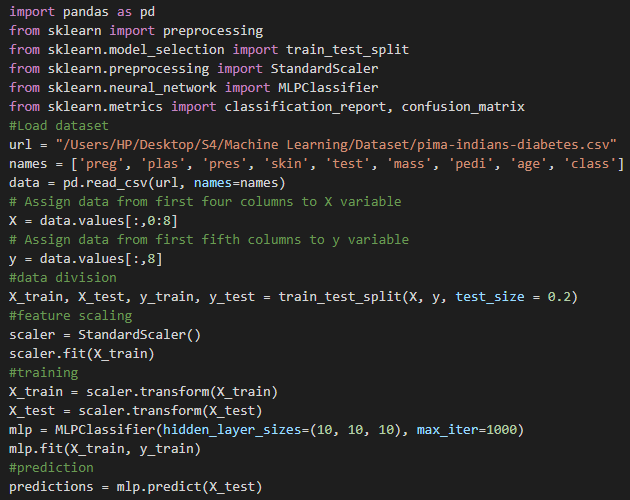
* SVM:



*\*Précision de l’entrainement* :

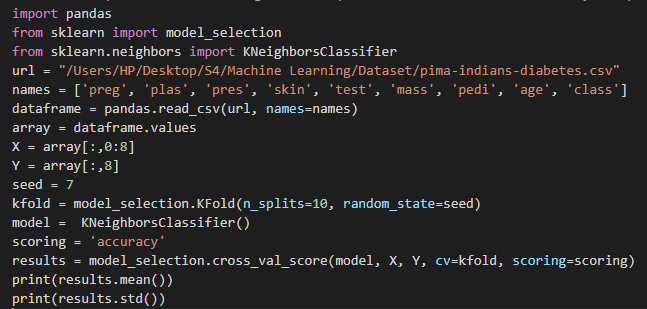


* ANN:



3- Evaluation de chaque modèle avec les métriques suivants :

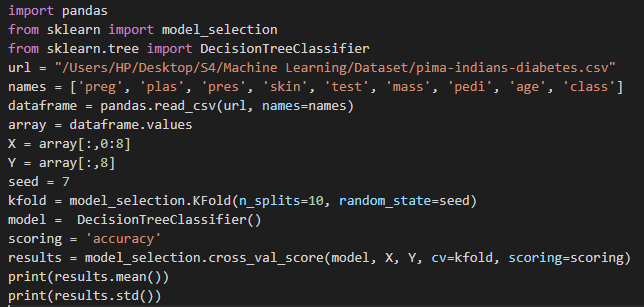
* *Classification Accuracy :*
* **KNN :**



Résultat :



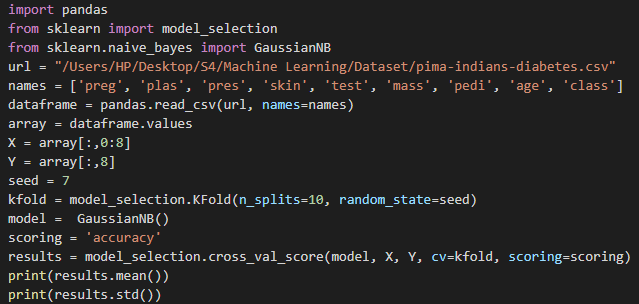
* **Decision Tree:**



Résultat :



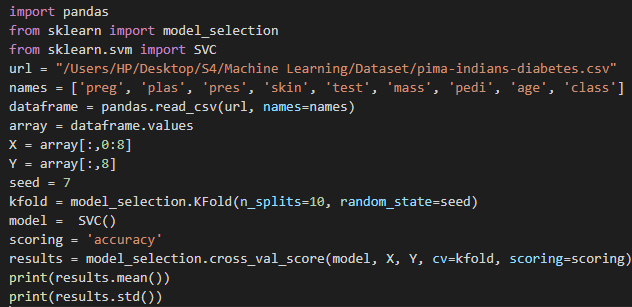
* **Naive Bayes:**



Résultat :



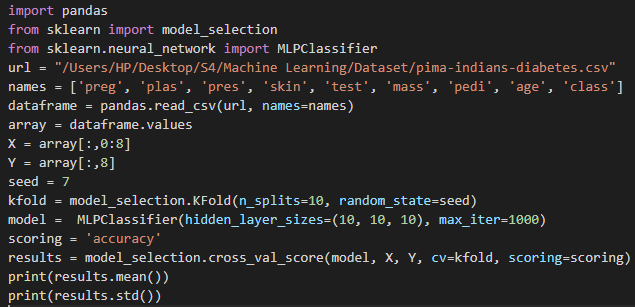
* **SVM:**



Résultat :



