Chapitre 2

Chaîne du processus Data Science

Plan du chapitre

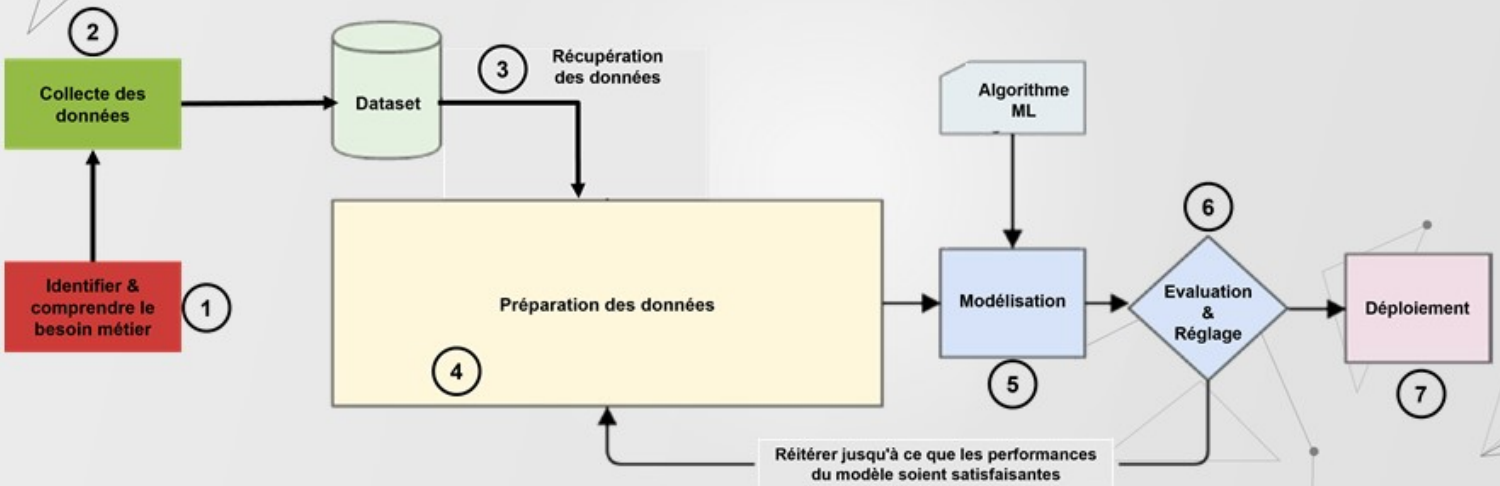
1. Introduction
2. Identifier & comprendre le besoin métier
3. Collecte des données
4. Préparation des données
5. Modélisation
6. Evaluation & Réglage
7. Déploiement

Introduction

- La Data Science joue un rôle central dans la prise de décisions
- Au cœur de cette discipline, la "Chaîne du processus Data Science" représente un ensemble structuré d'étapes cruciales permettant de passer de la compréhension du besoin métier à la mise en œuvre efficace de modèles prédictifs
- Le déroulement de la "Chaîne du processus Data Science" est plutôt itératif que linéaire. Il s'appuie sur un ensemble de phases répétées plusieurs fois au besoin
- Chaque étape de cette chaîne contribue à donner une compréhension approfondie et à extraire des connaissances exploitables

Chaîne du processus Data Science

Afin de mener à bien un projet Data Science, il est très important de suivre toutes les étapes du cycle de vie afin d'assurer le bon fonctionnement du projet :

