Rapport du projet de Data Mining

Sous thème :

**Prédiction du turnover des employés**

Filière : Cycle d’ingénieur « Génie informatique »

Module : Système d´Information d´Aide à la Décision

**CONCEPTION ET REALISATION D’UNE APPLICATION DE GESTION** des courriers

**Préparé par : Encadré par :**

LAACHIR Meriem Pr. Sabiri Mohamed

2023 /2024

# Dédicace

Nous dédions ce travail :

A nos très chers parents, dont l’amour, la compréhension et le soutien ont toujours éclairé nos chemins.

A nos sœurs et frères pour leurs encouragements.

A tous nos chers amis.

A ceux qui se dévouent sans cesse pour nous éclairer la voie et les immenses horizons du savoir et dont la vocation mérite largement notre respect.

Et à vous chers lecteurs.

Et enfin à toute personne ayant contribué de près ou de loin à la réalisation de ce travail.

# Remerciements

En préambule à ce mémoire on remercie Dieu qui nous a aidés et nous a donné la patience et le courage durant ces longues années d’étude.

Notre chère professeur **Sabiri Mohamed** pour ses efforts d’encadrement, pour son temps consacré et surtout pour ses conseils judicieux lors de la période de préparation du projet.

Nos remerciements s’adressent également à l’ensemble du corps professoral du cycle d’ingénieur : Génie informatique, d’abord pour la qualité de leurs enseignements théoriques et professionnels et puis après pour leurs conseils et orientations durant toute notre période de formation.

Et pour terminer, on tient à remercier l’ensemble du personnel du corps administratif et tous ceux parmi eux qui nous ont soutenus et aidé de loin et de près durant notre cursus.

**Sommaire :**

Table des matières

[Dédicace 2](#_Toc155386335)

[Remerciements 3](#_Toc155386336)

[Introduction 6](#_Toc155386337)

[CHAPITRE 1 : CONTEXTE GENERAL DU PROJET 7](#_Toc155386338)

[1. Introduction 7](#_Toc155386339)

[2. Problématique et objectifs du projet 7](#_Toc155386340)

[3. La solution proposée 8](#_Toc155386341)

[Objectifs 8](#_Toc155386342)

[Méthodologie 9](#_Toc155386343)

[4. Conclusion 9](#_Toc155386344)

[CHAPITRE 2 : FONDEMENTS DE LA FOUILLE DE DONNÉES (DATA MINING) 10](#_Toc155386345)

[1. Introduction 10](#_Toc155386346)

[2. Définition de l'Informatique Décisionnelle 10](#_Toc155386347)

[3. Processus de la Fouille de Données 11](#_Toc155386348)

[3.1 Collecte de Données : 11](#_Toc155386349)

[3.2Prétraitement : 11](#_Toc155386350)

[3.3 Exploration des Données : 11](#_Toc155386351)

[3.4 Modélisation : 11](#_Toc155386352)

[3.5 Évaluation : 11](#_Toc155386353)

[3.6 Interprétation : 11](#_Toc155386354)

[4. Techniques de Data Mining pour la Gestion du Personnel 11](#_Toc155386355)

[4.1 Exploration des Techniques 12](#_Toc155386356)

[4.2 Alignement avec les Objectifs du Projet 12](#_Toc155386357)

[4.3 Application Pratique 12](#_Toc155386358)

[5. Les Principes fondamentaux 12](#_Toc155386359)

[6. Outils et Technologies Utilisés dans le Projet : 13](#_Toc155386360)

[6.1 Environnement Anaconda: 13](#_Toc155386361)

[6.2 Jupyter: 14](#_Toc155386362)

[6.3 Weka: 14](#_Toc155386363)

[6.4 Gestion des Fichiers .csv : 15](#_Toc155386364)

[6.5 Algorithmes et Techniques de Data Mining : 15](#_Toc155386365)

[7. Conclusion 16](#_Toc155386366)

[CHAPITRE 3 : MISE EN ŒUVRE DE LA BUSINESS INTELLIGENCE 16](#_Toc155386367)

[1. Introduction 16](#_Toc155386368)

[2. Les outils mise en œuvre 17](#_Toc155386369)

[3. Processus ETL (Extract, Transform, Load) : 18](#_Toc155386370)

[3.1 Talend : 18](#_Toc155386371)

[3.2 SQL Server Integration Services (SSIS) : 21](#_Toc155386372)

[4. Conclusion 23](#_Toc155386373)

[CHAPITRE 4 : ANALYSE ET INTERPRÉTATION DES RÉSULTATS 24](#_Toc155386374)

[1. Introduction : 24](#_Toc155386375)

[2. Outils et Technologies Employés : Machine Learning (ML) : 24](#_Toc155386376)

[3. Analyses, Rapports et Visualisations : 24](#_Toc155386377)

[4. Résultats des Algorithmes : 26](#_Toc155386378)

[5. Conclusion : 28](#_Toc155386379)

[Conclusion 29](#_Toc155386380)

# Introduction

Dans le cadre du module "Fouille de Données Data Mining", je me lance seul dans un projet captivant qui constitue une étape cruciale de ma troisième année en cycle d'ingénieur en Génie Informatique. Ce projet s'inscrit dans le domaine de Data Mining, et son objectif est de réaliser une analyse approfondie du turnover des employés au sein d'une organisation.

Choisi de manière délibérée par moi-même, ce projet représente une opportunité stimulante pour mettre en œuvre mes compétences techniques et approfondir mes connaissances en matière de prise de décision basée sur l'information. L'objectif principal est d'appliquer les techniques de fouille de données pour analyser les données liées au turnover des employés, permettant ainsi aux entreprises d'anticiper les départs potentiels de leur personnel.

Au fil de ce rapport, je vais présenter de manière détaillée le contexte spécifique de ce projet, son importance stratégique dans le domaine du Data Mining appliqué à la gestion des ressources humaines. Nous allons explorer les différentes phases du projet, depuis l'analyse des besoins jusqu'à la mise en œuvre d'une solution de prédiction du turnover des employés. Ce projet offre une opportunité unique de concilier théorie et pratique, en déployant des outils avancés pour comprendre les tendances du départ des employés.

La rétention des employés talentueux est cruciale pour assurer la stabilité et la croissance d'une organisation. Comprendre les raisons du départ des employés est essentiel, et c'est dans cette optique que ce projet s'attache à exploiter les capacités de Data Mining pour offrir des analyses approfondies et des prévisions pertinentes en matière de turnover des employés.

# CHAPITRE 1 : CONTEXTE GENERAL DU PROJET

Le projet que je mène se focalise sur la prédiction du turnover des employés en exploitant les techniques de fouille de données (Data Mining). Avant d'approfondir les détails, il est essentiel de définir le concept de prédiction du turnover dans le contexte spécifique de ce projet.

## Introduction

Dans cette ère moderne de l'analyse des données, les entreprises sont confrontées à la nécessité croissante de tirer des enseignements significatifs de leurs vastes ensembles de données. Mon projet, axé sur la prédiction du turnover des employés à l'aide de techniques de fouille de données, s'inscrit dans cette perspective dynamique de l'informatique décisionnelle.