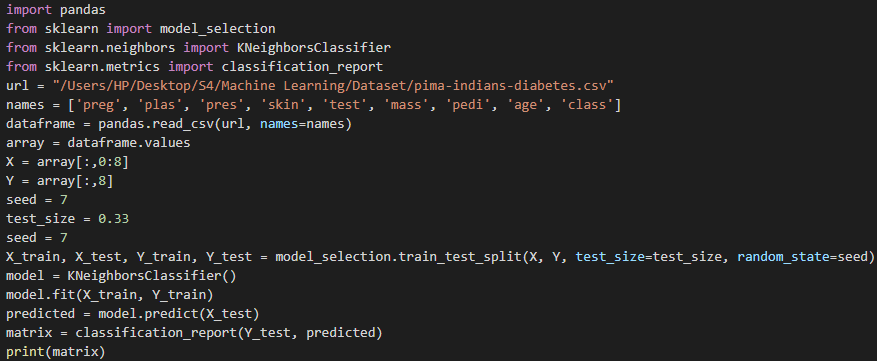
* **ANN:**



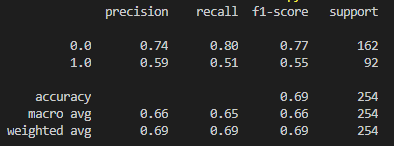
Résultat :



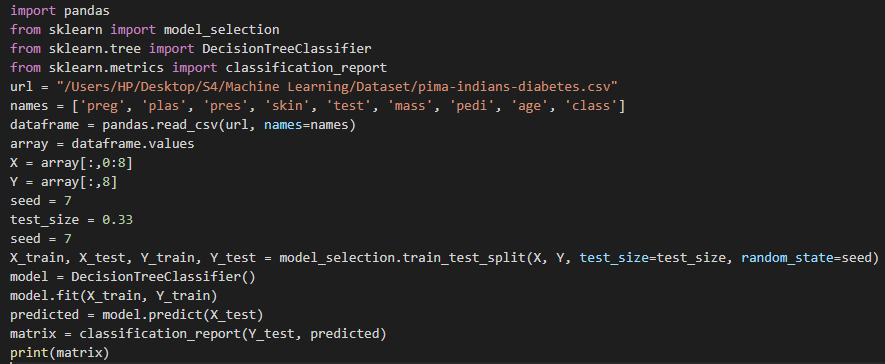
* *Classification Report:*
* **KNN :**



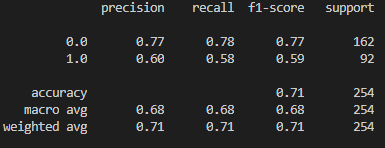
Résultat :



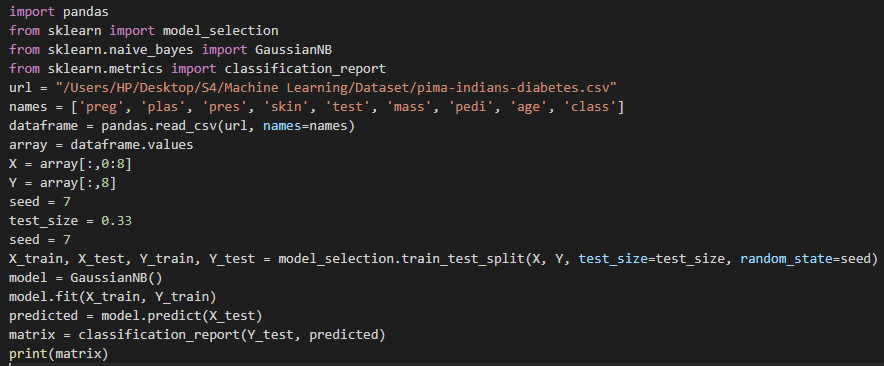
* **Decision Tree:**



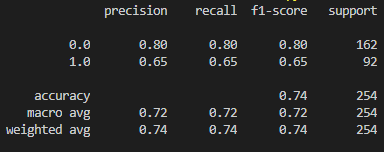
Résultat :



