Projet final - Analyse des données d'AGROBALYSE

KABONGO BUZANGU Ben

Prédiction des groupes d'aliments des ingrédients

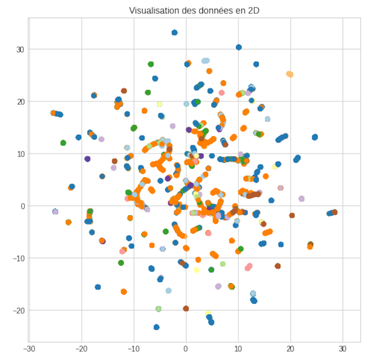

Figure: Données en fonction du groupe d'aliments

En connaissant différents indicateurs d'impacts environnementaux des différents ingrédients, est-il possible d'en prédire le groupe d'aliment ?
On dispose de 11 groupes d'aliments différents. Après suppression des attributs jugés non pertinents pour l'apprentissage (les codes, les noms français, etc.), normalisation des données, nous avons cherché à prédire avec les classifieurs **KNN** et d'**arbre de décision numérique** le groupe d'aliments depuis la table des ingrédients.
Par 10 itérations de **validation croisée**, nous obtenons les performances suivantes, en apprentissage et en test :

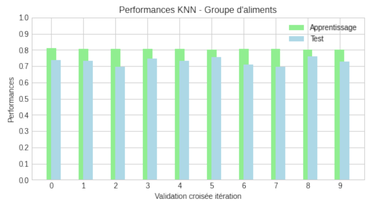

Figure: KNN : Groupe d'aliments

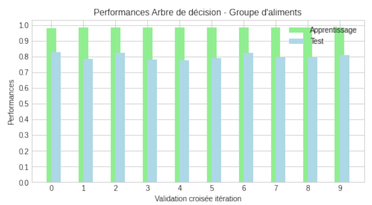

Figure: Arbre de décision : Groupe d'aliments

Prédiction de la note DQR

Les valeurs de DQR sont réelles. Pour notre cas, nous ne souhaitons pas effectuer ici de régression linéaire, mais une classification multi-classe. Nous les avons donc découpées en 5 intervalles correspondant à nos différentes classes à prédire.

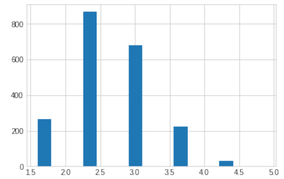

Figure: DQR

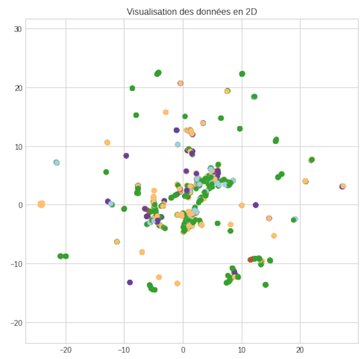

Figure: Données en fonction de la note DQR

Est-ce possible de prédire la fiaibilité des informations pour chaque exemple ?
Nous avons utilisé un classifieur **Random Forest**, avec 5 arbres de décision, un nombre d'attributs divisé de moitié pour chacun des arbres et une proportion d'exemples fixée à 0.5 pour chaque arbre. Et avons tenté de prédire la note DQR pour chaque exemple de la table synthèse, en supprimant les colonnes non pertinentes et les exemples dont la valeur DQR n'était pas un nombre.
Par 10 itérations de **validation croisée**, nous obtenons les performances suivantes :

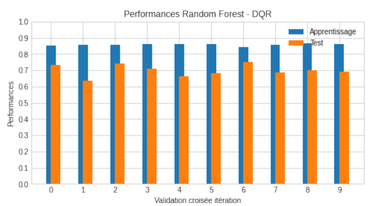

Figure: Random forest : DQR

Régression sur le score EF

Le score **Score unique EF** est une combinaison linéaire de certains indicateurs de nos différents datasets. Il a donc été pertinent pour nous de prédire ce score, en utilisant la régression.
Afin de faire de la régression polynomiale, nous avons utilisé le principe de kernelisation avec un noyau polynomial.
Nous avons fait cette étude sur le dataset **synthèse**, après suppression des colonnes non pertinentes, des exemples abberants et normalisation des données, nous avons entrainé notre modèle et observé l'évolution de la fonction de coût au fil des itérations. Nous avons utilisé la méthode des **moindres carrés** pour notre fonction coût.

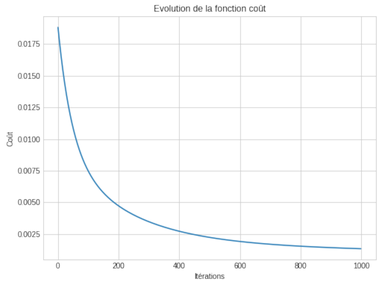

Figure: Régression : Score Unique EF

Le **coefficient de détermination** de notre modèle est de **0.886** ; étant assez proche de 1, nous pouvons conclure que notre modèle est performant.

K-Moyennes sur le dataset des étapes

Nous avons appliqué l'algorithme des K-Moyennes sur les données du dataset **étapes**.
Dans un premier temps, nous avons fixé la valeur de k à 11, en référence au nombre de groupes d'aliments.
Ci-dessous, une visualisation sur 2 dimensions des résultats obtenus :

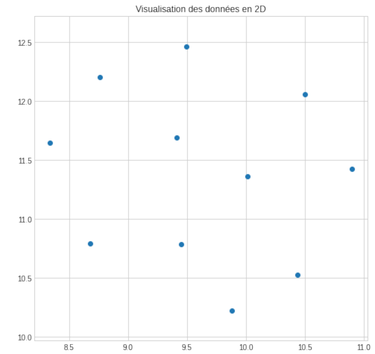

Figure: K = 11 - Visualtion des centres : Dataset étapes

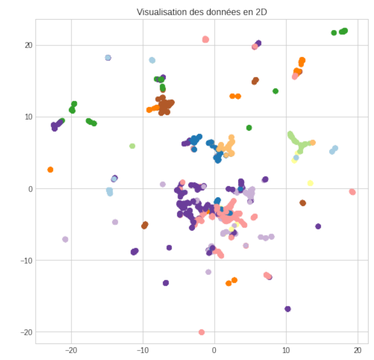

Figure: K = 11 - Visualtion des données : Dataset étapes

L'**index de Dunn** de la partition obtenue est de **0.0266**. Nous avons ensuite évalué les performances de différentes tailles de partitions, en fonction de leur index de Dunn.

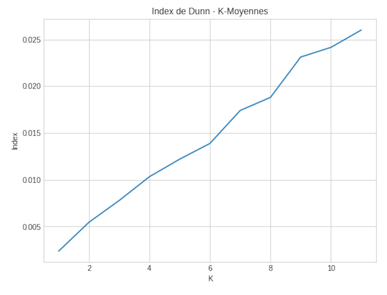

Figure: Evaluation de différentes partitions : Dataset étapes