## Problématique et objectifs du projet

La problématique centrale de mon projet réside dans la complexité de prédire le turnover des employés au sein d'une organisation. Ce défi crucial vise à garantir une gestion efficace des ressources humaines, mais il est confronté à une diversité de facteurs influençant le départ des employés.

En premier lieu, les motifs de départ sont multiples, allant de la satisfaction au travail à la durée passée dans l'entreprise, en passant par les accidents du travail, les promotions récentes et le salaire. Prévoir avec précision nécessite une analyse approfondie de ces variables.

Par ailleurs, la dimension temporelle ajoute une complexité supplémentaire. Les tendances de départ peuvent varier en fonction des saisons, des cycles économiques et d'autres événements temporels. Anticiper ces variations est crucial pour des stratégies de rétention efficaces.

De plus, la diversité des profils au sein de l'organisation complique davantage la prédiction. Chaque employé est unique, avec des caractéristiques personnelles, professionnelles et des aspirations spécifiques. Les prévisions de départ doivent donc prendre en compte cette diversité.

En raison de ces multiples facteurs et de la nécessité de comprendre les motifs individuels de départ, la variabilité des prévisions entre différents départements ou équipes est inévitable. Cela souligne l'importance d'une solution globale qui peut tirer parti de l'ensemble des données historiques pour fournir des prévisions précises et cohérentes sur l'ensemble de l'organisation.

En résumé, la complexité de la problématique réside dans la conciliation de divers facteurs, de la satisfaction au travail à la dimension temporelle, pour parvenir à des prévisions de départ fiables et uniformes. Ceci contribuera à une gestion optimale du capital humain au sein de l'organisation.

Cette complexité est d'autant plus cruciale dans un contexte où la rotation du personnel représente un défi majeur. Lorsqu'un employé quitte l'entreprise, cela entraîne une perte de productivité et des coûts significatifs de remplacement. Les entreprises cherchent ainsi à anticiper ces départs pour mettre en place des mesures préventives. Mon projet propose une approche d'apprentissage automatique, pour exploiter les données sur la main-d'œuvre et anticiper le turnover du personnel avant qu'il ne se produise. Cette approche s'inscrit dans la tendance actuelle où le "big data" devient un outil précieux pour comprendre et anticiper les dynamiques au sein des organisations.

## La solution proposée

### Objectifs

Le projet a pour objectif de mettre en place une solution innovante basée sur l'apprentissage automatique pour anticiper avec précision les départs des employés au sein de l'organisation. Les principales aspirations de ce projet sont les suivantes :

1. **Création d'un Modèle Prédictif :** Le développement d'un modèle prédictif robuste est au cœur de notre solution. Ce modèle utilisera des techniques avancées d'apprentissage automatique pour analyser les données historiques et générer des prédictions précises sur les départs potentiels des employés.
2. **Identification des Facteurs Clés :** Une partie essentielle de la solution consiste à identifier les facteurs déterminants du taux de désabonnement des employés. À travers une analyse approfondie, nous chercherons à comprendre quels éléments influent le plus sur la décision des employés de quitter leur poste.

### Méthodologie

1. **Collecte et Prétraitement des Données :** La première étape consistera à recueillir et à prétraiter les données liées aux employés. Cela inclura des informations telles que la satisfaction au travail, la durée passée dans l'entreprise, les promotions récentes, les accidents du travail, et d'autres variables pertinentes.
2. **Développement du Modèle :** En utilisant des techniques d'apprentissage automatique telles que la régression logistique, les machines à vecteurs de support (SVM), ou d'autres algorithmes adaptés à la nature prédictive du problème, nous créerons un modèle capable de prédire le départ des employés.
3. **Validation du Modèle :** Le modèle sera rigoureusement testé et validé sur des ensembles de données distincts pour s'assurer de sa précision et de sa généralisation.
4. **Analyse des Facteurs d'Attrition :** En parallèle, une analyse approfondie des facteurs d'attrition sera effectuée. Cela impliquera l'utilisation de techniques statistiques et d'exploration de données pour identifier les variables clés qui influent sur le taux de désabonnement.
5. **Intégration et Déploiement :** Une fois le modèle et l'analyse des facteurs prêts, ils seront intégrés dans un système global. Le déploiement permettra aux gestionnaires et aux responsables des ressources humaines d'accéder à des informations précieuses pour prendre des décisions informées.

Cette approche orientée vers les résultats vise à fournir à l'organisation une solution proactive pour la gestion du personnel, minimisant ainsi l'impact négatif des départs et optimisant la rétention des employés talentueux.

## Conclusion

En conclusion du chapitre introductif, nous avons posé les bases essentielles pour comprendre la nature et la portée de notre projet de prédiction du turnover des employés. L'introduction a fourni un aperçu global du contexte dans lequel s'inscrit cette initiative, tandis que la problématique a éclairé les défis complexes auxquels les organisations sont confrontées en matière de gestion des ressources humaines.

Les objectifs clairement définis de notre projet visent à créer un modèle prédictif précis pour anticiper les départs des employés et à identifier les facteurs déterminants liés au taux de désabonnement. Ces objectifs guideront notre parcours tout au long du projet, fournissant une feuille de route claire pour les étapes à venir.

La solution proposée, bien que brièvement abordée dans ce chapitre, constitue le cœur même de notre démarche. Elle représente une approche novatrice et technologiquement avancée pour aborder les enjeux cruciaux de la rétention des employés. Ce projet offre ainsi une opportunité unique de repenser les pratiques actuelles en matière de gestion des ressources humaines.

En poursuivant notre exploration, les chapitres suivants détailleront les fondements théoriques, la méthodologie adoptée, et les résultats obtenus. Chaque étape du projet sera minutieusement examinée pour garantir la compréhension complète du processus et la pertinence des conclusions.

Ainsi, le chapitre suivant plongera dans la théorie de l'informatique décisionnelle appliquée à la problématique du turnover des employés, jetant ainsi les bases conceptuelles nécessaires pour notre approche analytique.

# CHAPITRE 2 : FONDEMENTS DE LA FOUILLE DE DONNÉES (DATA MINING)

## Introduction

Ce deuxième chapitre se consacre aux bases de la fouille de données (Data Mining) en relation avec notre projet de prédiction du turnover des employés. Cette introduction jettera les bases en soulignant l'importance de la fouille de données dans le contexte de notre analyse prédictive.

## Définition de l'Informatique Décisionnelle

La fouille de données, également connue sous le terme de Data Mining, est une discipline informatique qui vise à découvrir des modèles significatifs, des structures et des tendances cachées au sein de grands ensembles de données. Cette approche s'appuie sur des techniques statistiques, mathématiques et informatiques avancées pour extraire des informations exploitables à partir de données brutes.

## Processus de la Fouille de Données

Le processus de fouille de données comprend plusieurs étapes cruciales qui guident le chercheur à travers l'analyse des données brutes pour en extraire des informations significatives. Voici les principales phases de ce processus :

### Collecte de Données :

* Rassembler des ensembles de données pertinents en fonction des objectifs de la fouille.
* Assurer la qualité et la fiabilité des données collectées.

