RAPPORT

STAGE : EXERCICE

PROPOSITION DE STAGE: Development of a generic NIRS calibration pipeline using deep learning and model ensembling: application to some reference datasets

January 9, 2020

Réalisé par :

ADOUM OKIM BOKA

CONTENTS

1 INTRODUCTION

Ce projet vise à mettre en place un modèle machine learning pour faire une prédiction dans un problème de régression linéaire multiple. pour mener à bien ce mini-projet, j'ai choisi pour des raison d'une ou d'autres, développer un modèle basé sur le **réseau de neurones artificiels (ANN)** et un modèle de **régression linéaire multiple** .

2 OBJECTIFS DU TRAVAIL

L'objectif de ce travail n'est pas nécessairement d'avoir la meilleure prédiction mais de montrer comment nous abordons et structurons un problème.
Nous notons que: Pour des raisons de manque des informations profondes sur le jeux de données et de l'objectif cité ci-haut, nous avons supposé que les features ont plus ou moins une corrélation avec les valeurs à prédire. Alors nous n'avons pas fait une analyse en composant (ACP).

3 RÉALISATION PRATIQUE, EXPÉRIMENTATIONS

Cette partie concerne la mise en œuvre de notre mini-projet. Nous présenterons quelques résultats sous des captures d'écrans pour le preuve de fonctionnement de ces modèles.

3.1 DATA SET: DONNÉES DE L'EXERCICE

Le jeux de données est composé des valeurs continues. il compte 1154 colones et 162 lignes. il contient deux fichiers, à savoir Xcal.csv, Ycal.csv.

3.1.1 PRÉ-VISUALISATION DES DONNÉES

Ci-dessous une vue partiel de "Xcal" composé des variables indépendants et "Ycal" qui est une variable dépendante.

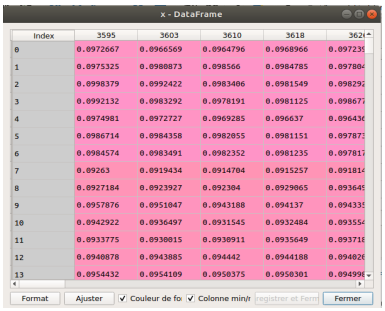

Figure 3.1 Xcal, les variables indépendantes

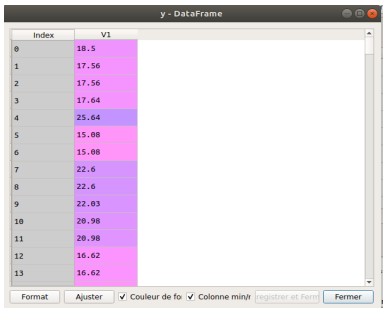

Figure 3.2 Ycal, variable dépendante

3.2 RÉSEAU DE NEURONE ARTIFICIEL

A la fin de l'entraînement de notre modèle, ci-dessous les deux graphes qui schématisent le graphique de la perte d'entraînement par rapport à la perte de validation et le graphique de la précision d'entraînement par rapport à la précision de la validation sur le nombre d'époques. le code source est le fichier **ciradstageexercice_ANN**.

3.2.1 GRAPHIQUE DE LA PERTE D'ENTRAÎNEMENT PAR RAPPORT À LA PERTE DE VALIDATION SUR LE NOMBRE D'ÉPOQUES

Cela nous aidera à prendre des décisions éclaires sur le choix de l'architecture de notre modèle à faire.

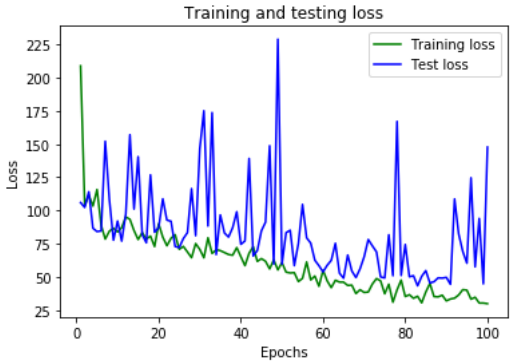

Figure 3.3 la perte d'entraînement par rapport à la perte de validation

3.2.2 GRAPHIQUE DE LA PRÉCISION D'ENTRAÎNEMENT PAR RAPPORT À LA PRÉCISION DE LA VALIDATION SUR LE NOMBRE D'ÉPOQUES

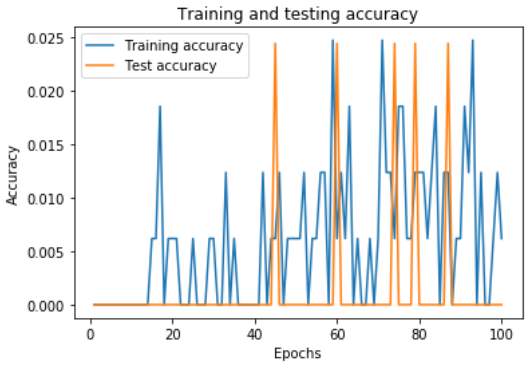

Figure 3.4 la précision d'entraînement par rapport à la précision de la validation sur le nombre d'époques

