EXAMEN FINAL :

INF6333-01 Éléments d'intelligence artificielle appliquée

Prof :

Cretu Ana-Maria

Réalisé par :

EMNA KRAIEM

Code permanant :

KRAE18369909

1. Définition du problème :

Le défi ici réside dans la prédiction de la spontanéité d'une personne en se référant à 34 caractéristiques et la cible extraites du fichier allFeaturesProcessedTraining.csv parmi lesquelles :

- 4 traits physiologiques (P1 à P4),
- 4 traits visuels associés aux yeux (Y1 à Y4),
- 26 traits faciaux (M1 à M26) caractéristiques.

Il s'agit d'un problème de régression supervisée en raison de la continuité de la variable cible (Spontaneity)

2. Justification de la technique d'apprentissage choisie :

Afin de résoudre ce souci, j'ai expérimenté différents modèles de régression tels que la régression de Ridge (vu en cours), les machines de support vectoriel (SVM), la régression linéaire, KNN (k-nearest neighbors) vu qu'il est un modèle non paramétrique efficace pour détecter les relations locales dans les données et des techniques plus avancées telles que le Xgboost. J'ai sélectionné ces modèles en raison de leur aptitude à gérer des données complexes et multidimensionnelles.

Prétraitement des données :

Les mesures suivantes ont été déjà mises en place afin d'assurer la qualité des données :

- Les zéros ont été utilisés pour remplacer les valeurs manquantes, une méthode simple mais efficace pour les données normalisées.

- L'échelle Min-Max a été utilisée pour mesurer les caractéristiques afin d'assurer des valeurs comprises entre 0 et 1.
- Les données ont été normalisées en utilisant le Z-score afin d'obtenir une moyenne de 0 et un écart-type de 1, ce qui est crucial pour les modèles sensibles à l'échelle tels que SVM.
- J'ai supprimé la colonne M19 vu qu'elle ne montre aucune corrélation et n'a aucun effet ce qui permet d'optimiser les performances et de réduire le surapprentissage et j'ai cherché les caractéristiques fortement corrélées pour pouvoir supprimer une seule et garder une autre.

3.Évaluation :

Les résultats des modèles ont été mesurés en utilisant des mesures classiques et spécifiques au domaine, telles que :

- L'erreur moyenne au carré (MSE) : évalue la différence entre les prédictions et les valeurs réelles.
- La moyenne des erreurs absolues, plus intuitive et moins sensible aux grandes erreurs, est mesurée par le MAE (Mean Absolute Error).

Évalue la proportion de variance expliquée par le modèle en utilisant le coefficient de détermination R^2.

- La corrélation linéaire entre les prédictions et les vraies valeurs est mesurée par le coefficient de corrélation de Pearson.
- La concordance entre les valeurs prédites et réelles est évaluée par le coefficient de corrélation de concordance (CCC), une métrique essentielle dans ce domaine.

4. Comparaison des modèles :

En se référant aux résultats obtenus, le modèle Gradient Boosting obtient les scores les plus élevés grâce à un MSE très bas (0.0030), un MAE minimal (0.0391) et un coefficient de détermination R2 élevé (0.9843), ce qui démontre une excellente aptitude à expliquer la variance des données cibles.

Le modèle Random Forest, en revanche, bien qu'il soit efficace, affiche un MSE plus élevé (0.1893), un MAE plus élevé (0.2994) et un R2 nettement inférieur (0.1218).

De plus, la corrélation de Pearson du modèle Gradient Boosting est supérieur à celle du Random Forest, ce qui suggère une relation linéaire plus élevée avec les valeurs cibles.

D'après ces résultats, il semble que le modèle le plus adapté pour résoudre ce problème soit le Gradient Boosting.

5. Recommandations :

Malgré les résultats intéressants du modèle sélectionné, il existe différentes approches supplémentaires que je les suggère à mon amie qui pourraient éventuellement améliorer les performances du modèle :

Par exemple, on peut ajuster les paramètres en utilisant des méthodes comme GridSearchCV ou RandomizedSearchCV afin de modifier les paramètres essentiels tels que n_estimators, max_depth, min_samples_split, min_samples_leaf et max_features. Cela favoriserait la recherche de la combinaison idéale afin d'améliorer les performances.

Modèle	MSE	MAE	R^2
Gradient Boosting (Initial)	0.0047	0.0526	0.9756
Gradient Boosting (Ajusté)	0.0034	0.0409	0.9825

Sélection des caractéristiques ou réduction de la dimensionnalité :

On peut éliminer les caractéristiques peu pertinentes en utilisant des méthodes telles que le RFE ou l'analyse de variance des caractéristiques.

Vérification croisée approfondie :

On peut effectuer une validation croisée avec plus de plis (par exemple, cv=10) afin de diminuer la variance des résultats.

Utilisation des techniques de suréchantillonnage (comme SMOTE) :

Pour optimiser les hyperparamètres et augmenter les données, outre que ça on peut utiliser d'autres méthodes de normalisation autre que le z_score comme le Robust Scaling.

Une autre recommandation à ajouter c'est d'essayer d'entraîner le modèle sur 100% des données de train(allFeaturesProcessedTraining.csv) vu qu'ils sont très peu et essayer de les tester sur le jeu de données de test (allFeaturesProcessedTesting).