QU'EST CE QU'UN DATASET

X : Features / Colonnes / Inputs

Y : Target / Variables / Labels

X

22	rouge	France	32.6	1/10/20
30	vert	Allemagne	7.2	9/09/20
42	bleu	∅	44.4	10/09/19

Y

Yes
No
Yes

* Les lignes représentent des "exemples"
* Chaque ligne de X est associée à une ligne de Y

Supervised Learning

The Mission

1. À partir d'un DATASET LABELLISÉ... → Isoler les Features des TARGETS... → Pour obtenir X et Y

2. À partir de X et Y → Créer un Modèle f → Tel que
   - f(X) = prédictions
   - prédictions ≈ Y

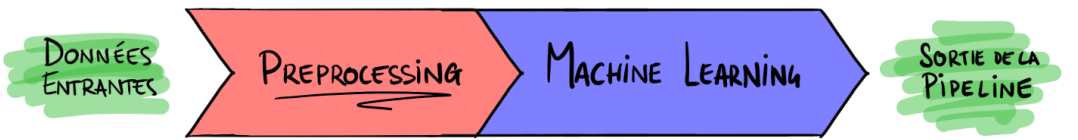
The Mission → L'objectif est d'obtenir un **modèle F**, qui lorsqu'on lui présente des **données inédites** provenant de l'environnement de **production**, fassent des **prédictions** qui se révèlent **correctes** suffisamment souvent pour être utilisable en production

On apelle ça la Généralisation ↑ à connaître

BREAK INTO.AI

Les grandes étapes du PREPROCESSING

Preprocessing : Les étapes préalables au traitement (=processing) par le modèle de Machine Learning



- Data Cleaning
- Data Transform
- Feature Engineering
- Dimensionality Reduction

Toutes ces étapes ne sont pas "obligatoires" ...

Il est même rare de toutes les utiliser en même temps...

EN FAIT... Ce qu'il faut RETENIR C'EST QUE...

LE PREPROCESSING A 2 "GRANDS" OBJECTIFS :

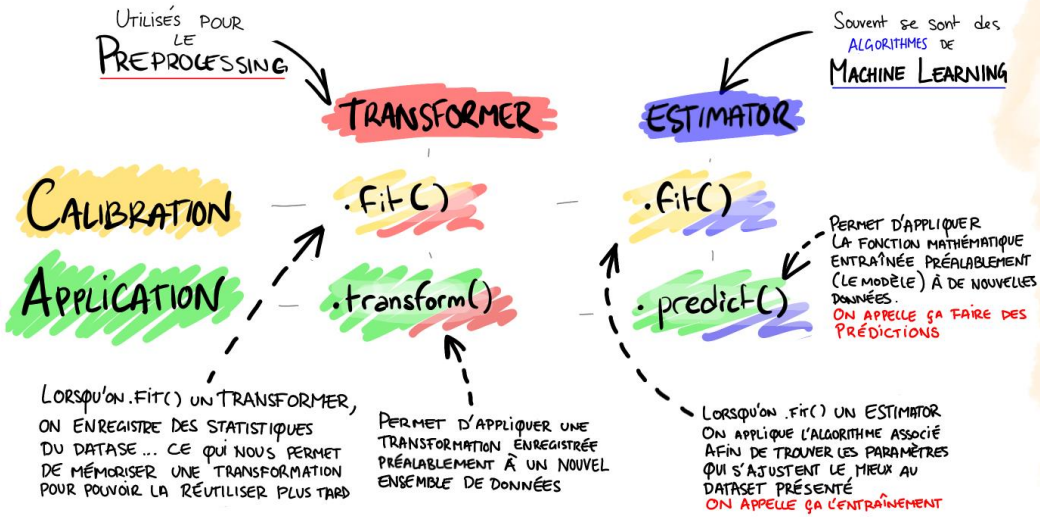
* Rendre les données entrantes compatibles avec nos modèles de Machine Learning
* Transformer les données pour améliorer les performances prédictives de nos Modèles

(et seulement cette étape là est Obligatoire)

BREAK INTO.AI

The Scikit-Learn A.P.I.

TRANSFORMER vs. ESTIMATOR



BREAK INTO.AI

COIN VOCABULAIRE

LA DIFFÉRENCE ENTRE ALGORITHME ET MODÈLES DE MACHINE LEARNING

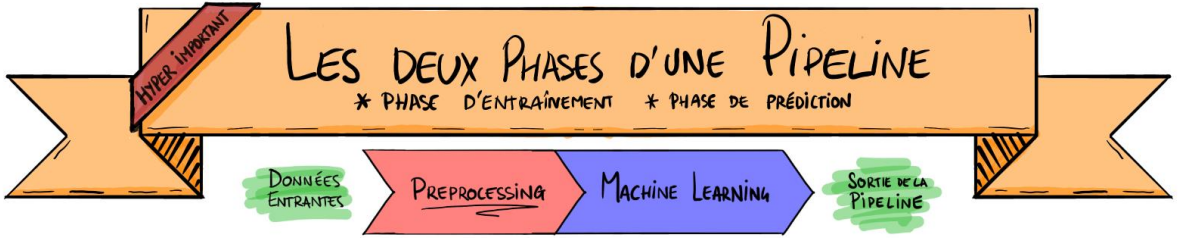
UN ALGORITHME DE M.L. EST UNE RECETTE DE CUISINE (procédure) POUR CRÉER UN MODÈLE DE ML.

