Rapport de l'atelier : ANN régression linéaire

Filière : Modélisation et sciences des données (MSD1)

Sujet :

Prédiction des charges de d'assurance de maladie

Réalisé par :

Mohcine MOURCHID

Ziad OULD BOUYA

Sous la direction de :

MR. Ali IDRI

Table des matières

Prédiction des charges de d'assurance de maladie 1
1. Introduction 3
- Contexte 3
- Objectif 3
- Représentation des variables 3
2. Description de Matériels 4
- Data-Set 4
- Techniques utilisées 4
- Mesures de performance 5
3. Experimental design 7
❖ Description et Visualisation 7
❖ Prétraitement des données 7
❖ Méthodologie 9
Conclusion 10

1. Introduction

• Contexte

Une compagnie d'assurance maladie ne peut gagner de l'argent que si elle perçoit plus qu'elle ne dépense en soins médicaux pour ses bénéficiaires. D'autre part, même si certaines conditions sont plus répandues pour certains segments de la population, les coûts médicaux sont difficiles à prévoir puisque la plupart de l'argent provient de conditions de patients rares.

L'historique des prédictions de prix montre que les auteurs ont largement utilisé les techniques d'apprentissage automatique dans ce domaine.de manière extensive. Le domaine de la santé ne fait pas exception à la règle, puisque les prix des médicaments sont prédits à l'aide de données relatives à la santé

• Objectif

Dans l'étude donnée, nous avons un problème de régression dans lequel on doit créer un model qui va permettre de prédire avec précision les coûts d'assurance en se basant sur les données des individus, notamment l'âge,bmi, le fait qu'il fume ou non, etc. En outre, nous déterminerons également quelle est la variable la plus importante qui influence les coûts d'assurance. Ces estimations pourraient être utilisées pour créer des tables actuarielles qui fixent le prix des primes annuelles plus ou moins élevées en fonction des coûts de traitement prévus. Il s'agit d'un problème de régression.

• Représentation des variables

Age **:** âge du premier bénéficiaire

Sex : sexe de l'assureur, féminin, masculin

BMI : Indice de Masse Corporelle, permettant de comprendre les poids corporels relativement élevés ou faibles par rapport à la taille, indice objectif de poids corporel (kg / m^2) utilisant le rapport taille/poids, idéalement 18,5 à 24,9

Children **:** nombre d'enfants couverts par l'assurance maladie, nombre de personnes à charge

Smoker **:** fumer ou pas

Region **:** la zone résidentielle du bénéficiaire aux États-Unis, nord-est, sud-est, sud-ouest, nord-ouest.

Charges **:** frais médicaux individuels facturés par l'assurance maladie

2. Description de Matériels

- Data-Set

Dans ce projet, on utilise une base de données kaggle, Le jeu de données contient 1338 lignes et 7 colonnes. Les colonnes présentes dans l'ensemble de données sont 'age',' sex','bmi', 'children', smoker', 'region', et 'charges '. La colonne charges est la colonne cible et les autres sont des colonnes indépendantes. Les colonnes indépendantes sont celles qui prédiront le résultat.

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

- Techniques utilisées

a) Régression linéaire

L'analyse de régression linéaire sert à prévoir la valeur d'une variable en fonction de la valeur d'une autre variable. La variable dont vous souhaitez prévoir la valeur est la variable dépendante. La variable que vous utilisez pour prévoir la valeur de l'autre variable est la variable indépendante.

Ce type d'analyse estime les coefficients de l'équation linéaire, impliquant une ou plusieurs variables indépendantes, qui estiment le mieux la valeur de la variable dépendante. La régression linéaire consiste en la détermination d'une droite ou d'une surface qui réduit les écarts entre les valeurs de sortie prévues et réelles.

b) Réseau de neurones

Un réseau neuronal artificiel (ANN) est un modèle mathématique souvent utilisé pour l'apprentissage machine / profond . Il contient un ensemble interconnecté de neurones artificiels . En règle générale, un ANN est un système adaptatif qui ajuste sa structure en fonction des données externes ou internes qui circulent dans le réseau pendant le processus d'apprentissage. Les réseaux de neurones actuels sont des outils de modélisation de données non linéaires. Ils sont généralement utilisés pour modéliser des relations délicates ou des bases de données très grandes.

c) Grid search

La grille de recherche est utilisée pour trouver la meilleure *hyperparamètres* d'un modèle qui donne les prédictions les plus "précises".

d) Cross validation

L'une des méthodes de validation des modèle, cette méthode est basée sur l'utilisation des différentes données d'entraînement pour qu'on puisse avoir une idée claire sur la performance du modèle

• Mesures de performance

L'étape essentielle de tout modèle d'apprentissage automatique consiste à évaluer la précision du modèle. L'Erreur quadratique moyenne, l'Erreur absolue moyenne, l'Erreur quadratique moyenne moyenne et les métriques R-carré ou Coefficient de détermination sont utilisées pour évaluer les performances du modèle dans l'analyse de régression.

