Dataverse4.0 Hackathon

Défi de Classification Virale

Note Technique

Organisé par DATAVERSE 4.0

Thème : *De la Donnée à l'IA – Déverrouiller le Futur*

Dates de l'événement : 8–9 décembre 2025

Lieu : ENSA Khouribga, Maroc

Ce document fournit tous les détails techniques, la structure du dataset, la livraison par étapes et la méthodologie d'évaluation pour le hackathon Dataverse4.0.

1 Présentation du Défi

Objectif. Construire un modèle d'apprentissage supervisé capable de prédire si un post sur les réseaux sociaux sera *viral* à partir des métadonnées disponibles (plateforme, métriques d'engagement, date/heure) et des informations textuelles (pseudo légende). Les participants doivent produire des pipelines robustes et reproductibles, inclure des visualisations, fournir un rapport de modèle dans le notebook, et expliquer clairement leurs choix de feature engineering et de modélisation.

2 Description du Dataset

Le dataset comprend environ 5,000 posts anonymisés provenant de plusieurs plateformes et régions. Chaque enregistrement correspond à un post unique et contient des informations sur l'engagement, le contexte et le texte.

Colonnes

- **Post_ID** – Identifiant unique pour chaque post.
- **Pseudo_Caption** – Légende textuelle associée au post (à utiliser pour NLP / embeddings).
- **Post_Date** – Date de publication (AAAA-MM-JJ).
- **Platform** – Nom de la plateforme (ex. TikTok, YouTube, Instagram).
- **Hashtag** – Hashtag principal lié au post.
- **Content_Type** – Type de contenu (ex. Vidéo, Reel, Short, Tweet).
- **Region** – Région géographique où le contenu est publié.
- **Views** – Nombre total de vues.
- **Likes** – Nombre total de likes.
- **Shares** – Nombre total de partages.
- **Comments** – Nombre total de commentaires.
- **Viral** – Label cible (1 = Viral, 0 = Non Viral). *(présent uniquement dans le fichier d'entraînement)*

- **Remarque sur les métriques d'engagement :** Toutes les métriques d'engagement (Vues, Likes, Partages, Commentaires) sont collectées dans les 24 heures suivant la publication du post. Les participants doivent les considérer comme des indicateurs précoces de viralité. L'utilisation de métriques après la fenêtre de 24 heures ou des totaux cumulés finaux n'est pas autorisée pour garantir l'équité de la prédiction.

Les participants recevront :

- un **fichier d'entraînement** contenant toutes les features et le label `Viral`,
- un **fichier de validation public** avec uniquement les features (pour l'inférence).

3 Phases du Hackathon (30 Heures)

Partie 1 – Analyse Exploratoire des Données (15 pts)

- Inspecter les valeurs manquantes et la qualité globale des données.
- Résumer les métriques d'engagement (vues, likes, partages, commentaires).
- Visualiser les relations clés (ex. vues vs viral, plateforme vs viral).

Évaluation : qualité des insights, clarté des visualisations et exactitude du prétraitement.

Partie 2 – Feature Engineering (20 pts)

- Encoder les variables catégorielles (`Platform`, `Region`, `Content_Type`).
- Extraire des features temporelles depuis `Post_Date` : jour de la semaine, mois, année.
- Garder les transformations reproductibles (**seeded, scriptées**).

Évaluation : pertinence, originalité et robustesse des features créées.

Partie 3 – Développement et Optimisation du Modèle (35 pts)

- Entraîner plusieurs modèles supervisés : Régression Logistique, Arbre de Décision, Random Forest, XGBoost, CatBoost, SVM, etc.
- Optimiser principalement le **macro F1** (le dataset n'est pas parfaitement équilibré).
- Inclure un **rapport de modèle dans le notebook : métriques pour tous les modèles, hyperparamètres et justification du choix du modèle final**.

Évaluation : performance prédictive, interprétabilité, stabilité et qualité du rapport.

Partie 4 – Intégration des Embeddings (10 pts)

- Générer des embeddings de phrases depuis `Pseudo_Caption` en utilisant Sentence Transformers.
- Optionnel : réduction de dimensionnalité (ex. PCA à 50–100) avant fusion avec les features tabulaires.
- Montrer l'amélioration de performance après ajout des features textuelles.

