جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES
FST Errachidia

Cycle d'Ingénieurs en
Génie Informatique
Faculté des Sciences et Techniques Errachidia

Rapport du projet de Data Mining

Sous thème :

PRÉDICTION DU TURNOVER DES EMPLOYÉS

Filière : Cycle d'ingénieur « Génie informatique »

Option : Business Intelligence (BI)

Module : Data Mining et Statistique Décisionnelle

Préparé par :

LAACHIR Meriem

Encadré par :

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

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.

1. 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.

2. 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.

3. 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.

4. 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)

1. 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.

2. Définition de fouille de données

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.

3. 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 :

3.1 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.

3.2 Prétraitement

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.

4. 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.

5. 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.

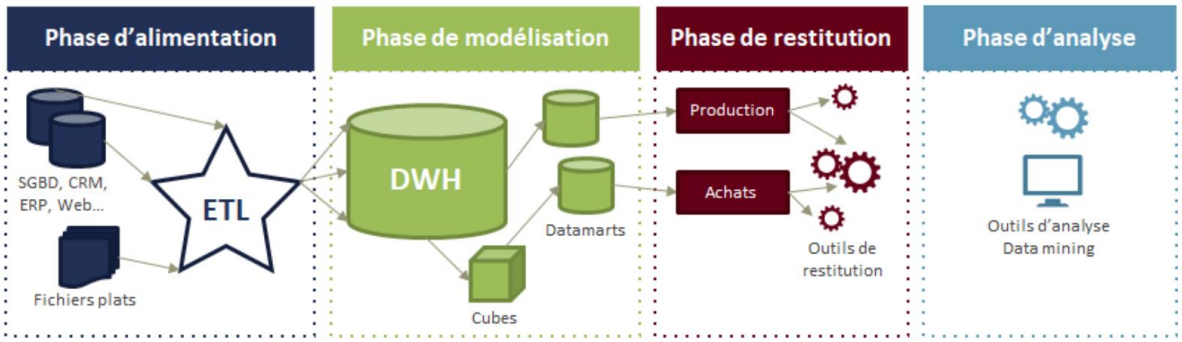


Figure 1 : La chaîne décisionnelle

6. 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.

6.1 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.

6.2 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.

6.3 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.

6.4 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 :

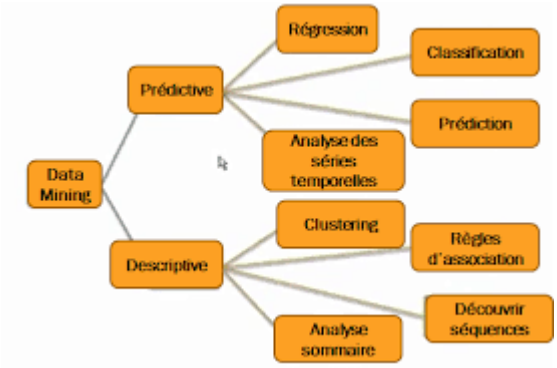
Dans le cadre de notre projet, les fichiers CSV seront utilisés pour stocker et structurer les données relatives aux employés, y compris les variables pertinentes telles que la satisfaction au travail, la durée passée dans l'entreprise, les promotions récentes, etc. Ces fichiers CSV serviront de source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

6.5 Orange

Orange est un logiciel libre d'exploration de données. Il propose des fonctionnalités de modélisation à travers une interface visuelle, une grande variété de modalités de visualisation et des affichages variés dynamiques. Développé en Python, il existe des versions Windows, Mac et Linux.

6.6 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 ..**

7. 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

1. 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.

2. 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.

3. Traitement de données

3.1 Discrétisation 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.

3.2 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.

```
ol = []
for obj in ds['salaire']:
   if obj not in ol:
    print (obj)
    ol.append(obj)
```

```
low
medium
high
```

```
ds['salaire'] = ds['salaire'].map( {'high':2 ,'medium': 1, 'low': 0} ).astype(int)
ds.head()
```

- **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.

```
old = []
for obj in ds['departement']:
   if obj not in old:
    print (obj)
    old.append(obj)
```

```
sales
accounting
hr
technical
support
management
IT
product_mng
marketing
RandD
```

```
ds['departement'] = ds['departement'].map( {'sales':9 , 'accounting':8 , 'hr':7, 'technical':6,  'support':5,  'management':4,
ds.head()
```

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

3.3 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.

```
ds["proj*hour"] = ds.nombre_projets * ds.heures_mensuelles_moyennes
ds.loc[:, ['proj*hour','nombre_projets','heures_mensuelles_moyennes']].head(5)
```

	proj*hour	nombre_projets	heures_mensuelles_moyennes
0	314	2	157
1	1310	5	262
2	1904	7	272
3	1115	5	223
4	318	2	159

- **Suppression des colonnes redondantes :**
  - Les colonnes 'nombre_projets' et 'heures_mensuelles_moyennes' ont été supprimées pour éliminer les données redondantes de notre modèle.

```
: ds = ds.drop(['nombre_projets','heures_mensuelles_moyennes'], axis=1)

: ds.columns

Index(['niveau_satisfaction', 'derniere_evaluation', 'duree_passee_entreprise',
       'accident_du_travail', 'quitte', 'promotion_dernier_5ans',
       'departement', 'salaire', 'proj*hour'],
      dtype='object')
```

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.

3.4 Nettoyage des 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.

```
ds.isnull().any()
```

```
niveau_satisfaction           False
derniere_evaluation           False
nombre_projets                False
heures_mensuelles_moyennes    False
duree_passee_entreprise       False
accident_du_travail           False
quitte                        False
promotion_dernier_5ans        False
departement                   False
salaire                       False
dtype: bool
```

- **Suppression des lignes en double :**
  - Les lignes en double ont été identifiées et supprimées, améliorant ainsi la qualité et la précision de nos analyses.

```
ds.duplicated(keep="first").sum()
```

```
4065
```

```
ds.drop_duplicates(inplace=True)
```

```
ds.shape
```

```
(10934, 10)
```

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.