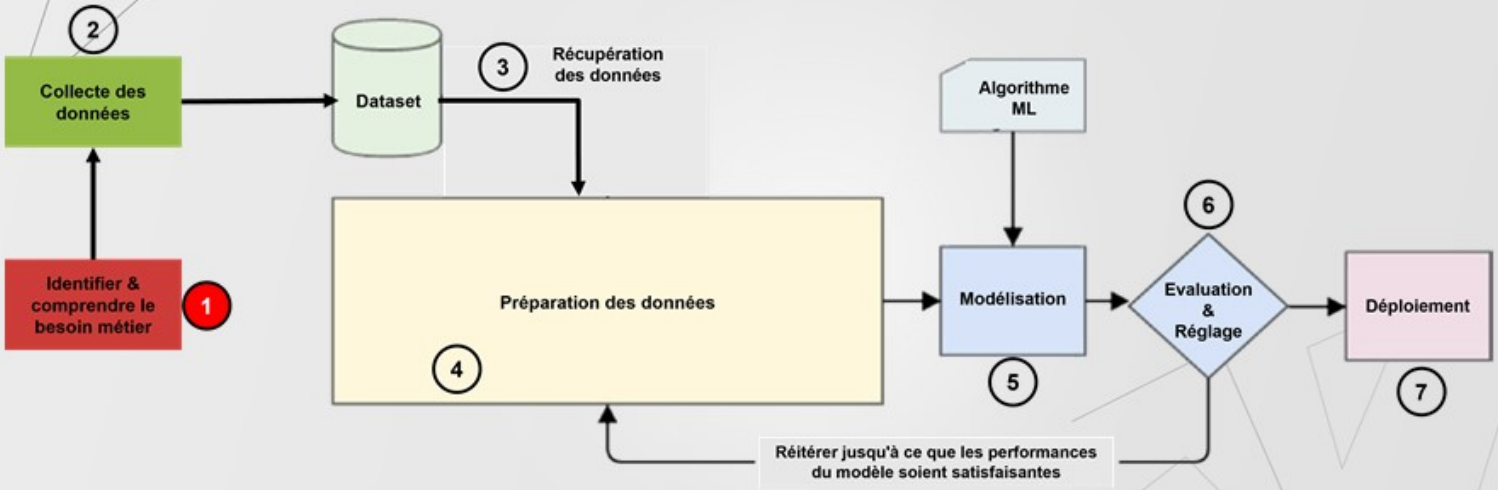


01

Identifier & comprendre le besoin métier

Chaîne du processus Data Science

Etape 1:



Chaîne du processus Data Science

Etape 1: Identifier & comprendre le besoin métier

Cette étape consiste à comprendre le besoin métier, les différentes spécifications, exigences et priorités.

1. **Compréhension du Problème :**
   - Il s'agit de définir clairement et précisément le problème que la data science vise à résoudre. Cela implique une collaboration étroite avec les parties prenantes métier.
   - Identifier les objectifs métier et les résultats attendus.

     Qu'est-ce qu'on cherche à accomplir ou à améliorer grâce à l'analyse des données ?

Chaîne du processus Data Science

Etape 1: Identifier & comprendre le besoin métier

2. **Analyse des Spécifications et des Exigences :**
   - Examiner les spécifications détaillées et les exigences du projet. Quelles sont les contraintes, les paramètres et les critères de succès ?
   - Comprendre les attentes en termes de performances, de précision, de délais, etc.
3. **Priorités et Contraintes :**
   - Déterminer les priorités des différentes tâches et exigences. Quelles sont les parties du projet qui ont une importance critique ?
   - Identifier les contraintes potentielles liées aux ressources, au budget, ou à d'autres facteurs.

Chaîne du processus Data Science

Etape 1: Identifier & comprendre le besoin métier

4. **Établissement d'une Communication Clé :**
   - Établir une communication claire et ouverte avec les parties prenantes pour garantir une compréhension mutuelle des besoins et des attentes.
   - Clarifier toute ambiguïté et s'assurer que l'équipe de data science a une vision cohérente avec les objectifs métier.
5. **Documentation :**
   - Documenter soigneusement toutes les informations recueillies, les hypothèses faites, et les décisions prises. Cela servira de référence tout au long du projet.