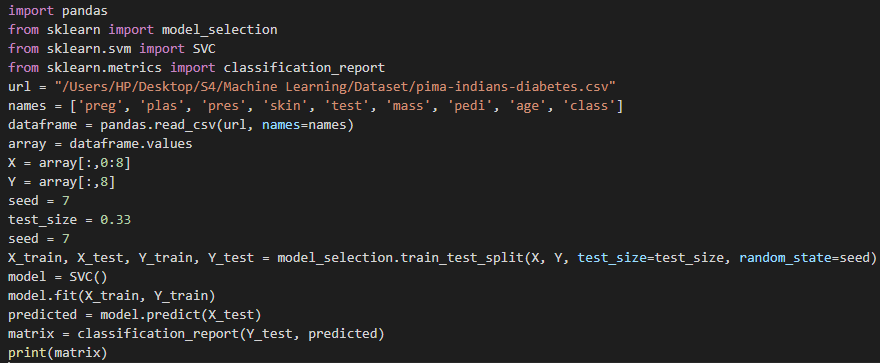
* **Naive Bayes:**



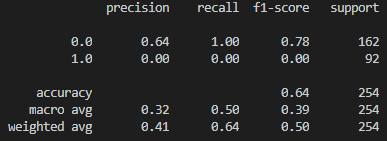
Résultat :



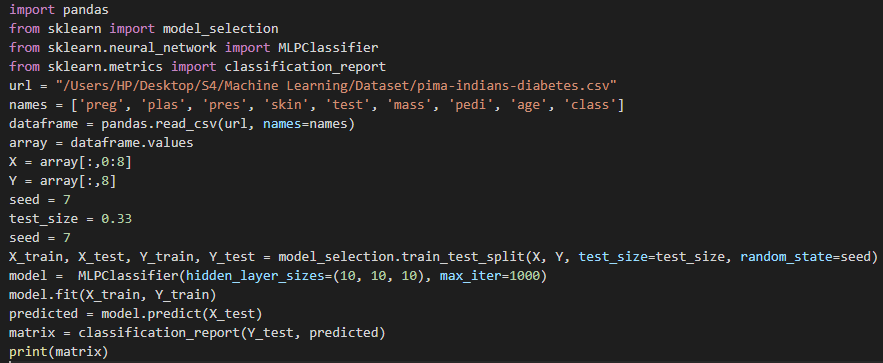
* **SVM:**



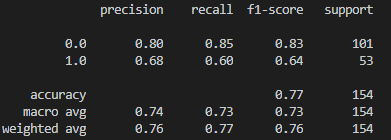
Résultat :



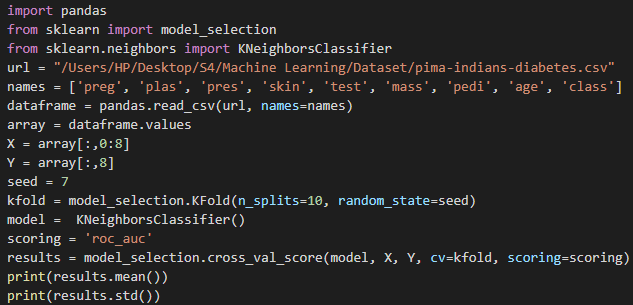
* **ANN:**



Résultat :



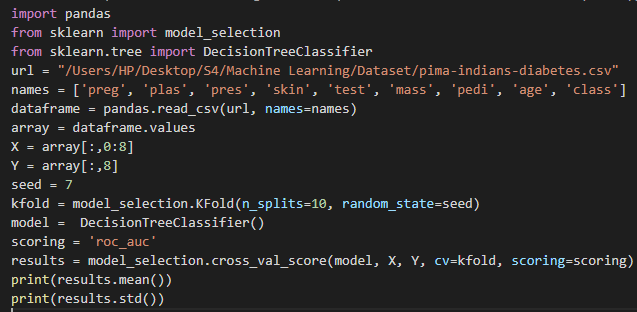
* *Area Under ROC Curve :*
* **KNN :**



Résultat :



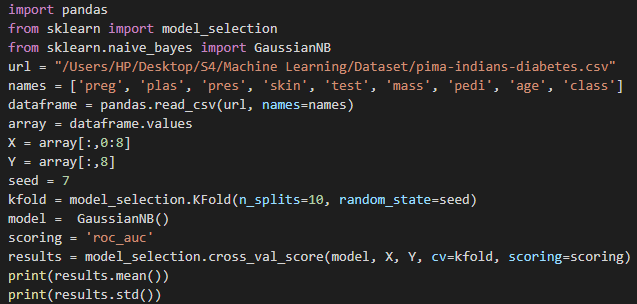
* **Decision Tree:**



Résultat :



