

Apprentissage supervisé : Classification

Mini-projets



Cours du Pr. Abdellah AZMANI
Département Génie Informatique
Cours pour les étudiants des Masters
et des Cycles d'Ingénieurs



Université Abdelmalek Essaadi
Faculté des Sciences et Techniques
Tanger



Machine Learning

Apprentissage supervisé : Classification

Consignes générales

➤ Objectif du mini-projet

Mettre en œuvre un processus complet de classification supervisée, depuis des données imparfaites jusqu'à l'évaluation de plusieurs modèles, en appliquant cette démarche à un seul projet choisi par chaque groupe et en justifiant chaque étape de prétraitement, de modélisation et d'analyse des résultats

Mini-projets :

- **Mini-projet 1** : Classification de l'occupation du sol
 - Jeu de données : *geo_lulc_data.csv*
- **Mini-projet 2** : Classification des zones à risque d'inondation
 - Jeu de données : *geo_flood_risk_data.csv*
- **Mini-projet 3** : Classification de zones urbaines selon leur densité
 - Jeu de données : *geo_urban_density_data.csv*
- **Mini-projet 4** : Classification de l'aptitude des sols agricoles
 - Jeu de données : *geo_agri_aptitude_data.csv*
- **Mini-projet 5** : Classification des zones selon leur sensibilité environnementale
 - Jeu de données : *geo_sensibilite_env_data.csv*
- **Mini-projet 6** : Classification de la stabilité des terrains (risque de glissement)
 - Jeu de données : *geo_stabilite_terrain_data.csv*

Machine Learning

Apprentissage
supervisé :
Classification

Consignes
générales

➤ Travail demandé

1. Compréhension des données

- Comprendre le contexte métier
- Identifier les variables explicatives et la variable cible
- Analyser la structure du jeu de données

2. Analyse exploratoire et visualisation des données (EDA)

- Réaliser une **visualisation des données avant le prétraitement** afin de :
 - comprendre la distribution des variables, géographique
 - identifier les valeurs manquantes, aberrantes et les déséquilibres de classes
 - ...
- Réaliser une **visualisation des données après le prétraitement** afin de :
 - vérifier l'impact des traitements appliqués
 - comparer l'état des données avant/après nettoyage
 - valider la cohérence des transformations effectuées
 - ...

Machine Learning

Apprentissage
supervisé :
Classification

Consignes
générales

➤ Travail demandé

3. Prétraitement des données

- Harmonisation des noms de colonnes et des modalités
- Détection et traitement des valeurs manquantes
- Suppression des colonnes trop manquantes et des lignes quasi vides
- Détection et suppression des doublons
- Normalisation / standardisation des variables numériques
-

4. Modélisation

- Séparation des données en jeu d'entraînement et de test (Train/Test)
- Implémentation de plusieurs modèles de classification (ex. : Régression Logistique, KNN, Arbre de Décision, Random Forest, SVM, etc.)
- Utilisation initiale des modèles avec les hyperparamètres par défaut de scikit-learn (ex. : *KNN avec `KNeighborsClassifier()` sans modification*)

https://scikit-learn.org/stable/supervised_learning.html

- Optimisation des hyperparamètres des modèles utilisés afin d'améliorer les performances (tester plusieurs configurations d'un même modèle et sélectionner la meilleure)

Machine Learning

Apprentissage supervisé : Classification

Consignes générales

➤ Travail demandé

5. Évaluation des modèles

- Calcul des métriques de performance : Matrice de confusion, Accuracy, Precision, Recall et F1-score
- Comparaison des performances :
 - entre les différents modèles de classification
 - entre les modèles avant et après l'optimisation des hyperparamètres

6. Analyse et interprétation

- Identifier le meilleur modèle
- Justifier les choix effectués
- Analyser les erreurs de classification

⚠ Règles importantes

- × Pas d'imputation de la variable cible
- ✓ Les variables spatiales sont utilisées uniquement pour les cartes et les visualisations, et ne sont pas intégrées dans les modèles
- ✓ Même jeu de données et même split pour tous les modèles
- ✓ Justification obligatoire de chaque décision de prétraitement