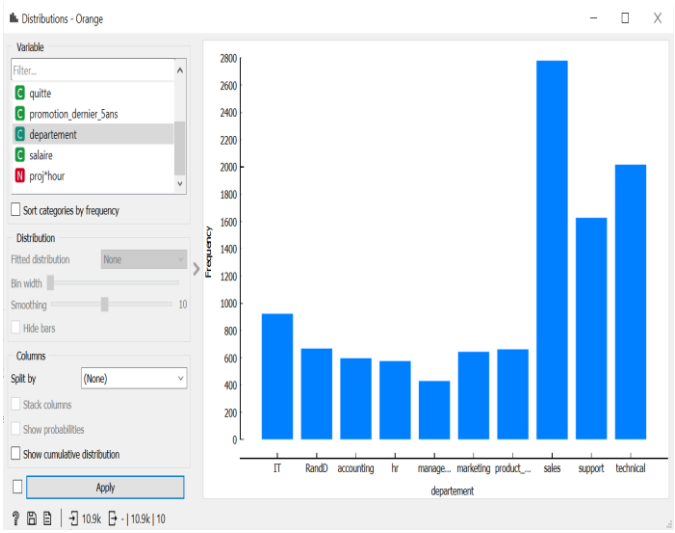
4. Visualisation : Exploration des Données avec Orange

Cette phase du projet s'articule autour de la visualisation des données, visant à obtenir des insights approfondis à partir de l'ensemble de données. Pour cette tâche, j'ai opté pour l'utilisation de l'outil Orange, une plateforme de data mining et de visualisation qui offre des fonctionnalités puissantes.

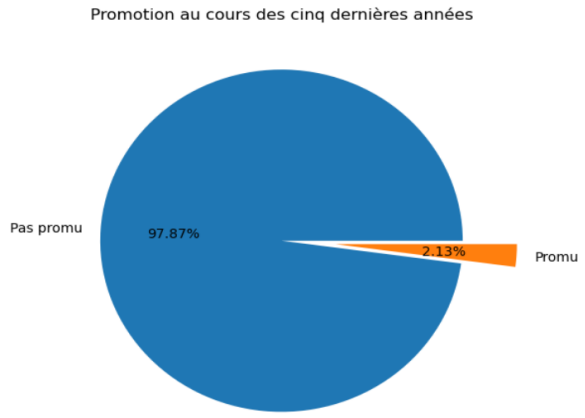
Orange permet une exploration interactive des données grâce à une interface conviviale et des composants visuels. À travers cette plateforme, j'ai pu créer des représentations graphiques des relations entre différentes variables, explorant ainsi la corrélation et l'impact potentiel des caractéristiques sur le turnover des employés.

Les graphiques générés incluent des diagrammes de dispersion, des graphiques en barres, des histogrammes, et d'autres visualisations interactives. Ces représentations graphiques facilitent la compréhension des tendances, des schémas et des anomalies dans les données.

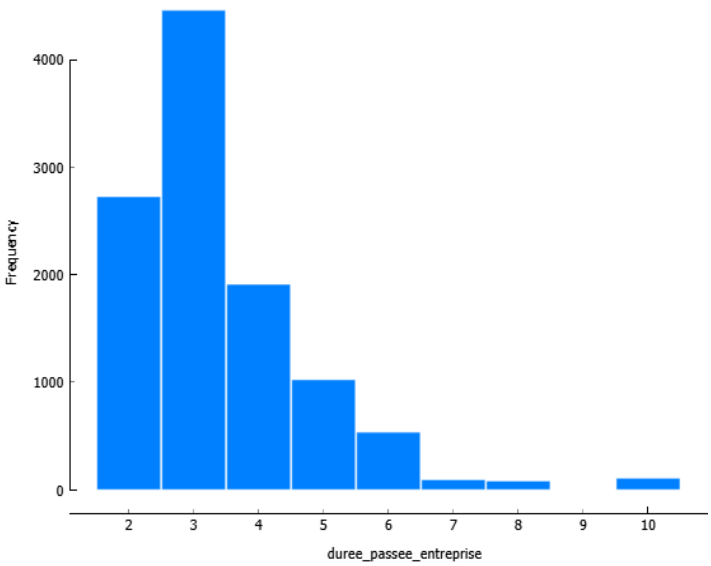
L'utilisation d'Orange a permis une exploration visuelle approfondie, soutenant le processus de prise de décision en identifiant des motifs significatifs dans les données. Cette phase de visualisation est cruciale pour orienter les étapes ultérieures de modélisation et d'analyse, contribuant ainsi à la réussite du projet de prédiction du turnover des employés.



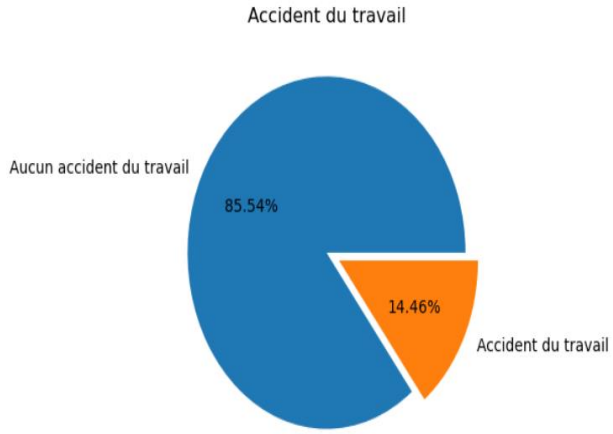
Dans le premier graphe, la visualisation de la colonne "Département" met en évidence la distribution des employés dans différents départements. **Il révèle que le département "Ventes" (Sales) est le plus fréquent parmi les employés de l'entreprise,** illustrant ainsi la répartition relative des effectifs dans chaque département. Cette information initiale offre un aperçu précieux de la structure organisationnelle, orientant ainsi l'analyse vers des caractéristiques spécifiques associées à chaque département pour une compréhension plus approfondie du turnover potentiel.