- L'erreur absolue moyenne représente la moyenne de la différence absolue entre les valeurs réelles et prédites dans l'ensemble de données. Il mesure la moyenne des résidus dans l'ensemble de données.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Where,

$\hat{y}$ — predicted value of y

$\bar{y}$ — mean value of y

- L'erreur quadratique moyenne représente la moyenne de la différence au carré entre les valeurs d'origine et prédites dans l'ensemble de données. Il mesure la variance des résidus.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

- L'erreur quadratique moyenne est la racine carrée de l'erreur quadratique moyenne. Il mesure l'écart type des résidus.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

- ❖ Le coefficient de détermination ou R au carré représente la proportion de la variance de la variable dépendante qui est expliquée par le modèle de régression linéaire. Il s'agit d'un score sans échelle, c'est-à-dire, que les valeurs soient petites ou grandes, la valeur de R au carré sera inférieure à un.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

3. Experimental design

❖ Description et Visualisation

Le but de cette étape est d'avoir une idée sur la distribution de notre jeu de données, bien comprendre les valeurs de chacune de ses variables qui peuvent être qualitatives ou quantitatives.

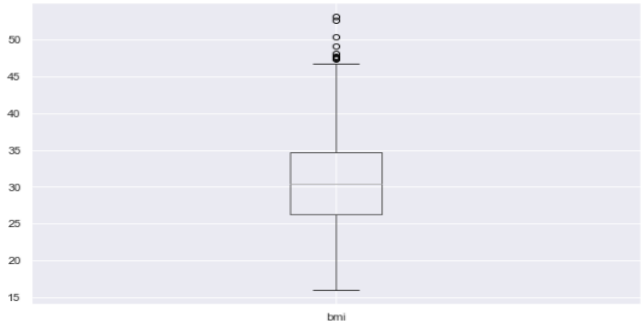
La visualisation des données désigne le fait de représenter visuellement ses data pour pouvoir déceler et comprendre des informations, les données brutes étant difficilement interprétables et exploitables. Ce processus se fait par des outils analytiques spécifiques et se matérialise par des tableaux (type Excel), des graphiques, des cartes visuelles ou même des infographies regroupées dans des Dashboard (tableaux de bord).

❖ Prétraitement des données

➢ Traitement des valeurs aberrantes

Les valeurs aberrantes sont les valeurs qui semblent différentes des autres valeurs dans les données.

Box plot est un outil visuel pour identifier les valeurs aberrantes. Le box plot est l'une des nombreuses façons de visualiser la distribution des données et il représente les valeurs q1 (25e percentile), q2 (50e percentile ou médiane) et q3 (75e percentile) des données ainsi que (q1-1,5*(q3-q1)) et (q3+1,5*(q3-q1)). Les valeurs aberrantes, le cas échéant, sont représentées par des points situés au-dessus et au-dessous du graphique.

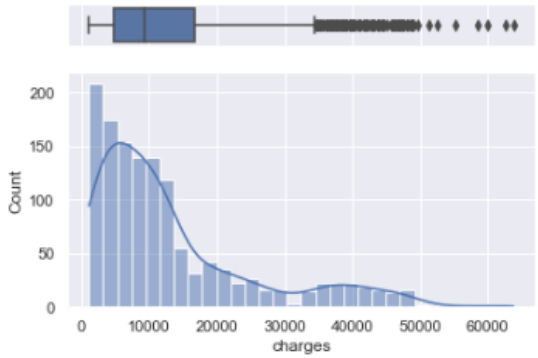


Il y a donc 9 valeurs aberrantes, ce qui représente environ 0.67% de toute la population. La décision est de les supprimer.

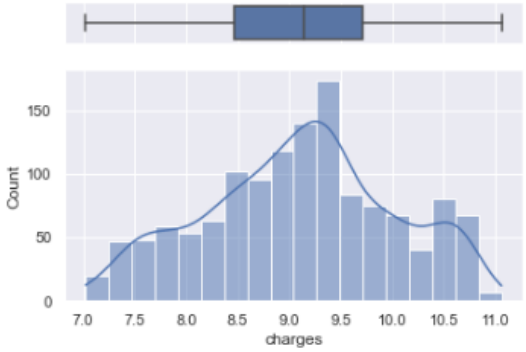
➢ Transformation of Target Variable

Nous étudions le comportement et la tendance de la variable cible ou de la variable de réponse du problème, c'est-à-dire "charges". Les résultats de cette étape nous aideront également à transformer la variable pour l'adapter au mieux à nos modèles.

Ceci est important pour garantir que les prédictions du modèle sont plus précises. Dans les algorithmes de régression, il est nécessaire que les résidus suivent une distribution normale.



Before transformation



After transformation

➢ Encodage des données

L'encodage est le fait de rendre les valeurs qualitatives des valeurs quantitatives, on le fait par rendre à chaque classe un code généralement (0,1,2…).

Notre base de données contient 3 variables categorielle : 'sex', 'smoker','region'. Le problème avec les variables qualitatives c'est que Certains algorithmes peuvent travailler directement avec des données catégorielles. Mais de nombreux algorithmes d'apprentissage automatique ne peuvent pas fonctionner directement sur les données des étiquettes. Ils exigent que toutes les variables d'entrée et les variables de sortie soient numériques.