### 3.2Prétraitement :

* Nettoyer les données en supprimant les valeurs manquantes, les doublons et les erreurs.
* Transformer les données pour les rendre compatibles avec les exigences du modèle.
* Normaliser les données pour éliminer les biais liés à l'échelle ou à la mesure.

### 3.3 Exploration des Données :

* Effectuer une analyse exploratoire pour comprendre la distribution des données.
* Identifier des tendances, des schémas ou des anomalies potentielles.
* Visualiser les données pour obtenir des perspectives intuitives.

### 3.4 Modélisation :

* Choisir et appliquer des algorithmes de fouille de données appropriés.
* Former le modèle en utilisant des ensembles d'entraînement.
* Ajuster les paramètres du modèle pour améliorer la performance.

### 3.5 Évaluation :

* Évaluer la qualité du modèle en utilisant des ensembles de test indépendants.
* Mesurer la précision, la sensibilité et la spécificité du modèle.
* Identifier et corriger les éventuels problèmes de surajustement ou de sous-ajustement.

### 3.6 Interprétation :

* Interpréter les résultats de la fouille de données en fonction des objectifs définis.
* Extraire des informations significatives et pertinentes.
* Formuler des recommandations ou des décisions basées sur les découvertes.

Le processus de fouille de données est itératif, et des ajustements peuvent être apportés à chaque étape en fonction des résultats obtenus. Il vise à transformer des données brutes en connaissances exploitables, fournissant ainsi des informations cruciales pour la prise de décision.

## Techniques de Data Mining pour la Gestion du Personnel

Dans cette section, nous nous pencherons sur les techniques de Data Mining spécifiques qui peuvent être appliquées à notre projet axé sur la gestion du personnel. Ces techniques jouent un rôle crucial dans l'analyse des données liées au roulement des employés, contribuant ainsi à l'atteinte de nos objectifs.

### **4.1 Exploration des Techniques**

Dans cette première sous-section, nous nous concentrerons sur l'exploration des techniques fondamentales de Data Mining applicables à notre projet de gestion du personnel.

a. Classification : Nous examinerons comment la classification, en attribuant des étiquettes prédéfinies en fonction de caractéristiques spécifiques, peut être un outil puissant pour classifier les employés en groupes tels que "à faible risque de départ" ou "à haut risque de départ".

b. Régression : Nous analyserons la régression en tant que technique visant à établir une relation mathématique entre des variables indépendantes et dépendantes. Dans notre contexte, la régression pourrait prédire le moment probable du départ d'un employé en fonction de facteurs tels que la satisfaction au travail, la rémunération, etc.

c. Clustering : La sous-section abordera le clustering, qui regroupe les données similaires en ensembles distincts. Dans la gestion du personnel, cela pourrait être appliqué pour identifier des groupes d'employés partageant des caractéristiques communes, permettant une personnalisation des approches de rétention.

### **4.2 Alignement avec les Objectifs du Projet**

Nous établirons un lien explicite entre les techniques explorées et les objectifs spécifiques de notre projet de gestion du personnel. Cette alignement vise à garantir que les techniques choisies contribuent de manière significative à l'identification des facteurs clés du taux de désabonnement des employés.

### **4.3 Application Pratique**

Cette sous-section illustrera la mise en pratique des techniques explorées par le biais d'exemples concrets liés à notre projet. Les exemples souligneront la pertinence de ces techniques et leur impact potentiel sur les résultats, offrant ainsi une vision concrète de leur application dans le contexte de la gestion du personnel.

## Les Principes fondamentaux

Cette partie centrale du chapitre approfondira les principes fondamentaux de la fouille de données. Nous examinerons en détail les principaux aspects liés aux algorithmes de classification, de régression, de regroupement et d'association.

Nous aborderons les concepts clés, tels que :

* **Algorithmes de Classification :** Nous détaillerons les algorithmes utilisés pour classer les données en catégories distinctes, offrant ainsi une base solide pour la compréhension de la classification dans le contexte de la gestion du personnel.
* **Algorithmes de Régression :** L'accent sera mis sur les algorithmes qui établissent des relations mathématiques entre des variables indépendantes et dépendantes, essentielles pour prédire le moment probable du départ d'un employé.
* **Algorithmes de Clustering :** Nous explorerons les algorithmes impliqués dans le regroupement des données similaires en ensembles distincts, facilitant ainsi l'identification de groupes d'employés partageant des caractéristiques communes.
* **Algorithmes d'Association :** La sous-section se penchera sur les algorithmes qui identifient les relations et les associations entre différentes variables, fournissant des informations précieuses pour la gestion du personnel.

En fournissant une compréhension approfondie de ces principes fondamentaux, cette partie jettera les bases nécessaires pour la mise en œuvre réussie des techniques de fouille de données dans notre projet de gestion du personnel.

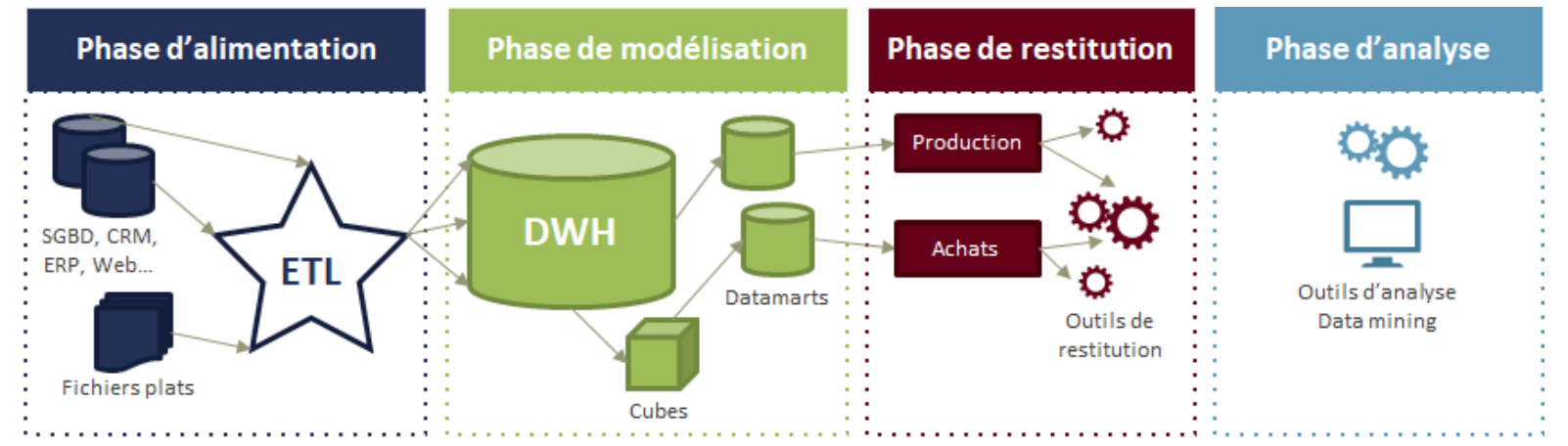


Figure 1 : La chaîne décisionnelle

## Outils et Technologies Utilisés dans le Projet :

La réalisation du projet d'informatique décisionnelle s'appuie sur une sélection stratégique d'outils et de technologies adaptés aux besoins spécifiques de collecte, de traitement, d'analyse et de présentation des données. Voici une présentation des principaux outils et technologies utilisés dans ce contexte particulier.

