Groupe 1 : Activity Log Monitoring

Membres :

* Maher Gasmi
* Zied Dridi
* Marwa Ayari
* Chems Marsaoui
* Hend Hedhli
* Zeineb Zarrouk

Introduction :

Dans cette partie pour chaque objectif on récupère les données à partir de notre Datawarehouse via une connexion avec PostgreSQL si ce dernier concerne les données fournies par l’entreprise, sinon pour les 2 derniers objectifs on utilise les fichiers csv qui contiennent les données externes récupérées à travers le WebScrapping principalement et le questionnaire.

Objectifs :

1. Déterminer les relations entre le temps d’exécution des jobs, le volume de données et le nombre de composants d’un job

* On a établi la connexion avec le datawarehouse et on a récupérer les données dont on a besoin

Une image contenant table

Description générée automatiquement

* Une description du maximum, minimum, moyenne … de nos données

Une image contenant table

Description générée automatiquement

* Normalisation des données

Une image contenant table

Description générée automatiquement

* Etude de corrélation entre les variables : graphiquement et coefficient de Spearman et p-value

Une image contenant texte

Description générée automatiquement

Une image contenant graphique

Description générée automatiquementUne image contenant graphique

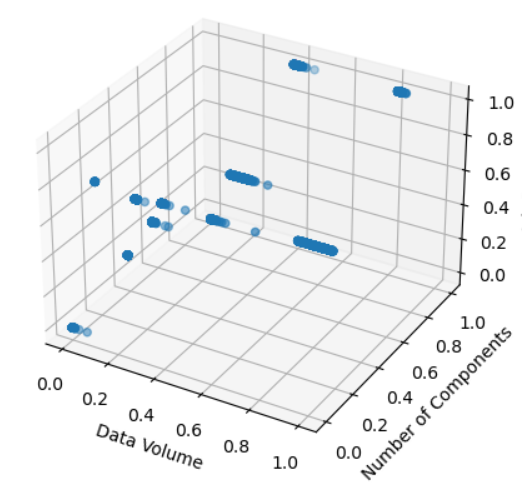
Description générée automatiquement

La variable Job\_Duration admet une corrélation non linéaire positive avec les 2 variables Data\_Volume et Nbr\_Components

1. Prédire le temps d’exécution d’un job à partir du volume des données et du nombre de ses composants

* Régression linéaire multiple :

Visualisation des données du train



Le modèle de prédiction



Visualisation des données de prédiction

Une image contenant graphique

Description générée automatiquement

La valeur de R² de 0,8557298876103305 signifie que 85,57 % de la variance de la variable dépendante (Job\_Duration) peut être expliquée par les variables indépendantes (Data\_Volume et Nbr\_Components) dans le modèle de régression linéaire.

Dans l'ensemble, une valeur de R² de 0,855 et une valeur de MSE de 0,0104 suggèrent que le modèle convient bien aux données et est un bon prédicteur de la durée de l'emploi donnée par le volume de données et le nombre de composants.



1. Déterminer la typologie des jobs en fonction du temps de leur exécution, du volume de données et du nombre des composants

* CAH

Données normalisées

Une image contenant table

Description générée automatiquement

Dendrogramme

Une image contenant graphique

Description générée automatiquement

Segmentation

Une image contenant graphique

Description générée automatiquement

1. Prédire la typologie d’un job à partir du temps de son exécution, de son volume de données et du nombre de ses composants

* Arbre de décision

Visualisation des données de train

Une image contenant graphique

Description générée automatiquement

Visualisation de la prédiction

Une image contenant graphique

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Modèle avec une précision de 99,88%, ce qui indique qu'il est capable de classer avec précision les données.

R² -> 1 donc le modèle est significatif

La matrice de confusion montre que le modèle n'a fait aucune erreur pour les deux premières variables et une seule erreur pour le nombre de composants.

* Modèle fiable

1. Faire une analyse émotionnelle des utilisateurs de Talend
2. Classer les produits de Talend par leurs popularités