3.2.3 TABLEAU AJUSTÉ DE $Y_test pour validation et y\_pred pour les valeurs prédites$

Vérifie la différence entre la valeur réelle et la valeur prévue, cas ANN

Figure 3.5 Tableau ajusté de y_test pour validation et y_pred pour les valeurs prédites

3.3 RÉGRESSION LINÉAIRE MULTIPLE (RLM)

Le modèle de régression linéaire multiple est l'outil statistique le plus habituellement mis en œuvre pour l'étude de données multidimensionnelles. nous avons de même développer ce modèle en deux phases (Traitement de données et création du modèle). Code source est le fichier **ciradstageexercice_regression_lineaire_multiple**

3.3.1 TABLEAU AJUSTÉ DE $Y_test pour validation et y\_pred pour les valeurs prédites$

Vérifie la différence entre la valeur réelle et la valeur prévue, cas de RLM.

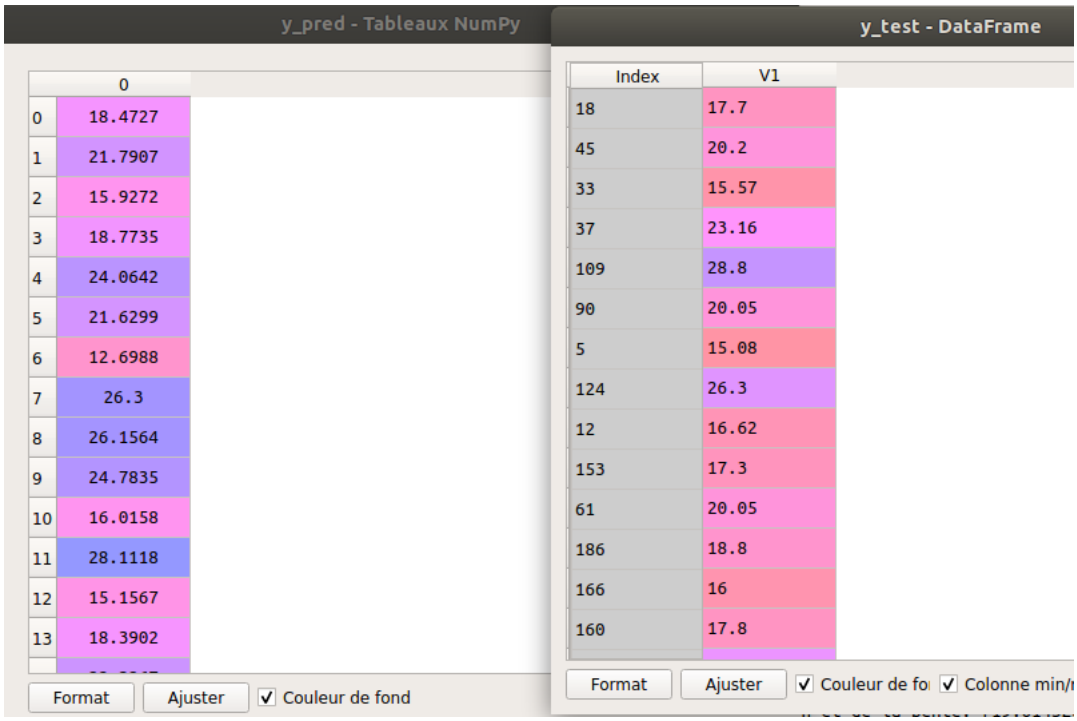

Figure 3.6 Tableau ajusté de y_test pour validation et y_pred pour les valeurs prédites

3.3.2 ÉVALUATION DE LA PERFORMANCE DE L'ALGORITHME

Nous pouvons évalué notre modèle avec partant des fonctions suivantes: Erreur quadratique moyenne, Écart quadratique moyen et Erreur absolue moyenne. Les résultats sont ci-dessous:

- Erreur moyenne absolue: 3.33071890231197;
- Erreur quadratique moyenne: 19.45591519799133;
- Racine carrée de Erreur quadratique moyenne: 4.4108859878703885.

4 CONCLUSION

Nous avons construit deux modèles de prédictions dont l'un est basé sur **réseau de neurones artificiels (ANN)** et l'autre est basé sur **régression linéaire multiple (RLM)** .

Des nombreux facteurs peuvent avoir contribué à des bonnes prédictions pour nos modèles:

- La quantité de données : nous avons besoin d'une énorme quantité de données pour obtenir la meilleure prédiction possible;
- Hypothèses de recueil de données: nous avons fait l'hypothèse que ces données ont une relation linéaire, mais cela pourrait ne pas être le cas;
- features (caractéristiques de "Xcal"): les caractéristiques que nous avons utilisées peuvent ne pas tous avoir une forte corrélation avec les valeurs que nous essayions de prédire "Ycal". Donc une **analyse en composant (ACP)** peut être nécessaire.

Les deux codes sources de ces différents modèles "ANN" et "RLM" sont en fichier joint.

5 RÉFÉRENCES

[1] https://journals.openedition.org/bmsap/4463
[2] https://www.lrde.epita.fr/ sigoure/cours_ReseauxNeurones.pdf
[3] http://eric.univ-lyon2.fr/ ricco/cours/cours/Regression_Lineaire_Multiple.pdf