Chaîne du processus Data Science

Etape 1: Identifier & comprendre le besoin métier

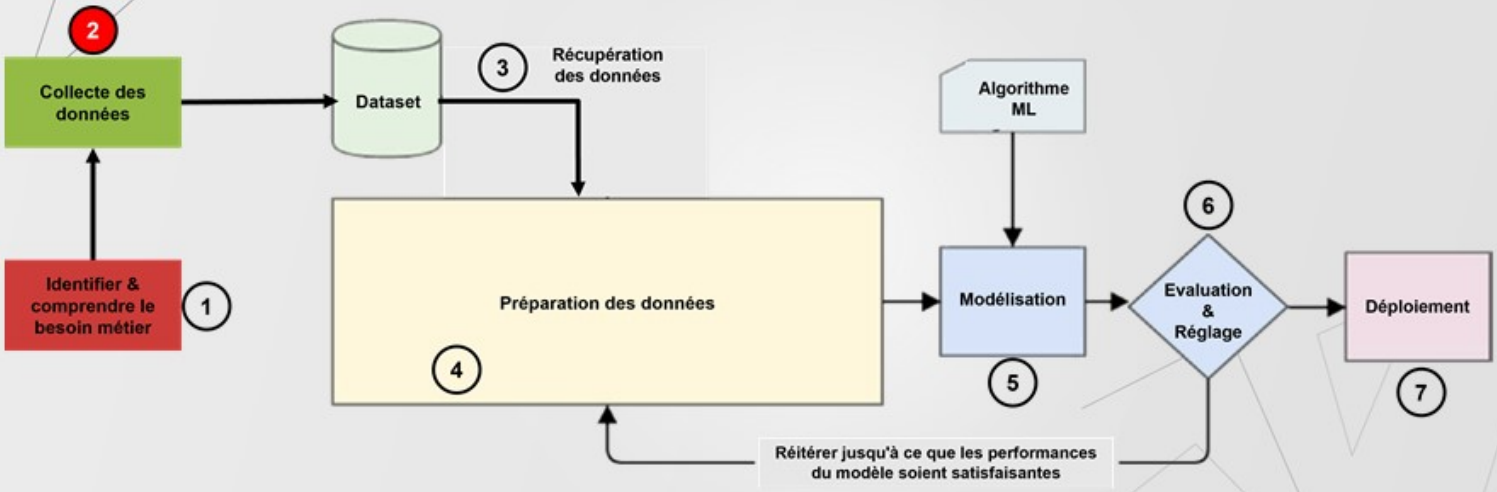
6. **Validation Continue :**
   - Il est important de valider régulièrement la compréhension du problème avec les parties prenantes. Les besoins métier peuvent évoluer, et une communication continue est cruciale.

02

Collecte des données

Chaîne du processus Data Science

Etape 2 : Collecte des données



Chaîne du processus Data Science

Etape 2 : Collecte des données

L'étape de collecte des données dans le domaine de la Data Science implique l'acquisition de données provenant :

- de différentes sources :
  - entrepôts de données
  - Web
  - capteurs,
  - …
- de différents types :
  - Numériques
  - Textuelles
  - Images/Graphes
  - Séquences
  - …

Chaîne du processus Data Science

Etape 2 : Collecte des données

Voici une approche plus détaillée :

1. **Identification des Sources de Données :**
   - Recenser toutes les sources potentielles de données, y compris les entrepôts de données internes, les sources externes, les données web, les capteurs, etc.
2. **Extraction des Données :**
   - Mettre en place des procédures pour extraire les données de ces sources. Cela peut impliquer l'utilisation d'outils ETL (Extract, Transform, Load) pour intégrer les données depuis différentes sources.

   a. **Web Scraping :**
      - Si les données proviennent du web, le web scraping peut être utilisé pour extraire des informations à partir de pages web. Cela peut être réalisé en utilisant des bibliothèques et des outils adaptés.

Chaîne du processus Data Science

Etape 2 : Collecte des données

b. Utilisation des API :

- Les API permettent aux développeurs d'accéder aux fonctionnalités et aux données d'une application ou d'un service e manière structurée
- Cette approche ouvre la porte à une diversité de sources de données ce qui élargit considérablement le champ des possibilités pour les professionnels de la Data Science
- Que ce soit pour comprendre les tendances du marché, ou accéder à des données météorologiques en temps réel, ou des plateformes en ligne aux bases de données gouvernementales les API jouent un rôle crucial dans l'acquisition de données pertinentes et actualisées
- De nombreuses plateformes en ligne (comme Twitter, Google ou Facebook) fournissent des API pour permettre aux développeurs de suivre les interactions sur les médias sociaux

Chaîne du processus Data Science

Etape 2 : Collecte des données

c. Enquêtes :

- Les enquêtes représentent une méthode essentielle de collecte d'informations directes auprès d'un groupe cible.
- L'utilisation d'outils en ligne tels que Google Forms ou SurveyMonkey a considérablement simplifié le processus de création, de distribution et de collecte de données

d. Collecte de Données à partir de Capteurs :

- Si les données proviennent de capteurs, il est nécessaire de mettre en place des mécanismes pour collecter ces données en temps réel ou périodiquement.