### ****Environnement Anaconda:****



Dans le cadre de notre projet de prédiction du turnover des employés, l'adoption d'Anaconda en tant qu'environnement de développement revêt une importance cruciale. Anaconda offre une suite complète d'outils dédiés à la science des données et à la fouille de données, ce qui en fait un choix optimal pour répondre à nos besoins spécifiques.

### ****Jupyter:****



Jupyter est une plateforme open-source qui permet la création et le partage de documents interactifs appelés "notebooks". Ces notebooks peuvent contenir du code, des équations, des visualisations et du texte narratif, offrant ainsi un environnement complet pour la programmation, l'analyse de données et la création de rapports.

Dans le contexte de notre projet de prédiction du turnover des employés, Jupyter constitue un outil essentiel pour l'exploration des données, la modélisation et la présentation des résultats de manière interactive et accessible.

### ****Weka****:



Weka est un logiciel open-source d'apprentissage automatique (machine learning) et de fouille de données. Le nom "Weka" est dérivé de l'oiseau kiwi de Nouvelle-Zélande, qui est connu pour être curieux et apte à apprendre de nouvelles choses, symbolisant ainsi l'approche d'apprentissage automatique du logiciel.

Principales caractéristiques de Weka :

1. **Large Bibliothèque d'Algorithmes :** Weka propose une vaste bibliothèque d'algorithmes d'apprentissage automatique pour la classification, la régression, le clustering, l'association, et plus encore. Cela offre aux utilisateurs une flexibilité dans le choix des méthodes appropriées pour leurs tâches spécifiques.
2. **Interface Graphique Intuitive :** Weka est doté d'une interface graphique conviviale qui facilite l'exploration des données, la configuration des modèles et l'évaluation des performances. Cela en fait un choix populaire, en particulier pour les utilisateurs débutants en apprentissage automatique.
3. **Fonctionnalités de Prétraitement :** Weka propose divers outils de prétraitement des données, notamment la normalisation, la transformation, la sélection de caractéristiques, etc. Ces fonctionnalités sont cruciales pour préparer les données avant de les soumettre aux algorithmes d'apprentissage automatique.
4. **Intégration avec Java :** Weka est développé en Java et offre une intégration étroite avec ce langage de programmation. Cela permet aux utilisateurs d'exploiter la puissance de Java tout en travaillant avec Weka.
5. **Outils de Visualisation :** Weka propose des outils de visualisation pour aider les utilisateurs à comprendre les données, à explorer les résultats des modèles, et à interpréter les performances des algorithmes.

Dans le contexte de notre projet de prédiction du turnover des employés, Weka sera utilisé pour appliquer des algorithmes d'apprentissage automatique aux données, explorer les modèles générés, et évaluer la performance des prédictions. Son interface graphique conviviale le rend adapté à l'expérimentation et à l'analyse exploratoire des données.

### Gestion des Fichiers .csv :

Les fichiers au format CSV (Comma-Separated Values) jouent un rôle essentiel dans notre projet de prédiction du turnover des employés. Ces fichiers sont des fichiers texte où les données sont organisées sous forme de table, avec chaque ligne représentant une entrée distincte et les valeurs séparées par des virgules.

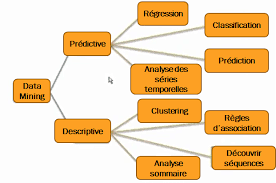


Utilisation dans le Projet :

Dans le cadre de notre projet, les fichiers CSV seront utilisés pour stocker et structurer les données relatives aux employés, y compris les variables pertinentes telles que la satisfaction au travail, la durée passée dans l'entreprise, les promotions récentes, etc. Ces fichiers CSV serviront de source de données pour l'application des techniques de Data Mining et d'apprentissage automatique.

La gestion efficace des fichiers CSV implique la lecture, la modification, et la sauvegarde de ces fichiers. Des outils adaptés seront utilisés pour effectuer ces opérations, assurant ainsi la qualité et l'intégrité des données nécessaires à notre analyse prédictive du turnover des employés.

### Algorithmes et Techniques de Data Mining :

Les algorithmes et techniques de Data Mining occupent une place centrale dans notre projet de prédiction du turnover des employés. Ces méthodes analytiques avancées nous permettront d'extraire des modèles significatifs à partir des données existantes, facilitant ainsi la prévision des départs potentiels des employés. Voici une présentation des principales techniques que nous allons exploiter :**Classification, Régression, Clustering, Analyse des Associations ..**

## Conclusion

En conclusion de ce chapitre, nous avons établi les fondements essentiels de la fouille de données, définissant son rôle crucial dans le domaine de la gestion du personnel. Nous avons exploré le processus de la fouille de données, mettant en lumière ses différentes étapes, de la collecte à l'interprétation des résultats. Ces étapes, telles que la collecte, le prétraitement, l'exploration, la modélisation, l'évaluation et l'interprétation, constituent le socle méthodologique sur lequel reposera notre projet de prédiction du turnover des employés.

De plus, nous avons approfondi notre compréhension des techniques spécifiques de Data Mining applicables à notre projet, soulignant leur alignement avec nos objectifs de gestion du personnel. L'exploration des techniques, leur alignement avec nos objectifs, ainsi que des exemples concrets d'application pratique, ont été présentés pour illustrer leur pertinence et leur impact potentiel.

Enfin, nous avons présenté les principes fondamentaux de la fouille de données, fournissant ainsi une base solide pour la suite de notre projet. L'introduction aux outils et technologies, tels qu'Anaconda, Jupyter, Weka, la gestion des fichiers .csv, et les algorithmes de Data Mining, nous offre un aperçu des ressources à notre disposition.

Ce chapitre pose ainsi les bases nécessaires pour aborder la mise en œuvre de la Business Intelligence dans le chapitre suivant, contribuant ainsi à la réalisation de notre projet ambitieux de prédiction du turnover des employés.

# CHAPITRE 3 : MISE EN ŒUVRE DE LA DATA MINING

## Introduction

Dans cette section, nous amorçons le chapitre en mettant en lumière l'importance cruciale de la mise en œuvre de la Data Mining dans notre projet. Nous présenterons brièvement les objectifs spécifiques de cette phase, mettant en avant le rôle central de la collecte, du traitement et de la visualisation des données dans le processus de prédiction du turnover des employés.

## Collecte de données :

Dans cette étape fondamentale de notre mise en œuvre, nous débutons par l'exploration détaillée des différentes sources de données pertinentes pour notre projet de prédiction du turnover des employés. Nous nous assurons de définir des méthodes rigoureuses visant à garantir la qualité et la pertinence des données collectées, mettant un accent particulier sur la diversité des variables essentielles liées au roulement du personnel.