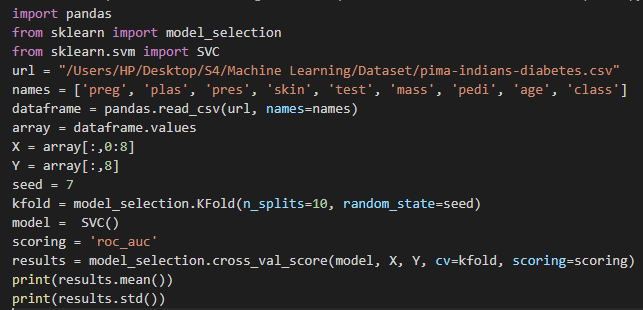
* **Naive Bayes:**



Résultat :



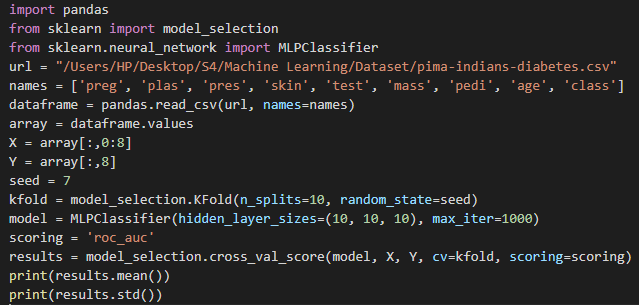
* **SVM:**



Résultat :



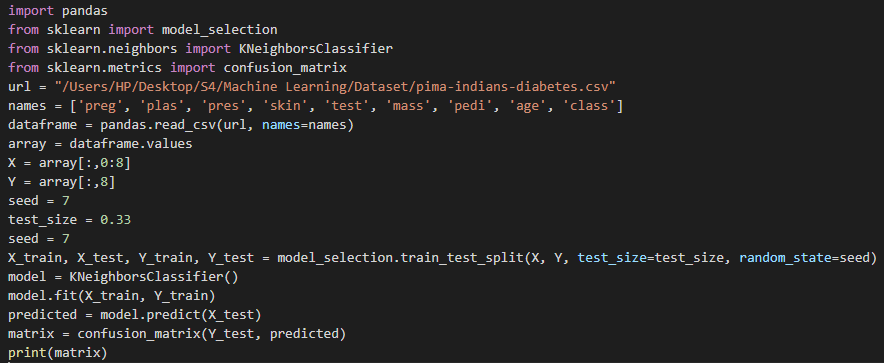
* **ANN:**



Résultat :



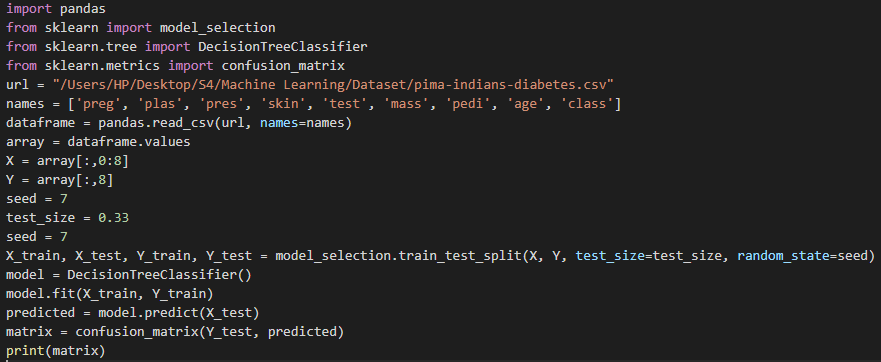
* *Confusion Matrix:*
* **KNN :**



Résultat :



