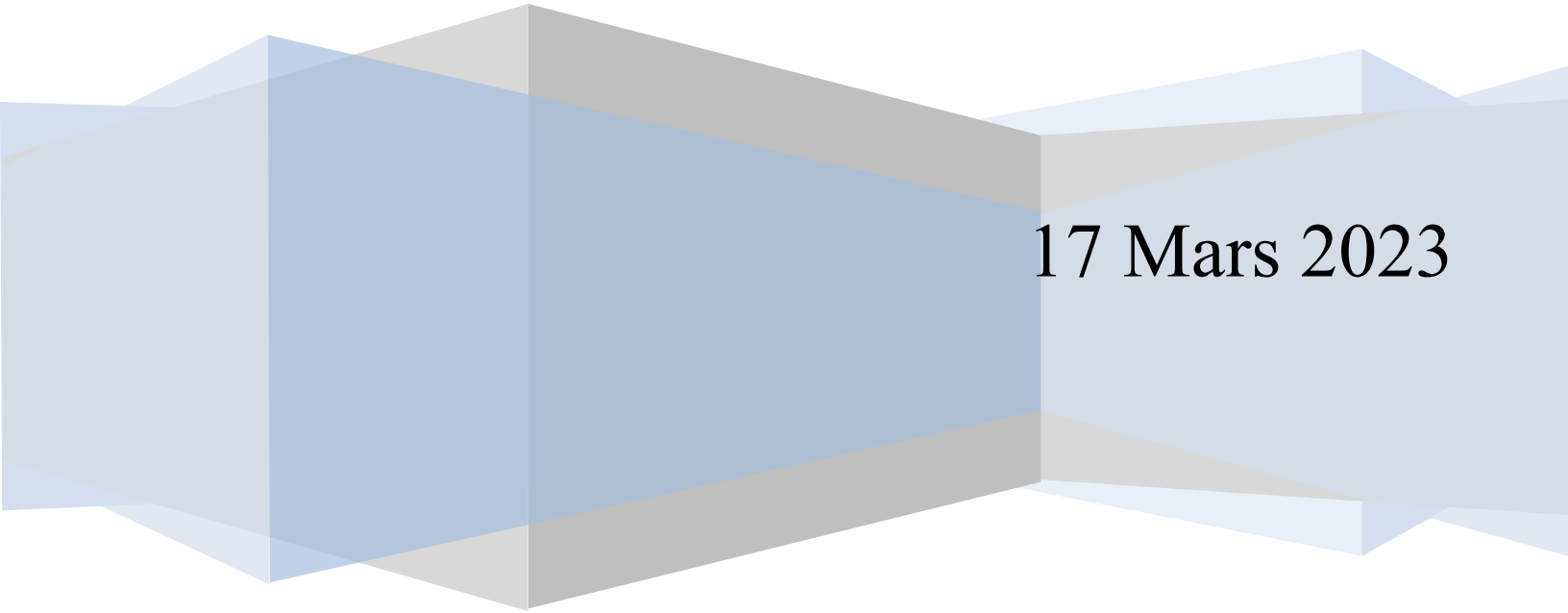
Université Paris Cité

Programmation Web

Objectif: Exploration de données et apprentissage de modèles supervisés avec RShiny.

Rose-Milca Cenat

Master I, Apprentissage Machine pour la science des données.



17 Mars 2023

MODEL LEARNING (Mammographic Mass)

Description :

Cette application fait l'étude complète du jeu de données : Mammographic Mass.

Cet ensemble de données peut être utilisé pour prédire la gravité (bénigne ou maligne) d'une masse mammographique à partir des attributs BI-RADS et de l'âge de la patiente.

Il comporte 6 attributs: 1 goal field, 1 non-predictive, 4 predictive attributes:

1. BI-RADS assessment: 1 to 5 (ordinal, non-predictive!)
2. Age: patient's age in years (integer)
3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4. Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6. Severity: benign=0 or malignant=1 (binominal, goal field!).

Utilisation :

Pour utiliser l'application vous devez charger les données de Mammographic Mass, l'extension du fichier doit être .CSV ou .data. Si vous importer d'autres jeux de données ou un autre format vous aurez une erreur.

C'est dû au fait que le fichier que nous avons utilisé ne contenait pas les entêtes (noms de colonnes). De ce fait, nous les avons ajoutées et n'avons pas traités d'autres jeux de données. Toute l'application est basée sur les données de Mammographic Mass.