Machine Learning

Apprentissage
supervisé :
Classification

Consignes
générales

➤ Livrables attendus

1. Notebook Jupyter :

- ❖ Le notebook doit être clair, structuré et commenté, et contenir obligatoirement :
 - Titre du projet
 - Noms des membres du groupe
 - Description du jeu de données
 - Analyse exploratoire et visualisation des données
 - visualisations avant le prétraitement
 - visualisations après le prétraitement
 - Détail du prétraitement des données
 - Implémentation des modèles de classification
 - Évaluation et comparaison des modèles
- ❖ Chaque cellule de code doit être accompagnée d'un commentaire ou d'une interprétation (***cellule Markdown ou commentaire dans le code***)
- ❖ Nom du fichier obligatoire : projetX_geo_info.ipynb (X= {1, 2, 3, 4, 5})

Le notebook doit s'exécuter sans erreur du début à la fin

Machine Learning

**Apprentissage
supervisé :
Classification**

**Consignes
générales**

➤ Livrables attendus

1. Notebook Jupyter :

Bonus :

- Menu interactif permettant de choisir :
 - le modèle de classification
 - les hyperparamètres du modèle
 -
- Interaction utilisateur via des entrées clavier (input)
- Visualisations supplémentaires (courbes, graphiques comparatifs) + Interprétation
- Test du modèle sur de nouvelles données
- Commentaires et interprétation détaillée des résultats
- Notebook bien structuré et lisible
- ...

Machine Learning

Apprentissage
supervisé :
Classification

Consignes
générales

➤ Livrables attendus

2. Rapport synthétique (PDF) (un seul rapport par groupe)

Le rapport doit être clair, structuré et analytique (max. 20 pages) et ne doit pas se limiter à un copier-coller du notebook Jupyter. Il se compose des sections suivantes :

- **Page de garde** (Titre du projet, Noms des membres du groupe, Filière / année, ...)
- Table de matière, Listes des figures et des tableaux, des abréviations, ...
- **Introduction** : Contexte du projet, domaine d'application, objectif, ...
- **Description du Data**: Présentation des variables, explication des variables, ...
- **Prétraitement des données** : Problèmes identifiés (valeurs manquantes, doublons, incohérences), solutions appliquées et justification des choix, Impact du prétraitement sur les données, ...
- **Modélisation** : Présentation et explication des modèles de classification utilisés, les hyperparamètres des modèles ...
- **Évaluation et résultats** : Métriques utilisées, tableau comparatif des performances, comparaison avant / après optimisation, ...
- **Discussion** : Analyse des résultats obtenus, Forces et limites des modèles, Difficultés rencontrées, ...
- **Conclusion et perspectives** : Synthèse du travail, Améliorations possibles

Mini-projet 1 : Classification de l'occupation du sol (Land Use / Land Cover)

- **Data** *geo_lulc_data.csv*
- **Objectif** : Classer des zones géographiques selon leur type d'occupation du sol
- **Variables explicatives**
 - **NDVI** : indice de végétation (couverture végétale)
 - **altitude_m** : altitude du terrain (mètres)
 - **pente_pct** : pente du terrain (%)
 - **dist_route_km** : distance aux routes (km)
 - **texture_sol** : indice de texture du sol
 - **chlorophyll_index_lab** : indice de chlorophylle issu d'analyses ponctuelles
- **Variables spatiales**
 - **latitude** : coordonnée géographique
 - **longitude** : coordonnée géographique
 - **zone_id** : identifiant de la zone spatiale
- **Variable cible**
 - **classe_lulc** : type d'occupation du sol (*Forêt, Urbain, Agricole, Eau*)

Mini-projet 2 : Classification des zones à risque d'inondation

- **Data** *geo_flood_risk_data.csv*
- **Objectif** : Identifier le niveau de risque d'inondation d'une zone.
- **Variables explicatives**
 - **altitude_m** : altitude du terrain
 - **pente_pct** : pente (%)
 - **distance_riviere_m** : distance au cours d'eau
 - **precip_mm** : cumul des précipitations (mm)
 - **perm_sol** : perméabilité du sol
 - **debit_station_amont** : débit mesuré à une station hydrométrique
- **Variables spatiales**
 - **latitude** : coordonnée géographique
 - **longitude** : coordonnée géographique
 - **zone_id** : identifiant de la zone spatiale
- **Variable cible**
 - **niveau_risque** : niveau de risque d'inondation (*Faible, Moyen, Élevé*)