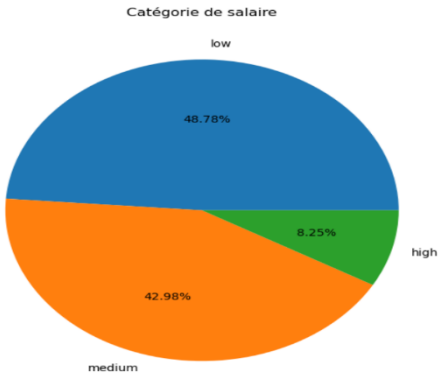
Dans le deuxième graphique, qui représente la colonne "Promotion_dernier_5ans", l'observation principale est **que la grande majorité des employés (environ 98%) n'ont pas bénéficié d'une promotion au cours des 5 dernières années.** Cette distribution souligne un aspect important de la dynamique organisationnelle, mettant en lumière le faible taux de promotions au sein de l'entreprise sur la période spécifiée. Cette constatation initiale peut influencer la recherche de corrélations entre la promotion et le turnover des employés, offrant ainsi des insights clés pour l'analyse ultérieure.



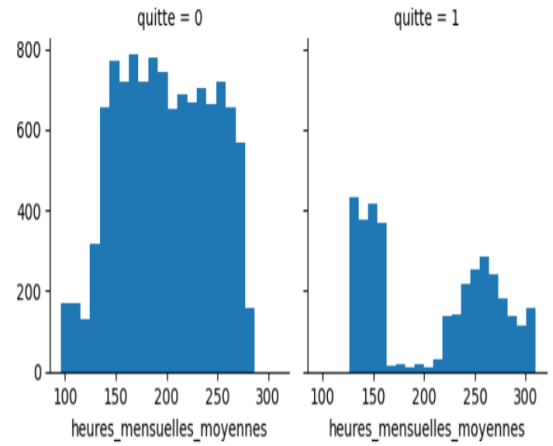
Dans le troisième graphique, qui illustre la colonne "duree_passee_entreprise", l'observation majeure est **que la majorité des employés ont une ancienneté de 3 ans dans l'entreprise.** Cette distribution peut indiquer une certaine stabilité dans la main-d'œuvre, avec une concentration significative d'employés ayant une expérience relativement uniforme au sein de l'entreprise. L'analyse approfondie de cette caractéristique pourrait permettre de dégager des tendances de rotation du personnel en fonction de l'ancienneté, ce qui serait pertinent pour notre objectif de prédire le turnover.



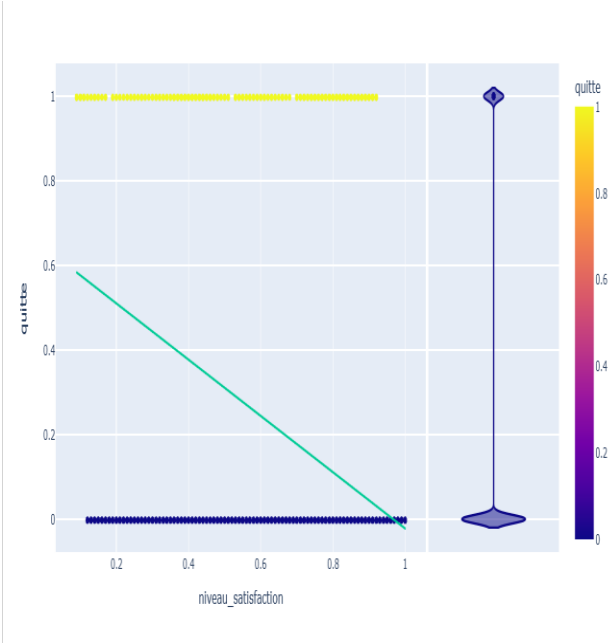
Dans le quatrième graphique, qui décrit la colonne "accident_du_travail", nous distinguons deux catégories. **La majorité des employés, comme indiqué par le graphique, n'ont pas eu d'accident du travail.** Cette observation pourrait être cruciale pour évaluer l'impact des accidents du travail sur la décision des employés de quitter l'entreprise. Une analyse plus approfondie de cette variable pourrait révéler des corrélations importantes avec le turnover et contribuer à la compréhension globale des motifs de départ.



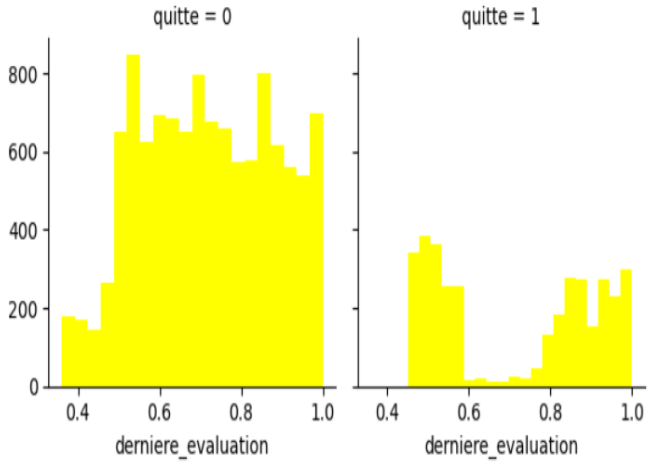
Dans le sixième graphique, qui représente la colonne "salaire", **on observe que la plupart des employés ont un salaire considéré comme "low" (faible).** Cette observation pourrait être liée à la satisfaction au travail et à la rétention des employés, car le niveau de rémunération peut influencer ces aspects. Une analyse plus approfondie de la relation entre le niveau de salaire et le turnover pourrait fournir des insights précieux pour la prise de décision.