Chaîne du processus Data Science

Etape 2 : Collecte des données

e. Types de Données :

- Gérer différents types de données, y compris les données numériques, textuelles, images, graphiques, séquences, etc. Chaque type de données peut nécessiter des techniques de collecte spécifiques.

3. Stockage des Données :

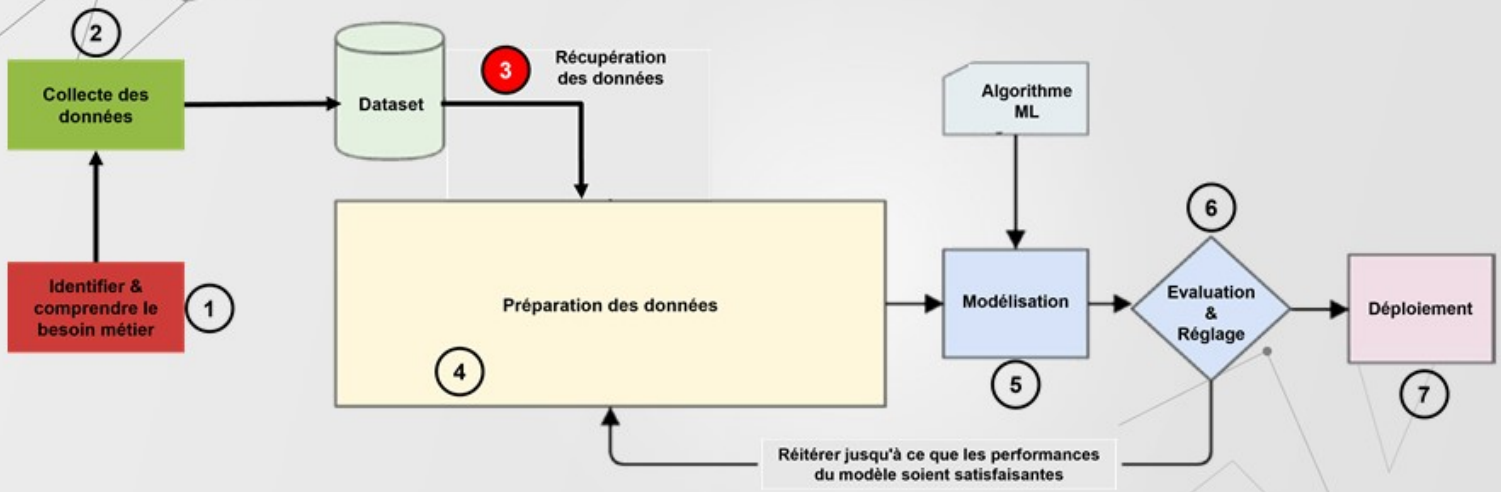
- Après l'extraction, les données doivent être stockées dans un fôrmat adapté. Les entrepôts de données, les bases de données relationnelles, ou d'autres systèmes peuvent être utilisés en fonction des besoins.

