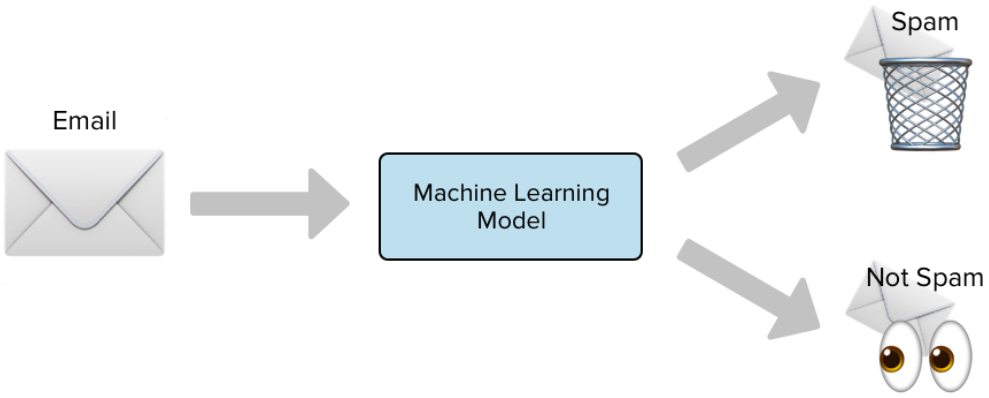
TP 5 : Détection de spam avec la régression logistique (RL)

Qu'est-ce qu'un email spam ?

- Un email spam est un email non sollicité et non pertinent, envoyé en grands lots vers des boites emails d'utilisateurs.
- Le but du spammer sont divers : les publicités pour les sites produits/Web, messages en chaines, assurer un gain rapide d'argent, la pornographie, l'usurpation d'identités, etc.



Exemples :

TrustedBank™

Dear valued customer of TrustedBank,

We have recieved notice that you have recently attempted to withdraw the following amount from your checking account while in another country: $135.25.

If this information is not correct, someone unknown may have access to your account. As a safety measure, please visit our website via the link below to verify your personal information:

http://www.trustedbank.com/general/custverifyinfo.asp

Once you have done this, our fraud department will work to resolve this discrepency. We are happy you have chosen us to do business with.

Thank you,
TrustedBank

Member FDIC © 2005 TrustedBank, Inc.

Filtres anti-spams

- Les filtres de spams basés sur l'analyse de texte vérifient l'existence/absence de certains mots ou de symboles.

- Dans un email, la présence de mots, tels que: héritage, viagra, loterie, dollars, etc., et de symboles tels que: '$', '¥', '€', '!', etc., augmentent la probabilité d'un spam.
- Ces probabilités sont estimées à partir d'un ensemble d'apprentissage "D" contenant des emails étiquetés.
- Les filtres peuvent faire des erreurs. Idéalement, les filtres doivent s'adapter et s'améliorer avec le temps.

Le jeu de données (Dataset) « Spambase » :

1) L'ensemble d'entrainement est créé par Mark Hopkins et al. de Hewlett-Packard Labs.

https://archive.ics.uci.edu/ml/datasets/Spambase

2) L'ensemble contient 4601 emails. Chaque email possède 57 valeurs attributs reflétant les propriétés de l'email. Parmi ces attributs, on a :
   - 48 sont des fréquences de certains mots.
   - 06 sont des fréquences de certains caractères.
   - 03 comptent la longueur de chaines non interrompues.

Exercice :

1. Étudier la régression logistique **(LogisticRegression)** sur Python.

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

2. Utiliser le classificateur de Bayes (CB) et la régression logistique (RL) pour classer les données de spams en se basant sur le dataset **spambase**.
3. Utiliser une validation croisée pour calculer l'erreur de classification (ex. moyenne de 10 validation en retenant à chaque fois 10% de données pour la validation). Comparer les résultats obtenus par les deux modèles implémentés (à savoir le CB et la RL).