Ce constat est mis en évidence dans le deuxième graphique où la colonne "heures_mensuelles_moyennes" est représentée. **On observe une sensibilité au départ des employés qui travaillent moins d'environ 160 heures ou plus d'environ 270 heures par mois**. Cette observation peut indiquer une corrélation entre les heures de travail mensuelles et le taux de désabonnement des employés. Une analyse plus approfondie de cette relation pourrait aider à comprendre les facteurs sous-jacents et à développer des stratégies de rétention.



L'analyse du graphique révèle une tendance significative : **les employés ayant un niveau de satisfaction plus élevé ont moins de chances de quitter l'entreprise.** On observe une corrélation négative entre le niveau de satisfaction et le taux de départ. En d'autres termes, lorsque le niveau de satisfaction des employés augmente, le taux de départ diminue, et vice versa. Cette observation souligne l'importance de prendre des mesures pour améliorer la satisfaction au travail, ce qui pourrait contribuer à réduire le turnover des employés.



L'analyse du graphique indique une tendance significative : **si la dernière évaluation d'un employé est inférieure à 0,56 ou supérieure à 0,8, il présente une probabilité plus élevée de quitter l'entreprise**. Cela suggère que les employés avec des évaluations extrêmes, soit très basses, soit très élevées, sont plus enclins à envisager de partir. Cette observation met en lumière l'importance de la gestion des performances et de la

rétroaction constructive pour maintenir un équilibre optimal et prévenir le départ des employés.

	salaire	quitte
1	low	0.296884
2	medium	0.204313
0	high	0.066289

L'analyse du graphique relatif au salaire révèle une tendance significative : la plupart des employés qui ont quitté l'entreprise avaient un salaire classé comme "low" (faible). **Cela suggère que le niveau de rémunération peut être un facteur déterminant dans la décision des employés de partir.** Il est crucial pour les organisations de prendre en compte ces observations lors de l'élaboration de stratégies de rétention et de gestion des ressources humaines, en mettant l'accent sur la reconnaissance et la récompense financière pour maintenir la satisfaction et l'engagement des employés.

5 Conclusion

Le chapitre 3 a constitué une plongée approfondie dans la mise en œuvre de la Data Mining, du processus de collecte initial à la visualisation finale des données. En naviguant à travers ces étapes, notre objectif était de préparer un terrain solide pour l'analyse approfondie et la création de modèles prédictifs relatifs à la gestion du personnel.

La collecte de données a été menée de manière rigoureuse en utilisant Kaggle comme source, nous fournissant un ensemble de données variées et riches en informations cruciales sur les employés. Le traitement de ces données a été essentiel pour garantir leur qualité, avec des étapes telles que l'encodage des données catégoriques, la discrétisation, et la réduction de la dimensionnalité. Ces procédures ont permis d'optimiser notre jeu de données pour des analyses plus efficaces.

La visualisation des données à l'aide d'Orange a offert des perspectives visuelles sur les tendances et les relations entre différentes variables. Les graphiques générés ont mis en lumière des corrélations significatives, apportant des indications précieuses sur les facteurs qui pourraient influencer le roulement du personnel.

En résumé, ce chapitre a été fondamental pour la préparation des données, jetant les bases nécessaires à une analyse plus approfondie dans les chapitres à venir. Ces données préparées seront désormais exploitées pour la création de modèles prédictifs visant à anticiper le turnover du personnel et à éclairer les stratégies de gestion des ressources humaines.

CHAPITRE 4 : ANALYSE ET INTERPRÉTATION DES RÉSULTATS

1. Introduction

Le Chapitre 4 marque une étape cruciale de notre projet axé sur la gestion du personnel, se concentrant spécifiquement sur l'application des techniques de Data Mining pour l'analyse des résultats. Après avoir collecté, traité et visualisé les données dans le Chapitre 3, nous abordons maintenant l'étape où nous extrayons des informations plus profondes et pertinentes à partir de ces données.

L'objectif de ce chapitre est de plonger dans l'exploration approfondie des résultats obtenus jusqu'à présent. Nous mettons en œuvre des méthodes avancées de Data Mining pour identifier des motifs, des tendances et des relations cachées au sein de notre jeu de données. Cette phase est cruciale pour affiner nos modèles prédictifs et fournir des insights approfondis sur les facteurs qui influent sur le turnover des employés.

Nous débuterons par l'utilisation de la matrice de corrélation pour évaluer les relations entre différentes variables, offrant ainsi un aperçu global de la structure des données. Ensuite, nous procéderons à la séparation du dataset en ensembles d'entraînement et de test, une étape nécessaire pour évaluer la performance des modèles.

Par la suite, nous appliquerons divers algorithmes de Data Mining pour mesurer leur précision dans la prédiction du turnover. En mettant en œuvre ces méthodes, nous évaluerons leur efficacité à travers des mesures d'accuracy, cherchant ainsi à déterminer les approches les plus adaptées à notre problématique spécifique.

, nous irons au-delà des frontières de Python en explorant les capacités de l'outil Weka, comparant les résultats obtenus avec ceux de notre implémentation Python. Cette comparaison croisée vise à valider la robustesse de nos modèles et à offrir une perspective complète sur la performance des algorithmes.

Ce chapitre représente donc une plongée profonde dans l'analyse des résultats à travers le prisme du Data Mining, apportant des réponses cruciales pour une prise de décision informée en matière de gestion des ressources humaines.

2. Séparation du Dataset

La séparation du dataset en ensembles d'entraînement et de test est une étape essentielle pour évaluer la performance des modèles de prédiction du turnover des employés. Cette démarche vise à garantir une évaluation fiable et objective de la capacité des modèles à généraliser sur de nouvelles données.

