Synthèse introduction à la data intelligence

Chapitre 1 :

Une donnée c’est quoi ?

Une donnée est une information brute, qui peut être présenté sous différentes formes (textes,chiffres,mesures, etc…). Une donnée est stockée et utilisée de différentes manières.

Les différents types de données :

Donnée structurée :

Ce sont des données organisées en tableaux, en lignes et colonnes, ce qui facilite la manipulation et l’analyse.

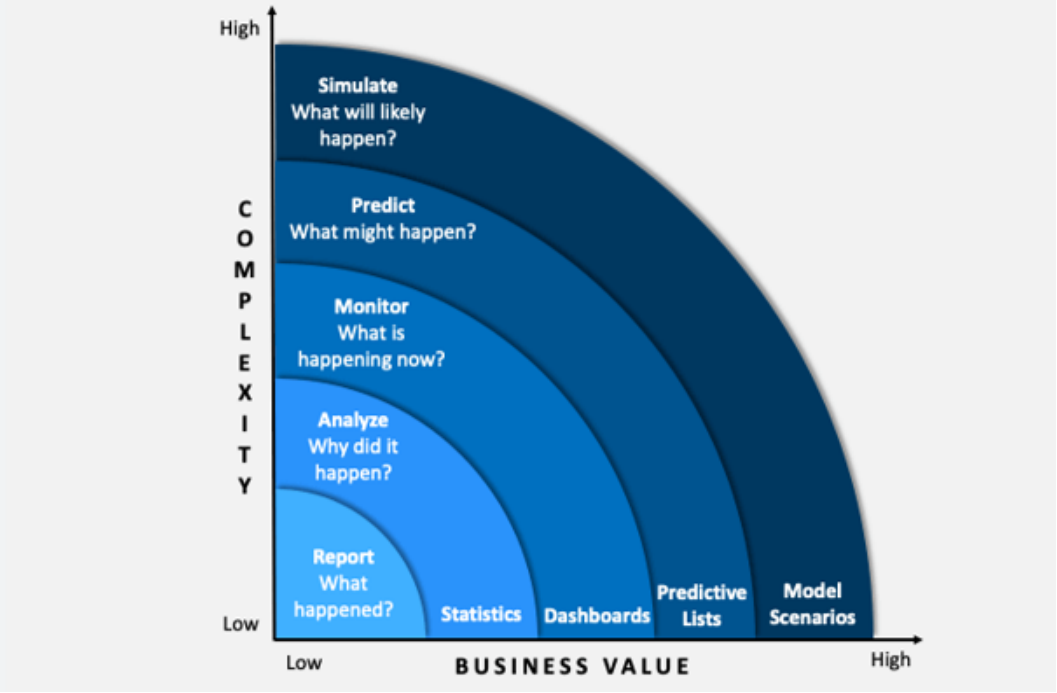
Donnée non structurées :

Ce sont des données qui ne suivent pas un format fixe.

Donnée semi-structurées :

Entre les données structurées et non structurées, elles ont une certaine organisation mais n’ont pas une structure rigide.

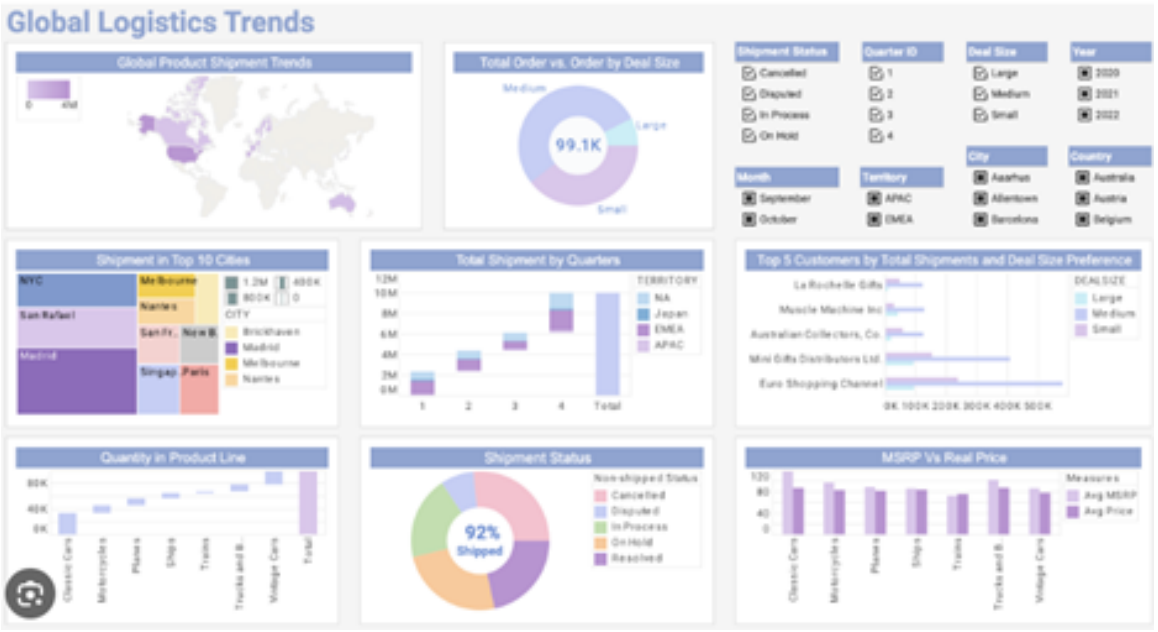
Rapports - statistiques & analyse :