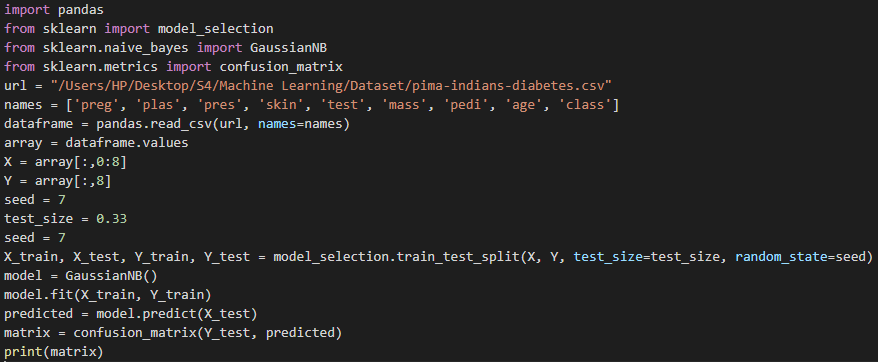
* **Decision Tree:**



Résultat :



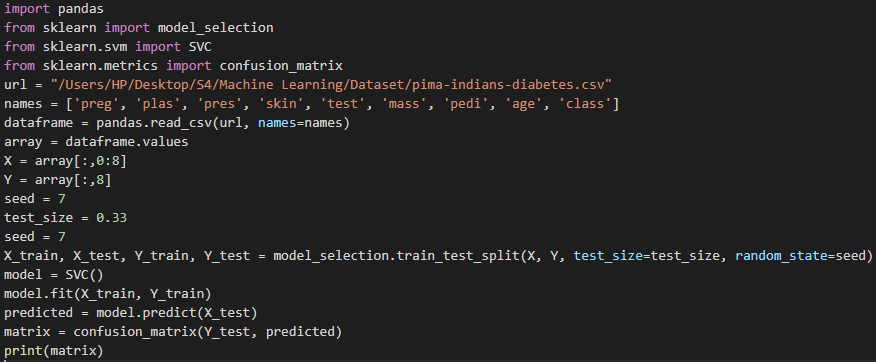
* **Naive Bayes:**



Résultat :



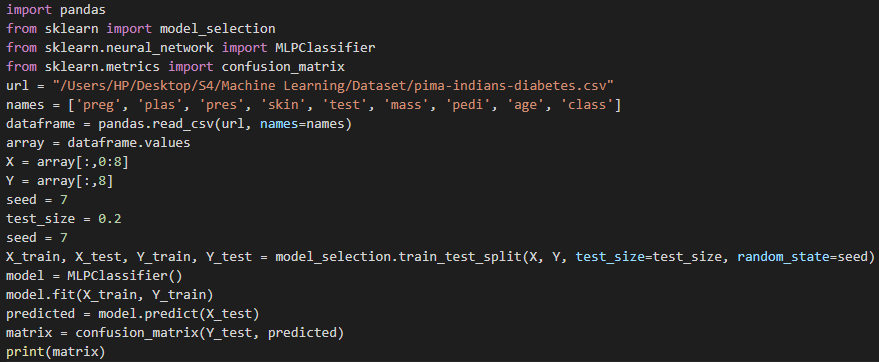
* **SVM:**



Résultat :



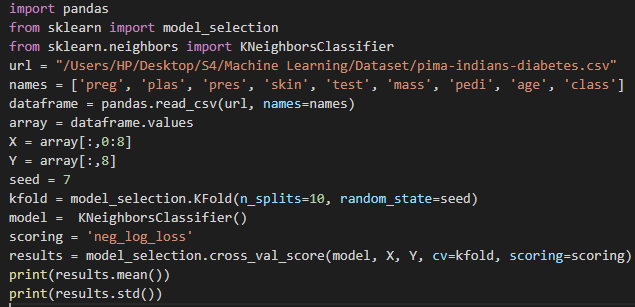
* **ANN:**



Résultat :



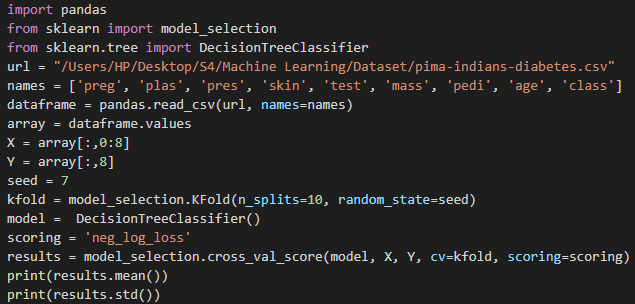
* *Logarithmic Loss :*
* **KNN :**



Résultat :