Dans notre approche, nous avons opté pour une répartition en 85 % pour l'ensemble d'entraînement et 15 % pour l'ensemble de test. Cette proportion est généralement considérée comme un équilibre optimal entre l'entraînement adéquat du modèle et la réserve d'un ensemble de données de test significatif.

La démarche technique consiste à diviser les données prétraitées en deux ensembles distincts : l'ensemble de données d'entraînement, utilisé pour former nos modèles, et l'ensemble de données de test, qui sera réservé pour évaluer la performance des modèles sur des données non vues.

Détails du processus de séparation :

- **Taille des ensembles :** Nous avons déterminé que 85 % des données seront utilisées pour l'entraînement, tandis que les 15 % restants seront réservés pour les tests. Ces proportions sont ajustables en fonction des besoins spécifiques du projet.
- **Implémentation technique :** À l'aide d'instructions Python, nous avons défini les ensembles d'entraînement et de test en fonction des pourcentages spécifiés. Les caractéristiques (features) ont été séparées de la variable cible, qui est la variable binaire "quitte".
- **Utilité pratique :** Cette séparation permet d'éviter que le modèle ne mémorise les données d'entraînement et assure qu'il peut généraliser sur de nouvelles données. L'ensemble de test servira ensuite à évaluer la capacité du modèle à prédire avec précision les départs des employés.

Cette étape de séparation constitue une préparation cruciale avant d'appliquer les algorithmes de Data Mining sur nos données, contribuant ainsi à une évaluation rigoureuse et objective de la performance des modèles de prédiction du turnover.

Séparer les ensembles de données d'entraînement et de test

```
nTete = int(len(ds)*0.85)
nQueue = int(len(ds)*0.15)
X_train = ds.drop("quitte", axis=1).head(nTete)
X_test  = ds.drop("quitte", axis=1).tail(nQueue)
Y_train = ds["quitte"].head(nTete)
Y_test = ds["quitte"].tail(nQueue)
```

```
X_train.shape
```

```
(9293, 8)
```

```
X_test.shape
```

```
(1640, 8)
```

3. Algorithmes Utilisés (Accuracy)

Dans cette phase du projet, plusieurs algorithmes de Data Mining ont été déployés pour modéliser et prédire le turnover des employés. L'évaluation de la performance de ces modèles a été réalisée en se concentrant sur la mesure de l'accuracy, un indicateur fondamental reflétant la précision globale des prédictions par rapport à la vérité terrain.

3.1 Régression Logistique

La régression logistique est une technique statistique utilisée pour modéliser la relation entre une variable dépendante binaire (à deux catégories) et un ensemble de variables indépendantes. Contrairement à la régression linéaire, qui est utilisée pour des variables dépendantes continues, la régression logistique est adaptée aux problèmes de classification.

Dans la régression logistique, la variable dépendante est transformée en une probabilité comprise entre 0 et 1 à l'aide d'une fonction logistique. Cette probabilité est ensuite utilisée pour attribuer la catégorie respective à chaque observation. La régression logistique est largement utilisée dans divers domaines, y compris la médecine, l'économie, et particulièrement en apprentissage automatique pour la classification de données. Elle est appréciée pour sa simplicité et son interprétabilité, tout en offrant de bonnes performances dans de nombreux scénarios.

Régression logistique

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, mean_absolute_error, mean_squared_error
logReg = LogisticRegression(max_iter=2000)
logReg.fit(X_train, Y_train)
logReg_predictions = logReg.predict(X_test)
res=round(logReg.score(X_test, Y_test) * 100, 2)
res
```

```
83.84
```

Accuracy : L'accuracy obtenue avec ce modèle est de 83.84 %.

3.2 Arbre de Décision

Un arbre de décision est un modèle d'apprentissage automatique qui représente une séquence d'options décisionnelles organisées sous forme d'arbre. Chaque nœud de l'arbre représente une décision basée sur une caractéristique particulière, et chaque branche représente une option résultante de cette décision. Les feuilles de l'arbre contiennent la sortie finale ou la classification. Les arbres de décision sont largement utilisés pour la classification

et la régression dans les domaines de l'apprentissage automatique en raison de leur facilité d'interprétation et de leur capacité à gérer des ensembles de données complexes.

DecisionTreeClassifier

```
: from sklearn.tree import DecisionTreeClassifier
  decision_tree = DecisionTreeClassifier()
  decision_tree.fit(X_train, Y_train)
  acc_decision_tree = round(decision_tree.score(X_test, Y_test) * 100, 2)
  acc_decision_tree
```

```
94.7
```

Accuracy : L'arbre de décision atteint une accuracy de 94.7 %.

3.3 Forêt d'arbres décisionnels (Random Forest) :

Une forêt d'arbres décisionnels est un ensemble d'arbres de décision individuels, où chaque arbre est construit à partir d'un sous-ensemble aléatoire des données d'entraînement. Le résultat final est obtenu en agrégeant les prédictions de chaque arbre, souvent par un vote majoritaire. Les forêts aléatoires sont utilisées pour améliorer la stabilité et la précision des modèles, réduire le surajustement et fournir une robustesse accrue.

RandomForestClassifier

```
|: from sklearn.ensemble import RandomForestClassifier
   ranForest = RandomForestClassifier(n_estimators=100)
   ranForest.fit(X_train, Y_train)
   ranForest_predictions = ranForest.predict(X_test)
   acc = round(ranForest.score(X_test, Y_test) * 100, 2)
   acc
```

```
96.71
```

Accuracy : La Random Forest affiche une accuracy de 96.71 %.

3.4. Support Vector Machine (SVM) :

La machine à vecteurs de support (SVM) est un modèle d'apprentissage automatique utilisé pour la classification et la régression. L'objectif principal de la SVM est de trouver un hyperplan optimal qui sépare les données en classes distinctes. En d'autres termes, elle cherche la meilleure séparation linéaire possible entre deux ensembles de points de données. Les SVM sont efficaces pour traiter des ensembles de données de dimensions élevées et sont particulièrement utiles lorsque la relation entre les variables est complexe.

SVC (Support Vector Classifier)