Mini-projet 3 : Classification de zones urbaines selon leur densité

- **Data** *geo_urban_density_data.csv*
- **Objectif** : Classer les zones urbaines selon leur densité et usage
- **Variables explicatives**
 - **densite_batiments** : densité des bâtiments
 - **densite_population** : densité de population (par km²)
 - **taux_impermeabilisation** : taux de surfaces imperméabilisées (%)
 - **distance_centre_ville_km** : distance au centre urbain
 - **hauteur_moyenne_batiments** : hauteur moyenne des bâtiments
 - **wifi_density_probe** : densité estimée de signaux Wi-Fi dans la zone
- **Variables spatiales**
 - **latitude** : coordonnée géographique
 - **longitude** : coordonnée géographique
 - **zone_id** : identifiant de la zone spatiale
- **Variable cible**
 - **classe_urbaine** : classe de densité urbaine (*faible, moyenne, élevée*)

Mini-projet 4 : Classification de l'aptitude des sols agricoles

- **Data** *geo_agri_aptitude_data.csv*
- **Objectif** : Classifier les parcelles agricoles selon leur niveau d'aptitude à la production
- **Variables explicatives**
 - **pente_pct** : pente du terrain (%)
 - **humidite_sol** : niveau d'humidité du sol
 - **humidite_sol** : niveau d'humidité du sol
 - **precip_mm** : cumul des précipitations (mm)
 - **temp_moy_C** : température moyenne (°C)
 - **texture_sol** : texture du sol
 - **ph_sol_labo** : pH mesuré en laboratoire
- **Variables spatiales**
 - **latitude** : coordonnée géographique
 - **longitude** : coordonnée géographique
 - **parcelle_id** : identifiant de parcelle agricole
- **Variable cible**
 - **aptitude_sol** : aptitude du sol (Faible, Moyenne, Élevée)

Mini-projet 5 : Classification des zones selon leur sensibilité environnementale

- **Data** *geo_sensibilite_env_data.csv*
- **Objectif** : Classifier des zones géographiques selon leur niveau de sensibilité environnementale
- **Variables explicatives**
 - **ndice_vegetation** : couverture végétale
 - **distance_eau_m** : proximité des ressources en eau
 - **pente_pct** : pente du terrain (%)
 - **pression_humaine** : indice de pression anthropique
 - **qualite_sol** : indice global de qualité du sol
 - **indice_biodiversite_labo** : indice issu d'inventaires écologiques
- **Variables spatiales**
 - **latitude** : coordonnée géographique
 - **longitude** : coordonnée géographique
 - **zone_id** : identifiant de la zone spatiale
- **Variable cible**
 - **sensibilite_env** : sensibilité environnementale (*Faible, Moyenne, Élevée*)

Mini-projet 6 : Classification de la stabilité des terrains (risque de glissement)

- **Data** *geo_stabilite_terrain_data.csv*
- **Objectif** : Classifier les zones géographiques selon leur niveau de stabilité du terrain afin d'identifier les zones à risque de glissement
- **Variables explicatives**
 - **pente_pct** : pente du terrain (%)
 - **altitude_m** : altitude du terrain (mètres)
 - **texture_sol** : type / texture du sol
 - **humidite_sol** : niveau d'humidité du sol
 - **distance_faille_km** : distance aux failles géologiques (km)
 - **couverture_vegetale** : taux de couverture végétale
 - **indice_geotech_labore** : indice géotechnique issu de tests ponctuels en laboratoire
- **Variables spatiales**
 - **latitude** : coordonnée géographique
 - **longitude** : coordonnée géographique
 - **zone_id** : identifiant de la zone spatiale
- **Variable cible**
 - **stabilite_terrain** : niveau de stabilité du terrain (*stable, moyennement stable, instable*)