UNIVERSITÉ TÉLUQ

VALIDATION DU CHOIX OU DE LA PROPOSITION DU PROJET

PRÉSENTÉ À

FATIMA BENSALMA

COMME EXIGENCE PARTIELLE

DU COURS

PROJET EN SCIENCE DES DONNÉES SCI 1402

PAR

GABRIEL FLEURENT-THIFFAULT

22307023

VALIDATION DU CHOIX OU DE LA PROPOSITION DU PROJET

1 AVRIL 2025

Table des matières

Proposition claire du projet

Premièrement, ce projet se base sur le jeu de données UCI Heart Disease Data trouvé sur kaggle (Kaggle, 2025).

Cet ensemble de données est caractérisé de 16 colonnes de différents types : soit un ensemble dit multivarié. Effectivement, il est possible de constater les 4 types suivants :

Booléens :

1. Exang. Angine induite par l'exercice
2. Fbs; Si glycémie à jeun > 120 mg/dl
3. Sex; Genre de l'individu

Catégories :

1. Cp; Type de douleur thoracique
2. Origin; Lieu de l'étude
3. Thal; Thalassémie
4. Restecg; Résultats électrocardiographiques au repos
5. Slope; Pente du segment ST d'exercice de pointe

Entiers :

1. Age; Âge du patient
2. Ca; Nombre de vaisseaux majeurs (0-3) colorés par fluoroscopie
3. Chol; Cholestérol sérique (en mg/dl)
4. Id; Id unique
5. Num l'attribut prédit
6. Thalach; Fréquence cardiaque maximale atteinte
7. Trestbps; Pression artérielle au repos (en mm Hg à l'admission à l'hôpital)

Nombres flottants :

8. Oldpeak. Dépression du segment ST induite par l'exercice par rapport au repos

En s'appuyant sur cet ensemble et de facto les attributs qui le constituent, le principal objectif est d'identifier les individus qui ont des problèmes de cœur de ceux qui n'en ont pas.

Le deuxième objectif se veut être expérimental et consiste à diagnostiquer et découvrir des informations qui pourraient être pertinentes dans la recherche, afin de mieux comprendre le problème.

Pour l'objectif principal, mon hypothèse est que les patients plus âgés et ayant un taux de cholestérol élevé ont plus tendance à avoir des problèmes de cœur.

Méthodologie à suivre

Premièrement, je devrai faire un ETL pour m'assurer que je travaillerai avec des données fiables.

1. Pour extraire le fichier .csv, j'utiliserai *Spark*.
2. Pour ce qu'il s'agit des transformations, je viendrai m'assurer de la qualité des informations reçues, ainsi, je viendrai préparer les données avant l'analyse. Différentes techniques pourront être requises : l'unification des intitulés, identification des données aberrantes, normalisation, etc. (Chikhaoui, 2025).
3. Une fois la nouvelle table chargée, l'analyse pourra débuter.

Un modèle de classification MLLIB via Spark sera créé. La classification sera faite avec une régression logistique ; puisque nous voulons prédire la probabilité qu'une personne ait des problèmes de cœur et ainsi nous utiliserons des variables dépendantes discrètes, soit oui ou non. De plus, l'évaluation de la performance avec les paramètres sélectionnés sera également faite, afin de déterminer la précision (Chikhaoui, 2025).

Finalement, le modèle pourra être déployé et un graphique sera créé également avec Spark. Du même coup, l'hypothèse pourra être confirmée ou infirmée (Chikhaoui, 2025).

Bibliographie

Chikhaoui, B. (2025). Module 1 Section 1.3 [Notes fournies dans le cours SCI 1017]. https://m2.teluq.ca/mod/page/view.php?id=137216

Chikhaoui, B. (2025). Module 3 Section 3.4 [Notes fournies dans le cours SCI 1017]. https://m2.teluq.ca/mod/page/view.php?id=137159

Kaggle. (2025). UCI Heart Disease Data. Repéré le 25 mars 2025 à https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data