03

Récupération des données

Chaîne du processus Data Science

Etape 3 : Récupération des données



Chaîne du processus Data Science

Etape 3 : Récupération des données

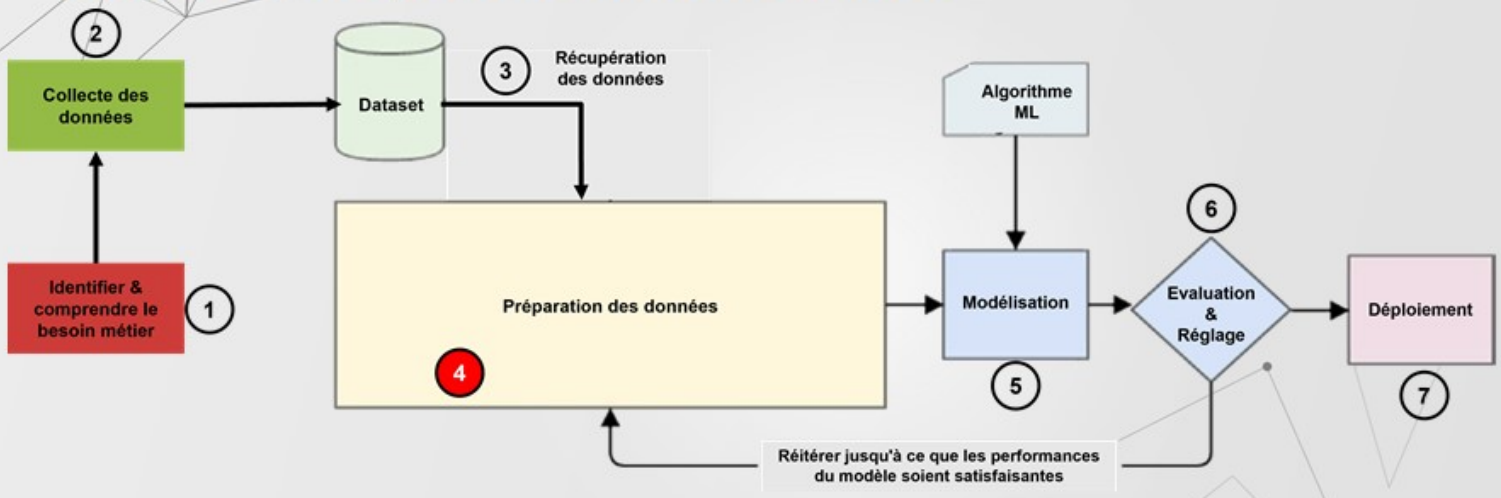
- La Récupération des données représente une étape importante dans tout projet de Data Science
- L'acquisition de données est simple. Cela peut impliquer d'importer directement des données à partir de fichiers CSV, de feuilles de calcul Excel ou de bases de données SQL, …
- La meilleure technique de collecte de données dépend des besoins spécifiques, de la nature du projet et du type de données dont on a besoin

04

Préparation des données

Chaîne du processus Data Science

Etape 4 : Préparation des données



Chaîne du processus Data Science

Etape 4 : Préparation des données

a. Exploration

- Faire un inventaire des données
  - Typologie: Numériques , Temporelles, Textes, Binaires, ...
  - Variables catégorielles, discrètes ou continues
  - Nombre d'observations (nombre de lignes)
  - Nombre de caractéristiques/features/variables (nombre de colonnes)
- Détecter les anomalies:
  - Outliers (valeurs aberrantes)
  - Valeurs manquantes
  - Corrélations

Chaîne du processus Data Science

Etape 4 : Préparation des données

b. Nettoyage

Préparer les features/variables afin qu'elles soient utilisables par des algorithmes du ML:

- Remplacer ou supprimer les valeurs manquantes/ aberrantes
- Transformer des données (variables) au format numérique