*Sources de données :*

Pour ce projet, nous avons opté pour l'utilisation de Kaggle, une plateforme renommée fournissant des ensembles de données de haute qualité pour le développement et l'entraînement de modèles d'apprentissage automatique. L'ensemble de données sélectionné contient 14 999 lignes et 10 colonnes, chaque ligne représentant un employé et chaque colonne une caractéristique spécifique.

*Variables clés du dataset :*

1. **Niveau de Satisfaction (niveau\_satisfaction) :** Variable binaire indiquant le niveau de satisfaction de l'employé (0 ou 1).
2. **Dernière Évaluation (derniere\_evaluation) :** Temps écoulé depuis la dernière évaluation de performance, mesuré en années.
3. **Nombre de Projets (nombre\_projets) :** Nombre de projets auxquels l'employé a participé.
4. **Heures Mensuelles Moyennes (heures\_mensuelles\_moyennes) :** Nombre moyen d'heures mensuelles consacrées au travail.
5. **Durée Passée dans l'Entreprise (duree\_passee\_entreprise) :** Nombre d'années passées par l'employé dans l'entreprise.
6. **Accident du Travail (accident\_du\_travail) :** Indique si l'employé a été impliqué dans un accident du travail.
7. **Départ (quitte) :** Variable cible, indique si l'employé a quitté ou non son lieu de travail (0 pour non, 1 pour oui).
8. **Promotion au Cours des 5 Dernières Années (promotion\_dernier\_5ans) :** Indique si l'employé a été promu au cours des cinq dernières années.
9. **Département (departement) :** Catégorie du département dans lequel travaille l'employé.
10. **Salaire (salaire) :** Niveau relatif de salaire, catégorisé en bas, moyen ou élevé.

Cette phase de collecte assure une base solide pour la construction et l'entraînement de notre modèle prédictif, avec une attention particulière portée à la représentativité et à la diversité des données.

## **Traitement** de données:

### Discretisation de données

Un aspect crucial du traitement de données est la discrétisation, une technique visant à convertir des variables continues en catégories discrètes. Dans notre projet, nous avons appliqué la discrétisation aux variables niveau\_satisfaction et derniere\_evaluation pour simplifier le modèle prédictif. Voici les étapes de discrétisation effectuées :

* **niveau\_satisfaction :**
  + Les valeurs inférieures ou égales à 0.25 ont été catégorisées comme 0.
  + Les valeurs comprises entre 0.25 (exclus) et 0.5 ont été catégorisées comme 1.
  + Les valeurs comprises entre 0.5 (exclus) et 0.75 ont été catégorisées comme 0.
  + Les valeurs supérieures à 0.75 ont été catégorisées comme 1.

Le résultat a été converti en type entier.

* **derniere\_evaluation :**
  + Les valeurs inférieures ou égales à 0.56 ont été catégorisées comme 0.
  + Les valeurs comprises entre 0.56 (exclus) et 0.80 ont été catégorisées comme 1.
  + Les valeurs supérieures à 0.80 ont été catégorisées comme 0.

Le résultat a été converti en type entier.

Ces étapes de discrétisation simplifient la complexité des données continues, permettant ainsi une meilleure interprétation et intégration dans nos modèles de prédiction du turnover des employés.

### Encodage de données

Dans cette phase du traitement des données, nous nous concentrons sur l'encodage des variables catégoriques, en particulier les colonnes 'salaire' et 'departement'. Ces deux variables doivent être converties en valeurs quantitatives pour être utilisées efficacement par les algorithmes d'apprentissage automatique.

* **Encodage de la colonne 'salaire' :**
  + Les valeurs 'high', 'medium', et 'low' de la colonne 'salaire' ont été transformées en valeurs numériques.
  + 'high' a été encodé comme 2, 'medium' comme 1, et 'low' comme 0.
  + Le résultat a été converti en type entier.
* **Encodage de la colonne 'departement' :**
  + Les différents départements ont été mappés à des valeurs numériques distinctes.
  + Chaque département a reçu une valeur numérique de 0 à 9.
  + Le résultat a été converti en type flottant.

Ces opérations d'encodage permettent de représenter de manière quantitative les informations catégoriques, facilitant ainsi l'utilisation de ces variables dans nos modèles prédictifs.

### Réduction de la dimensionnalité

Au cours de cette étape, nous identifions et traitons les corrélations importantes entre les variables, en particulier entre les colonnes 'nombre\_projets' et 'heures\_mensuelles\_moyennes'. La forte corrélation entre ces deux variables offre une opportunité de réduire la dimensionnalité et d'optimiser nos données.

* **Création d'une nouvelle fonctionnalité :**
  + Une nouvelle fonctionnalité, appelée 'proj\*hour', a été créée en multipliant les valeurs de 'nombre\_projets' par 'heures\_mensuelles\_moyennes'.
  + Cette nouvelle fonctionnalité capture la relation combinée entre le nombre de projets et le nombre d'heures mensuelles moyennes.
* **Suppression des colonnes redondantes :**
  + Les colonnes 'nombre\_projets' et 'heures\_mensuelles\_moyennes' ont été supprimées pour éliminer les données redondantes de notre modèle.

Cette approche de réduction de la dimensionnalité contribue à simplifier notre ensemble de données tout en préservant l'information importante liée à ces variables fortement corrélées.

### Discretisation de données

L'étape de nettoyage des données est essentielle pour assurer la qualité et la fiabilité de notre ensemble de données. Dans le cadre de cette phase :

* **Vérification des données manquantes :**
  + Aucune valeur manquante n'a été détectée dans notre ensemble de données, garantissant ainsi l'intégrité des informations.
* **Suppression des lignes en double :**
  + Les lignes en double ont été identifiées et supprimées, améliorant ainsi la qualité et la précision de nos analyses.

Ces mesures visent à garantir que notre ensemble de données est exempt de données manquantes et de doublons, établissant ainsi une base solide pour la suite de notre analyse et de notre modélisation.

### Discretisation de données

## Conclusion

En résumé, le choix de combiner l'utilisation de SQL Server Integration Services (SSIS) avec d'autres outils tels que Talend a été délibéré et fructueux dans notre quête pour établir des solutions ETL robustes. Cette approche diversifiée a permis une gestion efficiente des données au sein de notre environnement SQL Server, offrant ainsi une flexibilité et une adaptabilité nécessaires à la complexité de notre projet d'analyse des ventes chez Rossmann.

L'utilisation conjointe de SSIS et de Talend a apporté des avantages significatifs, en exploitant les forces spécifiques de chaque outil. Talend a été privilégié pour sa facilité d'utilisation et sa polyvalence dans la préparation initiale des données, tandis que SSIS a brillé dans la mise en œuvre des tâches ETL plus avancées au sein de notre infrastructure SQL Server.

Cette approche intégrée a assuré une intégration homogène des données, garantissant la qualité et la cohérence nécessaires pour l'analyse approfondie des ventes chez Rossmann. En exploitant au mieux ces outils complémentaires, notre équipe a pu construire des pipelines ETL performants, prêts à répondre aux défis changeants du traitement des données dans le domaine dynamique de l'analyse commerciale.

