Rapport projet machine learning : Classification

Données

Source :

L'ensemble de données a été créé par Athanasios Tsanas (tsanasthanasis@gmail.com) et Max Little (littlem@physics.ox.ac.uk) de l'Université d'Oxford, en collaboration avec 10 centres médicaux aux États-Unis et Intel Corporation qui a mis au point le dispositif de télésurveillance permettant d'enregistrer les signaux vocaux. L'étude originale a utilisé une série de méthodes de régression linéaires et non linéaires pour prédire le score des symptômes de la maladie de Parkinson du clinicien sur l'échelle UPDRS.

Informations :

Cet ensemble de données est composé d'une série de mesures biomédicales de la voix de 42 personnes atteintes de la maladie de Parkinson à un stade précoce, recrutées pour un essai de six mois d'un dispositif de télésurveillance pour le suivi à distance de l'évolution des symptômes. Les enregistrements ont été réalisés automatiquement au domicile des patients.

Les colonnes du tableau contiennent le numéro du sujet, l'âge du sujet, le sexe du sujet, l'intervalle de temps depuis la date de recrutement de base, l'UPDRS moteur, l'UPDRS total et 16 mesures biomédicales de la voix. Chaque ligne correspond à l'un des 5 875 enregistrements vocaux de ces personnes. L'objectif sera de prédire de prédire le sexe de l'individu en fonction de ses caractéristiques.

Informations sur les attributs :

subject# - Entier qui identifie chaque sujet de manière unique

age - Âge du sujet

sex - Sexe du sujet

test_time - Temps écoulé depuis le recrutement dans l'essai. La partie entière correspond au nombre de jours écoulés depuis le recrutement.

motor_UPDRS - Score UPDRS moteur du clinicien, interpolé linéairement.

total_UPDRS - Score UPDRS total du clinicien, interpolé linéairement.

Jitter(%),Jitter(Abs),Jitter:RAP,Jitter:PPQ5,Jitter:DDP - Plusieurs mesures de la variation de la fréquence fondamentale.

Shimmer,Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,Shimmer:APQ11,Shimmer:DDA - Plusieurs mesures de la variation de l'amplitude

NHR,HNR - Deux mesures du rapport entre le bruit et les composantes tonales de la voix

RPDE - Mesure de la complexité dynamique non linéaire

DFA - Exposant d'échelle fractale du signal

PPE - Mesure non linéaire de la variation de la fréquence fondamentale

Source : http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

Travail à Faire :

La tâche de classification pour cette base de données est de prédire le sexe des sujets atteints de la maladie de Parkinson à partir de leurs mesures vocales et d'autres caractéristiques biomédicales disponibles. Voici une description plus détaillée de la tâche de classification :

But de la classification : Prédire le sexe des patients atteints de la maladie de Parkinson à partir de données vocales et démographiques.

Variables d'entrée :

Mesures biomédicales de la voix (telles que le Jitter, le Shimmer, le NHR, le HNR, le RPDE, le DFA, le PPE).

Caractéristiques démographiques (telles que l'âge).

Autres variables pertinentes disponibles dans l'ensemble de données (telles que le temps écoulé depuis le recrutement, les scores UPDRS moteur et total).

Variable de sortie : Sexe du sujet (encodé comme 0 pour homme et 1 pour femme).

En somme, la classification consiste à construire un modèle capable de prédire le sexe des patients atteints de la maladie de Parkinson à partir de leurs mesures vocales et d'autres caractéristiques disponibles dans l'ensemble de données.

Rétention de caractéristiques :

Les caractéristiques :

1. **subject (0.286851)** : Cette caractéristique a la corrélation la plus élevée avec "sex", mais elle ne semble pas être une caractéristique pertinente pour la classification du sexe, car elle est simplement un identifiant unique pour chaque sujet et ne contient pas d'information pertinente sur le sexe.

2. **nhr (0.168170)** : Cette caractéristique a une corrélation significative avec "sex" et pourrait être pertinente pour le modèle de classification.

3. **jitter_ppq5 (0.087995), jitter_rap (0.076718), jitter_ddp (0.076703)** : Ces caractéristiques mesurent différentes variations de la fréquence fondamentale et ont des corrélations modérées avec "sex".

4. **shimmer_apq5 (0.064819), shimmer (0.058736), shimmer_db (0.056481), jitter (0.051422)** : Ces caractéristiques mesurent différentes variations d'amplitude et ont des corrélations plus faibles avec "sex".

5. **age (-0.041602), total_updrs (-0.096559), ppe (-0.099901), jitter_abs (-0.154645), rpde (-0.159262), dfa (-0.165113)** : Bien que ces caractéristiques aient des corrélations négatives avec "sex", elles ont également une corrélation significative et pourraient être pertinentes pour le modèle de classification.

Les caractéristiques retenues après étude de corrélation:

1. **nhr**
2. **jitter_ppq5**
3. **jitter_rap**

4. **jitter_ddp**
5. **shimmer_apq5**
6. **shimmer**
7. **shimmer_db**
8. **jitter**
9. **age**
10. **total_updrs**
11. **ppe**
12. **jitter_abs**
13. **rpde**
14. **dfa**

Cependant, il est recommandé de faire une analyse plus approfondie et de consulter des experts médicaux pour comprendre **l'importance clinique** de ces caractéristiques en tenant compte de l'interprétabilité : Étant donné que ces données sont médicales, il peut être important de choisir des caractéristiques interprétables et de **l'importance clinique** : car certaines caractéristiques peuvent avoir une importance clinique élevée dans le contexte médical.

Documentations :

- Little MA, McSharry PE, Hunter EJ, Ramig LO (2009), « Adéquation des mesures de dysphonie pour la télésurveillance de la maladie de Parkinson », IEEE Transactions on Biomedical Engineering, 56(4):1015-1022.

- Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection", BioMedical Engineering OnLine 2007, 6:23 (26 juin 2007).