c. Transformation

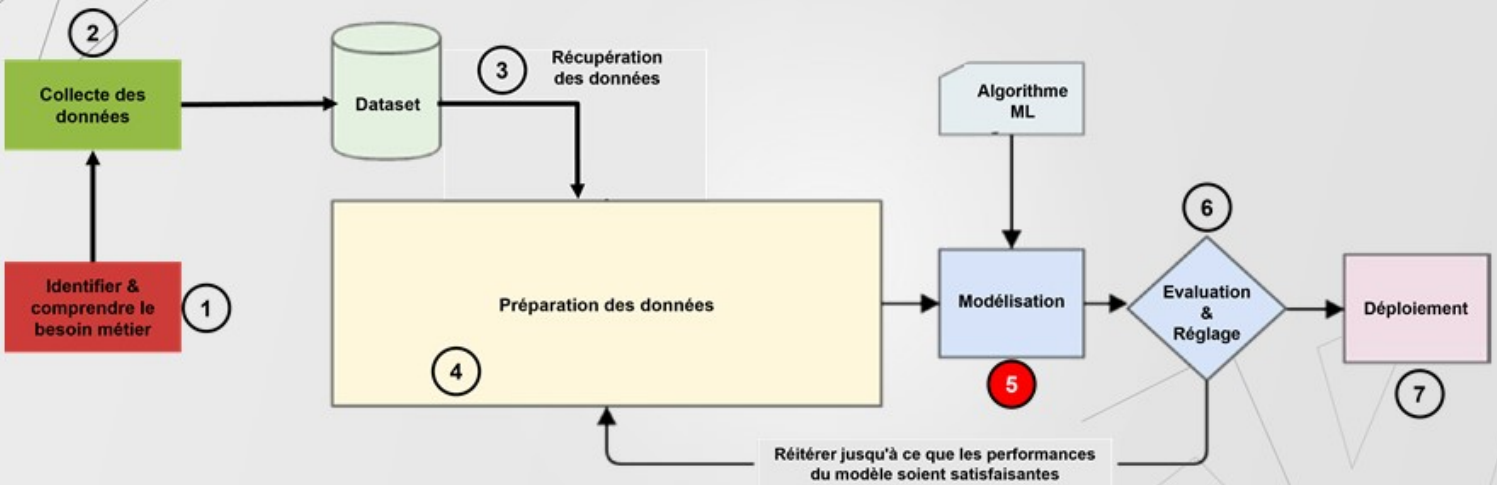
Modifier les données brutes pour les rendre plus adaptées à l'analyse. Cela permet d'améliorer la qualité des données, de normaliser les valeurs, et de créer de nouvelles caractéristiques

05

Modélisation

Chaîne du processus Data Science

Etape 5 : Modélisation



Chaîne du processus Data Science

Etape 5 : Modélisation

La modélisation consiste à utiliser les algorithmes pour extraire des schémas, identifier des tendances et faire des prédictions à partir des données collectées et préparées.

Les étapes de la modélisation impliquent souvent la sélection d'algorithmes et la division des données en:

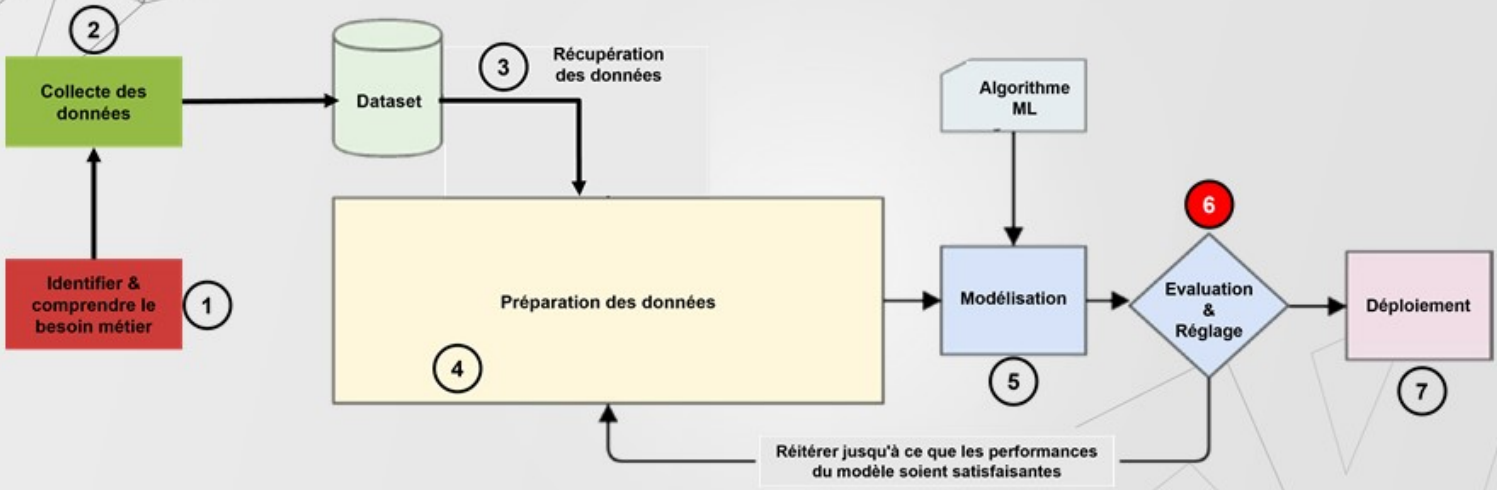
- **Données d'entrainement :** sous-ensemble destiné à l'apprentissage d'un modèle.
- **Données de test :** sous-ensemble destiné à l'évaluation du modèle. Ce jeu de données ne doit en aucun cas être utilisé dans les données d'entrainement.

