Introduction:

Les données traitées dans ce rapport sont obtenues en se basant sur le choix et préférences alimentaires des étudiants. Cet ensemble de données comprend des informations sur les choix alimentaires, la nutrition et les préférences des enfants et d'autres informations provenant des étudiants de l'Université de Mercyhurst à Pennsylvania. Il y a 126 réponses d'étudiants. Les données sont brutes et non nettoyées.

Questions à traitées :

Quelle est l'importance de l'information nutritionnelle pour les étudiants d'aujourd'hui ? Les enfants de parents qui cuisinent sont-ils plus susceptibles de faire de meilleurs choix alimentaires que les autres ? Quelle sont les descripteurs qui influencent sur la note (GPA) ? et Comment obtenir une bonne note ?

Cette dernière question est illustré par une application GUI dans le notebook 'Food_choice_App'. Elle marche bien localement mais le notebook crashe soit disant qu'il n'y a pas les 'plugins' QT nécessaires, pourtant le module PyQt5 est bien installé via la commande '! pip install'.

Cet ensemble de données comprend un certain nombre de questions ouvertes, telles que : Quel est votre aliment de réconfort préféré ? Quelle est votre cuisine préférée ? Qui pourrait bien fonctionner pour le traitement du langage naturel (NLP).

Outils et Méthodologie :

L'objectif principal de ce travail est de mettre en pratique tout / le maximum possible des connaissances acquises lors du cours de M. Vigon, en l'occurrence, traitement des données. Ainsi, la méthodologie consiste à revoir, mettre en œuvre ou réutiliser et appliquer les méthodes vues en cours pour ce cas d'étude.

Travail sur la NLP :

Ce type de donné est abordé en s'appuyant sur le cours de M. Vigon et les sites web mentionnés en dessous. Il s'agit d'étude qui se concentre sur les interactions entre le langage humain et les ordinateurs s'appelle le traitement du langage naturel ou PNL, en abrégé. Il se situe à l'intersection de l'informatique, de l'intelligence artificielle et de la linguistique informatique (Wikipédia).

Le travail est principalement fait sur trois volets :

- Supprimer la ponctuation
- Rendre les lettres en minuscules

- Supprimer les numéros

Le résultat de ce travail sera des données propres et organisées dans deux formats de texte standard :

- ✓ Corpus - un recueil de textes
- ✓ Matrice des termes du document - comptage des mots sous forme de matrice

Nous avons seulement exécuté les étapes de nettoyage communes ici et le reste pourra être fait plus tard pour améliorer nos résultats.

Plan :

- Commencer par une question principale : C'est quoi le modèle qui prédit la note selon le jeu de donné disponible ?
- Nettoyer les données
- L'exploration des données
- Appliquer des techniques
- Résultat et conclusion

Bases de pandas : Nettoyer les données

Obtenir les données :

Les données sont disponibles sur le lien suivant : https://www.kaggle.com/borapajo/food-choices .

Préparer les données :

Ce projet passe par une étape nécessaire à tout projet de traitement des données : le nettoyage des données. Le nettoyage des données est une tâche longue et fastidieuse, mais très importante. Gardez à l'esprit que "des déchets entrants sont des déchets qui sortent". En introduisant des données incomplètes dans un modèle, nous obtiendrons des résultats qui n'ont aucun sens.

- Plus précisément, nous allons traverser :
- Obtenir les données - dans ce cas, nous allons extraire les données d'un site web
- Nettoyage des données - nous allons passer en revue les techniques de prétraitement de texte les plus courantes
- Organiser les données - nous organiserons les données nettoyées de manière à ce qu'elles puissent être facilement introduites dans d'autres algorithmes

Nettoyage des données :

Lorsqu'il s'agit de données numériques, le nettoyage des données consiste souvent à supprimer les valeurs nulles et les données en double, à traiter les valeurs aberrantes, etc. Pour les données textuelles, il existe des techniques courantes de nettoyage des données, également appelées techniques de prétraitement du texte.

