Said MOHAMED SEGHIR
Mohamed SELMI
Amir BEN MALLEM

SUJET : HCV
GROUPE 14

Description et justification de la méthodologie

Après avoir effectué la description des données lors de la précédente partie, nous allons passer au choix des différents algorithmes à utiliser pour entrainer et tester notre modèle ainsi qu'aux différentes méthodes d'évaluation des modèles.

Avant de passer à l'entrainement des différents algorithmes choisis, nous devons passer par le pré-traitement des données afin de rééquilibrer la distribution des données entre les différentes catégories.

Nous devons aussi numériser les variables qualitatives (ex. Sexe) et corriger les valeurs aberrantes.

Choix des algorithmes

Après avoir étudié la nature de notre problème on constate qu'on a affaire à un problème de classification multi classes supervisée, en prenant en compte la taille des données à disposition, nous avons opté pour ces trois différents algorithmes :

- Knn : Cet algorithme se base sur le jeu de données pour faire la prédiction et n'a pas besoin de modèle prédictif. Par contre il doit garder en mémoire l'ensemble des observations pour effectuer les prédictions, la taille de notre jeu d'entraînement n'est pas vraiment grande donc il est intéressant de tester cette méthode.

Pour évaluer notre modèle nous allons varier K. Généralement plus K est grand plus notre prédiction est fiable, mais il faut faire attention à ne pas tomber dans l'erreur de sur-apprentissage.

Notre algorithme a besoin d'une fonction pour calculer les distances entre les plus proches voisins. Le choix de cette fonction est lié au jeu de données dans notre cas nous avons choisi la distance Euclidienne.

- SVM : Contrairement au Knn le SVM a besoin d'un modèle prédictif, nous avons décidé de retenir cet algorithme car il répond à notre problématique de classification, il est aussi adapté à notre jeu de données de petite taille.

Pour améliorer notre modèle nous pouvons faire varier ses hyper-paramètres, la constante C et gamma afin de contrôler la marge "margin" ou le noyau.

- Régression logistique : Cette méthode est très utilisée en médecine et a prouvé son efficacité dans la caractérisation de sujets malades par rapport à des sujets sains.

Nous avons choisi la régression logistique et non linéaire car notre variable expliquée est qualitative.

De plus, notre sujet étant un problème de classification non binaire, nous allons utiliser une extension de la méthode : la régression logistique multinomiale (multinomial logistic regression ou multiclass LR). Nous avons également remarqué lors de l'étude statistique que la colinéarité entre les variables n'était pas forte, ce qui est une hypothèse de la méthode.

Méthode d'evaluation des modèles

Cross validation : Au vu de la taille de notre dataset l'utilisation de la méthode de cross validation est nécessaire afin d'exploiter chaque donnée à la fois pour l'apprentissage et le test de notre modèle. Et ce en divisant les données en k sous-ensemble, chaque bloc sera utilisé pour entrainer le modèle k-1 fois.

Optimisation des hyper paramètres

Grid search : Cet algorithme permet d'évaluer le modèle pour chaque combinaison de paramètres afin d'obtenir la combinaison optimale, en comparant les scores des différentes combinaisons.

Comparaison des modèles

Afin de comparer les trois méthodes nous allons utiliser des métriques extraites de la matrice de confusion de chaque modèle. Les métriques à considérer sont :

- La précision.
- Le rappel.
- Le score.
- L'exactitude.