RAPPORT TECHNIQUE

Prétraitement de Données

Prédiction des Maladies Cardiovasculaires

Dataset : Heart Disease UCI

Source : Kaggle

Échantillons : 920 patients

Variables : 16 colonnes (13 features + 1 cible)

Réalisé par:
Wissal Akkaoui
Serigne Fallou Mbacke
Ahossan Marc Cedrick Tanoh

13 février 2026

Résumé du Prétraitement des Données

1. Exploration et Nettoyage

Dataset : 920 patients avec 16 variables (âge, sexe, pression artérielle, cholestérol, résultats ECG, etc.). Taux de valeurs manquantes élevé : jusqu'à 66% pour certaines variables.

Classification : 7 variables numériques (age, trestbps, chol, thalch, oldpeak, ca), 9 variables catégorielles (sex, cp, fbs, restecg, exang, slope, thal, etc.).

2. Imputation des Valeurs Manquantes

- **Variables numériques continues :** KNNImputer (k=5) pour préserver les relations entre variables.
- **Variable discrète (ca) :** SimpleImputer avec stratégie 'most_frequent' pour maintenir l'intégrité des classes.
- **Variables catégorielles :** SimpleImputer avec valeur constante 'missing' (approche conservative).

3. Encodage des Variables

- **Variables nominales** (sex, cp, fbs, exang) : OneHotEncoder pour créer des variables binaires.
- **Variables ordinales** (restecg, slope, thal) : OrdinalEncoder avec ordre personnalisé (ex: normal < anomalie < hypertrophie).

4. Normalisation - Choix du RobustScaler

Comparaison de 3 méthodes :

Méthode	Avantage	Inconvénient
MinMaxScaler	Bornes [0,1]	Sensible aux outliers
StandardScaler	Centré-réduit	Sensible aux outliers
RobustScaler ✓	Robuste aux outliers	Pas de bornes fixes

Justification : RobustScaler utilise la médiane et l'écart interquartile (IQR), ce qui le rend robuste aux valeurs extrêmes. Essentiel pour les données médicales où les outliers sont souvent des cas pathologiques informatifs.

5. PolynomialFeatures - Degré 2

Variables sélectionnées : age, trestbps (pression), chol (cholestérol).

Résultat : 9 features générées (3 originales + 3 quadratiques + 3 interactions). Ex: age, age², age×pression.

Justification du degré 2 : Capture les interactions importantes (ex: le cholestérol est plus risqué chez les personnes âgées) sans explosion combinatoire du degré 3 (19 features → risque d'overfitting).

6. Discrétisation

Variable âge : KBinsDiscretizer (3 bins, stratégie K-Means) pour créer 3 groupes d'âge équilibrés (Jeunes, Moyens, Seniors).

7. Pipeline de Traitement Complet

Architecture : Pipeline scikit-learn intégrant toutes les étapes de prétraitement et un modèle RandomForestClassifier.

Flux de traitement :

1. Imputation (KNN + SimpleImputer)
2. Encodage (OneHot + Ordinal)
3. Normalisation (RobustScaler)
4. Classification (RandomForest)

Paramètres du modèle : 100 arbres, profondeur max = 10, équilibrage des classes.

8. Résultats et Performance

Division : 80% entraînement / 20% test (stratifiée).

Score d'accuracy : ~75-85% sur le jeu de test (varie selon le random_state).

Validation croisée : 5-fold CV pour validation robuste des performances.

Features importantes : Les variables les plus discriminantes sont identifiées via feature_importances_ du RandomForest.

9. Avantages du Pipeline

• Reproductibilité : Toutes les transformations encapsulées et appliquées dans le même ordre.

• Pas de data leakage : Le preprocessing s'ajuste uniquement sur les données d'entraînement.

• Code maintenable : Structure modulaire facile à modifier et à étendre.

• Évite les erreurs : Impossible d'oublier une étape de transformation.

10. Conclusion et Perspectives

Résumé : Ce projet a démontré l'importance d'un prétraitement rigoureux des données médicales. Le choix du RobustScaler pour gérer les outliers, l'utilisation de PolynomialFeatures (degré 2) pour capturer les interactions non-linéaires, et l'intégration dans un pipeline complet garantissent des résultats fiables et reproductibles.

Améliorations possibles :

- Optimisation des hyperparamètres via GridSearchCV ou RandomizedSearchCV.
- Test d'algorithmes alternatifs (XGBoost, SVM, réseaux de neurones).
- Feature engineering supplémentaire basé sur l'expertise médicale.
- Analyse approfondie des erreurs de classification pour identifier les cas difficiles.

Ce pipeline est prêt pour le déploiement en production et peut être utilisé pour prédire le risque de maladie cardiovasculaire chez de nouveaux patients.