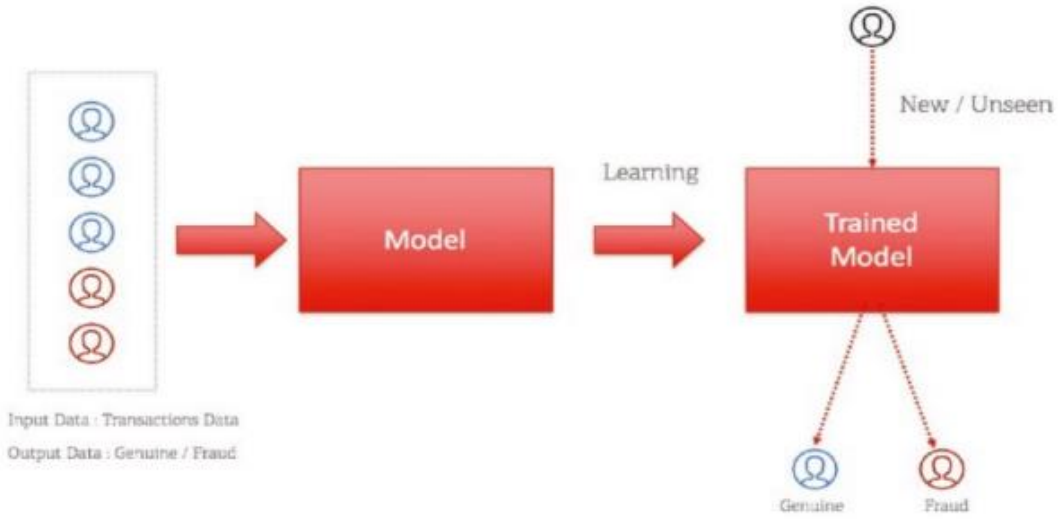
Chapitre 6: Apprentissage automatique supervise

A. Notes de cours

L'apprentissage automatique peut être divisé en quatre grandes catégories :

- l'apprentissage automatique supervisé : le processus d'apprentissage est supervisé, car l'algorithme d'apprentissage automatique utilisé corrige ses prédictions en fonction de la sortie réelle. les étiquettes ou la sortie correctes sont déjà connues pendant le modèle phase de formation, et, par conséquent, l'erreur peut être réduite en conséquence. Son processus est le suivant :



La catégorisation de l'apprentissage supervisé est basée sur les types de sorties ou de variables cibles utilisées pour la prédiction :

• Régression : utilisée lorsque la valeur cible qui est prédite est de nature continue ou numérique

• Classification : utilisée si la variable cible est une valeur discrète ou de nature catégorielle.

On parle de classification binaire lorsque la cible ou la variable de sortie ne contient que deux catégories au maximum.

on parle de classification multi-classe lorsque la cible ou la variable de sortie contient plus de deux catégories

La régression linéaire fait référence à la modélisation de la relation entre un ensemble de variables indépendantes et les variables de sortie ou dépendantes (numériques).

- l'apprentissage automatique non supervisé et, dans une moindre mesure
- l'apprentissage automatique semi-supervisé
- l'apprentissage automatique par renforcement.

Le modèle linéaire généralisé (GLM) est une version avancée de la régression linéaire qui considère que la variable cible a une distribution d'erreur autre qu'une distribution normale préférée.

On exécuter le GLM pour plusieurs distributions, telles que : 1. Binomial 2. Poisson 3. Gamma 4. Tweedie

L'algorithme de régression de l'arbre de décision peut être utilisé à la fois pour la régression et la classification. Il est assez puissant pour bien ajuster les données, mais présente le risque élevé de surajuster parfois les données. Les arbres de décision contiennent plusieurs divisions basées sur l'entropie ou les indices de Gini.