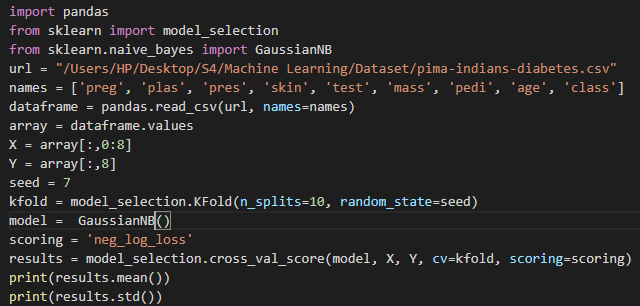
* **Decision Tree:**



Résultat :



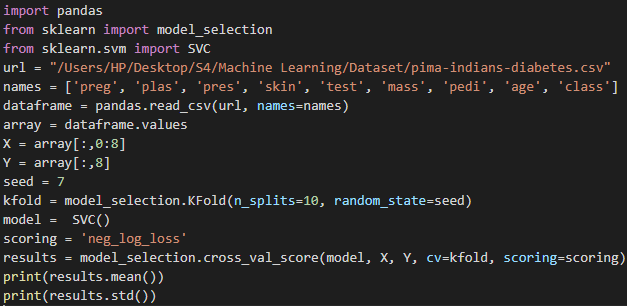
* **Naive Bayes:**



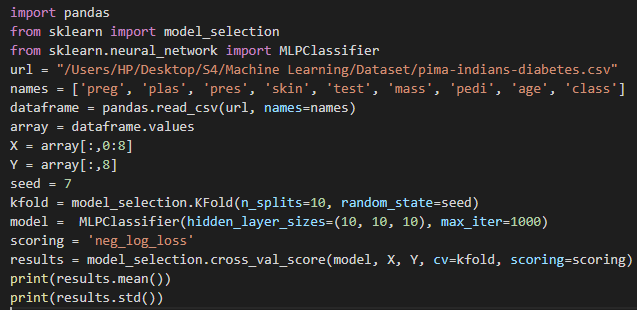
Résultat :



* **SVM:**



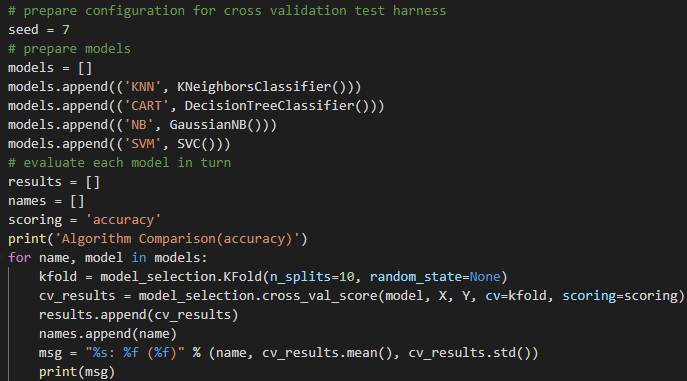
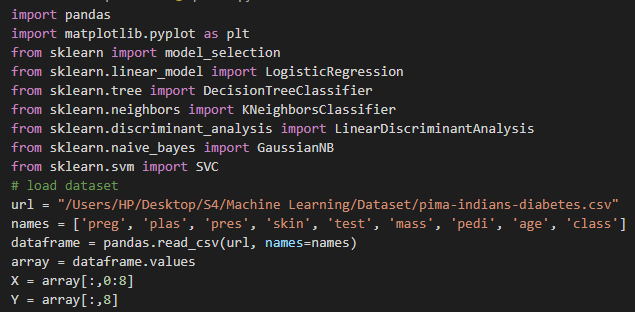
* **ANN:**



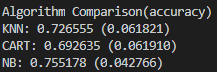
Résultat :



4. Comparaison de la performance des 5 algorithmes avec Spot-checking :



La comparaison :

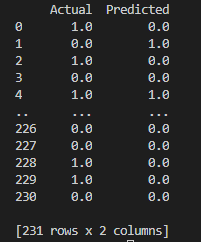


5- Prédiction les données du data set de test en utilisant les 5 modèles :

* **KNN :**



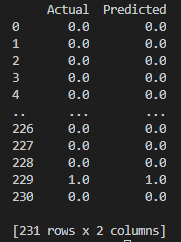
Prédiction :



* **Decision Tree :**



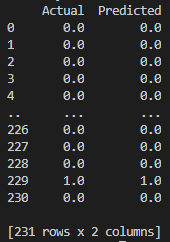
Prédiction :



* **Naive Bayes :**



Prédiction :



* **SVM :**



Prédiction :

