

Oui, tu peux tout à fait utiliser Dataiku en même temps que les outils et projets que tu explores pour développer ton profil hybride et approfondir tes compétences en machine learning ! Dataiku est une plateforme collaborative qui s'intègre bien avec des outils comme Excel, Python, et des projets de machine learning, ce qui en fait un complément intéressant à ton plan d'ensemble. Voici une analyse et des ajustements pour intégrer Dataiku dans ton approche, en tenant compte de ton niveau intermédiaire et de tes objectifs sur les deux prochaines semaines (29 mai - 12 juin 2025).

Compatibilité de Dataiku avec ton plan

Dataiku est une plateforme qui permet de combiner :

- **Excel** : Importation et traitement initial des données via des fichiers Excel.
- **Python** : Utilisation de scripts Python pour des analyses avancées, des modèles de machine learning, et des recettes personnalisées.
- **Machine Learning** : Automatisation de la modélisation (AutoML), création de pipelines, et déploiement de modèles.

Cela correspond parfaitement à ton profil hybride et à ton désir d'approfondir le machine learning. Voici comment l'intégrer :

Avantages de Dataiku

- **Interface visuelle** : Tu peux utiliser des recettes no-code/low-code pour préparer les données, ce qui te fait gagner du temps par rapport à un traitement manuel dans Excel ou Python.
- **Intégration Python** : Tu peux écrire des scripts Python dans des notebooks ou recettes, exploitant scikit-learn, pandas, etc., pour tes projets de machine learning.
- **Gestion des datasets** : Importe facilement des datasets (ex. Mall Customer Segmentation, Boston Housing) et les visualise dans un flux (Flow).
- **Déploiement** : Permet de déployer des modèles (ex. API, tableaux de bord) sans configuration complexe, ce qui est un plus pour ton portfolio.
- **Collaboration** : Idéal pour simuler un environnement d'entreprise, utile pour une alternance.

Limites à considérer

- **Courbe d'apprentissage** : Même avec une interface intuitive, il faudra quelques heures pour te familiariser (installation, configuration d'un environnement Python).
- **Version communautaire** : La version gratuite (Community Edition) a des restrictions (ex. pas de support complet pour tous les exports ou déploiements avancés).
- **Dépendance** : Si tu te relies trop sur Dataiku, tu risques de moins pratiquer le code Python pur, ce qui pourrait limiter ton apprentissage technique.

Ajustement du plan d'ensemble avec Dataiku

Voici une version révisée de ton plan sur deux semaines, en intégrant Dataiku pour les projets sélectionnés. Les tâches sont adaptées pour tirer parti des fonctionnalités de Dataiku tout en maintenant une pratique active de Python et Excel.

Semaine 1 : Bases du machine learning et exploration hybride (29 mai - 4 juin 2025)

- **Jour 1 : Préparation et installation de Dataiku (29 mai, jeudi)**
 - **Heure** : 11:30 AM - 3:30 PM (4h).
 - **Tâches** :
 - Télécharge et installe Dataiku Community Edition.
 - Configure un environnement Python dans Dataiku (inclus pandas, scikit-learn, seaborn, matplotlib).
 - Importe les datasets (Mall Customer Segmentation, Boston Housing, Global Superstore) dans Dataiku via le Flow.
 - Explore l'interface (Flow, Lab, Visualisation).
 - **Délivrab**le : Environnement Dataiku prêt, datasets importés.

- **Jour 2 : Clustering - Segmentation client avec Dataiku (30 mai, vendredi)**

- **Heure** : 10:00 AM - 2:00 PM (4h).
- **Tâche** : Projet 9 (Doc 1) - Segmenter les clients.
 - **Étape 1 (1h)** : Nettoie les données dans Dataiku avec une recette "Prepare" (suppression des doublons, gestion des valeurs manquantes).
 - **Étape 2 (2h)** : Crée un modèle de clustering dans le Lab :
 - Utilise AutoML pour tester K-means.
 - Ajuste manuellement avec un script Python (K-means via scikit-learn).
 - Visualise les clusters dans Dataiku (scatter plot).
 - **Étape 3 (1h)** : Exporte les résultats et crée un graphique avec seaborn.
- **Délivrable** : Script Python, graphique des clusters, sauvegarde sur GitHub.

- **Jour 3 : Régression - Prédiction des prix avec Dataiku (31 mai, samedi)**

- **Heure** : 10:00 AM - 2:00 PM (4h).
- **Tâche** : Projet 10 (Doc 1) - Prédire les prix des maisons.
 - **Étape 1 (1h)** : Analyse exploratoire dans Dataiku (statistiques, corrélations via Visualisation).
 - **Étape 2 (2h)** : Modélisation :
 - Utilise AutoML pour une régression linéaire initiale.
 - Écris un script Python pour une régression multiple (scikit-learn), évalue avec R² et RMSE.
 - **Étape 3 (1h)** : Visualise (prédictions vs. réelles avec matplotlib).
- **Délivrable** : Script Python, graphique des prédictions.

