Contrôle TP C

A l'issue de l'évaluation, vous devez déposer dans **dépôt évaluation TP ,** un fichier **zip** contenant :

- Votre code python commenté (script .**py**)
- Le ou les fichiers Excel utilisé(s) avec les tableaux/transformations demandées (.**xlsx**)
- Le document EvalA.docx enregistré au format **pdf** (avec vos réponses)

Partie I

Les données portent sur les variétés rouges et blanches du vin portugais « Vinho Verde ».

Référence : Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

Vous commencerez par travailler sur le fichier ***winequality-red - C.csv*** concernant les variétés « rouge » du Vinho Verde.

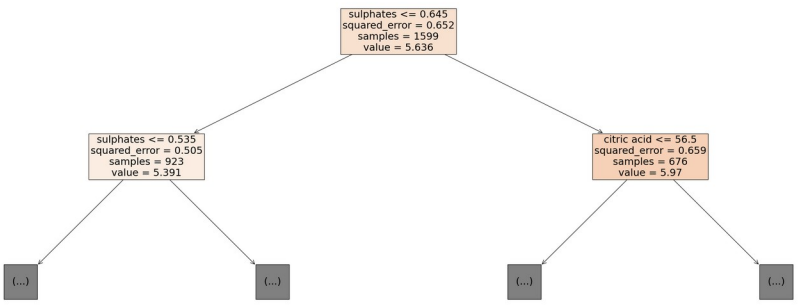
Les variables que nous avons retenues dans ce fichier :

citric acid : acide citrique – variable quantitative – de 0 à 1,66
residual sugar : sucre résiduel – variable quantitative – de 0,6 à 65,8
sulphates : sulfates – variable quantitative – de 0,22 à 1,08
alcohol : degré d'alcool - variable qualitative – modalités : « <10 », « [10,11[» , « [11,12[», « >12 »
quality : score d'évaluation de la qualité du vin – variable quantitative – de 3 à 8
avis : directement lié au score précédent (quality) – variable qualitative – modalités : bon (si quality >6), pas bon (sinon)

Nous allons, à partir de ces données, construire des arbres de décision/régression.
Les variables explicatives sont les variables **citric acid**, **residual sugar**, **sulphates, alcohol, quality, avis** .
Les variables cibles sont **quality** (arbre de régression) et **avis** (arbre de décision)

A - Arbre de régression (6 points)

1. Importer les données sous **Python** et construire un arbre de régression, la variable cible étant la variable **quality**.
   Collez ci-dessous une image de la première division (racine de l'arbre et deux premiers fils)

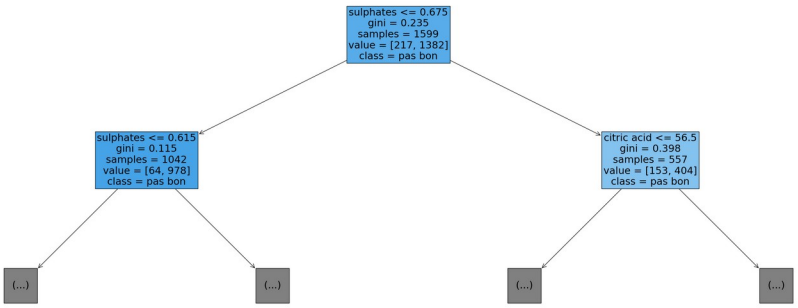


2. Sous **Excel :**
   Un tableau, sous Excel, permet de vérifier les valeurs obtenues sur les trois nœuds de l'arbre collé ci-dessus.
   a. Collez ci-dessous une image de ce tableau
   b. Expliquez comment vous l'avez obtenu
   c. Calculer la variance résiduelle associée à cette première division

B - Arbre de décision (6 points)

Nous travaillons sur les mêmes données

1. Importer les données sous **Python** (si vous ne l'avez pas déjà fait) et réaliser un arbre de régression, la variable cible étant la variable **avis**.
   Collez ci-dessous une image de la première division (racine de l'arbre et deux premiers fils)



2. Sous **Excel,**
   Deux tableaux, sous Excel, permettent de vérifier les valeurs obtenues sur les trois nœuds de l'arbre collé ci-dessus.
   a. Collez ci-dessous une image de ces tableaux
   b. Expliquez comment vous les avez obtenus
   c. Calculer le gain de pureté de Gini associé à cette division

Partie II

Les données portent maintenant sur les employés d'une entreprise : ***employes5C.csv***

Les variables qualitatives ont déjà été codées.

Liste des variables :

Attrition : variable cible, indiquant si l'employé a quitté (« oui ») volontairement l'entreprise
Age : âge de l'employé
Deplacements : variable qualitative relative aux déplacements professionnels de l'employé

0 Pas de déplacement professionnel
1 Peu de déplacements professionnels
2 Déplacements professionnels fréquents

Departement : variable qualitative relative au service dans lequel travaille l'employé

1 Ressources humaines
2 Recherche & Développement
3 Service commercial

DistanceDomicile : variable quantitative, indiquant la distance domicile/travail de l'employé

NiveauFormation : variable qualitative correspondant au niveau d'étude de l'employé

1 *Brevet*
2 Bac
3 Licence/Bachelor
4 Master
5 Doctorat

Genre : variable qualitative relative au genre de l'employé

1 Homme
2 Femme

StatutMarital : situation de l'employé

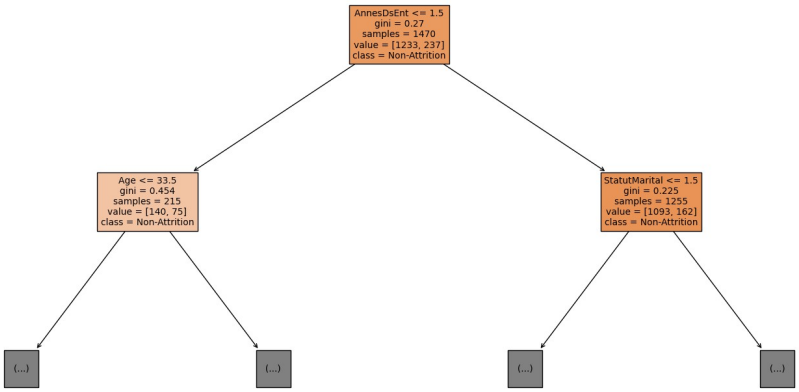
1 Célibataire
2 Divorcé
3 Marié

AnnesDsEnt : nombre d'années dans l'entreprise, de l'employé

AnneesPromo : nombre d'années depuis la dernière promotion

A - Arbre de décision (8 points)

1. Importer les données sous **Python** et réaliser un premier arbre de décision de profondeur 3. Collez ci-dessous une image de la première division (racine de l'arbre et deux premiers fils)



2. Que peut-on dire du traitement des variables qualitatives nominales ?

   On est pas obligés de les convertir en variables numériques.

3. Sous Excel ou sur Python, transformer le tableau de données (dataframe) afin que les variables qualitatives nominales soient correctement traitées.
4. Faire un nouvel arbre (coller l'image de la première division)