

Démarche de travail en Machine Learning

Exploratory Data Analysis : comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation.

Checklist (non-exhaustive)

- Analyse de forme :
 - ✓ Identification de la target
 - ✓ Nombre de lignes et de colonnes
 - ✓ Types de variables
 - ✓ Identification des valeurs manquantes
- Analyse du fond :
 - ✓ Visualisation de la target (Histogramme/Boxplot)
 - ✓ Compréhension des différentes variables (Internet)
 - ✓ Visualisation des relations features-target (Histogramme/Boxplot)
 - ✓ Identification des outliers

Preprocessing : transformer le data pour le mettre dans un format propice au ML.

Checklist (non-exhaustive)

- ✓ Création du Train / Test sets
- ✓ Elimination des NaN : dropna(), imputation, colonnes « vides »
- ✓ Encodage (var. catégorielles => dummies)
- ✓ Suppression des outliers néfastes au modèle (à faire plutôt après avoir créé un 1^{er} modèle de Machine Learning)
- ✓ Feature Selection (dans un 1^{er} temps, éliminer les var. avec une variance nulle et les variables redondantes)
- ✓ Feature Engineering
- ✓ Feature Scaling (MinMaxScaler/StandardScaler/RobustScaler)

Modeling : développer un modèle de Machine Learning qui réponde à l'objectif final.

Checklist (non-exhaustive)

- ✓ Définir une fonction d'évaluation (que cherche t-on à mesurer ?)
- ✓ Entraînement de différents modèles
- ✓ Optimisation avec GridSearchCV
- ✓ (Optionnel) Analyse des erreurs et retour au Preprocessing/EDA
- ✓ Learning Curve (savoir si récolter plus de données permettrait d'améliorer notre modèle)
- ✓ Prise de décision