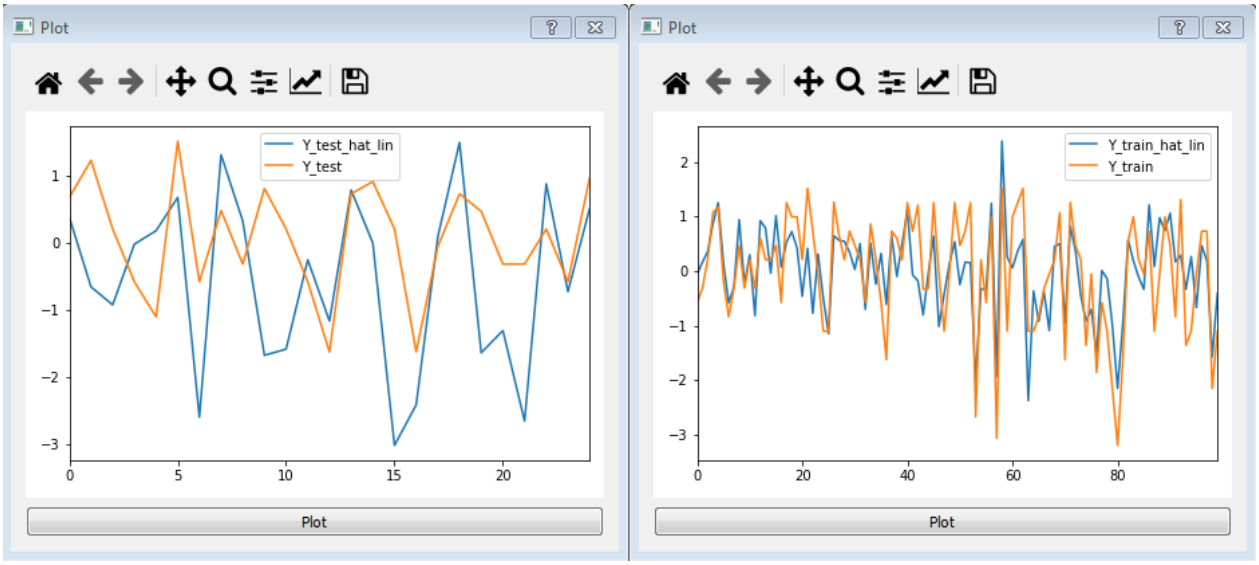
Modèles entrainés :

Régression linéaire :

Le jeu de donnés a été stratifié selon 5 classes donnée par le 'GPA' ce qui nous permis de bien mélanger les données et en tirer deux sous ensemble :

- Un jeu d'entrainement
- Et un jeu de test

Pour la validation on a utilisé la validation croisée car on n'a pas assez de donnés (125 lignes). L'Hyper-paramètre Alpha est fixé sur '0.5'. Voilà le résultat pour les données entrainée (à droite) et celles gardées pour le test (à gauche).

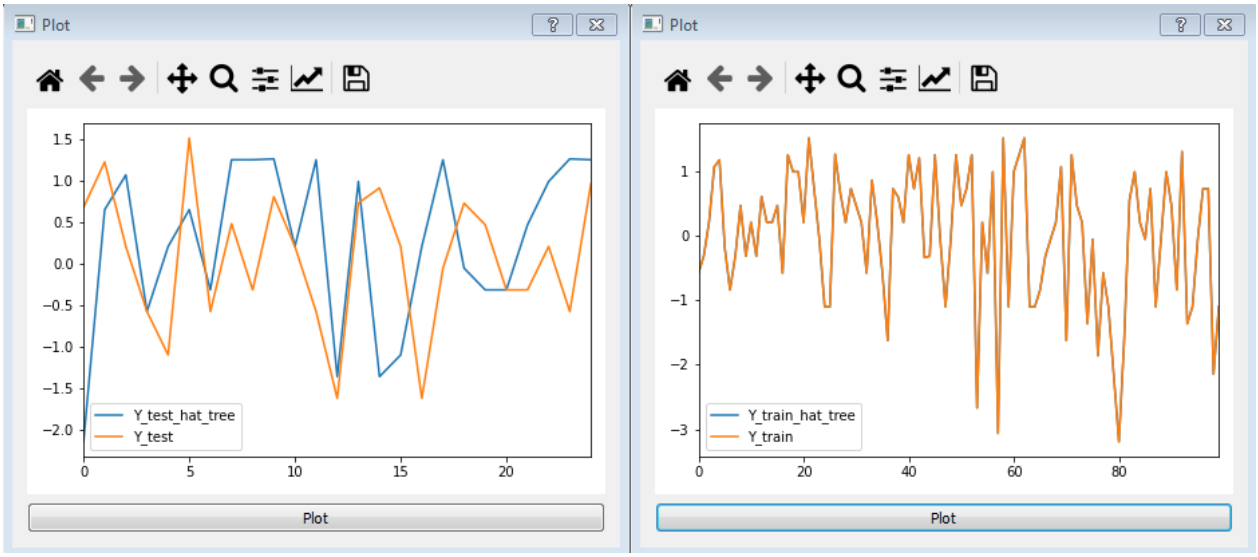


Ces résultats sont donnés par une interface graphique aussi développé le long de ce projet. Le modèle est sous-entraîné ce qui est dû au nombre limité de donnée. Il est aussi probable qu'il n'y a pas de corrélation linéaire avec les autres descripteurs.