# CHAPITRE 4 : ANALYSE ET INTERPRÉTATION DES RÉSULTATS

## Introduction :

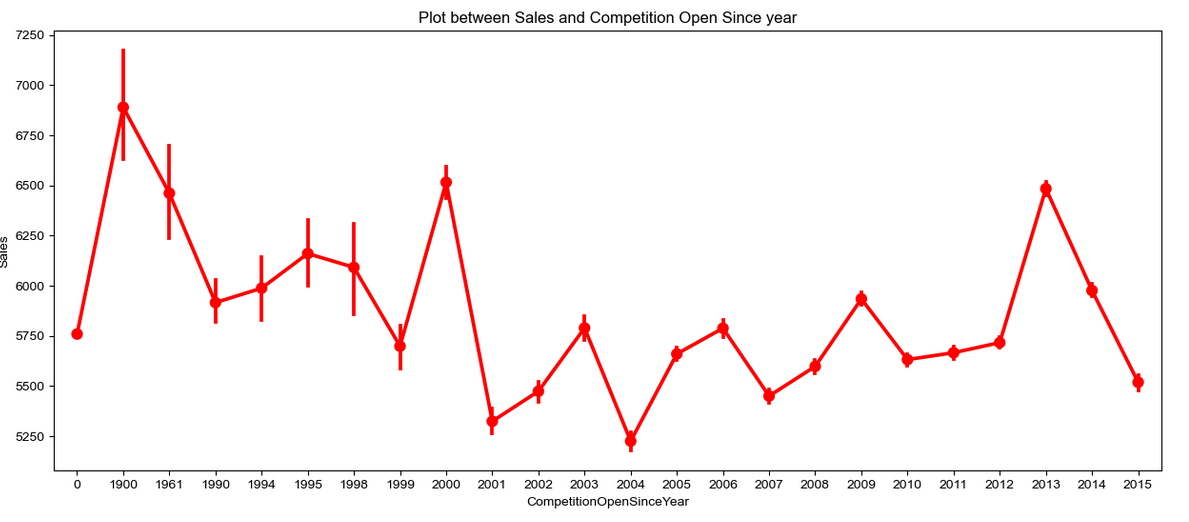
Dans cette section, nous explorerons les résultats de nos analyses en utilisant des algorithmes de Machine Learning, tels que DecisionTreeRegressor, RandomForestRegressor et AdaBoostRegressor. L'objectif est de comprendre comment ces modèles ont contribué à l'analyse des ventes de Rossmann et quels insights ils ont générés.

## Outils et Technologies Employés : Machine Learning (ML) :

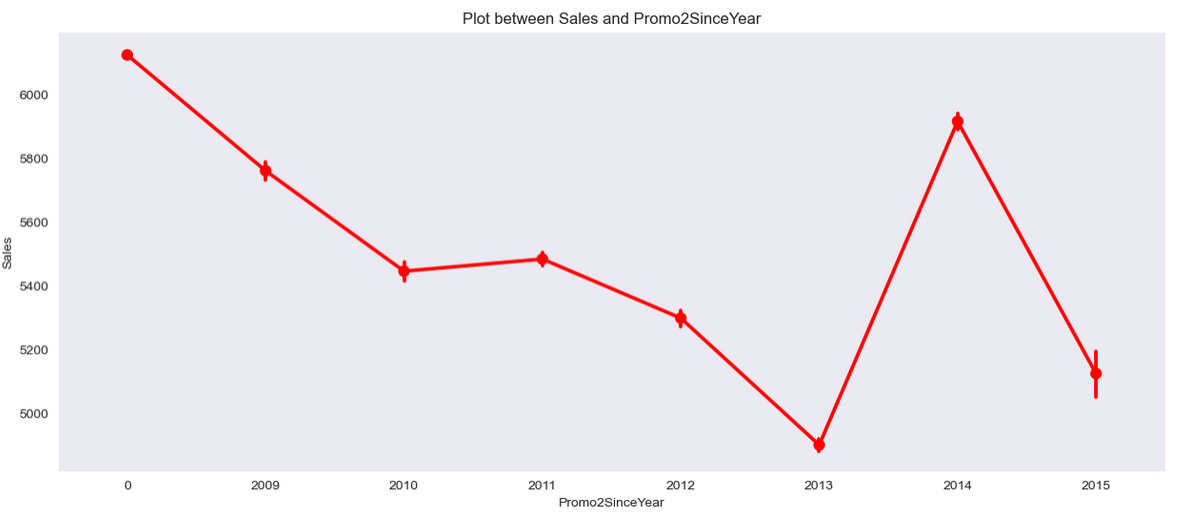
Nous introduirons l'intégration du Machine Learning (ML) dans notre projet d'analyse des ventes. Cette section expliquera en détail comment les outils tels que Talend et SSIS ont été employés pour intégrer des modèles de Machine Learning dans notre processus d'analyse. Nous mettrons en avant les choix stratégiques liés à l'utilisation du ML et son impact sur la qualité des résultats.

## Analyses, Rapports et Visualisations :

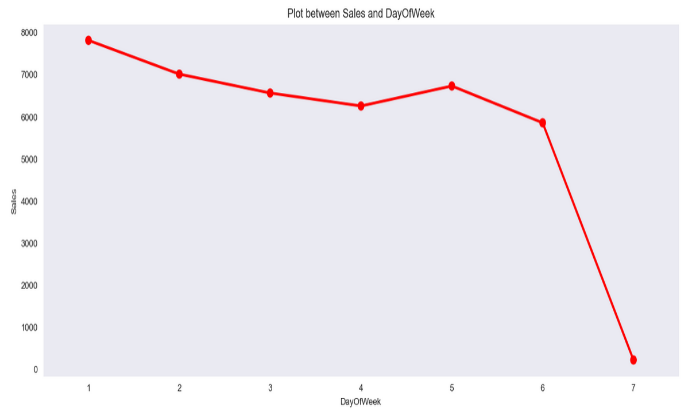
Nous présenterons les rapports les plus significatifs générés à partir des données traitées. L'accent sera mis sur les insights clés découverts, les tendances identifiées, et les métriques importantes pour la compréhension approfondie des performances de Rossmann.



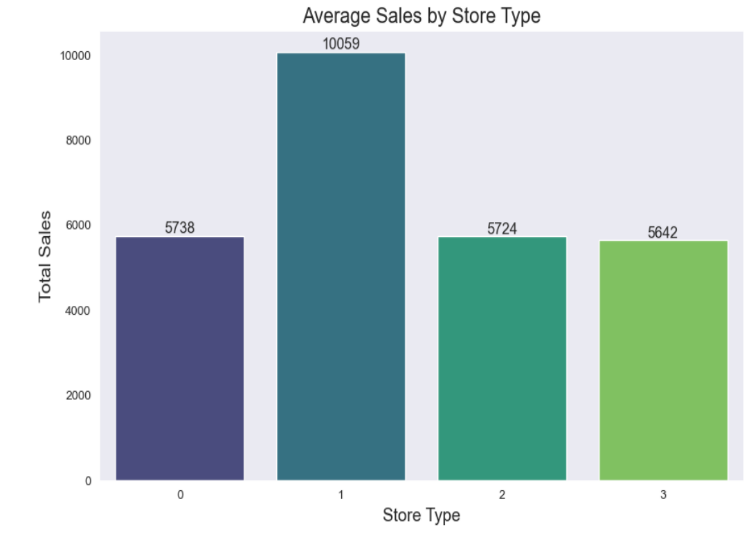
Ce graphique nous montre que les ventes ont été les plus élevées au cours de l'année 1900, car à cette époque le nombre de magasins était limité, la concurrence est donc très faible. Mais au fil des années, le nombre de magasins augmente, ce qui signifie que la concurrence a également augmenté en conséquence, d'où une baisse des ventes d'année en année .