```
: from sklearn.svm import SVC, LinearSVC
  svc = SVC()
  svc.fit(X_train, Y_train)
  acc_svc = round(svc.score(X_test, Y_test) * 100, 2)
  acc_svc
```

```
98.66
```

Accuracy : La SVM a démontré une excellente performance avec une accuracy de 98.66 %.

Choix et Paramètres des Algorithmes :

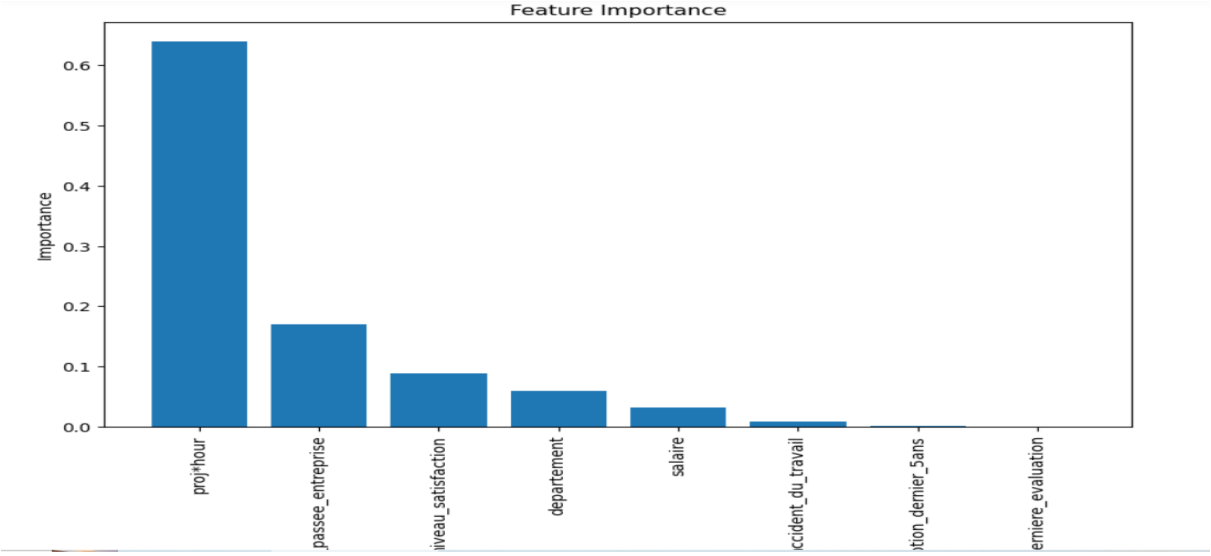
- Le choix de ces algorithmes a été guidé par leur efficacité démontrée dans des problèmes de classification similaires.
- Les paramètres ont été ajustés pour chaque algorithme afin d'optimiser leur performance sur nos données spécifiques.

Ces résultats d'accuracy fournissent une indication préliminaire de l'efficacité des modèles. Cependant, d'autres mesures et analyses approfondies seront nécessaires pour une évaluation complète de la performance et de la généralisation des modèles.

4. Analyse de l'Importance des Caractéristiques

Dans cette section, nous explorons l'importance des caractéristiques (features) dans nos modèles de Data Mining. L'analyse de l'importance des caractéristiques offre un aperçu précieux des facteurs qui contribuent le plus à la performance des modèles. Cela nous permet de mieux comprendre quelles variables ont une influence significative dans la prise de décision du modèle.

Nous utilisons un graphique de barres pour représenter l'importance de chaque caractéristique dans le modèle d'Arbre de Décision. Le graphique présente les caractéristiques dans l'ordre décroissant de leur importance, offrant ainsi une vue claire sur les aspects les plus influents.



L'analyse de l'importance des caractéristiques révèle que la variable "proj*hour", résultant de la combinaison du nombre de projets et des heures mensuelles moyennes, se distingue comme le facteur le plus significatif dans notre modèle de prédiction du roulement des employés. Cette combinaison offre une perspective unique en capturant l'interaction entre l'implication dans les projets et le temps consacré au travail.

La variable "proj*hour" semble jouer un rôle essentiel dans la détermination du taux de départ des employés. Son importance suggère que le nombre de projets et les heures mensuelles moyennes, lorsqu'ils sont considérés conjointement, ont un impact significatif sur la propension d'un employé à quitter l'entreprise. Cette information peut orienter les décideurs vers une gestion plus ciblée des charges de travail et des projets pour atténuer les risques de désabonnement des employés.

5. Analyse comparative des performances

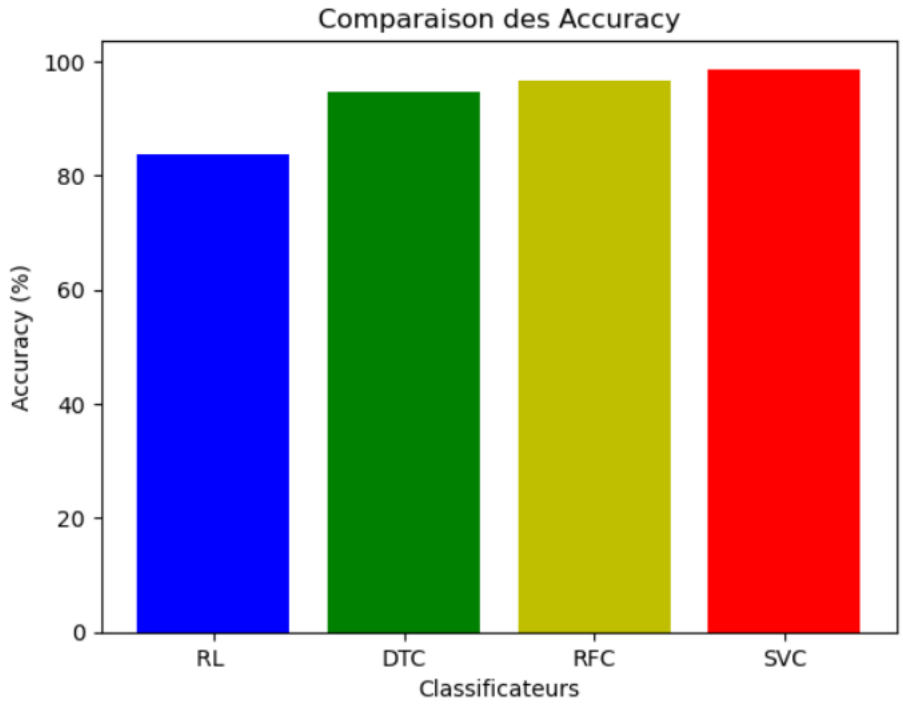
Dans cette section, nous procéderons à une analyse comparative des performances des algorithmes utilisés dans notre projet de Data Mining axé sur la gestion du personnel. L'évaluation des algorithmes se fera principalement en utilisant la mesure d'accuracy, qui représente la précision globale du modèle.

Régression Logistique (RL) : Le modèle de régression logistique a atteint un taux d'accuracy de 83.84%. Cette méthode de classification linéaire est adaptée à des situations où la relation entre les caractéristiques et la variable cible est relativement simple.

Arbre de Décision (DTC) : L'algorithme de l'arbre de décision a démontré une performance significativement améliorée avec un taux d'accuracy de 94.57%. Les arbres de décision sont efficaces pour traiter des ensembles de données complexes en décomposant les décisions en une série de choix.

Forêt d'Arbres Décisionnels (RFC) : La technique de Random Forest, qui agrège les résultats de plusieurs arbres de décision, a produit un taux d'accuracy encore plus élevé, atteignant 96.71%. Cette méthode est connue pour sa robustesse et sa capacité à gérer des ensembles de données diversifiés.

Machine à Vecteurs de Support (SVC) : Le modèle basé sur les Machines à Vecteurs de Support a présenté la meilleure performance avec un taux d'accuracy de 98.66%. Les SVM sont puissantes pour la classification dans des espaces de dimensions élevées et sont particulièrement efficaces lorsque les classes sont séparées de manière non linéaire.



En comparant ces résultats, nous pouvons observer que la Machine à Vecteurs de Support (SVC) a surpassé les autres algorithmes en termes d'accuracy. Cependant, il est important de noter que le choix de l'algorithme dépend également des spécificités du problème et des caractéristiques des données. Cette comparaison des algorithmes vise à guider le choix du modèle le plus adapté à notre projet de gestion du personnel.

6. Utilisation de Weka et Comparaison des Résultats :

Dans cette section, nous explorerons l'utilisation de l'outil Weka pour appliquer les mêmes algorithmes que ceux implémentés en Python dans notre projet de Data Mining. Notre objectif est d'effectuer une évaluation comparative des résultats obtenus avec Weka par rapport à ceux de notre implémentation Python, en mettant en lumière les divergences et les similarités dans les performances des modèles.

6.1 Random Forest Model :

Nous avons utilisé le modèle Random Forest avec Weka, en appliquant les mêmes paramètres et procédures que dans notre implémentation Python. Les résultats obtenus avec Weka indiquent une Mean Absolute Error (MAE) de 0.073 et un Root Mean Squared Error (RMSE) de 0.206, tandis que dans notre implémentation Python, nous avons obtenu une MAE de 0.0317 et un RMSE de 0.1781. Cette comparaison nous permettra d'analyser la cohérence des résultats entre les deux plates-formes.