Au sein de ce modèle, les deux premiers paliers de valorisation des données se concentrent sur la rétrospective, fournissant une vision sur le passé et un état des lieux de la situation actuelle. Cela sert a analyser et à comprendre les tendances historique pour prendre une décision plus éclairée dans l'avenir.

Tableaux de bord & monitoring :

Grâce à un tableau de bord, il est possible de surveiller en temps réel les données pour détécter tout problème ou événement important.



Le monitoring c’est une surveillance constante. L’objectif est de s’assurer que tout fonctionne comme prévu et d’identifier rapidement des problèmes potentiels.

Ont utilise les outils de l’informatique décisionnelle.

Machine Learning : prédiction - classification - séries temportelles…

Le Machine Learning est une sous catégorie de l’IA qui se concentre sur le développement de systèmes capables d’apprendre à partir des données sans être explicitement programmés.

Le machine learning est utilisé dans le but de prévenir l’avenir. Les données sont utilisées pour élaborer des modèles prédictifs sophistiqués.

Les systèmes de recommandation - amélioration de l’expérience client :

Ils analysent les données des clients pour générer des recommandations personnalisées. En utilisant des algorithmes avancés, ces systèmes peuvent anticiper les besoins et les désirs des clients.

Prévisions boursières ou météorologiques :

Dans les marchés boursiers, les algorithmes de Machine Learning permet de prévoir des mouvements futurs des marchés.

Dans les prévisions météorologiques, le ML est utile pour établir des prévisions à différentes échéances, qu'il s'agisse de prévisions court terme ou à long terme.

Prévision du prix d'une maison :

La régression pour prédire le prix d'une maison est un exemple classique d'application de la régression.

Deep Learning :

Le Deep Learning est une sous-catégorie du Machine Learning qui repose sur l'utilisation de réseaux de neurones artificiels. Le DL est utilisé par exemple pour la reconnaissance visuelle, la reconnaissance vocale, la traduction automatique etc…

Chapitre 2 :

Les données jouent un rôle essentiel dans le monde connecté. Tous les algorithmes sont basés sur les statistiques. L'analyse statistique des données massives permet de dégager des tendances, de découvrir des relations cachées et de prédire des tendances futures. L'analyse statistique permet de transformer des données brutes en connaissances exploitables.

Etapes de l'analyse des données :

1) Collecte des données

La collecte des données consiste à rassembler toutes les données pertinentes.

Ces données peuvent être collectées de différentes manières :

- En provenance de base de données
- En provenance de réseaux sociaux
- En provenance de capteurs
- En provenance d'enquêtes
- En provenance d'appareils photo, de caméras
- En provenance d'un site internet

2) Nettoyage des données :

Les données brutes sont souvent incomplètes, incorrectes ou incohérentes. Le nettoyage des données implique la gestion des valeurs aberrantes, la correction des erreurs et la gestion des données manquantes.

3) Exploration des données :

A ce stade les données sont explorées pour comprendre leur structure. Des statistiques, des graphiques et des visualisation sont utilisés pour révéler des tendances et des corrélations.

