Apprentissage Automatique 1

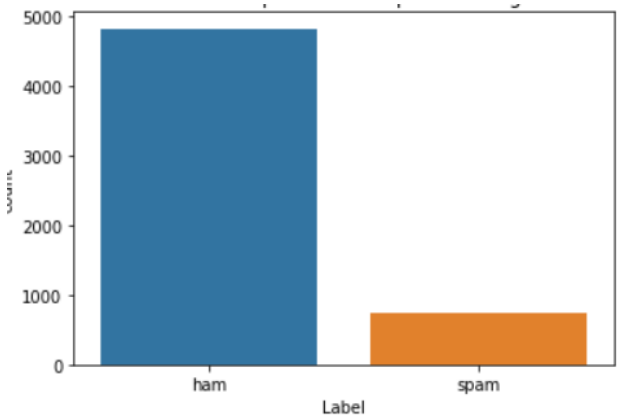
TP N°8– Random Forest

Exercice N°1 : Prédiction d'un email Spam

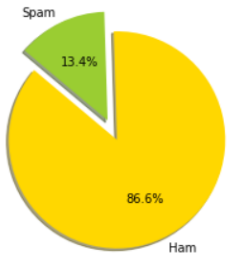
Le mot "spam" est un message indésirable. L'objectif de ce TP est d'utiliser l'algorithme de Random Forest pour prédire si un message est un spam ou non. On utilise un ensemble de *données* qui contient environ 5 572 emails spam et non spam.

Questions :

1. Importer les bibliothèques nécessaires.
2. Data Collection et Pre-Processing.
   - 2.1. Télécharger et afficher la base de données.
   - 2.2. Remplacer les valeurs nulles par une chaine de caractère vide.
   - 2.3. Afficher la base de données.
   - 2.4. Afficher le nombre de lignes et de colonnes.
   - 2.5. Remplacer spam et ham par une classification binaire :
     Spam : 0
     Ham :1
   - 2.6. Remplacer les champs messages et catégorie par :
     Message :X
     Catégorie : Y
   - 2.7. Afficher X et Y.
   - 2.8. Tracer la distribution de la base de données sous cette forme : utiliser la fonction **countplot**



2.9. Tracer la distribution de la base de données sous cette forme :



3. Diviser la base de donnée en deux partie (Testing (20%),Training(80%)).
4. Transformer le texte du mail en un vecteur pour l'utiliser dans l'algorithme de Random Forest : utiliser la fonction **TfidfVectorizer() ou CountVectorizer(),** n'oubliez pas de l'importer.
5. Normaliser le X_train_vecteur et X_test_vecteur.
6. Convertir y_train et y_test en integers.
7. Construire le modèle de Random Forest
8. Appliquer l'apprentissage automatique.
9. Calculer la prédiction pour X_train_vecteur, calculer l'accuracy score.
10. Calculer la prédiction pour X_test_Vecteur, calculer l'accuracy score.
11. Calculer la probabilité de la prédiction : exemple :

```
rfc = RandomForestClassifier(random_state=4)
y_train_pred  = rfc.predict(X_train_vecteur)
y_train_prob = rfc.predict_proba(X_train_vecteur)[:,1]
```

12. Afficher la matrice de confusion pour l'apprentissage : **confusion_matrix(y_train,y_train_pred)**
13. Calculer ROC AUC pour Train : **roc_auc_score(y_train, y_train_prob))**
14. Refaire les questions 11,12 et 13 pour la partie test.
15. Tester l'algorithme de Random Forest pour prédire si un mail est spam ou non :