UN MODÈLE DE M.L. EST UNE FONCTION MATHÉMATIQUE POUR FAIRE DES PRÉDICTIONS

LORSQU'ON CHOISIT UN ALGO ON CHOISIT EN PLUS DES "RÉGLAGES"... Ces "RÉGLAGES" choisis avant l'entraînement sont appelés des HYPER-PARAMÈTRES

LORSQU'UN ALGO ENTRAÎNE UN MODÈLE, IL CHERCHE L'ENSEMBLE DE PARAMÈTRES DU MODÈLE QUI EST LE MIEUX AJUSTÉ AU DATASET

Les HYPER-PARAMÈTRES et les PARAMÈTRES ONT UNE INFLUENCE SUR LES PRÉDICTIONS DU MODÈLE



Phase de Training

DATA → Étapes de Preprocessing (↑ HYPERPARAMÈTRES) → ALGORITHME (↑ HYPERPARAMÈTRES) → RÉSULTAT DE LA PHASE DE TRAINING =

- Étapes de Preprocessing Ajustées Préalablement
- Modèle Entraîné

NOTRE MISSION ÉTAIT DE CRÉER UN MODÈLE CAPABLE DE PRÉDIRE AUTOMATIQUEMENT. CE QUE NOUS AVONS ACCOMPLI

"TOUT ÇA POUR ÇA?"

Phase de Prédictions

LES RÉSULTATS DE LA PHASE D'ENTRAÎNEMENT SONT UTILISÉS DANS LA PHASE DE PRÉDICTION

DATA → Étapes de Preprocessing Ajustées Préalablement → Modèle Entraîné → RÉSULTAT DE LA PHASE DE PREDICTION = UNE PRÉDICTION

BREAK INTO.AI

1 TRAINING SUR TOUT LE DATASET

PROBLÈME : Aucun moyen de vérifier les résultats

STRATIFIÉ

2 SPLIT TRAIN / TEST

PROBLÈME : GROS FACTEUR CHANCE..
(changer la seed peut faire changer la mesure de performance)

3 K-FOLD CROSS VALIDATION

✓ PERMET D'OBTENIR UN SCORE FIABLE POUR UNE PIPELINE

⚠ Après avoir valider une pipeline, il faut la ré-entraîner sur la totalité du DATASET

PROBLÈME : PEUT SE RÉVÉLER TRÈS COÛTEUX EN TEMPS / PUISSANCE DE CALCUL

BREAK INTO AI

NOTRE MISSION

RECHERCHE DE LA MEILLEURE PIPELINE

1) Créez un schéma de crossvalidation

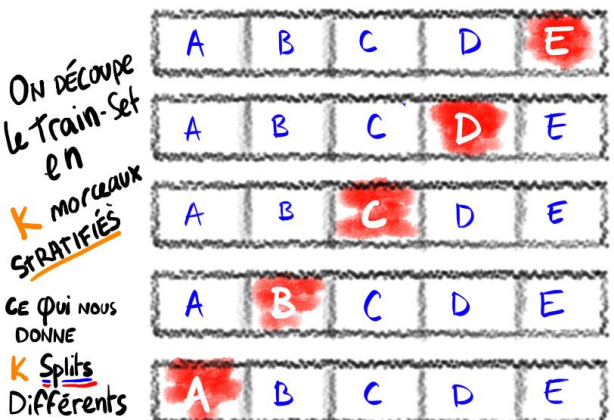
2) TESTEZ UNE GRANDE VARIÉTÉ DE PIPELINE
(↳ ≠ ALGOS, ≠ PREPROCESSING, ≠ HYPER-PARAMÈTRES)

3) SÉLECTIONNER LA PIPELINE avec le meilleur score de C.V.

4) RÉ-ENTRAÎNEZ LA PIPELINE SUR LA TOTALITÉ DES DONNÉES DISPONIBLES POUR L'ENTRAÎNEMENT

Le petit schéma de la K-Fold Cross Validation

ici K=5



ON ENTRAÎNE LA PIPELINE K-Fois (1 fois par Split)

Train Sets	Valid Sets
A + B + C + D	E
A + B + C + E	D
A + B + D + E	C
A + C + D + E	B
B + C + D + E	A

Score de Cross-Validation

ce qui nous donne

ON FAIT LA **MOYENNE** DES SCORES SUR CHAQUE VALIDATION SET

BREAK INTO.AI