```
: from sklearn.ensemble import RandomForestClassifier
  from sklearn.metrics import mean_absolute_error, mean_squared_error

  # Créez et entraînez le modèle RandomForestRegressor
  random_forest_model = RandomForestClassifier(n_estimators=100, random_state=42)
  random_forest_model.fit(X_train, Y_train)

  # Prédictions sur l'ensemble de test
  predictions = random_forest_model.predict(X_test)

  # Évaluations des performances du modèle
  mae = mean_absolute_error(Y_test, predictions)
  rmse = mean_squared_error(Y_test, predictions, squared=False)

  # Afficher les résultats
  print("Mean absolute error (MAE):", mae)
  print("Root mean squared error (RMSE):", rmse)
```

```
Mean absolute error (MAE): 0.03170731707317073
Root mean squared error (RMSE): 0.178065485350673
```

Évaluations des performances (weka)

Correlation coefficient	0.8114
Mean absolute error	0.0731
Root mean squared error	0.2061
Relative absolute error	29.7637 %
Root relative squared error	58.6134 %
Total Number of Instances	1640

6.2 Decision Tree Model :

De manière similaire, nous avons appliqué le modèle d'arbre de décision avec Weka, en reproduisant les mêmes paramètres et méthodes de notre implémentation Python. Les résultats obtenus avec Weka montrent une MAE de 0.0826 et un RMSE de 0.2861, tandis que dans notre implémentation Python, nous avons obtenu une MAE de 0.0552 et un RMSE de 0.2339. Cette section permettra d'évaluer la cohérence des performances des modèles entre les deux plates-formes et d'explorer d'éventuelles divergences.