À partir de ce graphique, nous avons vu que les ventes des magasins s'effectuent et que leur promotion se poursuit. Les ventes en 2013 et 2015 sont très faibles malgré la promotion. La raison peut être une concurrence accrue d'année en année.



À partir de ce graphique, nous obtenons des informations selon lesquelles les ventes maximales sont le jour 1, qui est le lundi, et elles diminuent jusqu'au jour 6, qui est le samedi. Comme la plupart des magasins sont fermés le jour 7, qui est le dimanche, les ventes sont fermées à zéro.

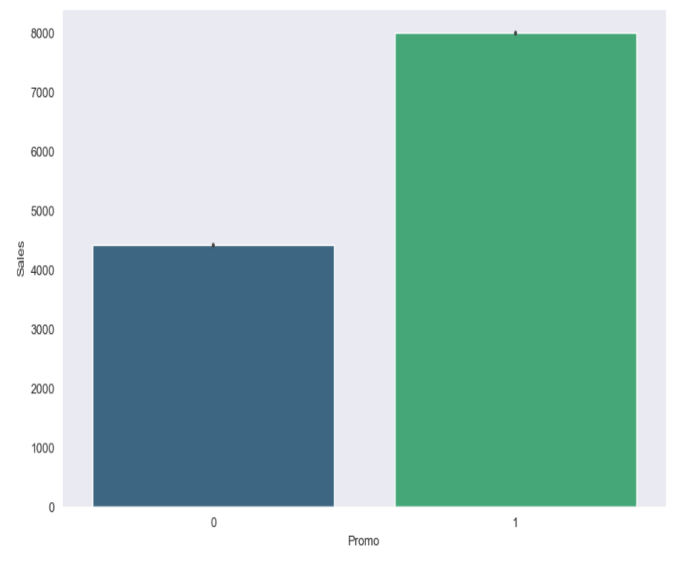


Le type de magasin « B » présente les ventes moyennes les plus élevées, ce qui indique qu'il s'agit en moyenne du type de magasin le plus rentable.

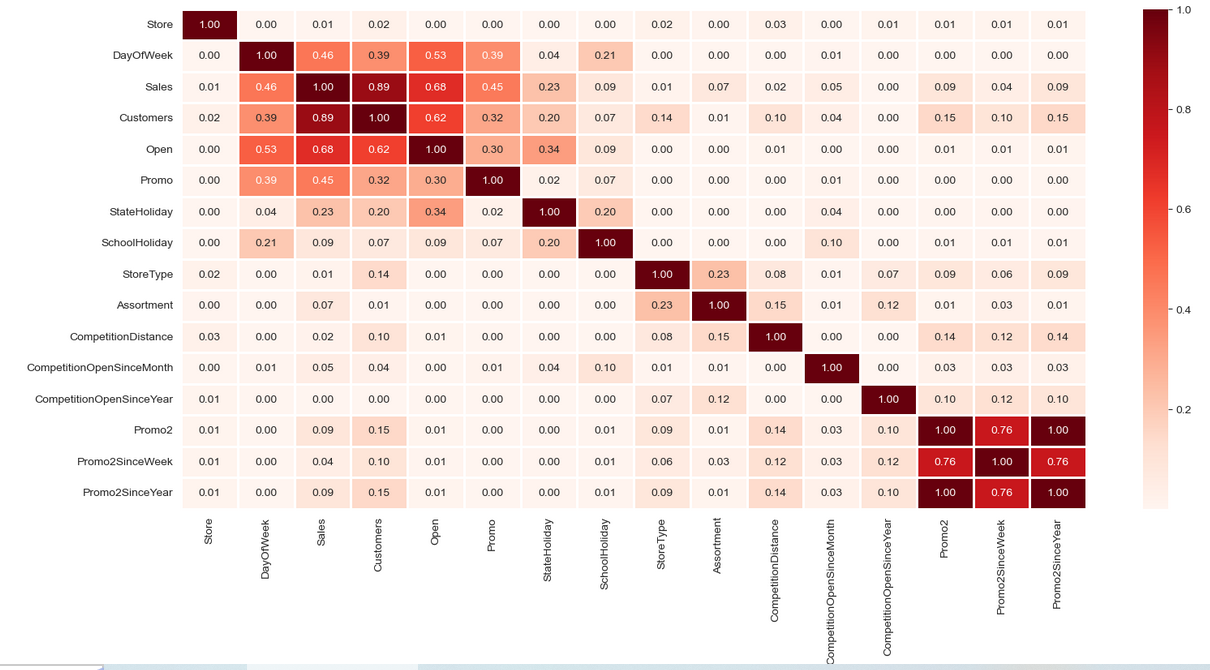
Les magasins de type « A » et « C » enregistrent également des ventes moyennes respectables.

Le type de magasin « D » présente les ventes moyennes les plus faibles parmi les catégories.

Cette analyse aide à comprendre les variations des performances commerciales selon les différents types de magasins, ce qui peut être utile pour la prise de décision stratégique.



Ici,0 représente le magasin qui n'a pas opté pour la promotion et 1 représente les magasins qui optent pour la promotion. Les magasins qui ont bénéficié de promotions ont des ventes élevées par rapport aux magasins qui n'ont pas bénéficié de promotions.

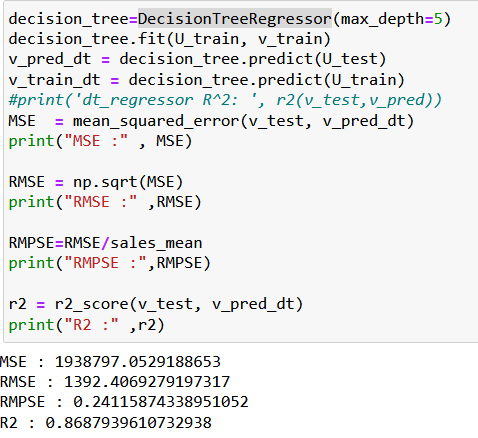
****

La couleur de la cellule indique la direction et la force de la corrélation : une corrélation positive est indiquée par une couleur chaude (comme le rouge) et une corrélation négative est indiquée par une couleur froide (comme l'orange). L'intensité de la couleur représente la force de la corrélation.