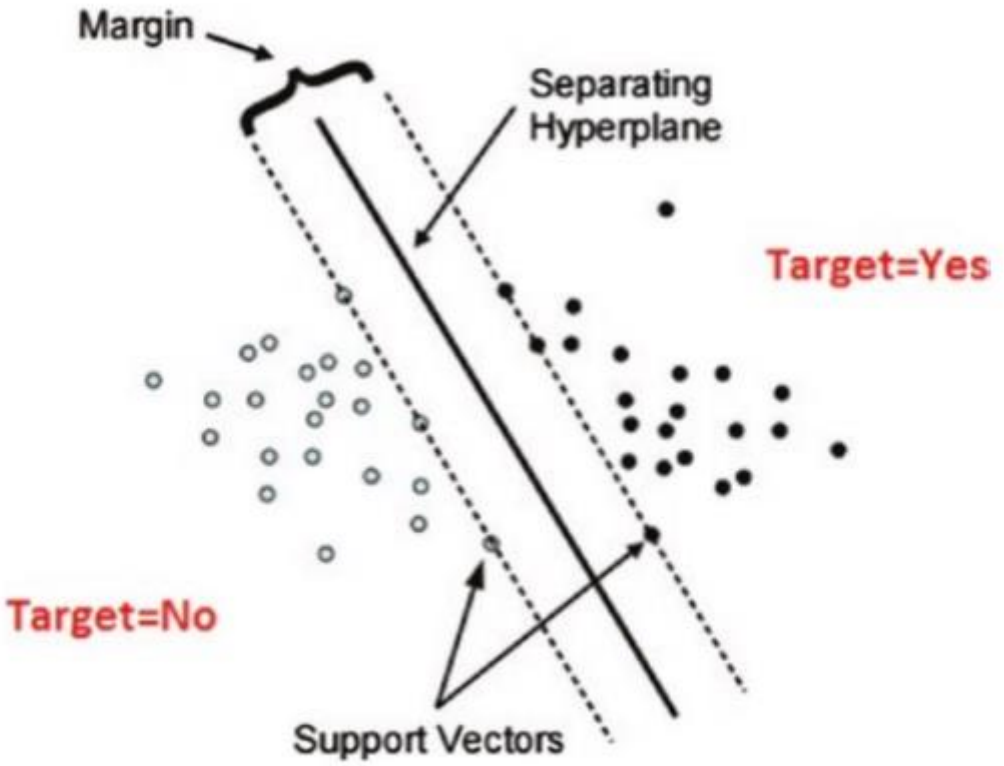
Random forest regressors : sont une collection de plusieurs arbres de décision individuels construits à l'aide de différents échantillons de données et c'est aussi une technique d'assemblage qui adopte une approche d'ensachage

A gradient-boosted tree (GBT) regressor est une technique d'assemblage, qui utilise l'amplification sous le capot. Une différence majeure entre le bagging et le boosting est que dans le bagging, les modèles individuels qui sont construits sont de nature parallèle, ce qui signifie qu'ils peuvent être construits indépendamment les uns des autres, mais dans le boosting, les modèles individuels sont construits de manière séquentielle.

La régression logistique est considérée comme l'un des modèles de référence, en raison de sa simplicité et de son interprétabilité. Sous le capot, c'est assez similaire à la régression linéaire

les arbres de décision peuvent être utilisés pour la classification ainsi que pour la régression. Support vector machines (SVMs) sont utilisées pour les tâches de classification, car elles trouvent l'hyperplan qui maximise la marge (distance perpendiculaire) entre deux classes. Toutes les instances et classes cibles sont représentées sous forme de vecteurs dans un

espace de grande dimension



Random Forest Classifier est une collection de plusieurs classificateurs d'arbre de décision. Il fonctionne sur le mécanisme de vote et prédit la classe de sortie qui a reçu le maximum de votes de tous les arbres de décision individuels.

On peut utiliser ParamgridBuilder et CrossValidator. les valeurs différentes dans la grille des paramètres pour trois hyperparamètres peuvent etre : maxDepth, maxBins, et numTree