4) Préparation des données :

Etape qui consiste à préparer les données pour l'analyse proprement dite. Il peut s'agir de normaliser ou formater les données, de réduire la dimensionnalité ou de créer de nouvelles variables plus pertinentes.

5) Choix des méthodes d'analyse :

En fonction de la nature des données et des objectifs de l'analyse, différentes méthodes peuvent être choisies : le monitoring, la régression, la classification, le clustering, l'analyse des séries temporelles.

6) Modélisation des données :

Construction de modèles statistiques ou d'apprentissage automatique pour extraire les informations utiles à partir des données. Exemples : modèles de Machine Learning, ajustement des courbes ou autres techniques d'analytiques.

7) Evaluation des modèles :

Etape cruciale : mesurer les performances du modèle et utiliser les métriques appropriées.

métrique ⇒ Elle représente le rapport entre le nombre de prédictions correctes et le nombre total d'échantillons d'entrée.

8) Interprétation - communication des résultats :

Les résultats sont interprétés pour tirer des conclusions qu'il faudra expliquer souvent à un public non technique.

9) Répéter et itérer :

Le processus est souvent itératif. Si de nouvelles données deviennent disponibles notamment, le processus peut être répété pour affiner les modèles et les conclusions.

10) Mise en oeuvre des recommandations :

Les résultats de l'analyse peuvent être à la base de prises de décision et de la mise en oeuvre d'actions concrètes.

Les différents types de données statistiques :

Les types de variables :

Il existe deux types de variables : les variables quantitatives et les variables qualitatives.

Les variables quantitatives :

Les variables quantitatives sont des nombres sur lesquels il est possible d'effectuer des opérations mathématiques. **Ces données peuvent être classées en deux groupes distincts :**

- **Les données quantitatives discrètes** : elles peuvent prendre un nombre fini de valeurs possibles, elles peuvent être énumèrées. (des notes entières, nombre de pattes d'un animal…)
- **Les données quantitatives continues** : ayant une infinité de valeurs possibles (temps de parcours, vitesse, taille, poids) **ATTENTION : Une variable discrète dont le nombre de valeurs différentes est élevé est assimilée à une variable quantitative continue.**

TIPS : pour savoir si c'est une quantitave discrète on doit se demander si le nombre de variété possible de valeurs n'est pas trop important et si une opération de type moyenne à un sens dans le contexte de l'étude statistique.

Les données qualitatives :

Les données qualitatives ont pour but de décrire des qualités propres aux données : est-ce un homme, une femme ? Est-ce un petit ou un grand ? **Ces données ne sont pas des nombres.**

Elles sont également réparties en deux groupes distincts :

- **Les données qualitatives ne pouvant pas être ordonnées :** Un homme, une femme, un chien ou chat…
- **Les données qualitatives pouvant être ordonnées :** petit,moyen,grand…

Mesure de tendance centrale :

Les mesures de tendance centrale servent à synthétiser la série statistique étudiée au moyen d'un petit nombre d'indicateurs. En d'autres termes c'est trouver un certain nombre de valeurs autour desquelles se regroupe l'ensemble des données.

Nombre d'observations :

La première chose à connaître lorsque nous étudions un ensemble d'observations c'est leur nombre. En connaissant le nombre d'observations attendues, il est possible de détecter plus rapidement si certaines données sont manquantes et si nous disposons d'assez de données pour permettre un apprentissage.

Moyenne arithmétique :

La moyenne est un indicateur qui permet de caractériser une série statistique.

La médiane :

La médiane représente la valeur au centre d'un ensemble de données lorsque celle-ci sont triées par ordre croissant. La médiane peut offrir une mesure de tendance centrale plus robuste que la moyenne qui peut être influencée par les valeurs extrêmes.

La médiane est particulièrement utile lorsque les données contiennent des valeurs aberrantes ou sont asymétriques car elle n'est pas autant affectée par ces situations que la moyenne.

Le mode :

Les mode est un autre indicateur de tendance centrale en statistique, quel que soit le type de variable y compris les qualitatives. Il représente la valeur qui apparait le plus fréquemment dans un ensemble de données. En d'autre termes, c'est la valeur qui a la fréquence la plus élevée. INFO : le mode n'est pas affecté par les valeurs extrêmes.

