|  |  |  |  |
| --- | --- | --- | --- |
| Session de : **Décembre 2022** | N° de la session : 1 | Enseignant : **P.A Bisgambiglia** | |
| Nom du diplôme : **Master 2 DFS** | | | Année : M2 |
| Nom de l’épreuve : **Introduction à la recherche opérationnelle**  A distribuer avec le sujet de : | | | Durée totale de l’épreuve : **3h** |

Calculatrice (FX 92 ou équivalente) autorisée : Non  Oui

1. Apprentissage supervisé : couts des assurances - 18 points

On souhaite évaluer le coût des assurances.

1. Analyse globale des données (**3 points**)

* *Utilisez la bibliothèque pandas pour récupérer les données.*
* *Quelle est la taille du DataSet, et de quel type sont les variables.*
* *Y a-t-il des valeurs manquantes dans le DataSet.*
* *A partir de l'analyse statistique sur les variables de type int et float que peut-on en conclure. Pensez-vous qu'il y ait des valeurs aberrantes.*
* *Identifiez pour chacune des variables catégorielles identifiées le nombre de valeurs différentes de ces variables. Y-a-t-il des données mal orthographiées.*

1. Type de problème et nettoyage des données (**3 points**)

* *Identifiez la* target *et le type de problème à traiter. Affichez comment se répartissent les prix des assurances.*
* *En regardant le DataSet sans faire d'analyse statistique pensez-vous que certaines variables puissent être supprimées.*
* *Identifiez où se trouvent les doublons. Pour identifier les doublons vous pourrez utiliser la fonction* duplicated *avec l’option* keep=False*.*
* *Supprimez maintenant les doublons du DataSet.*

1. Relations entre les features et la target (**3 points**)

* *En regardant les prix moyens des assurances en fonction des différentes valeurs des variables catégorielles identifiez quelles sont les valeurs qui semblent avoir une influence sur les prix. Vous pourrez utiliser les fonctions* groupby*,* means *et* value\_counts*.*
* *Analysez maintenant de manière plus fine l'impact des variables* sex *et* region *sur les prix moyens des assurances en fonction de l’état* smoker*.*
* *Comment évolue les coûts selon* l’âge*, du* bmi *ou du* nombre d’enfants *en fonction que les personnes soient des fumeurs ou pas. Vous pourrez utiliser pour cela soit la fonction* pairplot *pour visualiser toutes les distributions bivariées entre les variables non catégorielles, ou bien la fonction* scatterplot *pour chaque variable en fonction de la target.*
* *En analysant la relation* Age/Charges *que peut-on en déduire. Effectuez la même analyse de la relation* bmi/Charges*.*

1. Modification des features (**3 points**)

* *Transformer toutes les variables catégorielles en variables numériques. Vous pourrez utiliser soit un LabelEncodeur (pour encoder une variable) soit un OrdinalEncodeur pour un ensemble de variables.*
* *En affichant maintenant la matrice des corrélations quelles sont les variables qui ont le plus de lien avec la target. Est-ce que ces résultats confirment ce que vous avez constaté dans les parties précédentes.*
* *En ne conservant que les personnes qui fument, que deviennent les coefficients de corrélation. Y-a-t-il des différences.*
* *Reprenez le DataSet de départ et transformer les deux variables catégorielles* sex *et* smoker *en utilisant une transformation ordinale, et la variable* region *en utilisant une transformation One Hot. Vous pourrez utiliser pour dela la fonction* get\_dummies*.*

1. Apprentissage simple (**3 points**)

* *Découpez le DataSet en deux DataSet pour l'apprentissage et le test. Vous pourrez utiliser un découpage 80/20.*
* *Utiliser un modèle de régression linéaire :* LinearRegression() *de la bibliothèque* linear\_model*. Quels sont les résultats obtenus (*score*), vous afficherez également la différence entre les prédictions et les valeurs attendues via la fonction* mean\_squared\_error *de la bibliothèque* metrics*.*
* *En utilisant un affichage de type* scatterplot *entre les données attendues et les prédictions quels sont les coûts qui sont les moins bien estimés. Comment interprétez-vous cela.*
* *Quels sont les résultats si l'on supprime la variable* children*, puis les variables* children *et* region*.*
* *Quel est l'impact d'une transformation centrée réduite sur les données. Vous utiliserez pour cela utiliser la fonction* ScandarScaler *pour centrer et réduire les données du DataSet.*

1. Amélioration du modèle (**3 points**)

* *Utilisez maintenant un modèle* random forest *pour traiter le problème.*
* *Testez différentes valeurs [1 à 25] pour l’hyper paramètre* max\_depth*. Vous utiliserez une cross validation pour tester votre modèle et un découpage* cv=10*. Affichez les résultats obtenus via la fonction* plot*.*
* *Quel est la meilleure valeur pour la profondeur maximale. Evaluez le modèle avec cette valeur et affichez les valeurs attendues en fonction des valeurs obtenues avec* scatterplot*. Vous pourrez ajouter une couleur (paramètre* c*) en fonction de la variable* smoker *de X\_test.*
* *A partir des données obtenues via la fonction* feature\_importances *du modèle affichez l’importance des différentes variables sur le coût des assurances.*

1. Recherche opérationnelle - 5 points

*Résoudre le problème suivant :*

*Sous contraintes de*

* *Modifiez le problème pour obtenir une formulation standard. Quelles sont les variables hors base et les variables de la base. Quelles sont leurs valeurs.*
* *Quelle est la première valeur sortante.*
* *Calculez la nouvelle équation lorsque l’on change pour la première fois la base. Quel est le gain obtenu.*
* *Réalisez toutes les étapes tant que la fonction peut être maximisée. Quel est le gain. Quels sont les valeurs des quatre variables.*