Nous allons utiliser une technique appelée **Label Encoder** pour les colonnes 'sex' 'smoker. L'encodage des étiquettes consiste simplement à convertir chaque valeur d'une colonne en un nombre.

Le codage de la "région". En général, les variables catégorielles avec une grande variabilité sont mieux codées en utilisant OneHotEncoder et ainsi de suite. Mais dans ce cas, rien ne changera, car il n'y a pas d'ordre particulier dans lequel les régions seraient listées. Donc nous avons seulement utilisé le Label Encoder.

➢ Normalisation de données

La mise à l'échelle des caractéristiques est une méthode utilisée pour normaliser la plage de variables indépendantes ou les caractéristiques des données. La mise à l'échelle des fonctionnalités aide essentiellement à normaliser les données au sein de la même échelle. Nous avons utilisé la fonction fit_transform pour appliquer la mise à l'échelle des fonctionnalités sur les données d'apprentissage.

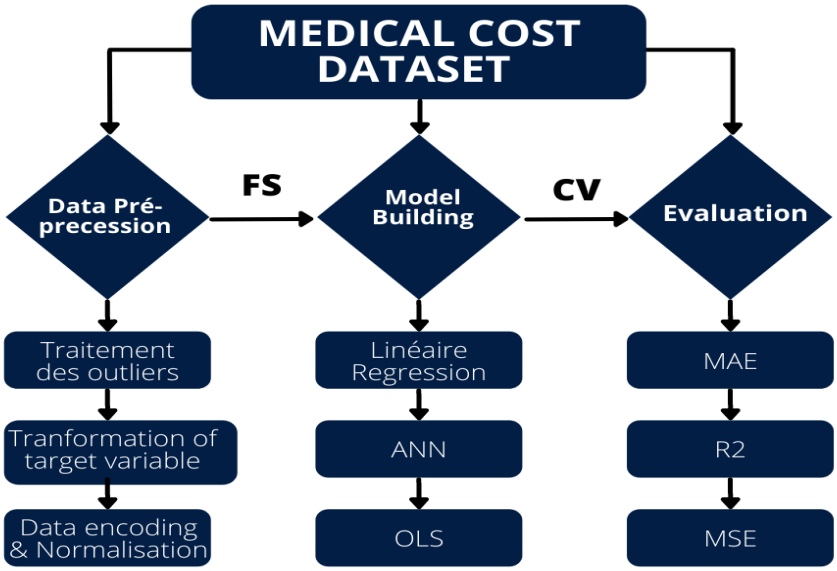
Pour la mise en échelle des on va la normaliser avec la méthode MinMax.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

❖ Méthodologie

La figue ci-dessous montre clairement la méthodologie suivie afin de realise toutes les evaluations empiriques, elle se centre sur 4 etape principales ;

1- L'importation du jeu de donnees Medical Cost.
2- Data pre-procession
3-



4. Analyse des résultats et discussion :

Dans cette partie, nous allons montrer et discuter les résultats des évaluations des techniques de Machine Learning sur la dataset de « medical cost ». Le tableau ci-dessous résume ces résultats, on a utilisé 3 mesures de performances (**mae** : mean absolute error, **mse** : mean squared error, **rmse** : root mean squared error), on remarque que ces performances ont des valeurs très petits donc les deux modèles performent bien sur notre base de données. Pour arriver aux résultats déclarés dans le tableau, on a utilisé Reseau de neurone, Reseau de neurone + Cross validation , Reseau de neurone + Grid search, OLS avec ces variables explicatives : « age, bmi, children, smoker, charges », on a choisi celles-ci grâce à la matrice de corrélation en éliminant les variable « sex,northeast, northwest, southeast, southwest» qu'ont des coefficients proche à 0.

Modèle	MAE	MSE	RMSE	R^2
RN	0 .056	0.008	0.091	0.846
RN + CV	0.053	0.009	0.095	0.818
RN + GS	0.060	0.010	0.099	0.824
OLS	0.073	0.013	0.115	0.76

On remarque que le meilleur modèle est Réseau de neurone.

Conclusion

Nous avons appris que le facteur le plus important pour prédire les dépenses médicales d'un individue est le comportement tabagique et l'âge, ce qui signifie que le tabagisme est mauvais pour la santé, comme nous le savons déjà et que cela augmente inévitablement les dépenses médicales parce qu'en raison du tabagisme, on est susceptible de tomber malade plus que les non-fumeurs.

Nous avons également constaté qu'avec l'augmentation de l'âge, il faut prendre plus de soins et de précautions pour sa santé car avec l'augmentation de l'âge, la santé devient fragile, donc ils vont faire des contrôles médicaux fréquents, susceptibles de tomber malade rapidement comme avec l'augmentation de l'âge l'immunité tombe donc ils adoptent des mesures pour rester en bonne santé en prenant des médicaments et en pratiquant certaines activités physiques comme le jogging, la marche, le yoga, ce qui entraîne une augmentation des dépenses médicales.