Mesure de dispersion :

La dispersion des données dans une série statistique se réfère à la manière dont les valeurs sont réparties autour d'une mesure de tendance centrale. Une mesure de dispersion permet de quantifier l'étandue des valeurs et à quel point les données sont dispersées ou regroupés.

L'étendue - valeur minimale - valeur maximale :

Connaître la valeur minimale et la valeur maximale d'un ensemble d'observation est toujours intéressant pour établir son étude.

Exemple :

Valeur min = 0 – valeur max = 10 donc l'étendue = 10

La variance - écart type - coefficient de variation.

La variance mesure la moyenne des carrés des écarts entre chaque valeur et la moyenne. L'écart-type est la racine carré de la variance. Plus la variance ou l'écart-type est élevé, plus les données sont dispersées.

Il est à noter que ces valeurs ne sont pas interprétables en elles-mêmes. Si l'ordre de grandeur des valeurs est de 1000, un écart-type de 2.63 est très petit ; si l'ordre de grandeur est de 2, cet écart-type est très grand. C'est la raison pour laquelle on le compare souvent à la moyenne ce qui donne le coefficient de variation

$$CV = (Ecarttype/moyenne) * 100$$

*100 pour avoir le coefficient de variation en pourcentage.

On considère en général que si ce coefficient est inférieur à 30%, la dispersion des valeurs autour de la moyenne est faible.

Graphique :

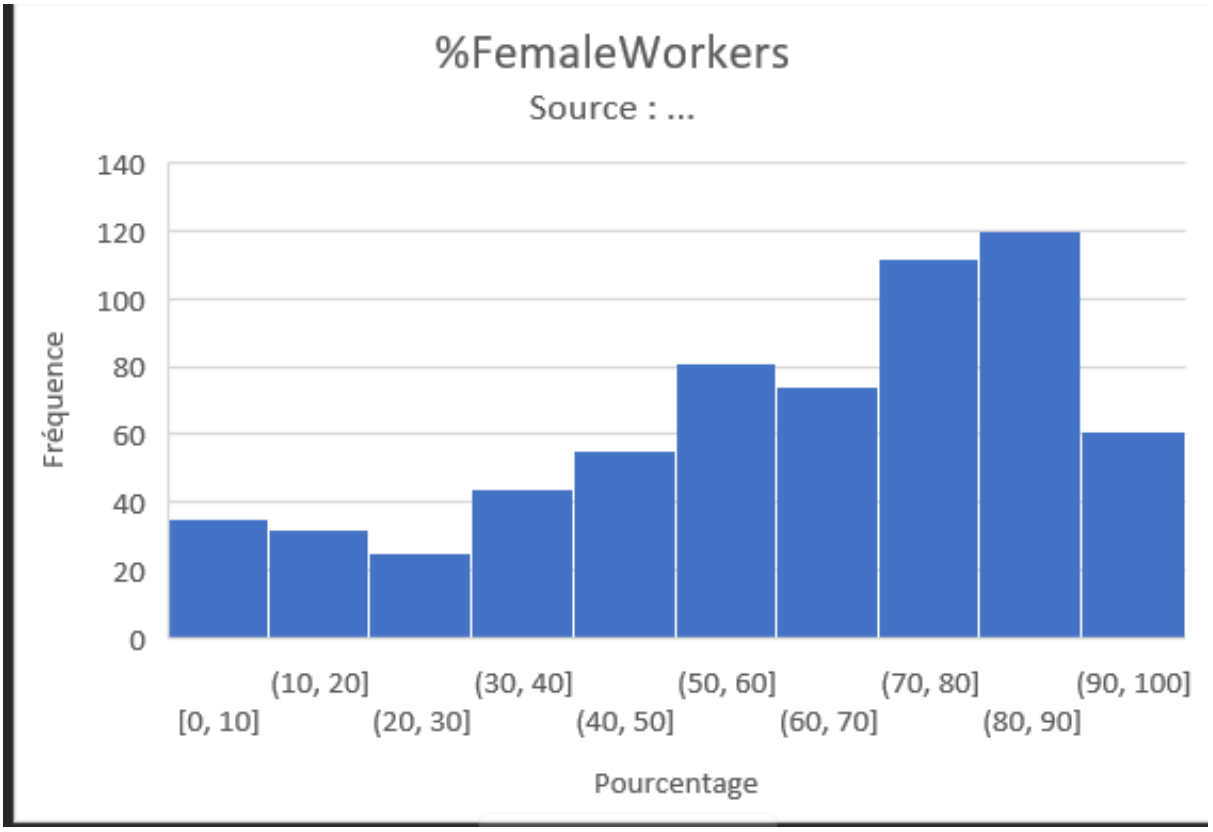
Un graphique doit toujours être accompagné :