Arbre de décision :

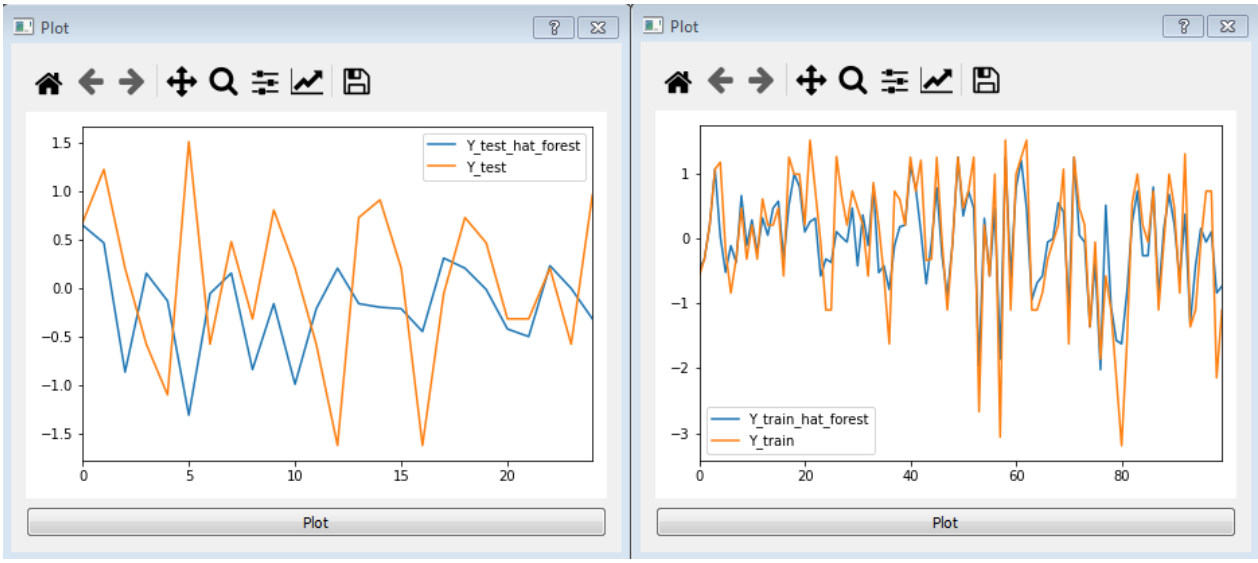
Les données sont stratifiées de la même façon, toujours en utilisant la bibliothèque scikit-learn. Le résultat est comme suite :

Les données de test á gauche et celles pour entrainer á droite. Cette fois on a clairement un surentraînement. Ceci est dû au grand nombre des descripteurs. Il va falloir régulariser les hyper-paramètres.



Forêt aléatoire :

Le modèle de la forêt aléatoire est aussi entrainé par le même jeu donné et aussi pour la validation. La différence et qu'on a effectué une régularisation d'hyper-paramètres afin d'éviter le surentrainement. Le résultat suivant donne par l'interface graphique 'Food choice' illustre visuellement un avantage pour cette méthode. Concrètement, l'erreur calculée par les données de test nous confirme la même chose.



Afin de trouver ces hyper-paramètres la méthode aléatoire a été choisie vu le nombre d'avantage qu'elle a. En effet, Les avantages de la recherche aléatoire sont :

- on peut interrompre ou reprendre à tout moment une recherche aléatoire. Alors qu'en interrompant une recherche en grille, on laisse des intervalles de valeur inexplorés.

- on a beaucoup de souplesse dans le choix du hyper-paramètre: on n'est pas obligé de les tirer de manière uniforme: On peut insister sur les valeurs qui nous paraissent les meilleures. On peut facilement ajouter des contraintes (tel hyper-paramètre doit être plus petit que tel autre ...)
- On peut travailler avec un grand nombre d'hyper-paramètre et un grand nombre de valeur. Dans ce cas, une recherche exhaustive serait trop couteuse.
- La recherche aléatoire a aussi un avantage caché: supposons qu'un premier paramètre a très peu d'importance (ils influent très peu les performances) et qu'un second paramètre a beaucoup d'influence. Bien entendu, cette information nous ait cachée (sinon on ne ferait varier que le second paramètre). Avec la recherche aléatoire, on va explorer beaucoup de valeurs différentes du second paramètre, donc on aura plus de chance de tomber sur une valeur optimale.

Conclusion :

Afin de visualiser plusieurs scenarios et répondre à plusieurs questions, une interface graphique pourrai automatiser et simplifier le choix du modèle. Car le modèle parfait ne peut être trouvé que si on le connaissait déjà ou du moins on fait des hypothèses pour qu'il soit parfait. ''Dans un article célèbre de 1996, David Wolpert a démontré que si on fait absolument aucune hypothèse sur les données, alors il n'y a aucune raison de préférer un modèle à un autre.''

Sources:

- https://www.kaggle.com/borapajo/food-choices
- https://irma.math.unistra.fr/~vigon/
- https://datascience.foundation/sciencewhitepaper/natural-language-processing-nlp-simplified-a-step-by-step-guide
- https://www.youtube.com/watch?v=xvqsFTUsOmc
- https://github.com/adashofdata/nlp-in-python-tutorial/blob/master/1-Data-Cleaning.ipynb