## Résultats des Algorithmes :

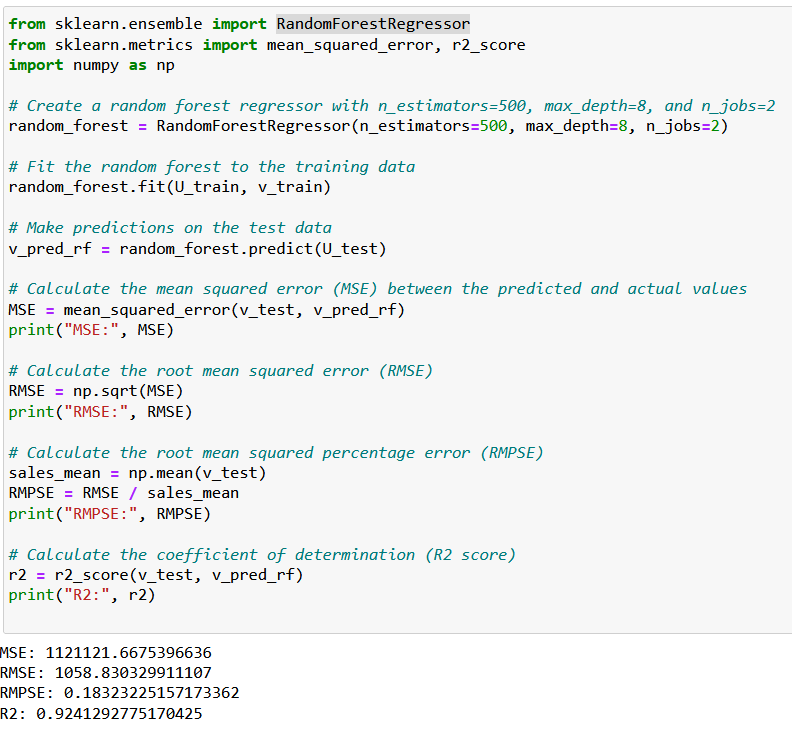
**DecisionTreeRegressor :**

Un modèle Decision Tree Regressor est un algorithme d'apprentissage supervisé utilisé pour la régression. Il utilise un arbre de décision où chaque nœud représente une caractéristique (ou attribut), chaque branche une décision basée sur cette caractéristique, et chaque feuille une valeur de sortie (réponse). Il divise l'espace des caractéristiques en segments pour modéliser les relations non linéaires entre les caractéristiques et la variable cible.



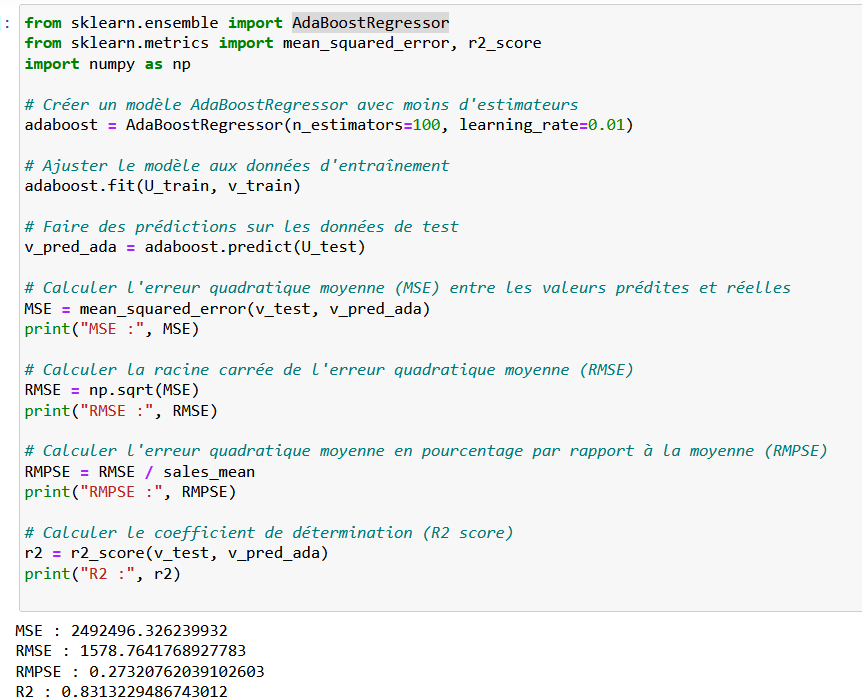
**RandomForestRegressor :**

Le modèle RandomForest Regressor est une extension du modèle Decision Tree Regressor. Il appartient à la famille des méthodes d'ensemble (ensemble learning). Il crée un ensemble de plusieurs arbres de décision, chacun formé sur un sous-ensemble aléatoire des données d'entraînement. Les prédictions résultent de la moyenne (pour la régression) des prédictions de chaque arbre, ce qui améliore la stabilité et la précision du modèle.



**AdaBoostRegressor :**

AdaBoost (Adaptative Boosting) Regressor est un autre algorithme d'ensemble learning. Il vise à améliorer la performance d'un modèle de faible qualité en lui assignant des poids adaptatifs. Il construit itérativement des modèles faibles, mettant davantage l'accent sur les instances mal prédites lors des itérations précédentes. Les modèles faibles sont pondérés et combinés pour former un modèle global robuste. AdaBoost Regressor est souvent utilisé pour des problèmes de régression.



## Conclusion :

Le chapitre se conclura par un résumé des principales conclusions tirées de l'analyse des résultats. Nous mettrons en évidence les éléments les plus pertinents et les enseignements clés qui peuvent orienter les décisions stratégiques de l'entreprise. De plus, nous aborderons les perspectives futures et les pistes d'amélioration pour continuer à renforcer l'efficacité de notre système d'analyse.

Cette structure claire et progressive facilitera la compréhension du lecteur et lui permettra de suivre logiquement le cheminement de l'analyse des données jusqu'aux conclusions finales.

# Conclusion

En conclusion, ce projet dédié à l'analyse des ventes des magasins Rossmann à l'aide de la Business Intelligence a démontré l'efficacité d'une approche intégrée mêlant des outils d'ETL tels que SSIS et Talend, des algorithmes de Machine Learning comme RandomForestRegressor, et des solutions de visualisation telles que Power BI.

L'utilisation de l'ETL a permis une gestion fluide des données, garantissant la qualité des informations extraites, transformées, et chargées dans notre système. L'impact des algorithmes de Machine Learning, en particulier RandomForestRegressor, s'est révélé significatif, générant des prévisions de vente précises avec un coefficient R2 élevé.

Power BI a joué un rôle essentiel en transformant ces résultats complexes en visualisations intuitives et informatives, facilitant ainsi la communication des insights tirés des données. Cette convergence d'outils a créé un environnement robuste pour l'analyse des ventes, permettant des prises de décision éclairées et une compréhension approfondie des performances des magasins.

Au-delà de l'aspect technique, ce projet souligne l'importance stratégique de la Business Intelligence dans le secteur commercial. Les informations obtenues ont le potentiel d'influencer positivement la prise de décision, d'optimiser les opérations et de positionner l'entreprise de manière compétitive sur le marché.

En somme, cette initiative a démontré que la synergie entre l'ETL, le Machine Learning et la Business Intelligence est un pilier puissant pour une analyse approfondie des données. Les enseignements tirés de ce projet offrent une base solide pour des initiatives futures, soulignant l'importance croissante de la Business Intelligence dans la gestion proactive des données et la réalisation d'objectifs commerciaux.