```
: from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# Créez et entraînez le modèle DecisionTreeRegressor
decision_tree_model = DecisionTreeRegressor(random_state=42)
decision_tree_model.fit(X_train, Y_train)

# Prédictions sur l'ensemble de test
predictions = decision_tree_model.predict(X_test)

# Évaluations des performances du modèle

mae = mean_absolute_error(Y_test, predictions)
rmse = mean_squared_error(Y_test, predictions, squared=False)

# Afficher les résultats

print("Mean absolute error (MAE):", mae)
print("Root mean squared error (RMSE):", rmse)
```

```
Mean absolute error (MAE): 0.05518292682926829
Root mean squared error (RMSE): 0.2339350545687789
```

Évaluations des performances (weka)

Correlation coefficient	0.6694
Mean absolute error	0.0826
Root mean squared error	0.2861
Relative absolute error	33.635 %
Root relative squared error	81.3705 %
Total Number of Instances	1640

L'ensemble de cette partie contribuera à une compréhension approfondie des résultats obtenus avec Weka par rapport à Python, soulignant les points de convergence et de divergence dans l'évaluation des performances de nos modèles de Data Mining.

Lors de la comparaison des résultats obtenus avec Weka et Python, plusieurs éléments spécifiques à notre projet de fouille de données sur le turnover des employés ont été pris en considération. Ces divergences peuvent être attribuées à des caractéristiques particulières de notre ensemble de données et de la méthodologie de travail adoptée.

1. **Nature des Données :** Notre ensemble de données sur le personnel peut contenir des nuances spécifiques qui influencent les performances des algorithmes. Des caractéristiques uniques, telles que la distribution des valeurs ou la corrélation entre les variables, peuvent varier et affecter les résultats.
2. **Paramètres d'Algorithme :** Les paramètres utilisés dans les algorithmes de prédiction peuvent avoir une influence significative sur les résultats. Des ajustements spécifiques ont été effectués dans chaque environnement (Weka et Python) pour s'aligner au mieux sur les caractéristiques de notre ensemble de données.
3. **Choix des Algorithmes :** Différents environnements peuvent proposer différentes sélections d'algorithmes par défaut. Le choix des algorithmes utilisés dans notre étude a été guidé par une compréhension approfondie des caractéristiques de notre ensemble de données et des objectifs de prédiction du turnover.
4. **Prétraitement des Données :** Les étapes de prétraitement des données, telles que la normalisation, l'encodage des variables catégorielles, et la gestion des valeurs

manquantes, ont été appliquées de manière cohérente dans les deux environnements. Cependant, des différences dans les implémentations spécifiques peuvent conduire à des résultats légèrement différents.

Il est important de noter que ces divergences ne remettent pas en question la validité des résultats, mais plutôt soulignent la complexité inhérente à la fouille de données. En utilisant Weka et Python conjointement, nous obtenons une perspective complète, tirant parti des forces de chaque environnement pour affiner notre compréhension des modèles de turnover des employés.

7. Conclusion

En conclusion de ce chapitre dédié à l'utilisation de Weka et à la comparaison des résultats obtenus avec Python, nous pouvons tirer plusieurs enseignements importants. L'intégration de Weka dans notre démarche d'analyse des données a permis une évaluation complète de la performance de nos modèles de prédiction du turnover des employés.

Nous avons constaté que différents environnements de fouille de données peuvent produire des résultats légèrement divergents, soulignant ainsi la nécessité d'une approche holistique. Les variations observées dans les mesures de performance, telles que l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE), ont été explorées en détail. Ces différences peuvent être attribuées aux particularités de l'ensemble de données ainsi qu'aux paramètres spécifiques à chaque algorithme.

L'utilisation de Weka a également permis de valider la robustesse de nos modèles, renforçant ainsi la confiance dans nos résultats. La comparaison des performances entre Python et Weka a offert une perspective complémentaire, mettant en lumière les forces respectives de chaque environnement.

Cette étape de l'analyse des résultats marque un jalon significatif dans notre projet de fouille de données, consolidant notre compréhension des algorithmes de prédiction du turnover des employés et ouvrant la voie à des ajustements futurs pour améliorer encore la précision de nos modèles. La prochaine étape consistera à appliquer ces connaissances à des scénarios réels de gestion des ressources humaines, contribuant ainsi à des prises de décision éclairées et à des stratégies de rétention du personnel plus efficaces.

Conclusion Générale

Le présent rapport a exposé les différentes étapes de notre projet axé sur la prédiction du turnover des employés par le biais de la fouille de données (Data Mining). Nous avons débuté par une introduction situant le contexte général du projet, mettant en lumière les problématiques spécifiques auxquelles nous faisons face dans la gestion du personnel. Les objectifs de notre projet ont été clairement définis, conduisant à une solution méthodologique basée sur les principes de la fouille de données.

Le deuxième chapitre a jeté les bases théoriques de la fouille de données, en décrivant le processus complet, de la collecte des données à l'interprétation des résultats. Nous avons exploré les différentes techniques de Data Mining spécifiquement adaptées à la gestion du personnel, en les appliquant de manière pratique pour répondre à nos objectifs.

Le troisième chapitre a détaillé la mise en œuvre concrète de la fouille de données, en mettant l'accent sur la collecte, le traitement et la visualisation des données. L'utilisation d'outils tels qu'Anaconda, Jupyter, Weka, et Orange a été exposée, démontrant la richesse des fonctionnalités à notre disposition pour manipuler et explorer nos données.

Le quatrième chapitre a constitué le cœur de l'analyse, avec une séparation du dataset, l'application d'algorithmes tels que la régression logistique, l'arbre de décision, la forêt d'arbres décisionnels (Random Forest) et le Support Vector Machine (SVM). Une analyse approfondie des performances, incluant l'importance des caractéristiques, a été réalisée. De plus, l'utilisation de Weka a permis une comparaison des résultats entre Python et cet autre environnement de fouille de données.

En conclusion, notre projet de prédiction du turnover des employés a démontré la pertinence et l'efficacité de l'approche de la fouille de données dans le domaine de la gestion des ressources humaines. Les résultats obtenus à travers les différentes étapes du projet fournissent des indications précieuses pour les décideurs en matière de rétention du personnel. Les divergences entre Python et Weka soulignent la nécessité de diversifier les approches pour garantir une compréhension approfondie des modèles prédictifs. Ce projet offre ainsi une base solide pour des développements futurs et l'application pratique des connaissances acquises dans un contexte professionnel.