- De la source des données + date de recensement
- Du titre de l’axe des abscisses et de l’unité éventuelle
- Du titre de l’axe des ordonnées et de l’unité éventuelle
- Du titre du graphique

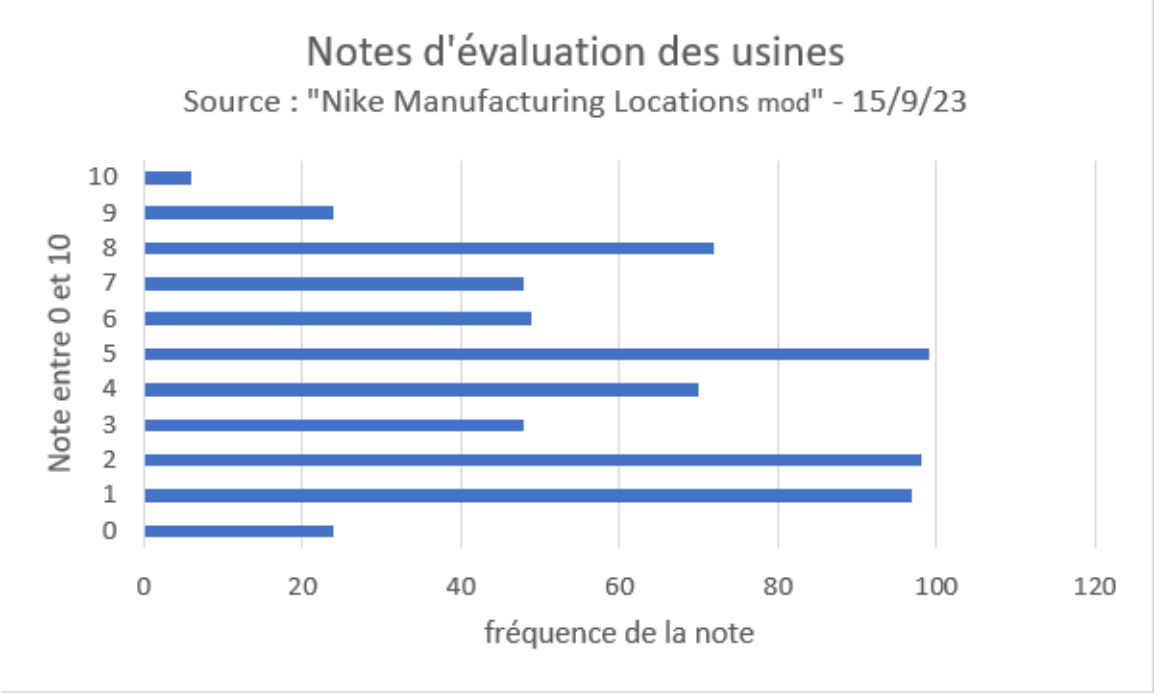
Parmi les graphique les plus couramment utilisés :

Variable quantitative continue - Histogramme :