Évaluation : qualité de l'intégration et contribution démontrée.

Partie 5 – Analyse du Meilleur Jour pour Poster (10 pts)

- Utiliser les features de date pour estimer les jours optimaux de publication par plateforme et type de contenu.
- Présenter les résultats avec des graphiques ou tableaux clairs.

Évaluation : exactitude, clarté et applicabilité.

Partie 6 – Explicabilité du Modèle (10 pts)

- Utiliser des outils d'importance des features (ex. SHAP, LIME, permutation importance) pour expliquer l'influence des features sur les prédictions.
- Identifier les variables les plus impactantes sur la viralité selon plateformes et régions.
- Visualiser les facteurs clés avec des graphiques de synthèse ou des heatmaps.

Évaluation : profondeur de l'analyse, clarté des explications et interprétabilité.

4 Livraison par Étapes et Exigences

Phase 1 – Soumission du Notebook (Obligatoire, 8 heures avant la fin du hackathon)

- Soumettre un seul notebook contenant :
  - Inclure toutes les étapes dans votre notebook :
    - Exploration des données (EDA, valeurs manquantes, statistiques descriptives)
    - Pipeline de création de features (feature engineering)
    - Entraînement et évaluation du modèle avec plusieurs modèles
    - Rapport du modèle uniquement dans les markdowns du notebook : métriques, hyperparamètres et justification du choix du modèle final
  - Générer un fichier CSV de prédictions avec votre modèle sélectionné contenant les colonnes : **Post_ID, Viral**

- **La soumission du notebook est obligatoire. Les équipes ne soumettant pas ne pourront pas poursuivre.**

5 Règles de Soumission et Timeline

- Tous les notebooks (.ipynb) doivent être soumis avec les cellules de sortie exécutées.
- Les participants peuvent **commencer à travailler sur les éléments de la Phase 2** (UI, structure de l'app, modèles améliorés) avant l'annonce des résultats de la Phase 1.
- L'évaluation de la Phase 1 est rapide. Les résultats seront annoncés dans les **30–40 minutes** suivant la date limite de soumission.

Seules les 5 meilleures soumissions de notebooks seront qualifiées pour la Phase 2.

Phase 2 – Application Web + Présentation

- Construire une application web interactive (stack au choix)
- Dashboard affichant les insights du dataset et interface de prédiction
- **Fonctionnalités bonus optionnelles :** upload admin, réentraînement, versionnage du modèle, explicabilité SHAP/LIME
- Présentation pptx

- Présentation de 10 minutes par équipe couvrant :
  - Insights issus de l'EDA et visualisations
  - Justification du choix du modèle (markdown + discussion)
  - Démo de l'application
  - Bonus optionnel

6 Résumé du Scoring

Catégorie	Description	Points
Partie 1	Analyse Exploratoire des Données	15
Partie 2	Feature Engineering	20
Partie 3	Développement et Optimisation du Modèle	35
Partie 4	Intégration des Embeddings	10
Partie 5	Meilleur Jour pour Poster	10
Partie 6	Explicabilité du Modèle	10
Total		**100**

Remarque : La phase 2 (Application Web et Présentation) comptera pour **40% du score total du hackathon**.

7 Traitement des Données et Reproductibilité

- Ajuster les composants de prétraitement (ex. `Imputer`, `Encoder`, `PCA` ou générateur d'embeddings) **uniquement sur l'ensemble d'entraînement**, puis appliquer (`transform`) sur les données de validation ou test ; ne jamais ré-ajuster sur les données de validation.
- Maintenir un ordre, un nom et un type de features cohérents entre entraînement et validation.
- Utiliser des seeds fixes pour toutes les opérations aléatoires (splits, échantillonnage, PCA, initialisation de modèle).
- Séparer clairement les sections "préparation des données" et "entraînement du modèle" dans le notebook pour la transparence.

Rappelez-vous : le code d'aujourd'hui peut être désordonné, mais demain, il pourrait devenir viral !