06

Evaluation & Réglage

Chaîne du processus Data Science

Etape 6 : Evaluation & Réglage



Chaîne du processus Data Science

Etape 6 : Evaluation & Réglage

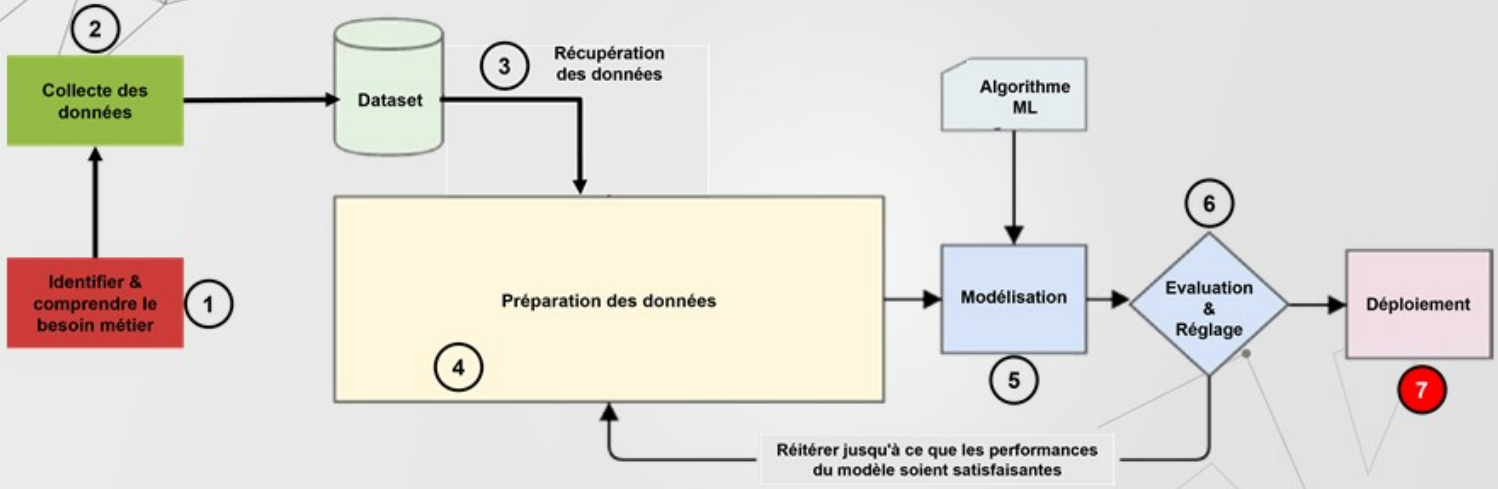
Ce processus itératif vise à perfectionner le modèle, à le rendre robuste face à de nouvelles données, et à garantir sa capacité à généraliser des tendances plutôt que de simplement mémoriser les exemples existants. Cela comprend:

- **La validation du modèle :** Observer les performances du modèle sur de nouvelles données
- **La mesure de performances :** Elle s'effectue à travers un ensemble de métriques et selon le type du problème.
- **Optimisation des hyperparamètres :** Afin d'obtenir des modèles avec les meilleurs résultats.

07
Déploiement

Chaîne du processus Data Science

Etape 7 : Déploiement



Chaîne du processus Data Science

Etape 7 : Déploiement

- Intégrer la solution: mettre en place une interface d' exécution
- Prendre en considération l'évolution des données sur les quelles est basé l'apprentissage
- Surveiller en fonction des prédictions et des résultats
- Créer des stratégies business