- **Jour 4 : Tableau de bord - Visualisation avec Dataiku (1er juin, dimanche)**
 - **Heure** : 10:00 AM - 2:00 PM (4h).
 - **Tâche** : Projet 7 (Doc 1) - Tableau de bord.
 - **Étape 1 (1h)** : Prépare les données dans Dataiku (recette "Prepare").
 - **Étape 2 (2h)** : Crée un tableau de bord dans Dataiku Dashboards (KPI, slicers).
 - **Étape 3 (1h)** : Exporte une capture d'écran.
 - **Délivrable** : Tableau de bord, capture d'écran.
- **Jour 5 : Révision et consolidation (2 juin, lundi)**
 - **Heure** : 10:00 AM - 1:00 PM (3h).
 - **Tâches** : Optimise les scripts, ajoute des commentaires, met à jour GitHub et CV.
 - **Délivrable** : Code optimisé, CV mis à jour.

Semaine 2 : Approfondissement du machine learning et projet avancé (5 juin - 12 juin 2025)

- **Jour 6 : Classification - Détection de défauts avec Dataiku (5 juin, jeudi)**
 - **Heure** : 10:00 AM - 2:00 PM (4h).
 - **Tâche** : Projet 11 (Doc 1) - Prédire les défauts.
 - **Étape 1 (1h)** : Analyse dans Dataiku (Visualisation).
 - **Étape 2 (2h)** : Modélisation :
 - Teste AutoML pour régression logistique.
 - Ajoute un script Python avec SMOTE et évaluation (AUC).
 - **Étape 3 (1h)** : Visualise (courbe ROC).
 - **Délivrable** : Script Python, courbe ROC.

- **Jour 7 : Systèmes de recommandation - Début avec Dataiku (6 juin, vendredi)**
 - **Heure** : 10:00 AM - 2:00 PM (4h).
 - **Tâche** : Projet 10 (Doc 2) - Systèmes de recommandation.
 - **Étape 1 (1h)** : Analyse dans Dataiku.
 - **Étape 2 (2h)** : Modélisation avec Python (Surprise pour SVD).
 - **Étape 3 (1h)** : Débute une interface Streamlit.
 - **Délivrable** : Script Python, interface partielle.
- **Jour 8 : Systèmes de recommandation - Finalisation (7 juin, samedi)**
 - **Heure** : 10:00 AM - 2:00 PM (4h).
 - **Tâche** : Finalise le projet.
 - **Étape 1 (1h)** : Ajoute filtre basé sur le contenu.
 - **Étape 2 (2h)** : Finalise Streamlit.
 - **Étape 3 (1h)** : Teste localement.
 - **Délivrable** : Système fonctionnel, capture d'écran.
- **Jour 9 : Révision et déploiement (8 juin, dimanche)**
 - **Heure** : 10:00 AM - 1:00 PM (3h).
 - **Tâches** : Optimise les modèles, ajoute une API Flask via Dataiku, met à jour le portfolio.
 - **Délivrable** : API (optionnel), portfolio.
- **Jour 10-11 : Projet avancé - Séries temporelles (9-10 juin, lundi-mardi)**
 - **Heure** : 10:00 AM - 2:00 PM (8h).
 - **Tâche** : Projet 12 (Doc 1) ou 6 (Doc 2).
 - **Étape 1 (2h)** : Analyse dans Dataiku.
 - **Étape 2 (4h)** : Modélise avec Prophet (script Python).
 - **Étape 3 (2h)** : Visualise.
 - **Délivrable** : Script, graphique.

- **Jour 12-13 : Finalisation (11-12 juin, mercredi-jeudi)**
 - **Heure** : 10:00 AM - 1:00 PM (6h).
 - **Tâches** : Compile les résultats, finalise CV, prépare une présentation.
 - **Délivrab**le : CV, portfolio, présentation.

Conseils pour utiliser Dataiku

- **Installation** : Commence aujourd’hui avec la Community Edition. Suis le guide d’installation.
- **Environnement Python** : Crée un code env avec les packages nécessaires (ex. via le panneau Administration > Settings > Code Envs).
- **Flux (Flow)** : Utilise le Flow pour organiser tes datasets et recettes, mais garde un script Python pour maîtriser les détails.
- **Apprentissage** : Consacre 1-2h à un tutoriel Dataiku (ex. <https://knowledge.dataiku.com/latest/courses/getting-started.html>) avant de plonger dans les projets.

Impact sur ton CV

- Ajoute : "Utilisation de Dataiku pour la préparation de données, la modélisation ML, et la visualisation."
- Met en avant : "Pipelines ML avec Dataiku et Python (clustering, régression, recommandation)."

Conclusion

Dataiku est un excellent ajout à ton plan. Il accélérera certaines tâches (préparation, visualisation) tout en te permettant de pratiquer le machine learning avec Python. Commence par l’installation aujourd’hui, et ajuste les heures si besoin. Si tu veux un guide détaillé pour configurer Dataiku ou un projet spécifique (ex. clustering dans Dataiku), dis-le-moi ! Que penses-tu de ce plan ajusté ?