Apres l'importation, vous pouvez faire l'exploration de vos données, si vous n'importez pas les données, certains champs lèveront des erreurs. Nous avons remplacées les valeurs manquantes, au début elles sont toutes remplacées par NA, pour permettre une meilleure visualisation des graphes et des pourcentages de valeurs manquantes, elles seront imputées plus tard.

Analyse Exploratoire :

Pour l'analyse unidimensionnelle, pour chaque variable, vous pouvez voir les graphes adaptés, selon son type soit catégoriel ou quantitatif.

Pour l'analyse bidimensionnelle, vous pouvez croiser les variables entre elles, et vous pouvez visualiser les graphes adaptés. Le graphe apparait en fonction des variables : Qualitative/Qualitative (que la variable Age avec elle même), Qualitative/Quantitative, Quantitative/Quantitative.

Les données ont été modifiées avant les graphes pour une meilleure visualisation, les valeurs 0, 1, etc. ont été remplacées par leurs vraies valeurs, par exemple pour Severity, 0 et 1 ont été remplacés par Benign et Malign.

Les variables qualitatives ont été converties en factor (variables nominales) et order(variables ordinales).

Classificateurs :

Nous utilisons trois classificateurs : Régression Logistique, Decision Tree et KNearest.

L'imputation des données manquantes a été faite avant d'utiliser les classificateurs avec la méthode KNN et la métrique Gower Distance.

Régression Logistique :

La régression logistique est multiple, nous avons utilisé la fonction step() pour la sélection des fonctionnalités, afin de trouver le meilleur ensemble de variables (AIC le plus bas). Ensuite nous avons utilisé les variables recommandées dans un nouveau modèle.

Nous avons fait un k-fold cross validation avec k=10 et mesuré la performance à l'aide d'une courbe ROC et d'une aire sous la valeur de la courbe (AUC). L'AUC est de 0.8708801, ce qui suggère que le modèle fonctionne très bien.

La Precision est de 0.8397566, ce qui signifie que sur toutes les observations que le modèle a prédites comme positives, 83,97% étaient effectivement positives. En d'autres termes, lorsqu'il a prédit que Severity était Malign, dans 83,97% des cas, il avait raison.

Le Recall est de 0.8023256, signifie que le modèle a correctement identifié 80,23% des cas Malign parmi tous les cas réellement Malign. Cela signifie que le modèle a manqué environ 20% des cas Malign.

Le F-Score de 0.8206145, ce qui indique que le modèle a une performance globale bonne.

Decision Tree

Nous exécutons le modèle basé sur toutes les variables, puis nous taillons l'arbre du nombre idéal de variables. Nous avons utilisé un *K-fold cross validation* avec k = 10, et fait une prune, afin de trouver le meilleur modèle.
L'AUC est de 0.8567699, ce qui suggère que le modèle fonctionne très bien.

La Precision est de 0.8648069

Le Recall est de 0.7810078

Le F-Score est de 0.8207739

Ces mesures d'évaluation sont assez similaires à celles du modèle de la Régression Logistique précédent en termes de F-Score, mais il y a une légère amélioration de la précision et une légère diminution du rappel. Cela peut indiquer que le modèle est plus précis dans la prédiction des cas positifs, mais qu'il peut manquer certains cas positifs.

K-nearest neigbour:

Nous parcourons chaque combinaison de variables possible et nous les testons avec une valeur de k égale à 4, 8, 16 et 20. Il ya donc un modèle pour chaque combinaison de variables pour 5 valeurs de k différentes.

La Precision est de 0.8463115

Le Recall est de 0.8003876

Le F-Score est de 0.8227092

Conclusion :

Si nous nous concentrons sur la Precision, le Recall, et le F-score, le modèle de Decision Tree semble être le meilleur choix, car il a la Precision la plus élevée tout en ayant un Recall et un F-score raisonnables. Toutefois, si nous privilégions la courbe ROC, le modèle de Régression Logistique a la plus grande surface sous la courbe (AUC), ce qui peut être un critère important.

Les features les plus importantes sont : Age, Shape et Margin selon les meilleurs modèles.

Density n'est pas une variable très pertinente et BI_RADS est non prédictive.

Ce projet se veut le plus complet que possible tous les points mentionnés dans le document ont été abordés sauf ceux qui concernent la présence possible d'outliers et du déséquilibre des classes.

Il n'ya qu'une seule donnée qui pouvait être considérée comme outliers dans BI_RADS, puisque BI_RADS n'est pas prédictive, nous nous avons passée de ce cas. Les classes n'étaient pas déséquilibrées. D'autres points comme error rate a été calculé mais n'est pas affiché sur l'interface.

L'application peut se perfectionner en devenant responsive et porter beaucoup plus d'attention dans l'exploration des données.