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

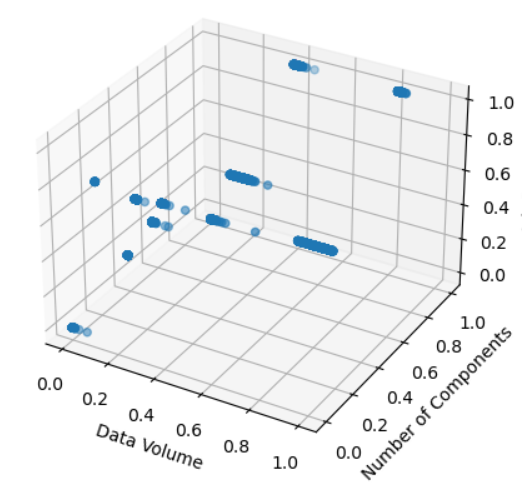
Description générée automatiquement

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.



* Réseaux de neurones:

On a commencé par éliminer les valeurs aberrantes (imputation)

Une image contenant graphique

Description générée automatiquement

Puis normalisation des données

Une image contenant table

Description générée automatiquement

Puis on a crée le modèle de réseaux de neurones (3 couches)

Une image contenant texte

Description générée automatiquement

Et enfin on a calculé les indicateurs

Une image contenant texte

Description générée automatiquement

* Random Forest

Une image contenant texte

Description générée automatiquement

Etude Comparative entre les 2 modèles :

|  |  |  |  |
| --- | --- | --- | --- |
|  | Régression linéaire multiple | Réseaux de neurones | Random Forest |
| R² | 0.85572 | 0.99976 | 0.999937 |
| MSE | 0,010419 | 1.81945e-05 | 4.12877e-06 |

**D'après les résultats obtenus, le modèle Random Forest semble être plus performant que le modèle de réseaux de neurones et le modèle de régression linéaire multiple pour prédire la variable Job\_Duration à partir des variables explicatives Data\_Volume et Nbr\_Components. En effet, le R² du modèle de régression linéaire est de 0.86 et son MSE est de 0.01, le R² du réseaux de neurones est de 0.9997 et son MSE est de** 1.81945e-05**, tandis que le modèle Random Forest a un R² de 0.9999 et un MSE de** 4.12877e-06

**, ce qui est nettement meilleur. Cela suggère que le modèle Random Forest est mieux capable de capturer les relations complexes entre les variables et de généraliser à de nouvelles données.**

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

* Random Forest

Entrainement du modèle puis test

Une image contenant table

Description générée automatiquement

Une image contenant graphique

Description générée automatiquement

Etude Comparative entre les 2 modèles :

|  |  |  |
| --- | --- | --- |
|  | Arbre de décision | Random Forest |
| accuracy | 0.997 | 1.0 |
| Matrice de confusion | Contient une erreur de prédiction | Aucune erreur |

=> Le 2ème modèle est plus fiable

1. Faire une analyse émotionnelle des utilisateurs de Talend

* Pour commencer voici le Word Cloud : on voit clairement que parmi les mots qui se répètent le plus on a des mots techniques comme (Talend, data, product ) et des mots positifs comme (good, easy, great, well)

Une image contenant calendrier

Description générée automatiquement

* Voici les données après l’analyse des sentiments à partir de la variable Text

Une image contenant texte

Description générée automatiquement

Une image contenant graphique

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Une image contenant graphique, diagramme circulaire

Description générée automatiquement

Les sentiments sont entre [-1,1] et on les a divisés en 3 catégories

La majorité des avis sont positifs par rapport à l’utilisation des produits Talend

* Voici les sentiments que dégage chaque nombre d’étoiles on constate que tous les sentiments positifs ont un nombre d’étoiles supérieur à 3

Une image contenant graphique

Description générée automatiquement

* Sentiment par Industrie : les sentiments positifs sont les plus commun pour toutes les industries

Une image contenant graphique

Description générée automatiquement

Une image contenant graphique

Description générée automatiquement

1. Classer les produits de Talend par leurs popularités

* Ordre de popularité des produits de Talend :

Une image contenant graphique

Description générée automatiquement

* La distribution des sentiments par produit

Une image contenant graphique

Description générée automatiquement

* Les mots 10 mots les plus utilisés pour parler de chaque produit

Une image contenant texte, lettre

Description générée automatiquement

* Fréquence des top 10 mots pour chaque produit

Une image contenant graphique

Description générée automatiquement

Une image contenant graphique

Description générée automatiquement