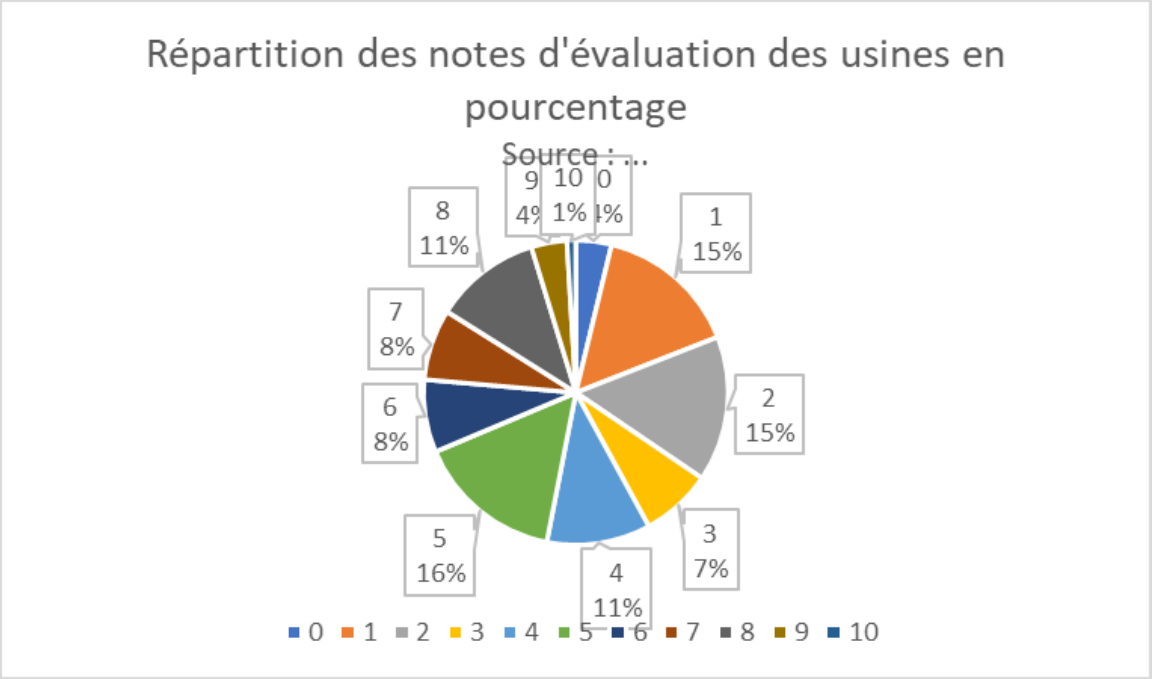
Graphique qui représente la distribution des données quantitative continues en regroupant les valeurs par intervalles sur l’axe horizontal et en affichang la fréquence ou le pourcentage sur l’axe vertical. Il permet de visualiser la forme de la distribution des données.



Variables quantitatives discrètes - Graphique en barres :



Variables quantitatives discrètes - Camembert :



Association de variables :

Corrélation linéaire - nuage de points :

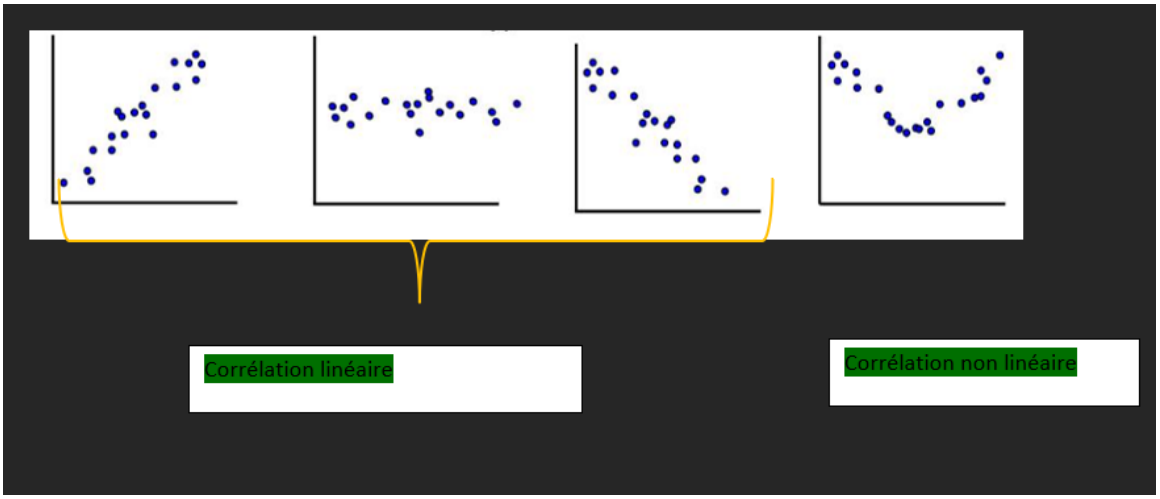


Tableau croisé dynamique :

Quelle est la moyenne des évaluations par région + graphique ?

Étiquettes de lignes	Moyenne de Eval
AMERICAS	3,90
EMEA	4,00
N ASIA	4,28
S ASIA	4,27
SE ASIA	4,37
Total général	**4,24**

Figure 5 : Vidéo TCD